HEAVY-TRAFFIC LIMITS VIA AN AVERAGING PRINCIPLE
FOR SERVICE SYSTEMS RESPONDING TO UNEXPECTED OVERLOADS

OHAD PERRY

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
under the Executive Committee of the
Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2010

ABSTRACT

Heavy-Traffic Limits via an Averaging Principle
for Service Systems Responding to Unexpected Overloads

Ohad Perry

This dissertation considers how two networked large-scale service systems, such as call centers, that normally operate separately, can help each other in face of an unexpected overload, caused by a sudden shift in the arrival rates. We assume that the time of the shift and the values of the new arrival rates are not known a-priori, and are hard to detect in real time. We also assume that staffing cannot be increased immediately.

We propose the *fixed-queue ratio with thresholds* (FQR-T) control, and show that it is optimal in a deterministic fluid approximation. The FQR-T control activates serving some customers from the other system when a ratio of the two queue lengths (numbers of waiting customers) exceeds a threshold. Two thresholds, one for each direction of sharing, automatically detect the overload condition and prevent undesired sharing under normal loads. After a threshold has been exceeded, the control aims to keep the ratio of the two queue lengths at a specified value.

To gain insight, we introduce an idealized X model, i.e., a stochastic model with two customer classes, each with its own dedicated service pool, containing a large number of agents. The agents in both pools are assumed to be crossed-trained, so that they are able to serve the other class, even if somewhat inefficiently. To set the important queue-ratio parameters, we consider an approximating deterministic fluid model. We determine queue-ratio parameters that minimize convex costs for this fluid model. Simulations show that the proposed queue-ratio control with thresholds, which uses no information about the new arrival rates during the overload, outperforms the optimal fixed partition of the servers when the new arrival rates are known.

We then consider the stochastic X model under our proposed FQR-T control, and prove that the fluid approximation, developed heuristically for the optimality analysis, holds as a many-server heavy-traffic fluid limit. In particular, under an appropriate fluid scaling, the processes describing the X system, i.e., the queue-length and service processes, converge to a deterministic fluid limit as the number of servers and arrival rates approach infinity. This fluid limit is characterized by an *ordinary differential equation* (ODE), coupled with a *fast-time-scale process* (FTSP). In proving the fluid limit we also achieve a *state-space collapse* (SSC) result, which allows us to develop diffusion refinements.

Proving convergence to the fluid limit is complicated because the limit involves a heavy-traffic *averaging principle* (AP). The X model, operating under FQR-T, is driven by a queue-difference stochastic process operating in a faster time scale than the other processes describing the system, thus achieving a time-dependent steady state instantaneously in the limit. Hence, for the limiting ODE, the queue-difference process is replaced by the long-run average behavior of the FTSP at each instant of time.

In addition to complicating the convergence proofs, the AP also makes standard ODE and dynamical-systems theory difficult to apply. First, the deterministic ODE is driven by a stochastic process, whose distributional characteristics determine the evolution of the solution to the ODE. Moreover, due to the AP and its resulting SSC, the ODE is not continuous in its full state space.

Nevertheless, we provide results about the existence and uniqueness of the solution to the ODE, prove that there exists a unique stationary point; and give easily verifiable conditions for the fluid limit to converge to its stationary point, which was used in our optimization analysis. We also provide an efficient numerical algorithm, based on matrix-geometric methods, for solving the ODE.

Contents

List of Figures

Acknowledgements

This thesis would not have been possible without the help and support of many people.

First and foremost, I would like to thank my advisor, Prof. Ward Whitt. Prof. Whitt is one of the "founding fathers" of the field of this thesis, and having him as my "academic father" is a great honor. Working with him made me confident that any research problem will be solved, no matter how hard and in which field of Probability. We spent many hours together in the past five years, and I will miss our almost daily interactions. I cannot imagine having a better advisor.

I would like to thank (in alphabetical order) Jose Blanchet, Costis Maglaras, Karl Sigman and Assaf Zeevi for agreeing to serve on my committee.

I am thankful to my office mates (in both offices...) who made working fun. I would like to thank all the friends I made in New York, and especially Itai Gurvich, with whom I had many insightful conversations. Without these friends, this work would not be nearly as enjoyable.

My wife Nomi and I are thankful to Rivka and Haim Rosenstein for being a family to us in New York. We will never forget their help, and will forever be grateful. We are also thankful to our family in Mexico for their support.

I would like to thank my family in Israel for being there for me, even from so far away. Spending so little time with my family in the past five years was the hardest part of my PhD. However, their love has always encouraged me, even in my hardest times.

Finally, I want to thank my wife Nomi. Ten years ago she left her family and friends in Mexico and fulfilled her dream to live in Israel. Nevertheless, she agreed to leave her new home once more, and follow me to New York. I dedicate this thesis to her.

New York, NY

December, 2009

Chapter 1

Introduction

The introduction consists of four parts: In §1.1, we quickly review the general motivation for our problem. In §1.2 we briefly describe our modeling approach and our contribution to the existing literature. In §1.3 we provide a short review of mathematical models that commonly appear in the literature and are relevant to our work. Finally, in §1.4 we briefly explain the mathematical methods employed in this dissertation, and our contribution to the existing mathematical-modeling literature of large many-server systems.

1.1 Motivation

One of the characteristics of an advanced economy is its large service sector. For example, in the United States, the service sector is responsible for about 80% of the nominal GDP and over 80% of the work force.

An important part of the service sector is the call-center industry. In the United States alone it employs more than 3.5 million agents (or 2.5% of the total workforce) [17, 70]. However, the importance of the call-center industry goes well beyond its size; It is estimated that call centers handle more than 70% of all business interactions.

Since labor-related costs comprise 60-80% of the overall operating budget of modern call centers [2], managers have to balance two conflicting objectives: on the one hand, they seek to minimize operating costs by reducing the number of agents to the possible minimum. On the other hand, they are required to provide some pre-specified levels of service (which can be measured in various ways, e.g., the proportion of customers that abandon, the average waiting time, the probability of delay in queue, etc.). These two objectives can be achieved in large call centers by staffing appropriately. However, this means that the arrival rates of customers (i.e., number of calls per unit time), must be known with a reasonable accuracy, where the forecasting and staffing decisions are being performed in advance. See §2 in [2], and §§3 and 6 in [26]. Since the call center operates in a random environment, with the arrival rates possibly larger than expected, at least during some time periods, it may become overloaded due to larger than expected arrival rates, so that the desired service levels cannot be met.

1.1.1 The Basic Research Problem: Overload Control

This dissertation considers how two networked large-scale service systems, such as call centers, that normally operate separately, can help each other in face of an unexpected overload. We assume that occasionally, for various reasons, there may be unforeseen surges in demand, going significantly beyond the usual stochastic fluctuations, and lasting for a significant period of time. A demand surge might occur because of a catastrophic event in emergency response, a system failure experienced by an alternative service provider, or an unanticipated intense television advertising campaign in retail. Such unexpected demand surges typically cause congestion that cannot be eliminated entirely. Since the demand surge is sudden and unexpected, it may not be possible to immediately change the staffing level.

However, there may be an opportunity to alleviate the congestion caused by the overload by getting help from another service system, which ordinarily operates independently. (For example, with the reduction of telecommunication costs, it is more and more common to have networked call centers, often geographically dispersed, even on different continents.) Such sharing is typically possible among different hospitals in a metropolitan area. It is often desirable to operate these service systems separately, but their connection provides opportunities, in particular, to provide assistance under overloads.

An important consideration is that we typically do not want sharing under normal loads. One reason is that it is easier to manage the different facilities separately, e.g., by maintaining clear accountability. Another reason is that the agents in each service facility may be less effective and/or less efficient serving the customers from the other system, because each requires specialized skills not required for the other. We want to consider the case in which serving the other class is possible, but that there are penalties for doing so. We will assume that the service rates are slower for non-designated agents.

The proposed overload control applies directly to separate service systems run by a single organization, but could also be adopted by two different organizations by mutual agreement. Our analysis provides useful information about the likely consequences of any agreement, which should facilitate making the agreement. Current practice for call centers (that we are aware of) is limited to sharing within a single organization, and then only manually or on a regular basis under normal loading. Load-balancing schemes used in practice are described in §5.3 of [26].

Thus, our goal is to develop a control to automatically detects when an overload has occurred (in either system, or in both) and then, before the staffing levels can be changed, reduce the resulting congestion by activating appropriate sharing from agents in the other system. We also want to prevent undesired sharing under normal loads. By focusing on this overload problem, we aim to contribute new insight into the longstanding question about

the costs and benefits of resource pooling; see §4.2 of [2] and references cited therein. Here we focus on a situation where we want to turn on and off the pooling.

To gain insight, we introduce an idealized X model, i.e., a stochastic model with two customer classes, each with its own dedicated service pool, containing a large number of agents. See figure 1.1. The agents in both service pools are assumed to be crossed-trained, so that they are able to serve customers from the other class, even if somewhat inefficiently.

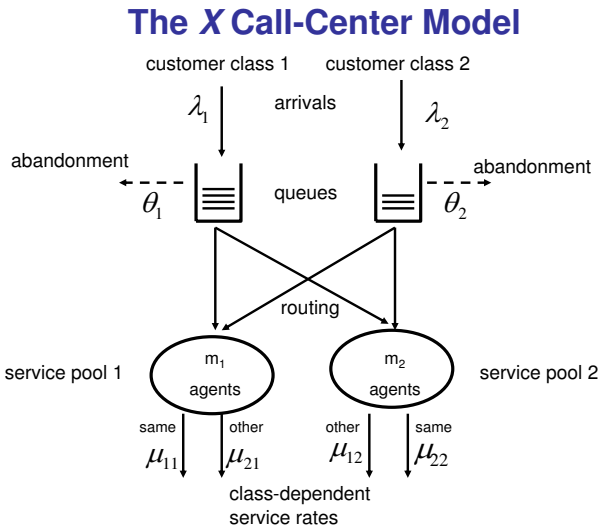


Figure 1.1: The X model

1.1.2 The Proposed Control: FQR-T

We now explain our proposed control, which we call *fixed-queue ratio with thresholds* (FQR-T). The purpose of FQR-T is to prevent sharing of customers as long as the two classes are not overloaded, and detect overloads quickly when they occur. These two objectives are achieved by placing two thresholds, $k_{1,2}$ and $k_{2,1}$, one for each queue. If queue i crosses its threshold $k_{i,j}$, $i, j = 1, 2$, then class i is considered to be overloaded.

In addition to the two positive thresholds $k_{1,2}$ and $k_{2,1}$, we introduce two ratio parameters $r_{1,2}$ and $r_{2,1}$. We then define two queue-difference stochastic processes

$$D_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t) \quad \text{and} \quad D_{2,1}(t) \equiv r_{2,1}Q_2(t) - Q_1(t), \tag{1.1.1}$$

where $Q_i(t)$ denotes the number of class-i customers waiting in queue at time t, $i = 1, 2$. As long as $D_{1,2}(t) < k_{1,2}$ and $D_{2,1}(t) < k_{2,1}$, we do not allow any sharing, i.e., we only let agents serve customers from their designated class. Thus, FQR-T is designed to permit sharing only in the presence of unbalanced overloads.

However, available pool-2 agents are assigned to class-1 customers when $D_{1,2}(t) \geq k_{1,2}$, provided that no pool-1 agents are still serving a class-2 customer. As soon as the first pool-2 agent is assigned to serve a class-1 customer, we drop the threshold $k_{1,2}$, but keep the other threshold $k_{2,1}$. (We could elect to add another threshold for the sharing; see §2.7.6.) Upon service completion, a newly available type-2 agent serves the customer at the head of the class-1 queue (the class-1 customer who has waited the longest) if $D_{1,2}(t) > 0$; otherwise the agent serves a customer from his own class. In this phase, pool-1 agents only serve class-1 customers. Only one-way sharing in this direction will be allowed until either the class-1 queue becomes empty or the other difference process crosses the other threshold, i.e., when $D_{2,1}(t) \geq k_{2,1}$. As soon as either of these events occurs, newly available pool-2 agents are only assigned to class 2 and the threshold $k_{1,2}$ is reinstated.

We can initiate sharing in the opposite direction when first $D_{2,1}(t) \geq k_{2,1}$ and there are no class-2 agents serving class-1 customers. At the first time both conditions are satisfied, we start sharing with a pool-2 agent serving a class-1 customer. When that first assignment takes place, we remove the threshold $k_{2,1}$ and again use the same procedure as before, but now with the ratio parameter $r_{2,1}$. In particular, a newly available type-1 agent serves the customer at the head of the class-2 queue if $D_{2,1}(t) > 0$; otherwise the agent serves a

customer from his own class.

Upon arrival, a class-i customer is routed to pool i if there are idle servers; otherwise the arrival goes to the end of the class-i queue. An arrival might increase the queue to a point that sharing is activated. Then the first customer in queue is served by the other class (presumably the agent that has been idle the longest, but we do not focus on individual agents).

The queue-difference stochastic processes in (1.1.1) will never provide any instantaneous motivation to have agents of both types simultaneously inefficiently serving the other class if $r_{1,2} \geq r_{2,1}$. That property will be satisfied when we apply a cost function to specify the ratio parameters in §2.3.2.

Our FQR-T control is appealing for several reasons. First, it is automatic and simple; we need not directly discover the arrival rates in order to find out when overloads occur, and then decide what amount of sharing should be done. Instead, FQR-T automatically detects the time the system becomes overloaded, and then automatically enforces the optimal ratio, by observing only the size of the two queues. It is easier to use the information about the queues, which is readily available, than to use information about the arrival rates, which is not readily available. Moreover, simulation experiments indicate that FQR-T performs better (produces lower expected costs) than fixing $Z_{i,j}$ at their optimal values, even with known arrival rates; see Figure 2.4.

1.2 Our Modeling Contribution

In this dissertation we contribute to the literature on overload (or congestion) control in queueing systems. There is a substantial literature studying controls that route (or assign) customers (or jobs) to servers, possibly exploiting thresholds, but many of these papers, like [11] and references therein, focus on single-server systems without customer abandonment,

whereas we focus on many-server systems with customer abandonment.

One feature of many-server systems with customer abandonment we will exploit is the rate at which the transient distribution approaches its steady-state limit: It tends to be much faster for many-server queues. In particular, the systems we consider tend to reach steady-state in a few mean service times; e.g., see (20) in [24] and (2.17) in [79]. (We will elaborate in §2.7.1.) Hence, in our analysis of performance during an overload incident, we approximate using the new steady state, determined by the new arrival rates (assumed constant). Customer abandonments ensure that the system remains stable.

We contribute to the call-routing problem for multi-class and multi-site call centers with skill-based routing; see §5 of [26] and §§2.3.3, 4.1, 4.2 of [2]. Others have proposed responding to stochastic fluctuations and unexpected overloads by modulating demand in different ways: (i) admission control, (ii) making delay announcements that may induce customers to leave, use a different service channel (e.g., email instead of voice), or call back later, and (iii) acting to reduce service times, e.g., by curtailing cross-selling activities; see §3 of [2] and [5].

In contrast, our work relates to the larger literature exploiting server flexibility (supply-side management). One approach is to have extra temporary servers available on short notice; see [12] and references therein. Instead, we propose using servers that are already working; i.e., we propose a form of resource pooling, which exploits cross training; see §4.2 of [2] and §5.1 of [26]. As should be anticipated, though, our control tends to be more effective in alleviating congestion (rather than just balancing the service degradation) when the less-loaded system actually has some slack. Our work draws on the queue-ratio control proposed in [29, 31], which applies to very general network topologies. Here we consider the relatively difficult X model, allowing sharing in both directions (as depicted in Figure 1.1), but our approach makes the model behave more like the N model, in which only one service pool can serve both classes (so that there is sharing in only one direction); see [69].

However, we make significant departures from the previous literature. First, we want resource sharing only in the presence of the unanticipated overload, and only in the proper direction, which depends on the nature of the overload. Hence, we turn on and off the sharing. Second, we regard the overload as a rare exceptional unanticipated event, rather than a stochastic fluctuation in demand. Thus, we think that it is inappropriate to perform a long-run steady-state analysis of system performance with alternating normal and overload periods (although that could be done). Instead, we focus on a single overload in isolation.

Since the system tends to be overloaded, even after sharing has been activated, system performance tends to be well approximated by deterministic fluid approximations, as in [79]. Our work also relates to the literature on arrival-rate uncertainty; see §4.4 of [26] and §2.4 of [2]. Arrival-rate uncertainty also tends to make deterministic fluid approximations remarkably accurate; e.g., see [10] and their previous papers with Harrison, and [82].

1.3 Mathematical Models

The most basic mathematical model of a call center is the $M/M/N$ (also known as the Erlang C) model. In this model there is one service pool having N agents, one class of customers and an infinite waiting room for customers in queue. The first 'M' stands for the assumed Markovian arrival process, i.e., a Poisson process, and the second 'M' stands for the Markovian service process, i.e., service times are assumed to be independent and identically distributed (i.i.d.) exponential random variables. Important extensions of the Erlang-C model are the $M/M/N/K$ model, having a finite buffer (waiting room) of size K, and the $M/M/N+M$ (Erlang-A), which incorporates customer abandonment. In this model, customers are assumed to be i.i.d. with exponential patience (the '+ M' stands for the Markovian abandonment process). In particular, the Erlang-A model assumes that each arriving customer has an exponential patience, and will abandon if he cannot enter service

before running out of patience. All the above models can be viewed as a variation of the $M/M/N/K+M$ model, having a single class of customers, single service pool with N agents, a buffer of size $K \leq \infty$ and customer abandonment (with the patience rate allowed to be infinite).

The $M/M/N/K+M$ model is attractive since it is relatively easy to analyze. Specifically, due to the Markovian assumptions, the queue-length process constitutes a birth-and-death (BD) process with state space $\{0,1,\ldots,N+K\}$, or $\mathbb{N} \equiv \{0,1,2,\ldots\}$ if $K=\infty$. Thus, closed-form expressions for several steady state quantities of interest are easy to derive and compute. (See §6.1 in [85] for exact analysis of the Erlang-A model.)

Why Heavy Traffic?

Even for the simple $M/M/N/K+M$ model a useful way to obtain insight is *heavy traffic* (HT) approximations. For example, the insight gained by considering the three regimes in §1.3.2 below is due to HT-limits considerations. Moreover, if the Markovian assumption is relaxed, exact analysis becomes much harder to carry out and often intractable. (However, results for the $M/M/N+G$ model, having a general abandonment distributions, are available. See §6 in [85] for a summary of these results.)

Even in fully Markovian models, exact analysis becomes too difficult to conduct once the dimension of the model increases. If we consider a system having more than one customer class and/or more than one service pool, then exact analysis becomes intractable and finding the optimal staffing and routing schemes becomes impractical. In these cases HT approximations become a valuable tool.

The X model considered in this thesis is a multidimensional generalization of the Erlang-A model. Although it is assumed to be Markovian, exact analysis becomes intractable when sharing is taking place under FQR-T. The transient analysis of our model proves to be hard, *even when we consider the deterministic HT fluid approximation* for the

stochastic systems. However, the stationary fluid-limit approximation for the X model under FQR-T is simple, and we use it in order to determine the control parameters and show that the control is optimal in the fluid limit.

1.3.1 The Conventional Heavy Traffic Regime

Heavy traffic limits were first proved by Kingman in [47, 48] as approximations for steady state distributions of a heavily loaded $G/G/1$ queue. The HT procedure was adapted to the multi-server settings, and extended to stochastic-process limits, by Iglehart and Whitt in [37, 38]. We refer to [78] for a literature review.

We now describe the standard HT limit for the $G/G/1$ queue. Let ρ denote the server utilization, i.e., the long-run proportion of time that the server is busy. If $\rho < 1$ but close to unity, then, although the system is stable, the queue length becomes arbitrarily large over large time intervals. Loosely speaking, under the right condition on ρ, the queue length process has fluctuations of order $1/(1-\rho)$ over time intervals of order $1/(1-\rho)^2$, when ρ is close to 1.

To make these statements rigorous, consider a sequence of $G/G/1$ systems indexed by $n \geq 1$, and let $\rho^n \equiv \lambda^n/\mu$, where λ^n denotes the arrival rate in system n and μ denotes the service rate, which is fixed for all systems. Let $Q^n(t)$ denote the queue-length process (number of customers in the system at time $t \geq 0$) in system n. Then, if

$$\sqrt{n}(1-\rho^n) \to \beta \in (-\infty, \infty) \quad \text{as } n \to \infty,$$

then

$$\frac{Q^n(nt)}{\sqrt{n}} \Rightarrow R(t) \quad \text{as } n \to \infty, \tag{1.3.1}$$

where $\Rightarrow$ stands for weak convergence (convergence in distribution, see [13], [78]), and

R is the well-studied reflected Brownian motion, which has an exponential steady state distribution. The convergence holds not only pointwise (at each time t); the full process $\{n^{-1/2}Q^n(nt) : t \geq 0\}$ converges in distribution to the process $\{R(t) : t \geq 0\}$ in an appropriate function space. A result of the type (1.3.1) is called a *functional central limit theorem* (FCLT) since the scaling is that of the central limit theorem, but the convergence takes place in a function space. It is a generalization of the basic FCLT in Donsker's theorem.

The limit in (1.3.1) is the basis of what is now known as "conventional HT". In the conventional HT, the number of servers in each station remains fixed, and the utilization of each station approaches one in an appropriate manner. Observe that, since the queue-length process becomes large while the service rate remains fixed, the waiting times of customers in queue also become very large. This phenomenon is true in general for systems in the conventional HT. In single-server systems (or when the number of servers in each station is fixed along the sequence) one must choose between having a highly utilized system with long waiting times, or a less utilized system with shorter waiting times in queues.

The conventional HT is inadequate for the study of large systems with many servers. As we mentioned above, call canters typically consist of service pools having a large number of agents. Also, as experience shows, customers in large call centers typically do not experience long waiting times, even when the utilization of the agents is close to 1. Thus, a different approach is needed in order to adequately analyze large systems.

1.3.2 The Many-Server HT Regime

As the name suggests, the many-server heavy-traffic (MS-HT) regime is concerned with large pools of servers. The first result is due to Iglehart in [36], who considered the $M/M/N$ (Erlang-C) model. In particular, to achieve a MS-HT limit, Iglehart considered a sequence of $M/M/N$ systems (it is natural to let the number of servers N be also the index

of the sequence), with a fixed service rate for all elements along the sequence, but with the arrival rates and number of servers growing to infinity. However, the number of servers N grows to infinity so fast, that the limit becomes equivalent to that of the Markovian infinite-server queue ($M/M/\infty$). In particular, a properly scaled sequence of stochastic processes, representing the number of customers in the system, converges weakly to an Ornstein-Uhlenbeck (OU) diffusion process.

Since the limit in this regime is indistinguishable from that of the $M/M/\infty$, if this regime is applied to the analysis of call centers, then we say that the system is operating under the *quality-driven* (QD) regime. The QD regime is usually not suitable for call-center analysis, as there are no customers waiting to be served, and no customer abandonment in the limit.

Note the difference between conventional HT and the infinite-server-type HT: In conventional HT, the probability that customers will be waiting in queue approaches 1 in the limit, while in the latter QD regime the probability of waiting approaches 0. Systems which are designed to operate such that the probability that a customer will wait to be served approaches one are said to operate in the *efficiency driven* (ED) regime, since, asymptotically, all agents are busy and all customers must wait to be served.

However, the QD and ED regimes are not the only MS-HT regimes; Consider a sequence of $G/M/N$ queues with arrival rate λ^N for the N^{th} element of the sequence, with $\lambda^N/N \to \lambda$ as $N \to \infty$, for some $\lambda > 0$. Also assume that the service rates satisfy $\mu^N \equiv \mu > 0$ for all $N \in \mathbb{N}$. For $t \geq 0$, let

$$\hat{Q}^N(t) \equiv \frac{Q^N(t) - N}{\sqrt{N}}, \quad N \in \mathbb{N}. \tag{1.3.2}$$

Observe that $\hat{Q}^N$ is centered about the number of servers N, so that $\hat{Q}^N(t) < 0$ indicates that there is idleness in the system, whereas $\hat{Q}^N(t) > 0$ indicates that there are customers

waiting to be served, at time t.

In [33], Halfin and Whitt observed that a non-degenerate diffusion limit can be obtained for the number-in-system process (1.3.2), with this limiting diffusion process fluctuating above and below zero. Specifically, assume that

$$\sqrt{N}(1-\rho^N) \to \beta > 0 \quad \text{as} \quad N \to \infty. \tag{1.3.3}$$

Condition (1.3.3) is equivalent to the *square-root safety staffing rule*, namely

$$\lambda^N/\mu = N - \beta\sqrt{N} + o(\sqrt{N}),$$

where $o(\sqrt{N})$ denotes any function $f : \mathbb{N} \to \mathbb{R}$, satisfying $f(N)/\sqrt{N} \to 0$ as $N \to \infty$. Then, under (1.3.3), if $\hat{Q}^N(0) \Rightarrow \hat{Q}(0)$, then

$$\hat{Q}^N \Rightarrow \hat{Q} \quad \text{as} \quad N \to \infty, \tag{1.3.4}$$

where $\hat{Q}$ is a well defined diffusion process, and the convergence takes place in an appropriate function space (i.e., the whole process $\hat{Q}^N$ converges to the diffusion process $\hat{Q}$).

In addition to the convergence result above, Halfin and Whitt [33] analyzed the steady state properties of the limiting diffusion process. They were able to show that the stationary distribution of the limiting diffusion is the limit of the scaled sequence of stationary distributions for the stochastic number-in-system processes. Thus, the steady-state probability of having to wait converges to a number strictly between zero and one, as opposed to the QD and ED regimes described above. The regime developed in [33] is now known the Halfin-Whitt regime, or the *quality and efficiency* (QED) regime, since it incorporates both the QD and the ED regimes. In the QED regime, high server utilization (the proportion of idle servers in the system is at most of order $1/\sqrt{N}$) is achieved together with short waiting

times, which are of order $1/\sqrt{N}$. As a consequence, one needs only $\beta\sqrt{N}$ "extra" service capacity in order to achieve any given level of service, for β in (1.3.3).

For service systems, modeling customer abandonment is important. For the the $M/M/N+M$ (Erlang-A) model, the same three different MS-HT limiting regimes, identified in [33], were shown to exist in [27] by the limit in (1.3.3). The regimes (i) ED, (ii) QED, and (iii) QD then occur, respectively, if the limit in (1.3.3) holds with (i) $\beta = -\infty$, (ii) $-\infty < \beta < \infty$, and (iii) $\beta = +\infty$. These three regimes also generalize to more complex queueing networks (and non-exponential service times).

Operating under the QED regime carries a risk; a well-operated call center is usually designed to have a utilization of $\rho \approx 0.95$. Hence, if the arrival rates are even slightly larger than expected, (or if the number of agents is smaller than planned) the system may encounter an unexpected overloaded. In such cases, the ED regime becomes the appropriate limiting approximation to consider, e.g., see [79], [82]. Moreover, systems are sometimes *designed* to operate under the ED regime, particularly if they are not revenue generating, in order to decrease the operating costs. Thus, in recent years there has been a growing interest in the HT ED limits. See, e.g., [27], [58], [79], [81] and [82]. The ED regime is often appropriate for service systems, and is nontrivial to analyze, since customers can abandon. Abandonment ensures that the system under consideration is stable.

Consider the $M/M/N+M$ model, and let $\rho^N \equiv \rho > 1$ for all $N \in \mathbb{N}$, so that the arrival rate to each system along the sequence is larger than its maximal service rate. In that case, the sequence of queue-length processes, centered about the number of servers (similar to the expression (1.3.2)) will diverge to infinity for each $t > 0$, since the queue length is order N larger than the number of agents, even though abandonment keeps each system along the sequence stable. If one is interested in obtaining FCLT, one should find a new argument to center about. It turns out the the centering argument in the ED regime is often of interest for its own right, and can be nontrivial to achieve. In order to find that

centering argument, we consider the limit of the sequence $\bar{Q}^N \equiv Q^N/N$. Let

$$\bar{Q} \equiv \lim_{N\to\infty} \bar{Q}^N. \tag{1.3.5}$$

(Assuming the limit in (1.3.5) exists.) Limits of the type (1.3.5) are called "fluid limits", since they tend to be continuous and deterministic processes. They are also called *functional laws of large numbers* (FLLN), since the scaling is that of the law of large numbers, and the limit describes the mean values of the random sequence. The fluid limit of the above Erlang-A model is relatively easy to establish; see [79]. In general, however, fluid limits can be hard to characterize. Moreover, they can prove to be a crucial step in the proof of the FCLT refinements, as in our case. See also [21].

1.3.3 Establishing HT Limits

There are several ways to prove that a sequence of stochastic processes converges in distribution. Here we briefly review the three most widely used methods in the HT literature. For background on the different methods see [13], [25] and [78].

We start with defining the space in which the stochastic processes under consideration exist; For a subinterval I of $[0,\infty)$ let $\mathcal{D}_k(I) \equiv \mathcal{D}([I,\mathbb{R}_k)$ be the space of all right-continuous $\mathbb{R}_k$-valued functions, with limits from the left everywhere, endowed with the Skorohod J_1 topology. Let $\mathcal{D}_k \equiv \mathcal{D}([0,\infty),\mathbb{R}_k)$, with $\mathcal{D} \equiv \mathcal{D}_1$. Let $\mathcal{C}_k(I) \subset \mathcal{D}_k(I)$ be the subspace of continuous functions in of $\mathcal{D}_k(I)$. Usually, the sequences of stochastic processes considered are random elements in $\mathcal{D}_k$, while the limits are typically in the subspace of continuous functions $\mathcal{C}_k$, in which case the J_1 topology coincides with the uniform topology.

The most widely used method in the HT literature uses the *continuous mapping theorem* (CMT). The CMT approach exploits established stochastic-process limits (usually

Donsker's theorem or a variant) to obtain new limits. In particular, if the sequence of processes considered can be represented as a continuous mapping from $\mathcal{D}_k$ to itself of processes whose limits are known, then the limit of the sequence can be characterized. The hard step is then showing that a given mapping is continuous. (Even addition is in general not a continuous function in $\mathcal{D}$). However, many useful functions, which are often needed in practice, were shown to be continuous in appropriate topologies. For many useful continuous functions (in different topologies) see §13 in [78].

The second method, which can be applied to Markov processes, is the operator semigroup approach. Convergence of the generators of Markov processes (in an appropriate sense; see [25]) implies the convergence of the corresponding semigroups, which in turn implies the convergence of the Markov processes. The general theory is hard to apply, and rarely used in the HT literature. (But see [71] for a queueing application). However, simplified versions of the theory were applied extensively. Specifically, if the elements in the sequence are *birth and death* (BD) processess, Stone's theorem [66] can often be applied. Stone's theorem reduces the problem of showing that the generators of the processes converge to showing that the sequence of infinitesimal means and variances converge. For queueing applications, see e.g., [27], [36], [79].

The third method is the compactness approach. Proving limit theorems in this method follows two steps: (i) showing that the sequence under consideration is pre-compact and (ii) uniquely characterizing the limit. In the function space $\mathcal{D}_k$ (endowed with the J_1 metric), pre-compactness is equivalent to sequential compactness, i.e., a sequence is pre-compact if each of its subsequences has a further converging subsequence, e.g., see [57]. The framework of weak convergence via the compactness approach was developed by Prohorov [60]. The direct half of Prohorov's theorem (Theorem 5.1 in [13]), applied for random elements of $\mathcal{C}_k$ and $\mathcal{D}_k$, essentially reduces to the Arzelà-ascoli characterization (and its variant to $\mathcal{D}_k$) of relative-compact sets in those function spaces.

We will use the compactness approach to prove the main result of this dissertation, namely the fluid limit of the overloaded X model under FQR-T. However, in our proofs, we will also make extensive use of the CMT. Moreover, the diffusion limits can be derived using the CMT, building on the fluid limit.

1.4 The Analytical Contribution

Chapters 3 and 4 are devoted to the mathematical analysis of the X model under FQR-T. In particular, Chapter 3 is dedicated to a dynamical-system-type study of an *ordinary differential equation* (ODE), which will be shown to arise as the fluid limit (FLLN) of the X model in Chapter 4 (with some proofs appearing in Chapter 5). Chapters 3 and 4 may seem out of order, because we establish properties of the limit before we prove the convergence to that limit. However, the order is appropriate because the properties of the limiting ODE and its solution play a key role in the proof of the limit theorems in Chapter 4.

In Chapter 3 we show that there exists a unique solution to the ODE over an interval $[0, \delta)$ for some $\delta > 0$. Conditions for extending this interval (typically all the way to infinity) are provided. We also prove that there exists a unique stationary point to the ODE. If the solution to the ODE exists on $[0, \infty)$, then it is shown that the solution must converge to its stationary point exponentially fast. Finally, we provide an efficient numerical algorithm, based on the matrix geometric method and the classical Euler forward algorithm, for solving the ODE.

In Chapter 4 we show that the sequence of overloaded X systems, operating under FQR-T, is pre-compact. We then show that the limit of every converging subsequence satisfies the three-dimensional ODE which was studied in Chapter 3. The uniqueness of the solution to the ODE over $[0, \delta)$ implies that the whole sequence of fluid-scaled processes converges

to the solution of the ODE. Proving convergence over the finite interval $[0, \delta)$ (no matter how small) is sufficient, since the convergence can be extended as long as the solution to the ODE is known to be unique.

An Averaging Principle

The main difficulty in establishing weak convergence via the compactness approach is usually in characterizing the limit. In our case, characterizing the limit is hard since FQR-T is driven by one of the queue-difference processes in (1.1.1), depending on which class receives help. When the sequence of fluid-scaled X models is considered, the queue-difference process is not being scaled and hence does not converge to a deterministic quantity due to the spatial scaling. However, this control-driving process operates in a different time scale than the fluid-scaled processes. In the limit, a complete separation of time scales is achieved, so that the queue-difference process converges to a (time-dependent) steady state at each instant of time. (Hence, it achieves a *long-run averaging* instantaneously, where the "long-run" is with respect to the fast time scale.) We refer to this fast long-run averaging phenomenon as an *averaging principle* (AP).

The AP of the queue-difference stochastic process also complicates the analysis of the limiting ODE. Since this process is not being scaled, it does not become deterministic in the limit. The ODE itself is deterministic only due to the AP. We call the stochastic process which drives the ODE the *fast-time-scale process* (FTSP), because at each time $t \geq 0$, the FTSP is replaced by its long-run average behavior. Now, since the FTSP determines the evolution of the ODE while, at the same time, the solution to the ODE determines the distribution of the FTSP, it may seem that the ODE cannot be fully analyzed. However, the separation of time scales allows for a complete analysis of the ODE, since the long-run behavior of the FTSP at each fixed time $t \geq 0$ is determined by the value of the solution to the ODE at the fixed time t.

The second complication is that the AP produces a singularity region in the state space, causing the ODE to be discontinuous in its full state space. Hence, both the convergence to the MS-HT fluid limit, and the analysis of the solution to the ODE depend heavily on the state space for the ODE, which is characterized in terms of the FTSP.

There are evidently only a few papers in the queueing literature involving averaging principles. Two notable papers are [19], which considers the diffusion limit of a polling system with zero switch-over times, and [35], which considers large loss networks under a large family of controls. Reference [35] is closely related to our work since it considers the fluid limits of such loss systems, with the control-driving process moving at a faster time scale than the other processes considered. However, the proof techniques here and in [35] are very different. In particular, the AP in [35] is proved via the martingale problem, building on [49].

We refer to [35] and [49] for a review of AP phenomenon in stochastic settings, and to [45] for AP-type arguments in deterministic dynamical systems. However, we note that our AP is very different than the settings considered in [45]. In particular, although our ODE is deterministic, the AP is stochastic in nature. In other words, our ODE *is driven by a stochastic process*. The ODE itself is deterministic since the stochastic process is, at each time $t \geq 0$, replaced by its long-run average behavior. This makes our ODE analysis an interesting combination of the dynamical-systems and probability theories.

State Space Collapse

As in many multi-class queueing networks, there is no underlying continuous mapping representing the queue-length process, e.g., see [16], [84]. However, often one can work with "cruder" processes, which do not include the exact interaction between the different processes considered. In that case, a CMT may be applicable for a lower-dimensional

("crude") process. That can be done by showing that the multidimensional process describing the system is experiencing a *state-space collapse* (SSC). That is, in the limit, the multidimensional processes exists in a lower-dimensional hyperplane of its space.

In our proof of the AP we also achieve a SSC result for the queue-length process. We show that the limit of the two-dimensional queue-length process exists in a one-dimensional hyperplane of $\mathcal{D}_2([0,\delta))$. Letting $\hat{Q}_i$ denote the diffusion limit of class-i queue, $i = 1, 2$, we have that

$$\hat{Q}_1(t) = r_{1,2}\hat{Q}_2(t) \quad t \in [0,\delta). \tag{1.4.1}$$

(Here $[0,\delta)$ denotes the maximal interval on which the fluid limit is known to exist. Typically we can take $\delta = \infty$.) Hence, we can analyze the sequence of the one-dimensional total queue-length processes $\hat{Q}^n_s \equiv \hat{Q}^n_1 + \hat{Q}^n_2$, Using (1.4.1), we deduce that

$$\hat{Q}_1 = \frac{r}{1+r}\hat{Q}_s, \quad \text{and} \quad \hat{Q}_2 = \frac{1}{1+r}\hat{Q}_s,$$

where $\hat{Q}_s$ is the diffusion limit of $\hat{Q}^n_s$. Of course, the same relation holds for the fluid-limit queues.

There is a large body of HT literature which includes SSC results, and we refer to [29] for references. We mention that a framework for proving SSC was developed by Bramson [16] in the conventional HT. His work was later extended by Dai and Tezcan [21] to the MS-HT QED regime. Gurvich and Whitt [29] proved that SSC holds for general network topologies operating under the queue-and-idleness ratio (QIR) family of controls. QIR aims to keep the queue length of each class and the proportion of idle servers at each pool at pre-specified ratios of the aggregated queue length and aggregated idleness in the system.

In closing we remark that we could not use Bramson's framework and its extensions to the MS-HT in [21] and [29]. First, these two references are concerned with the QED regime, whereas we are concerned with the ED regime. More importantly, the SSC result

is too crude for our needs. Knowing that SSC occurs in the HT limit, so that the queues exist in a one-dimensional hyperplane, is not sufficient for our purposes. To characterize the limit, we must also know the service process, i.e., how many customers from each class are being served in each service pool at every time $t \geq 0$. We thus need to consider the customer-server assignment process, which drives the control and depends on all processes. (In [29] this problem is avoided by assuming that the service rates are class or pool dependent. in cyclic networks, such as the X model. Thus the exact service process can be ignored in their settings.)

Chapter 2

Responding to Unexpected Overloads in Large-Scale Service Systems

In this chapter we elaborate on the X model and its main characteristics. We then derive a heuristic stationary fluid approximation (which will be justified rigorously in the next chapters) to analyze the system under the unexpected and unknown overloads.

Assuming that a convex holding cost is incurred on the two queues during overload incidents, we find the optimal server allocation in the heuristic stationary fluid approximation. We then propose the QR-T and FQR-T family of controls, which are argued to be superior to a fixed partition of the service pools when sharing is needed. Simulation experiments show that our control actually performs better than the fluid-optimal fixed server allocation, even if the arrival rates are known.

2.1 The Modelling Approach

The X model. As an idealized model of two separate service systems with the capability of sharing, we consider the X model, depicted in Figure 1.1 Chapter 1. The X model has

two homogeneous customer classes and two homogeneous agent pools. We assume that each customer class has a service pool primarily dedicated to it, but all agents are cross-trained, so that they can handle calls from the other class, even though they may do so inefficiently or ineffectively. Under normal loading (at or near forecasted arrival rates), we want each class to be served only by its designated agents, without any help from cross-trained agents in the other service pool. We assume that staffing has been performed in standard ways, so that the number of agents in each pool is adequate to meet performance targets at forecasted arrival rates. However, we also want to automatically activate sharing when there are unexpected unbalanced overloads, either when only one class is overloaded or when both classes are overloaded but one is much more overloaded than the other.

More specifically, we consider a fully Markovian model. Customers from the two classes arrive according to independent Poisson processes with arrival rates λ_1 and λ_2. There is a queue for each customer class, with customers from each class entering service in order of arrival. We assume that waiting customers have limited patience. A class-i customer will abandon if he does not start service before a random time that is exponentially distributed with mean $1/\theta_i$. There are two service pools, with pool j having m_j homogeneous servers working in parallel. The service times are mutually independent exponential random variables, but the mean may depend on both the customer class and the service pool. The mean service time for a class-i customer served by a type-j agent is $1/\mu_{i,j}$. Let the service times, abandonment times and arrival processes be mutually independent. Let $Q_i(t)$ be the number of class-i customers in queue and let $Z_{i,j}(t)$ be the number of type-j agents busy serving class-i customers, at time t. With the assumptions above, the stochastic process $(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2)$ becomes a six-dimensional continuous-time Markov chain, given any routing policy that depends on this six-dimensional state.

In this context, under normal loading we want each class served only by agents from its own designated service pool; i.e., we want $Z_{1,2}(t) \approx Z_{2,1}(t) \approx 0$ for all t. One possible

reason is that the value of service by agents from the other pool might be less, perhaps because they lack specialized skills. Another possible reason is that service by the cross-trained agents is less efficient; we might have the **strong inefficient-sharing condition**

$$\mu_{1,1} > \mu_{1,2} \quad \text{and} \quad \mu_{2,2} > \mu_{2,1}. \tag{2.1.1}$$

We examine the inefficient-sharing case. Throughout this chapter, we assume the **basic inefficient-sharing condition**

$$\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}. \tag{2.1.2}$$

Clearly, condition (2.1.1) implies condition (2.1.2). These conditions play a role in §2.3.2.

In this X-model setting with inefficient sharing, we suppose that an unexpected overload occurs at some unanticipated time that changes the arrival rates. We assume that we are unable to immediately change the staffing levels in response to that unexpected overload. However, we do have the option of allowing some of the cross-trained agents from the less-loaded service pool serve customers from the more overloaded customer class. In addition, we do not know the new arrival rates when the overload occurs. Thus we need to develop a control that depends on the system history; in some way we must discover that the arrival rates have indeed changed. That is challenging, because stochastic fluctuations under normal loading may make us think that the arrival rates have changed when in fact they have not. We illustrate with the following example.

Example 2.1.1. To illustrate, consider a symmetric model with forecasted arrival rates $\lambda_1 = \lambda_2 = 90$ per unit of time, where the mean service time for customers served by designated agents is $\mu_{1,1}^{-1} = \mu_{2,2}^{-1} = 1.0$, while the mean service time for customers served by agents from the other pool is $\mu_{1,2}^{-1} = \mu_{2,1}^{-1} = 1.25$. We measure time in units of mean service times by designated agents, which for discussion we take to be 5 minutes. Notice that condition (2.1.1) holds here: For all agents, the mean time required to serve the other class

is 25% greater than the mean time required to serve an agent's own class. Let customers abandon at rate $\theta_1 = \theta_2 = 0.4$.

Because serving the other class is less efficient, with these parameters it makes sense to operate the system as two separate systems. Following standard staffing methods for a single-class single-pool $M/M/m + M$ model, we may assign $m_1 = m_2 = 100$ agents to the two service pools. That makes the traffic intensities $\rho_1 \equiv \lambda_1/m_1\mu_{1,1} = \rho_2 = 0.90$, which we regard as normal loading. With this staffing, standard algorithms show that steady-state performance is quite good: 82% of the arrivals enter service immediately upon arrival without joining the queue, only 0.5% of the arrivals abandon, the average size of each queue is 1.1, and the expected conditional waiting time, given that the customer is served, is only 0.012 (about 3.6 seconds with a mean service time of 5 minutes).

Now suppose that, at some unanticipated time, the arrival rate for class 1 jumps to $\lambda_1 = 130$, while the arrival rate for class 2 remains at $\lambda_2 = 90$. If class 1 receives no help from pool 2, then class 1 experiences severe congestion. Assuming that the system reaches steady state after this shift in arrival rate (which does not take very long, approximately a few mean service times, as confirmed by simulations - see §2.7.1), almost all class-1 customers must wait before starting service, 23% of the class-1 customers abandon, the average size of the class-1 queue becomes 75, the expected conditional waiting time given that a class-1 customer is served is 0.65 (3.25 minutes).

If, as system managers, we were able to recognize that the class-1 arrival rate had shifted to 130, then we might elect to reassign some of the class-2 agents. For example, we might let 25 of the pool-2 agents be devoted to serving class 1. That increases the total service rate responding to the class-1 arrival rate of 130 from 100 to $100+(1/1.25)25 = 120$, while leaving a total service rate of $100 - 25 = 75$ to respond to the class-2 arrival rate of 90. Since sharing is inefficient, we must sacrifice 25 units of service rate for class 2 in order to gain 20 units of service rate for class 1.

Assuming that the two classes can be modelled as $M/M/m+M$ queues (which is only approximately correct for class 1 because its servers have become heterogenous), we can analyze the performance, e.g., by [80]. The pair of abandonment probabilities for the two classes changes from $(0.23, 0.005)$ to $(0.08, 0.17)$; the pair of mean queue lengths for the two classes changes from $(75, 1.1)$ to $(26, 38)$; and the pair of conditional expected waiting times given that the customer is served changes from $(0.65, 0.012)$ to $(0.205, 0.450)$ (1.03 minutes and 2.25 minutes, respectively). In this chapter we develop a control that responds in a similar way, but does so automatically without having to know that the arrival rates made that specific shift, and without making a fixed partition of the agents.

Analysis with a cost function. The advantage of such sharing, or any other control that produces similar sharing by the inefficient cross-trained agents, depends upon the cost of the congestion experienced. To assess that cost, we will assume that there is a cost function C, with $C(Q_1(t), Q_2(t))$ representing the expected cost rate incurred at time t if the vector of queue lengths at time t is $(Q_1(t), Q_2(t))$. If the overload incident takes place over the time interval $[a, b]$, then the expected total cost would be

$$C_T \equiv E\left[\int_a^b C(Q_1(t), Q_2(t))\, dt\right] = \int_a^b E[C(Q_1(t), Q_2(t))]\, dt. \tag{2.1.3}$$

We assume that the cost function C is convex and strictly increasing. The convexity explains why we might want to share when one class is much more overloaded than the other, no matter which class is overloaded.

In this context, our goal is to choose a routing policy, which may allow assignments to cross-trained agents, in order to achieve low (near-minimum) expected total cost for all possible overload incidents and resulting stochastic processes $(Q_1(t), Q_2(t))$, while producing only a negligible amount of sharing under normal loading. To define what we mean by an "overload incident," We can first specify an interval $[a, c]$ over which the arrival-rate

vector $(\lambda_1(t), \lambda_2(t))$ differs from the nominal vector. (We assume that the arrival process is a nonhomogeneous Poisson process with these new arrival rates.) However, we should also include an additional interval $[c, b]$ after time c to allow the vector queue-length $(Q_1(t), Q_2(t))$ to return to its nominal steady-state value. (Engineering judgement is required.) In our analysis, we simplify by restricting attention to scenarios, as in the example above, in which the pair of arrival rates (λ_1, λ_2) makes a sudden unexpected shift at some time, and remains at the new vector for a significant duration, so that the system reaches a new steady-state at the new arrival-rate vector. (Customer abandonment ensures that the system reaches steady state for any arrival-rate vector.) Our control applies more generally.

For such scenarios, we simplify by re-expressing our goal as minimizing the expected steady-state cost; i.e., we aim to minimize $C_T \equiv E[C(Q_1, Q_2)]$, where (Q_1, Q_2) is the vector of steady-state queue lengths associated with the new arrival-rate vector associated with the overload. We will use this steady-state overload framework to set the control parameters and demonstrate effectiveness, but the control applies to other overload scenarios. For this steady-state analysis to be effective, it is important that the system approaches the new steady state associated with the overload relatively quickly. As illustrated in the concrete example above, this tends to happen in a few mean service times. We discuss this important point further in §2.7.1.

In the context of Example 2.1.1, we might have a shift in arrival rates lasting five hours. It might not be possible to change the staffing in response, because it is in the middle of the same day. The initial transient period might last 3 mean service times or 15 minutes, which is 5% of the total overload incident. There might then be a recovery period lasting about 5 mean service times or 25 minutes, after which the system returns to steady state. For such overloads, the steady-state is evidently reasonable, and it is essential for tractability. Even with this simplifying approximation, the control problem for the stochastic system is very difficult. We will get an approximate solution only after exploiting a fluid approximation in

addition to this steady-state analysis; see §2.3.2. Even with that approximation, the analysis with a general increasing convex cost function gets complicated; see §2.3.2. However, as a byproduct, there is a very nice simple story (explicit formulas for everything), provided that we assume a separable quadratic power cost function; see Proposition 2.3.5.

2.2 The Proposed Control

We start by briefly reviewing the *fixed-queue-ratio* (FQR) *routing rule* from [29] and then we show that the FQR rule without thresholds can perform poorly with inefficient sharing, where the conditions in the theorems of [29] are violated. Then we introduce our proposed modification of FQR in order to treat unexpected overloads. It involves general queue-ratio functions, as in [31], and thresholds, one of each for each direction of sharing.

2.2.1 FQR and its Difficulties with Inefficient Sharing

With two queues, FQR can be implemented by considering a (weighted) *queue-difference stochastic process* $D(t) \equiv Q_1(t) - rQ_2(t)$, $t \geq 0$, where r is a single target-ratio parameter that management can set. With FQR for the X model, a newly available agent in either service pool serves the customer at the head of the class-1 (class-2) queue if $D(t) > 0$ ($D(t) < 0$), and serves the customer at the head of its own queue if $D(t) = 0$. The goal of FQR is to maintain a nearly constant queue ratio: $Q_1(t)/Q_2(t) \approx r$ throughout time. When $r = 1$, FQR coincides with serving the longer queue.

Under regularity conditions, the FQR control has two very desirable features for large-scale service systems, which makes it possible to reduce the multi-class multi-pool staffing-and-routing problem to the well-understood single-class single-pool staffing problem. First, if the required conditions are satisfied, then FQR tends to produce *state-space collapse* (SSC); i.e., for the X model, the two-dimensional queue-length vector $(Q_1(t), Q_2(t))$ tends

to evolve approximately as a one-dimensional process determined by the total queue length $Q_\Sigma(t) \equiv Q_1(t)+Q_2(t)$. In particular, $Q_i(t) \approx p_i Q_\Sigma(t)$ for $i = 1, 2$, where $p_1 = r/(1+r) = 1 - p_2$; e.g., see Figure 2.12 in §2.7.2. Moreover, it does so in a way such that all three stochastic processes - $Q_\Sigma(t)$, $Q_1(t)$ and $Q_2(t)$ - remain appropriately stable as $t \to \infty$. Indeed, [29] show that, under regularity conditions, FQR achieves SSC asymptotically in the quality-and-efficiency-driven (QED) many-server heavy-traffic limiting regime. Second, with FQR, it is possible to choose the ratio parameter r (or, equivalently, the queue proportions p_i) in order to determine the optimal level of staffing to achieve desired service-level differentiation; i.e., staffing costs are minimized subject to meeting class-dependent delay targets $P(W_i > T_i) = \alpha$; see 2.7.2 and [29]. [31] also showed how to staff to minimize convex costs under normal loading. In that case, the asymptotically optimal control in the QED regime is not FQR, but a state-dependent generalization: the *queue-and-idleness-ratio* (QIR) control. Our optimal queue ratios for the fluid model under overloading with convex costs are of the same state-dependent form.

However, in our setting, where service provided by non-designated agents is inefficient, neither FQR nor QIR, without the extra thresholds, is appropriate in normal loading, because they induce undesired sharing. Because of the inefficient sharing, the system is *not* work-conserving; sharing causes the required workload to increase. Indeed, the conditions in the key theorems of [29, 31] are violated. In fact, those conditions are actually needed to maintain stability. (However, for FQR without the thresholds, SSC is still achieved; the two queues explode together.)

Example 2.2.1. To illustrate, consider the X model with parameters $m_1 = m_2 = 100$, $\mu_{1,1} = \mu_{2,2} = 1.0$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\lambda_1 = \lambda_2 = 0.99$ and $\theta_1 = \theta_2 = 0.0$ (no abandonment). Since the traffic intensities are $\rho_i = \lambda_i/m_i\mu_{i,i} = 0.99$, the two separate systems without sharing are stable (with mean queue length 85 and mean waiting time 0.85). However, if we use FQR with $r = 1$, then inefficient sharing is generated, so that a significant

proportion of each agent pool is busy serving the other class. As a consequence, the arrival rate actually exceeds the service rate and the queue lengths diverge to infinity. Here, there still is SSC, but the two queue lengths diverge together.

This difficulty when FQR is applied inappropriately is illustrated by Figures 2.1 and 2.2. They show the sample paths of $Q_1(t)$ and $Z_{2,1}(t)$, starting empty, in one simulation run. After an initial transient period, the number of agents serving the other class fluctuates around $E[Z_{1,2}] = E[Z_{2,1}] \approx 39$, while the queue grows in an approximately linear rate; the simulation estimate is $E[Q_i(t)] \approx 6.8t$, $t \geq 0$. (These numerical values are estimated from multiple simulation runs. The confidence intervals are less than 1%.

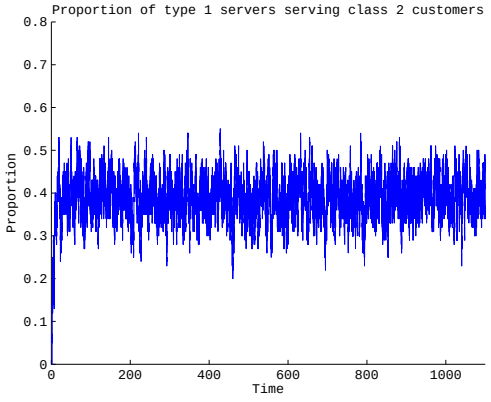


Figure 2.1: Sample path of $Z_{2,1}(t)$ for FQR

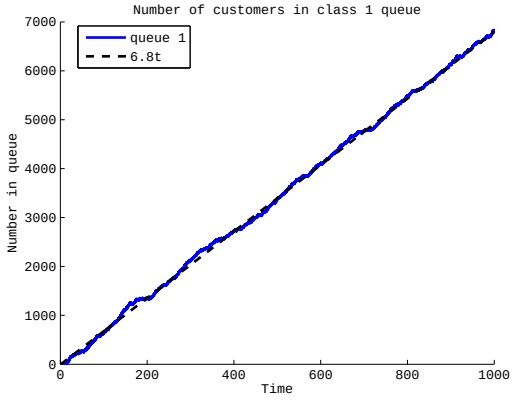


Figure 2.2: Sample path of $Q_1(t)$ for FQR

Customer abandonment necessarily prevents the queues from exploding. Even in the worst case, when all agents are dedicated to the wrong class, the system would be stable. However, there still is performance degradation, e.g., with $\theta_1 = \theta_2 = 0.2$ and $r = 1$ about 39% of the agents in each pool are busy serving customers from the other class which causes the queues to grow from 10, if there is no sharing, to 34. More details appear in §2.7.2.

2.2.2 The Proposed Control: FQR-T

Here is the lesson from the previous subsection: If we are going to use a queue-ratio control, then we need to take extra measures to prevent sharing under normal loading. First, we want to prevent simultaneous inefficient sharing in both directions. Hence, we restrict the routing to *one-way sharing* at any time: We do not allow a newly available type-2 agent to serve a waiting class-1 customer if there are any type-1 agents busy serving class-2 customers. And similarly in the other direction. (However, over time, the direction of one-way sharing may change; we are not considering the so-called N model, which only allows one-way sharing in one fixed direction.)

From cost considerations, discussed in §2.3, we want to allow different ratio parameters $r_{1,2}$ and $r_{2,1}$ for the different ways we may share. (In general, we may need more complicated ratio functions or, equivalently, sharing regions; see §2.3.2, especially Figure 2.3.) In order to permit sharing only in the presence of unbalanced overloads, we suggest *fixed-queue-ratio routing with thresholds* (FQR-T). In addition to the two ratio parameters $r_{1,2}$ and $r_{2,1}$, we introduce two positive thresholds $k_{1,2}$ and $k_{2,1}$. We then define two queue-difference stochastic processes

$$D_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t) \quad \text{and} \quad D_{2,1}(t) \equiv r_{2,1}Q_2(t) - Q_1(t). \tag{2.2.1}$$

As long as $D_{1,2}(t) < k_{1,2}$ and $D_{2,1}(t) < k_{2,1}$, we do not allow any sharing, i.e., we only let agents serve customers from their designated class.

However, available pool-2 agents are assigned to class-1 customers when $D_{1,2}(t) \geq k_{1,2}$, provided that no pool-1 agents are still serving a class-2 customer. As soon as the first pool-2 agent is assigned to serve a class-1 customer, we drop the threshold $k_{1,2}$, but keep the other threshold $k_{2,1}$. (We could elect to add another threshold for the sharing; see §2.7.6.) Once one-way sharing has been activated with pool 2 helping class 1, we use

ordinary FQR with ratio parameter $r_{1,2}$. Upon service completion, a newly available type-2 agent serves the customer at the head of the class-1 queue (the class-1 customer who has waited the longest) if $D_{1,2}(t) > 0$; otherwise the agent serves a customer from his own class. In this phase, pool-1 agents only serve class-1 customers. Only one-way sharing in this direction will be allowed until either the class-1 queue becomes empty or the other difference process crosses the other threshold, i.e., when $D_{2,1}(t) \geq k_{2,1}$. As soon as either of these events occurs, newly available pool-2 agents are only assigned to class 2 and the threshold $k_{1,2}$ is reinstated.

We can initiate sharing in the opposite direction when first $D_{2,1}(t) \geq k_{2,1}$ and there are no class-2 agents serving class-1 customers. At the first time both conditions are satisfied, we start sharing with a pool-2 agent serving a class-1 customer. When that first assignment takes place, we remove the threshold $k_{2,1}$ and again use FQR with one-way sharing, but now with the ratio parameter $r_{2,1}$.

Upon arrival, a class-i customer is routed to pool i if there are idle servers; otherwise the arrival goes to the end of the class-i queue. An arrival might increase the queue to a point that sharing is activated. Then the first customer in queue is served by the other class (presumably the agent that has been idle the longest, but we do not focus on individual agents).

The queue-difference stochastic processes in (2.2.1) will never provide any instantaneous motivation to have agents of both types simultaneously inefficiently serving the other class if $r_{1,2} \geq r_{2,1}$. That property will be satisfied when we apply a cost function to specify the ratio parameters in §2.3.2.

To illustrate how FQR-T performs in normal loading (heavy load, but not overloaded), we again consider Example 2.2.1 with abandonments at rate $\theta_i = 0.2$. We let $r_{1,2} = r_{2,1} = 1$, so that there is no change from FQR above, but now we add thresholds $k_{1,2} = k_{2,1} = 10$. The performance is greatly improved with FQR-T compared to FQR without

thresholds: $E[Z_{1,2}] = E[Z_{2,1}] \approx 2.0$ for FQR-T, while $E[Z_{1,2}] = E[Z_{2,1}] \approx 39$ for FQR. As a consequence, the performance for FQR-T is almost the same as without sharing. In particular, with FQR-T, the abandonment rate is slightly higher than without sharing (2.5% compared to 2.0%), but the average queue length is actually less $(9.4$ compared to $10.0)$. In fact, FQR-T can outperform no sharing with larger threshold values, because of the resource-pooling effect. For more details, see §2.7.2.

2.3 The Fluid Approximation for the Steady State

In order to obtain a tractable characterization of performance for FQR-T and find good queue-ratio parameters, we now introduce a deterministic fluid approximation. To describe the steady-state behavior of our model when there is no sharing, we first discuss the case of a single customer class served by a single service pool - the classical $M/M/m+M$ model, with arrival rate λ, individual service rate μ and abandonment rate θ. Afterwards we treat the more general X model.

2.3.1 One Class and One Pool

For the $M/M/m + M$ model, the approximating deterministic fluid model has been studied in [79] via many-server heavy-traffic limits. Here we will derive the simple steady-state formulas directly. We assume that input and output (which we call fluid) occurs deterministically at the specified rates. We think of the system as large and thus regard the number of customers and servers as continuous quantities as well. Thus, fluid arrives deterministically and continuously at constant rate λ. Fluid also is served and abandons deterministically and continuously at rates that are directly proportional to the number of busy servers and the queue length, respectively. If the "number" of busy servers is x, then fluid is served at rate $x\mu$; if the queue length is q, then fluid abandons at rate $q\theta$.

We say that the system is overloaded if the input rate exceeds the maximum possible total service rate. Given m servers, each working at rate μ, the maximum possible total service rate is $m\mu$. Thus the system is overloaded if $\lambda > m\mu$, and not overloaded otherwise. If the system is overloaded, then in steady state all servers will be busy and there will be a queue of waiting fluid, with content q, which can be determined simply be equating the rate in to the rate out, including customer abandonment: *rate in* $\equiv \lambda = m\mu + q\theta \equiv$ *rate out*. As an immediate consequence, we get $q = (\lambda - m\mu)/\theta$. If the system is not overloaded, i.e., if $\lambda \leq m\mu$, then there will be no queue. Then we can describe the steady-state via the amount of spare service capacity (number of idle servers), s, which again can be determined by equating the rate in to the rate out: *rate in* $\equiv \lambda = (m-s)\mu \equiv$ *rate out*. As an immediate consequence, we get $s = m-(\lambda/\mu)$. Without directly specifying whether or not the system is overloaded, we can write

$$q = \frac{(\lambda - m\mu)^+}{\theta} \quad \text{and} \quad s = \left(m - \frac{\lambda}{\mu}\right)^+, \tag{2.3.1}$$

where $(x)^+ \equiv \max\{x, 0\}$. We always have the *complementarity relation* $qs = 0$.

From the point of view of our analysis, we regard λ as an unknown parameter, but we consider the remaining parameters m, μ and θ as fixed and known. For any given λ, we can compute q and s as indicated above. With our overload control problem in mind, it is significant that we can recover λ from the pair (q, s), because we want to learn about λ by observing (q, s). If $q > 0$ and $s = 0$, then necessarily we are *overloaded*, and $\lambda = \theta q + m\mu$; if $q = 0$ and $s > 0$, then necessarily we are *underloaded* (which includes normally loaded), and $\lambda = (m - s)\mu$; if $q = 0$ and $s = 0$, then necessarily we are *critically loaded*, and $\lambda = m\mu$; we cannot have $q > 0$ and $s > 0$. For an overloaded fluid queue, λ is an increasing linear function of q; for an underloaded queue, λ is a decreasing linear function of s.

As discussed in [79], we can also describe the transient behavior of the fluid model and determine other performance measures. For example, if the fluid model is overloaded, then the associated approximate potential steady-state waiting time (virtual waiting time for a customer with infinite patience) is $w = \log(\lambda/m\mu)/\theta_1 = \log(\rho)/\theta_1$, where $\rho \equiv \lambda/m\mu$ is the traffic intensity, satisfying $\rho > 1$; see (2.26) of [79].

Note that an increasing convex function of w is an increasing convex function of λ for $\lambda \geq m\mu$. Since λ is a positive linear function of q under overloads, we see that an increasing convex function of w itself is a convex increasing function of q, as we have assumed in our optimization formulation. Similarly, the abandonment rate in the overloaded fluid model is $\theta q = \lambda - m\mu$, so the abandonment rate is an increasing linear function of q under overloads.

2.3.2 The Optimal Solution for the X Fluid Model

The X fluid model is a natural generalization of the single-class single-pool fluid model above. Now we have two deterministic arrival rates λ_1 and λ_2, one for each class, with the additional parameters $\{m_j, \theta_i, \mu_{i,j}; i = 1, 2; j = 1, 2\}$. Closely paralleling the discussion above, we will be characterizing the steady-state performance in terms of the quantities (Q_1, Q_2, S_1, S_2), where Q_i is the fluid content at the class-i queue, while S_j is the amount of spare capacity at pool j.

The steady-state behavior of the X fluid model depends on the number of agents from each pool assigned to (and actually busy serving customers from) each customer class, i.e., the deterministic vector $(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2})$, where $Z_{i,j}$ is the number of pool-j agents assigned to serve class-i customers, which is regarded as a continuous variable. To be legitimate assignments, we must have $Z_{i,j} \geq 0$ for all i and j with $Z_{1,1} + Z_{2,1} \leq m_1$ and $Z_{1,2} + Z_{2,2} \leq m_2$. Since these agents are actually busy serving customers, we must also have $\lambda_1 \geq Z_{1,1}\mu_{1,1} + Z_{1,2}\mu_{1,2}$ and $\lambda_2 \geq Z_{2,1}\mu_{2,1} + Z_{2,2}\mu_{2,2}$. Once we assign values to these variables $Z_{i,j}$, we reduce the X model to two single-class single-pool models. The

arrival rate for class i is λ_i, while the service rate for class i is $Z_{i,1}\mu_{i,1} + Z_{i,2}\mu_{i,2}$. Class i is then overloaded if and only if $\lambda_i > Z_{i,1}\mu_{i,1} + Z_{i,2}\mu_{i,2}$, in which case the steady-state fluid content in the class-i is

$$Q_i = \frac{\lambda_i - Z_{i,1}\mu_{i,1} - Z_{i,2}\mu_{i,2}}{\theta_i}. \tag{2.3.2}$$

If class i is not overloaded, then $Q_i = 0$. The spare capacity in pool j in steady state is $S_j = m_j - Z_{1,j} - Z_{2,j} \geq 0$, $j = 1, 2$.

In this X fluid model setting, for known arrival rates, our initial goal is to determine the minimum cost $C^*(\lambda_1, \lambda_2)$, which is the minimum of $C(Q_1(Z_{1,1}, Z_{1,2})), Q_2(Z_{2,1}, Z_{2,2}))$ for specified arrival-rate vector (λ_1, λ_2), which we denote simply by $C(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2})$, over all feasible fixed assignment vectors $(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2})$ in $\mathbb{R}^4$ with $Q_i \equiv Q_i(Z_{i,1}, Z_{i,2})$ defined in (2.3.2). We let the asterisk denote the optimal solution. (We do not consider more general controls.) We will apply the optimal solution to find the optimal state-dependent queue-ratio functions.

Let q_i be the queue length of class i and let s_i be the spare capacity in pool i when there is no sharing. They can be expressed as in (2.3.1), with formulas depending on i. In the fluid model, we regard the system as being in normal loading if neither queue is overloaded without sharing, i.e., if $q_1 = q_2 = 0$, but the amount of spare capacity is not too large. Since the cost function is increasing and convex, under normal loading we achieve the minimum cost by letting $Z_{1,2} = Z_{2,1} = 0$ (no sharing) to obtain $Q_i = 0$ for $i = 1, 2$. The unexpected overload means that either $q_1 > 0$ or $q_2 > 0$, or both. Henceforth we assume that to be the case.

The natural model state is (λ_1, λ_2), but an equivalent representation is (q_1, s_1, q_2, s_2), where we always have the complementarity relation $q_1 s_1 = q_2 s_2 = 0$. If $q_i > 0$, then $\lambda_i = m_i\mu_{i,i} + q_i\theta_i$; if $s_i > 0$, then $\lambda_i = (m_i - s_i)\mu_{i,i}$. This alternative representation implies that, for the X fluid model, we can determine the arrival rates by observing the

queue lengths and spare capacities.

Let $Z^*_{i,j}$ be the optimal value of the variable $Z_{i,j}$. We start by stating some basic propositions, which serve to simplify our X-fluid-model optimization problem. We first reduce the number of variables from four to two. The following is immediate.

Proposition 2.3.1. (no idle agents) *If we do not have $Q^*_1 = Q^*_2 = 0$, then there should be no idle agents, i.e., $S^*_j = 0$ or, equivalently, $Z^*_{1,j} + Z^*_{2,j} = m_j$ for $j = 1, 2$.*

As a consequence of Proposition 2.3.1, if $q_1 > 0$, $q_2 = 0$ and $s_2 > 0$, then necessarily $Z^*_{1,2} > 0$. Moreover, either $Z^*_{1,2} \geq s_2$ or $Q^*_1 = Q^*_2 = 0$.

We next show that inefficient sharing implies no two-way sharing.

Proposition 2.3.2. (one-way sharing) *Since the service rates satisfy the inefficient-sharing condition $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$ in* (2.1.2)*, it suffices to consider one-way sharing; i.e., $Z^*_{1,2}Z^*_{2,1} = 0$.*

Proof: Suppose that $Z_{1,2} > 0$ and $Z_{2,1} > 0$, so that we have sharing in both directions. It suffices to assume that $Q_1 > 0$ and $Q_2 > 0$. We will show that, for appropriate positive variables $x_{1,2}$ and $x_{2,1}$, if we replace $(Z_{1,2}, Z_{2,1})$ by $(Z_{1,2} - x_{1,2}, Z_{2,1} - x_{2,1})$, then both queue lengths will decrease until one of the variables $Z_{1,2} - x_{1,2}$ or $Z_{2,1} - x_{2,1}$ becomes 0 or both queues become empty. We define $x_{2,1}$ as an appropriate constant multiple of $x_{1,2}$, so that we have a single real variable. To do so, let $\gamma_i \equiv \lambda_i - Z_{i,1}\mu_{i,1} - Z_{i,2}\mu_{i,2} > 0$ for $i = 1, 2$. Then let $x_{2,1} \equiv \beta x_{1,2}$, where $\beta \equiv (\gamma_2\mu_{1,2} + \gamma_1\mu_{2,2})/(\gamma_2\mu_{1,1} + \gamma_1\mu_{2,1})$. Then we consider what happens as we increase $x_{1,2}$, assuming that β remains constant. Let $\Delta_i \equiv \theta_i(Q_i(0) - Q_i(x_{1,2}))$, where $Q_i(x_{1,2})$ denotes Q_i with the initial vector of sharing levels $(Z_{1,2}, Z_{2,1})$ replaced by $(Z_{1,2} - x_{1,2}, Z_{2,1} - \beta x_{1,2})$. Then

$$\begin{aligned} \Delta_1 &= x_{1,2}\gamma_1 \left(\frac{\mu_{1,1}\mu_{2,2} - \mu_{1,2}\mu_{2,1}}{\gamma_2\mu_{1,1} + \gamma_1\mu_{2,1}} \right) \\ \Delta_2 &= x_{1,2}\gamma_2 \left(\frac{\mu_{1,1}\mu_{2,2} - \mu_{1,2}\mu_{2,1}}{\gamma_2\mu_{1,1} + \gamma_1\mu_{2,1}} \right) \end{aligned} \tag{2.3.3}$$

Clearly, $\Delta_i \geq 0$ for both i if and only if inequality (2.1.2) holds. Moreover, from (2.3.2) and (2.3.3), we see that both queues become empty at the same level of $x_{1,2}$. Hence, we can decrease both variables $Z_{1,2}$ and $Z_{2,1}$ by increasing $x_{1,2}$ until one of these variables becomes 0 or both queue lengths simultaneously become 0. ∎

As a consequence of Proposition 2.3.2, we can re-express the basic optimization problem, first, in terms of two convex real-valued functions of a single real variable, $C_{1,2}$ and $C_{2,1}$, and second, in terms of a single combined convex function of a single real variable, C_c. Let 1_A be the indicator function of the set A; i.e., $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise.

Proposition 2.3.3. (single-variable functions) *Since the inefficient-sharing condition* (2.1.2) *holds, the optimal cost can be expressed as*

$$\begin{aligned} C^*(\lambda_1, \lambda_2) = C^*(q_1, s_1, q_2, s_2)) &= \min\{C_{1,2}(Z_{1,2}), C_{2,1}(Z_{2,1})\} \\ &= \min\{C_c(Z_{1,2} - Z_{2,1})\} \end{aligned} \tag{2.3.4}$$

over $Z_{1,2}$ *and* $Z_{2,1}$ *such that* $0 \leq Z_{1,2} \leq m_2$, $0 \leq Z_{2,1} \leq m_1$ *and* $Z_{1,2}Z_{2,1} = 0$, *where*

$$\begin{aligned} C_{1,2}(Z_{1,2}) &\equiv C_{1,2}(Z_{1,2}; \lambda_1, \lambda_2) \\ &\equiv C\left(\frac{(\lambda_1 - m_1\mu_{1,1} - Z_{1,2}\mu_{1,2})^+}{\theta_1}, \frac{(\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2})^+}{\theta_2}\right) \\ &\equiv C_{1,2}(Z_{1,2}; q_1, s_1 = 0, q_2, s_2) \\ &\equiv C\left(\frac{(q_1 - \mu_{1,2}Z_{1,2})^+}{\theta_1}, \frac{(q_2 - s_2\mu_{2,2} + \mu_{2,2}Z_{1,2})^+}{\theta_2}\right), \end{aligned} \tag{2.3.5}$$

$$\begin{aligned} C_c(Z_{1,2} - Z_{2,1}) &\equiv C_{1,2}(Z_{1,2} - Z_{2,1})1_{\{Z_{1,2} - Z_{2,1} \geq 0\}} + C_{2,1}(-(Z_{1,2} - Z_{2,1}))1_{\{Z_{1,2} - Z_{2,1} < 0\}} \\ &= C_{1,2}(Z_{1,2})1_{\{Z_{1,2} > 0\}} + C_{2,1}(Z_{2,1})1_{\{Z_{2,1} > 0\}} + C(q_1, q_2)1_{\{Z_{1,2} = Z_{2,1} = 0\}}, \end{aligned}$$

with q_i and s_i defined in (2.3.1)*, satisfying $q_1 s_2 = q_2 s_2 = 0$, and $C_{2,1}(Z_{2,1})$ defined analogously to $C_{1,2}(Z_{1,2})$ in* (2.3.5)*. The functions $C_{1,2}$ and $C_{2,1}$ are continuous strictly convex functions of the single real variables $Z_{1,2}$ and $Z_{2,1}$ over their domain. If, in addition, the stronger inefficient-sharing condition $\mu_{1,1} > \mu_{1,2}$ and $\mu_{2,2} > \mu_{2,1}$ in* (2.1.1) *holds, then C_c is also a continuous strictly convex function of the single real variable $Z_{1,2} - Z_{2,1}$ over the domain specified in Proposition 2.3.3.*

Proof: The representation is an immediate consequence of Proposition 2.3.2. Since $C_{1,2}(Z_{1,2})$ is the composition of a strictly convex function and a linear function, it is a strictly convex function of $Z_{1,2}$; e.g., p. 38 of [62]; similarly for $C_{2,1}(Z_{2,1})$. To establish the convexity of C_c, first assume that C is differentiable. It suffices to show that the derivative of C_c with respect to $Z_{1,2} - Z_{2,1}$, denoted by C_c', is nondecreasing. Existence and monotonicity of the derivative C_c' away from the boundary point $Z_{1,2} - Z_{2,1} = 0$ follows from the differentiability and convexity of $C_{1,2}$ and $C_{2,1}$, assuming that C is differentiable and convex. However, even if C is differentiable, the derivative of C_c need not exist at $Z_{1,2} - Z_{2,1} = 0$. It suffices to show that the left derivative is less than the right derivative at this point. The right derivative of C_c at 0, denoted by $C_c'^{+}(0)$, coincides with the derivative $C_{1,2}'(0)$, while the left derivative of C_c at 0, denoted by $C_c'^{-}(0)$, coincides with $-C_{2,1}'(0)$. Let C_i' denote the partial derivative of C with respect to its i^{th} coordinate at the argument $(q_1 - (s_1\mu_{1,1}/\theta_1), q_2 - (s_2\mu_{2,2}/\theta_2))$, which is positive because C is increasing. Then observe that

$$C_{1,2}'(0) \;=\; -C_1'\left(\frac{\mu_{1,2}}{\theta_1}\right) + C_2'\left(\frac{\mu_{2,2}}{\theta_2}\right) \quad \text{and} \quad -C_{2,1}'(0) = -C_1'\left(\frac{\mu_{1,1}}{\theta_1}\right) + C_2'\left(\frac{\mu_{2,1}}{\theta_2}\right)$$

Hence, $C_{1,2}'(0) \geq -C_{2,1}'(0)$, so that $C_c'^{+}(0) > C_c'^{-}(0)$ if the two inequalities in (2.1.1) hold. These relations can be extended to non-differentiable functions C by working with left and right derivatives. ∎

Corollary 2.3.1. (three intervals) *If the stronger inefficient-sharing condition* (2.1.1) *holds,*

then for each pair of arrival rates (λ_1, λ_2) *or initial state* (q_1, s_1, q_2, s_2) (*without sharing*), *there are two thresholds* $\zeta_{1,2} \geq \zeta_{2,1}$ *such that exactly one of the following occurs:*

$$\begin{aligned}
(i) \quad & Z^*_{1,2} > 0 \quad \textit{and} \quad Z^*_{2,1} = 0 \quad \textit{for} \quad Z_{1,2} - Z_{2,1} > \zeta_{1,2}, \\
(ii) \quad & Z^*_{2,1} > 0 \quad \textit{and} \quad Z^*_{1,2} = 0 \quad \textit{for} \quad Z_{1,2} - Z_{2,1} < \zeta_{2,1} \\
(iii) \quad & Z^*_{1,2} = Z^*_{2,1} = 0 \quad \textit{for} \quad \zeta_{2,1} \leq Z_{1,2} - Z_{2,1} \leq \zeta_{1,2}.
\end{aligned} \tag{2.3.6}$$

The value of Corollary 2.3.1 will be clear when we turn our attention to the queue ratio r below. We can apply Proposition 2.3.2 to get further simplification if there is initially spare capacity. Then, from the beginning, we know that we can only have sharing with help provided by the pool with spare capacity; i.e., if $q_1 > 0 > s_2$, then $Z^*_{1,2} > 0$ and $Z^*_{2,1} = 0$, so that it suffices to minimize $C_{1,2}(Z_{1,2})$.

It is natural to have the cost function C be smooth, in which case the optimal solution can be found by simple calculus. Proposition 2.7.1 concludes that, if the optimal solution found by calculus falls outside the feasible set, then the actual optimum value is obtained at the nearest boundary point.

It is easy to see that there is a one-to-one correspondence between the queue ratio $r \equiv Q_1/Q_2$ and the real variable $Z_{1,2} - Z_{2,1}$ used to specify the optimization problem in Proposition 2.3.3. That implies that there is a one-to-one correspondence between the fixed-agent-allocation optimization problem (choosing $Z_{1,2}$ and $Z_{2,1}$) and the (fixed) queue-ratio control problem (choosing state-dependent queue-ratio functions $r_{1,2}$ and $r_{2,1}$) in the fluid-model context. We establish it formally in §2.7.4.

Finally, we provide a basis for an efficient algorithm to determine the equivalent optimal controls. To do so, we effectively reduce the dimension from two to one by observing that special weighted sums of the queue lengths (and corresponding weighted sums of the arrival rates) are independent of the agent-assignment variables $Z_{1,2}$ and $Z_{2,1}$. We only

state the result for $Z_{1,2}$; the corresponding result for $Z_{2,1}$ is stated in §2.7.5. The proof is verification by direct computation, so we omit it. For understanding, it may be helpful to refer to Figure 2.3 in the next subsection.

Proposition 2.3.4. (constant weighted queue lengths) *Let*

$$a_{1,2} \equiv \frac{\mu_{2,2}\theta_1}{\mu_{1,2}\theta_2} \quad \textit{and} \quad \tilde{a}_{1,2} \equiv \frac{\mu_{1,2}}{\mu_{2,2}}. \tag{2.3.7}$$

Consider any initial state (λ_1, λ_2)*, or equivalently* (q_1, s_1, q_2, s_2)*, with* $s_1 = 0$*. Then*

$$\begin{aligned} w_{1,2} &\equiv a_{1,2}\left(\frac{\lambda_1 - m_1\mu_{1,1}}{\theta_1}\right) + \left(\frac{\lambda_2 - m_2\mu_{2,2}}{\theta_2}\right) = a_{1,2}q_1 + q_2 - \frac{s_2\mu_{2,2}}{\theta_2} \\ &= a_{1,2}Q_1(Z_{1,2}) + Q_2(Z_{1,2}) - \frac{S_2(Z_{1,2})\mu_{2,2}}{\theta_2} \end{aligned} \tag{2.3.8}$$

for all $Z_{1,2}$ *with* $0 \le Z_{1,2} \le m_2$*.*

Proposition 2.3.4 implies that the locus of all nonnegative queue-length vectors $(Q_1, Q_2) \equiv (Q_1(Z_{1,2}), Q_2(Z_{1,2}))$ associated with initial state (λ_1, λ_2), or equivalently (q_1, s_1, q_2, s_2), with $s_1 = 0$, is on the line $\{(Q_1, Q_2) : a_{1,2}Q_1 + Q_2 = w_{1,2}\}$ in the nonnegative quadrant. Thus, for any nonnegative constant $w_{1,2}$, the optimal queue-length vector (Q_1^*, Q_2^*) and the optimal queue-ratio $r_{1,2}^* \equiv Q_1^*/Q_2^*$ restricted to one-way sharing $(Z_{2,1} = 0)$ are the same for all initial states (q_1, s_1, q_2, s_2) with $s_1 = 0$ satisfying (2.3.8) provided that $q_1 \ge Q_1^*$. In that case, $a_{1,2}Q_1^* + Q_2^* = w_{1,2}$. That same optimal queue-length vector and optimal queue ratio holds for all arrival pairs (λ_1, λ_2) where $s_1 = 0$, $Z_{2,1} = 0$ and

$$\lambda_1 + \tilde{a}_{1,2}\lambda_2 = \tilde{w}_{1,2} \equiv \frac{\theta_1\theta_2 w_{1,2} + a_{1,2}\theta_2 m_1\mu_{1,1} + \theta_1 m_2\mu_{2,2}}{a_{1,2}\theta_2}. \tag{2.3.9}$$

And similarly for sharing in the other direction; see §2.7.5.

2.3.3 Computing the Optimal Queue-Ratio Functions

We now demonstrate how to numerically find the optimal state-dependent queue ratios $r^*_{1,2}$ and $r^*_{2,1}$ as functions of the fluid state (Q_1, S_1, Q_2, S_2). With the thresholds, this gives us a state-dependent *queue-ratio control with thresholds* (QR-T). To illustrate, we consider a (nonseparable) quadratic cost function of the form

$$C(Q_1, Q_2) = 3Q_1^2 + 2Q_2^2 + Q_1Q_2 + 10Q_1 + 5Q_2. \tag{2.3.10}$$

For any vector of arrival rates (λ_1, λ_2) we can assign one, and only one, point in the (Q_1, Q_2) plane, which represents the queue lengths associated with these arrival rates, when there is no sharing. To represent spare capacity, we allow negative values; i.e., $-Q_i$ is shorthand for $-S_i\mu_{i,i}/\theta_i$. (We actually plot $(Q_1 - S_1\mu_{1,1}/\theta_1, Q_2 - S_2\mu_{2,2}/\theta_2)$ even though the axes are simply labelled Q_i.)

We apply Proposition 2.3.4 to find the optimal queue ratios. We first consider when pool 2 helps class 1. To treat that case, we let $\lambda_2 = m_2\mu_{2,2}$, so that class 2 has no queue before pool 2 helps class 1. We then assume that $\lambda_1 > m_1\mu_{1,1}$ so that class 1 is overloaded. We then choose a large set of positive weighted arrival sums $\{\tilde{w}^1_{1,2}, \ldots, \tilde{w}^n_{1,2}\}$ and find the optimal queue ratio for each. In the first step, we let $\lambda_1 \equiv \tilde{w}_{1,2} - \tilde{a}_{1,2}\lambda_2$, using (2.3.9). We then write (2.3.10) as a function of $Z_{1,2}$, take its derivative and find the optimal $Z^*_{1,2}$. Plugging $Z^*_{1,2}$ in the queue equations gives us the optimal queue lengths (for the specific arrival rates), and the optimal queue ratio $r^*_{1,2}$. We repeat this for every $\tilde{w}^i_{1,2}$ to get the curve $1/r^*_{1,2}$ depicted in Figure 2.3. To find the curve $1/r^*_{2,1}$ we go through essentially the same procedure for $Z^*_{2,1}$.

Figure 2.3 simultaneously depicts the three optimal sharing regions in the two-dimensional state space and the two curves of optimal queue ratios. It was generated using Matlab on a system with the following parameters: $m_1 = m_2 = 100$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$

and $\theta_1 = \theta_2 = 0.3$. In addition, Figure 2.3 shows how to find the optimal queue ratio for

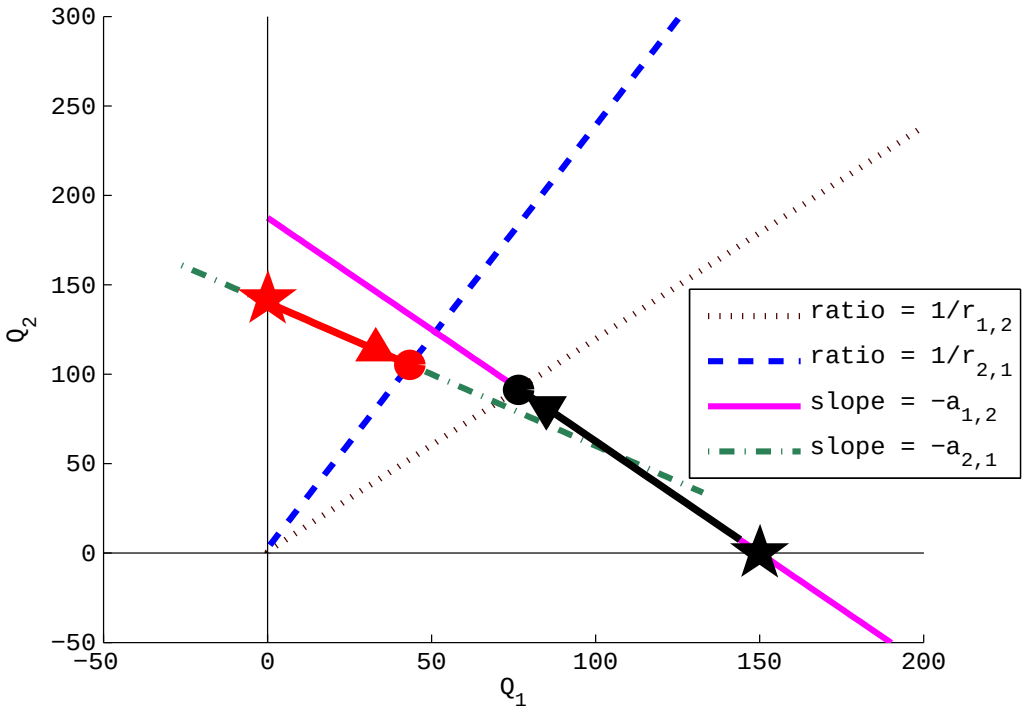


Figure 2.3: Curves of the optimal queue ratios for an X model

two possible initial queue lengths denoted by stars. When the initial queue-length vector is $(Q_1, Q_2) = (150, 0)$ (equivalently, $\lambda_1 = 145$ and $\lambda_2 = 100$), then the optimal queue-length vector is $(Q_1^*, Q_2^*) = (76.5, 91.8)$ and the optimal queue ratio is $r_{1,2}^* \equiv Q_1^*/Q_2^* = 0.83$. This optimal queue ratio is the intersection of the curve $1/r_{1,2}$ with the line with slope $-a_{1,2}$ that passes through $(150, 0)$ (the circle on the $1/r_{1,2}$ curve). When the initial queue-length vector is $(Q_1, Q_2) = (0, 150)$ (equivalently, $\lambda_1 = 100$ and $\lambda_2 = 145$), we get $r_{2,1}^* = 0.41$ and $(Q_1^*, Q_2^*) = (46.6, 112.8)$. The optimal queue ratio is also the intersection of the curve $1/r_{2,1}$ with the line with slope $-a_{2,1}$ that passes through $(0, 150)$ (the circle on the $1/r_{2,1}$ curve).

Both the $1/r_{1,2}^*$ and $1/r_{2,1}^*$ curves seem to be linear, although that is actually not quite the

case; the $r^*_{i,j}$'s are not constants for this cost function. For example, we already noted that, for $(\lambda_1, \lambda_2) = (145, 100)$ the optimal queue ratio is $r^*_{1,2} = 0.83$. If we change λ_1 to 110 then the optimal ratio becomes 0.80. For the other sharing direction, if $(\lambda_1, \lambda_2) = (100, 145)$, then $r^*_{2,1} = 0.41$, but if we change λ_2 to 110, then the optimal ratio changes to 0.38.

The fact the the two optimal-ratio curves are nearly linear in Figure 2.3 suggests that we can approximate the optimal queue-ratio function by fixed queue ratios, depending only on the direction of sharing; i.e., we can use FQR-T with only two values: one for $r_{1,2}$ and the other for $r_{2,1}$. In our example we may choose to use $r_{1,2} = 0.8$ and $r_{2,1} = 0.4$. The cost for using a nearly optimal ratio is very small in the fluid approximation, and even smaller in the stochastic system.

To understand when the optimal queue-ratio functions are nearly linear, as in the example above, and what the structure should be more generally, we investigate structured cost functions in §2.7.6. We obtain explicit analytical expressions in special cases. We focus on separable cost functions: $C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2)$, where each component cost function C_i is strictly convex, strictly increasing and twice differentiable. For example, we find that QR-T reduces to FQR-T exactly when $C_i(Q_i) = c_i Q_i^{n_i}$ with $n_1 = n_2$; the case $n_1 = n_2 = 2$ is close to (2.3.10).

Proposition 2.3.5. (explicit solution) *When $C(Q_1, Q_2) = c_1 Q_1^2 + c_2 Q_2^2$, FQR-T is optimal for the X fluid model with*

$$\begin{aligned} r^*_{1,2} &\equiv \frac{a_{1,2}c_2}{c_1} = \frac{c_2\mu_{2,2}\theta_1}{c_1\mu_{1,2}\theta_2}, \qquad r^*_{2,1} \equiv \frac{a_{2,1}c_2}{c_1} = \frac{c_2\mu_{2,1}\theta_1}{c_1\mu_{1,1}\theta_2}, \\ Z^*_{1,2} &= \frac{(c_1\mu_{1,2}\theta_1)(q_1 - (s_1\mu_{1,1}/\theta_1)) - (c_2\mu_{2,2}\theta_2)(q_2 - (s_2\mu_{2,2}/\theta_2))}{c_1\mu^2_{1,2}/\theta_1 + c_2\mu^2_{2,2}/\theta_2}, \\ Z^*_{2,1} &= \frac{c_2\mu_{2,1}\theta_2(q_2 - (s_2\mu_{2,2})/\theta_2) - c_1\mu_{1,1}\theta_1(q_1 - (s_1\mu_{1,1})/\theta_1)}{c_1\mu^2_{1,1} + c_2\mu^2_{2,1}}. \end{aligned} \tag{2.3.11}$$

In Proposition 2.3.5, the cost is specified by a single parameter: The ratio c_1/c_2 specifies the relative importance of the two queues. (The remaining parameter is equivalent to choosing the monetary units.) Finally, we caution that other cases (e.g., linear costs) can be quite different; see §2.7.6.

2.3.4 Application to the Stochastic Model

We can directly apply the QR-T control derived above to the stochastic model. Figure 2.3 identifies three sharing regions to apply to the stochastic process $(Q_1(t), S_1(t), Q_2(t), S_2(t))$ once sharing has been activated. There are two regions for each direction of sharing; e.g., if sharing has been activated with pool 2 helping class 1, available pool-2 agents serve class-1 customers when the queue-length vector falls in the lower right region, whereas there is no sharing in the other two regions. The way to share is described in §2.2.2.

2.4 Choosing the Thresholds

We now consider how to choose the thresholds $k_{1,2}$ and $k_{2,1}$. These thresholds have two important roles: First, they automatically detect when the system becomes overloaded and, second, they prevent unwanted sharing in normal loading. If the thresholds are too large, then the queues may not reach them during the overload. (Abandonments necessarily keep the queues from increasing without bound, even under overloads.) On the other hand, as discussed in §2.2.1, if the thresholds are too small, then sharing may be activated too often, so that we may get inefficient sharing.

Unfortunately, the fluid analysis cannot reveal the "right" size of the thresholds, since the fluid queues are empty under normal loading. We need to understand the extent of the stochastic fluctuations, something which is not captured by the fluid approximation. At

this point, it is convenient to apply many-server heavy-traffic limits to gain additional insight. To understand the general idea, it suffices to refer to established limits for the basic $M/M/n+M$ model, as in [27] and [79]. There, both fluid models and refined diffusion process models are obtained as limits as the scale increases, where scale is measured by the number n of servers. What is unusual here, though, is that we are simultaneously interested in the quality-and-efficiency-driven (QED) regime and the efficiency-driven (ED) or overloaded regime.

The QED regime is appropriate to describe normal loading, which is what prevails before the overload occurs, while the ED regime is appropriate to describe the overloaded system. In both cases, the arrival rate is allowed to grow as $n \to \infty$, while the service rate and abandonment rate are held fixed. The important insight is that the queue lengths tend to be of order $O(\sqrt{n})$ in the QED regime, as depicted by the diffusion limit in the QED regime, while the queue lengths tend to be of order $O(n)$ in the ED regime, as depicted by the fluid limit in the ED regime.

Thus, to prevent unwanted sharing when the system is normally loaded, we should choose the thresholds to be of size bigger than $O(\sqrt{n})$. That ensures that the weighted-queue-difference processes, $D_{1,2}$ and $D_{2,1}$, will not move above the thresholds by random fluctuations. On the other hand, we should choose the thresholds to be $o(n)$, so that the thresholds will be asymptotically negligible compared to the $O(n)$ fluid content. Then, asymptotically, they will be exceeded instantaneously when the overload occurs and they will not significantly alter the queue ratios. From this simple reasoning, we see that it suffices to have $\kappa_{i,j}^{(n)} = O(n^p)$ as $n \to \infty$ for $1/2 < p < 1$. (Incidentally, that scaling also makes the thresholds out of reach in normal loading in the law-of-the-iterated logarithm scaling of $(n \log\log n)^{1/2}$.)

This asymptotic analysis shows that the thresholds chosen in this way are asymptotically optimal, both during normal loading and during overload incidents. Asymptotically,

the thresholds will be exceeded negligibly often during normal loading; asymptotically, the thresholds do not alter the optimal average cost in the overload incidents. For the case of normal loading, we can apply the QED results in [27]; for the overload incidents, the ED results in [79] provide only heuristic support, because they apply only to the $M/M/n+M$ model. The ED fluid model for the X model with FQR-T is analyzed in the next two chapters.

Of course, we actually have a system with one fixed n. When we want to apply the theory to a real system, with a finite number of agents, it becomes hard to distinguish between $O(n)$ and $O(\sqrt{n})$. For example, if $n = 100$, then both $10 = 0.1n$ and $10 = \sqrt{n}$. Thus, as in any application of asymptotic results, we should numerically verify that the values chosen are appropriate, and refine them if necessary, for which we can use simulation. For example, for a system with 100 agents in each pool (and abandonment rates less than service rates), we found that $\kappa_{i,j} = 10$ is effective. We found that the performance is not too sensitive to the choice of the thresholds, provided that they are neither too small nor too large. We present simulation results for FQR-T under normal loading in §2.8.2, including a sensitivity analysis for the thresholds.

2.5 Simulation Experiments

Our analysis has been based on a fluid approximation of a stochastic system. It remains to show that the fluid approximation is suitably accurate for the stochastic system and that the optimal control for the fluid model works well in the stochastic system. For those purposes, we conduct simulation experiments.

2.5.1 Accuracy Of The Fluid Approximation

In this subsection we investigate the accuracy of the approximation. To show how the accuracy increases as the system becomes larger, we simulated three cases, each case represents an element in a sequence of queueing systems indexed by n, scaled to satisfy a many-server heavy-traffic limit in the ED regime as $n \to \infty$. We use the same fixed service and abandonment rates as before ($\mu_{i,i} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and $\theta_i = 0.3$). We consider a fixed queue ratio $r = 1$. We let the arrival rates be $\lambda_1 = 1.3n$ and $\lambda_2 = n$, when there are n agents in each service pool. The three cases we consider are $n = 25, 100, 400$. We let the thresholds for these three values of n be $k_{1,2} = k_{2,1} = 3, 10, 30$, respectively. (The thresholds were dropped when exceeded.)

Table 2.1 shows the results. Each result is the average of 5 independent simulation runs having $300,000$ arrivals in each run. The half-width of 95% confidence intervals, calculated using the t random variable with 4 degrees of freedom, are also given.

To show both the actual performance and the convergence to the fluid limit as n increases, we display both the direct values and the scaled values, dividing by n. Since the scaled values tend to be nearly independent of n, we witness the heavy-traffic fluid limit. We see that the approximations get better as n increases, but they are already not too bad when $n = 25$.

2.5.2 Comparing The Two Controls

In the fluid analysis, choosing the number of agents in each pool that are helping customers from the other class is equivalent to choosing the queue ratio, $r_{i,j}$. However, that is not the case in the actual stochastic system. With specified numbers of agents serving customers from the other class, the queue ratio fluctuates randomly. With specified queue ratios, the numbers of agents helping the other class fluctuates randomly. Moreover, with specified

	n=25		n=100		n=400	
perf. meas.	approx.	sim.	approx.	sim.	approx.	sim.
Q_1	13.9	13.5 ± 0.4	55.6	52.8 ± 1.2	222.2	216.7 ± 7.0
Q_1/n	0.56	0.54 ± 0.02	0.56	0.53 ± 0.01	0.56	0.54 ± 0.02
Q_2	13.9	15.7 ± 0.5	55.6	58.4 ± 1.2	222.2	223.1 ± 7.0
Q_2/n	0.56	0.63 ± 0.02	0.56	0.58 ± 0.01	0.56	0.56 ± 0.02
ratio	1.0	0.98 ± 0.02	1.0	0.90 0.00	1.0	0.96 0.00
$Z_{1,2}$	4.2	4.8 ± 0.2	16.7	17.7 ± 0.3	66.7	66.4 ± 2.2
$Z_{1,2}/n$	0.17	0.19 ± 0.01	0.17	0.18 ± 0.00	0.17	0.17 ± 0.01
Cost (thousands)	1.37	1.79 ± 0.01	19.35	20.28 ± 0.81	299.6	299.8 ± 19.2

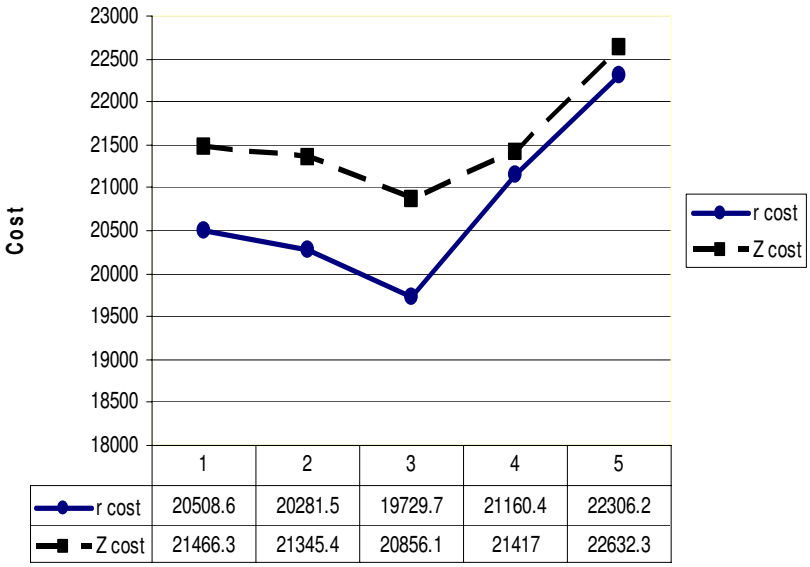
Table 2.1: A comparison of steady-state performance measures in the fluid approximation with corresponding simulation results for the Markovian X model. For each n, there are n agents in each pool, with $\lambda_1 = 1.3n$, $\lambda_2 = n$, $\mu_{i,i} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and $\theta_i = 0.3$. The thresholds $k_{1,2} = k_{2,1}$ are $3, 10, 30$ for $n = 25, 100, 400$, respectively.

numbers of agents serving customers from the other class, the two queue-length processes evolve independently. In sharp contrast, with specified queue ratios, the queue-length processes are strongly dependent, as in Figure 2.12. This suggests that there is a big difference between the two controls in a real, stochastic system. We thus expect the average cost under FQR-T to be different than the average cost when fixing $Z_{i,j}$. We conducted simulation experiments to compare the two controls.

To compare the two controls - FQR-T, and fixed $Z_{i,j}$ - we simulated a system with $m_i = 100$ agents in each pool, arrival rates $\lambda_1 = 130$ and $\lambda_2 = 100$, service rates $\mu_{1,1} = \mu_{2,2} = 1$ and $\mu_{1,2} = \mu_{2,1} = 0.8$, and abandonment rates $\theta_1 = \theta_2 = 0.3$. Since class 1 is overloaded,

we took $k_{2,1} = k_{1,2} = 10$, but once we go over the threshold $k_{1,2}$, we drop it, so that it becomes $k_{1,2} = 0$.

Figure 2.4 presents simulation results comparing the two average costs for five different cases: $(1)\, r_{1,2} = 1.2, Z_{1,2} = 15$, $(2)\, r_{1,2} = 1, Z_{1,2} = 17$, $(3)\, r_{1,2} = 0.83, Z_{1,2} = 19$ (optimal point), $(4)\, r_{1,2} = 0.6, Z_{1,2} = 22$ and $(5)\, r_{1,2} = 0.4, Z_{1,2} = 25$. For each point, we fixed the queue-ratio $r_{1,2}$, and used FQR-T with this ratio. For each such $r_{1,2}$, there is an equivalent $Z_{1,2}$ in the fluid equations. Since this $Z_{1,2}$ is not necessarily an integer, we rounded it up to the smallest integer larger than $Z_{1,2}$, i.e., we used $\lceil Z_{1,2} \rceil$ in each simulation of the fixed-$Z_{i,j}$ control. According to the fluid approximation, the optimal queue ratio is $r_{1,2} = 0.83$, and the respective optimal $Z_{1,2}$ is equal to 18.4, rounded up to 19 in the simulation experiments.



	1	2	3	4	5
r cost	20508.6	20281.5	19729.7	21160.4	22306.2
Z cost	21466.3	21345.4	20856.1	21417	22632.3

Figure 2.4: Cost of using FQR-T vs. fixed partition

For each case, we conducted 5 independent simulation runs using FQR-T, and 5 independent simulation runs with a fixed $Z_{1,2}$, each run with $300,000$ arrivals. The independent

replications make it possible to reliably estimate confidence intervals using the t statistic with 4 degrees of freedom. The large number of arrivals ensures that the transient behavior in the beginning of the simulation, before reaching steady state, does not affect the final simulation estimates of the steady-state averages. Additional simulation results are given in Table 2.2 in §2.8, including the half-width of 95% confidence intervals and a comparison of the simulation to the fluid approximation.

There are several interesting observations to be made: First, the r-cost curve lies significantly below the Z-cost curve, which shows that FQR-T is a superior control. At the optimal point for FQR-T, the average cost under FQR-T is about 5.4% smaller than the average cost under the fixed-$Z_{1,2}$ control.

Secondly, FQR-T tends to be a more robust control. Small changes in r do not produce large changes in the cost. Note that the largest r value here (1.2) is 3 times as large as the smallest r value (0.4), whereas the largest Z value here (25) is only 1.6 times as large as the smallest Z value (15). Moreover, the average costs when using FQR-T with $r_{1,2} = 1.2$ and $r_{1,2} = 1$ are still smaller than the cost of fixing $Z_{1,2}$ at its optimal value. For further discussion, see §2.8.

2.6 Conclusions

In this chapter we studied ways to respond to unexpected overloads in large-scale service systems. We considered the Markovian X model with two customer classes and two service pools, assuming that agents are more effective serving customers from their own class than customers from the other class, as specified by the inefficient-sharing conditions in (2.1.1) and (2.1.2). Thus we want negligible sharing under normal loads, but we want to activate sharing when there is an unexpected overload at an unanticipated time, without knowing what the new arrival rates will be.

The main ideas for analyzing the performance and determining appropriate queue-ratio functions for the *queue-ratio with thresholds* (QT-R) and *fixed-queue ratio with thresholds* (FQR-T) controls we propose are: (i) to use steady-state analysis and (ii) to apply an approximating deterministic fluid model. The QR-T and FQR-T controls proposed for the actual stochastic system are direct applications of the corresponding optimal controls derived for the fluid model in §2.3.2. We developed an algorithm to find the optimal queue-ratio curves for a general convex cost function in Proposition 2.3.4 and §2.3.3. The resulting QR-T control is easily understood as a partition of the state space into three sharing regions, as depicted in Figure 2.3, with two regions for each direction of sharing.

In Proposition 2.3.5 we also provided strong justification for FQR-T when the cost function has the form $C(Q_1, Q_2) = c_1 Q_1^2 + Q_2^2$ for some constants c_1 and c_2. In that case, we proved that FQR-T is optimal for the fluid model (i.e., the optimal QR-T control reduces to an instance of FQR-T) and exhibited the explicit optimal queue-ratio parameters. Then the optimal queue-ratio parameters depend on the cost function C only via the single parameter c_1/c_2, which succinctly captures the relative importance of the two queues. For other sharing regions, see §2.7.6.

The main ideas for gaining insight into appropriate threshold values were to apply (i) many-server heavy-traffic asymptotics and (ii) simulation. Heuristically, we showed that the thresholds should be asymptotically optimal simultaneously for periods of normal loading and for periods of overload. Asymptotically, no tradeoff need be made. The requirement is that the thresholds should be of order $O(n^p)$ as $n \to \infty$, where $1/2 < p < 1$ and n is the system scale factor. We used simulation to verify that the thresholds work well for given finite n.

Our FQR-T (or QR-T) control is appealing for several reasons. First, it is automatic and simple; we need not directly discover the arrival rates in order to find out when overloads

occur, and then decide what amount of sharing should be done. Instead, FQR-T automatically detects the time the system becomes overloaded, and then automatically enforces the optimal ratio, by observing only the size of the two queues. It is easier to use the information about the queues, which is readily available, than to use information about the arrival rates, which is not readily available. Moreover, simulation experiments indicate that FQR-T performs better (produces lower expected costs) than fixing $Z_{i,j}$ at their optimal values, even with known arrival rates; see Figure 2.4.

2.7 Supporting Material

In this section we present additional material supplementing the results in the main chapter. The topics are ordered as they arise in the chapter. In §2.7.1 we discuss the way the transient distribution approaches its steady-state limit, both at the beginning and the end of an overload incident. In §2.7.2 we provide additional discussion about the FQR and FQR-T controls, supplementing §2.2. In §2.7.3 we present additional details about the optimal solution for the deterministic fluid model during the overload, supplementing §2.3. Finally, in §2.8 we present additional simulation results about the performance of the control. In §2.8.1 we present a table of detailed simulation results supporting Figure 2.4. In §2.8.2 we present additional simulation results about the performance of FQR-T under normal loading. We perform a sensitivity analysis for the thresholds there.

2.7.1 Time To Reach Steady State

An important aspect of our QR-T and FQR-T controls is the transient behavior of the system. When the overload incident occurs, the system must shift from steady state under normal loading to steady state under the overload. Afterwards, at the end of the overload period, there is a recovery period, during which the system shifts back to the original steady state. From analysis and extensive simulations, we conclude that these two transient periods do not dominate, so that it is possible to use steady-state analysis as a reasonable approximation. In this section, we provide some supporting simulation results and discuss the supporting mathematical results.

Simulation Experiments We start by doing a simulation experiment of an overload incident, including all five regimes: (i) steady state before the overload, (ii) transition to new steady state at the beginning of the overload, (iii) new steady state under the overload, (iv) recovery period and (v) original steady state again after the overload.

Our example is based on Example 2.1.1 in the main chapter and the associated typical overload incident described at the end of §4.2. We assume that there is an overload incident that lasts 5 hours when the mean service times are 5 minutes. Given that we measure time in units of mean service times, the overload incident lasts 60 time units. Thus, we simulate the system over the time interval $[0, 150]$, and have the overload begin at time 80 and end at time 140. Thus, the initial transient begins at time 80, while the recovery period begins at time 140.

We consider a large system with $n = 400$ agents in each pool. For the normal loading, we let $\lambda_1 = \lambda_2 = 380$; for the overload during $[80, 140]$, we let $\lambda_1 = 520$, while $\lambda_1 = 380$ as before. As in Example 2.1.1, we let the mean service time for customers served by designated agents be $\mu_{1,1}^{-1} = \mu_{2,2}^{-1} = 1.0$, while the mean service time for customers served by agents from the other pool is $\mu_{1,2}^{-1} = \mu_{2,1}^{-1} = 1.25$. We let customers abandon at rate $\theta_1 = \theta_2 = 0.4$.

Since class 1 experiences the overload, we will have pool 2 helping class 1 during the overload incident. Typical sample paths of the processes $Z_{1,2}(t)$ and $Q_1(t)$ generated by simulation are shown in Figures 2.5 and 2.6 below. A dotted horizontal line depicts the steady-state fluid approximation during the overload. We do not show the other processes. From corresponding plots of $Q_2(t)$ and $Q_1(t)$, it is evident that they move together during the overload, reflecting state-space collapse, but they move independently during normal loading. From the displayed sample path, we see that the system indeed reaches a new steady state after a few mean service times, as claimed in the introduction.

To elaborate, we also show corresponding sample paths in Figures 2.7 and 2.8 with $n = 100$ agents in each pool. One important observation is that in both systems ($n = 100$ and $n = 400$) it takes less than 3 time units for the queues to hit their fluid value, denoted by the dotted horizontal lines. The recovery time, after the overload incident has ended, is also very short, and is about 2 time units for the queues in both systems.

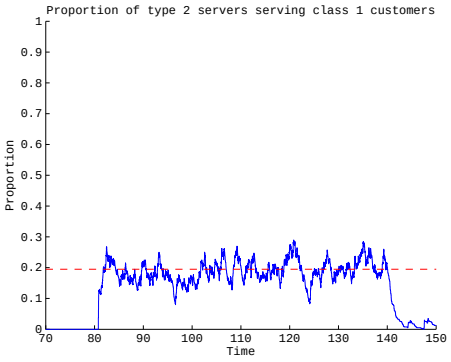


Figure 2.5: $Z_{1,2}(t)/400$ with overload over $[80, 140]$, $n = 400$.

Figure 2.6: $Q_1(t)$ with overload over $[80, 140]$, $n = 400$.

The story is different for the $Z_{1,2}(t)$ process. To make the connection between the two cases clear, we present the **proportion** of class-1 customers in pool 2 instead of the actual number, i.e., we show $Z_{1,2}(t)/n$ in Figures 2.5 and 2.7. First, when the overload begins at time 80, it takes some time until the queues hit the threshold $k_{1,2}$. That is the reason why $Z_{1,2}(t)$ starts growing a bit after time 80. It is interesting to see how our choice of the thresholds influences this delay. Recall that we choose the thresholds to be of order of size less than $O(n)$ but greater than $O(\sqrt{n})$; see §2.4 for more details. In these simulations, we took $\kappa_{i,j} = 20$ for $n = 400$ and $\kappa_{i,j} = 10$ for $n = 100$. This explains why in the $n = 400$ system it takes less time for $Z_{1,2}(t)$ to start increasing than in the $n = 100$ system: The thresholds are relatively smaller for the bigger system.

We also observe a difference between the two systems after the arrival rates return to normal at time 140. At this time, the $Z_{1,2}(t)$ processes start decreasing immediately and in a very fast rate. But now, service-pool 2 stops serving class-1 customers faster in the small system. Let $T_{1,2}$ be the time it takes for pool 2 to stop serving all class-1 customers after the end of the overload incident (after 140 in our example). As an approximation, we have

$$E[T_{1,2}] \approx \sum_{j=1}^{r} \frac{1}{j \cdot \mu_{1,2}},$$

where $r \equiv Z_{1,2}(140)$. Hence, the larger $Z_{1,2}(140)$ is, the longer it takes $Z_{1,2}(t)$ (or equivalently, $Z_{1,2}(t)/n$) to reach zero after the arrival rates shift back to normal. Yet, in both cases $Z_{1,2}(t)/n$ drops below 0.1 in about 2 time units, so that the total service rate in service-pool 2 is greater than λ_2 in 2 time units after the shift. In summary, we see that the transient period is relatively short, and a steady-state analysis is reasonable to apply.

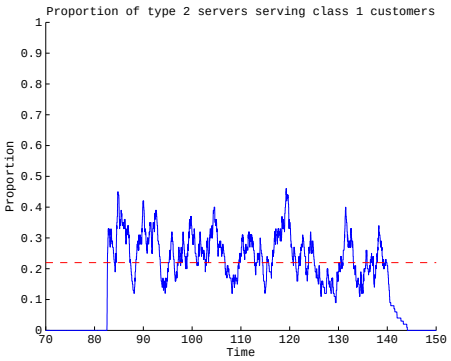


Figure 2.7: $Z_{1,2}(t)/100$ for FQR-T, with overload over $[80, 140]$, $n = 100$.

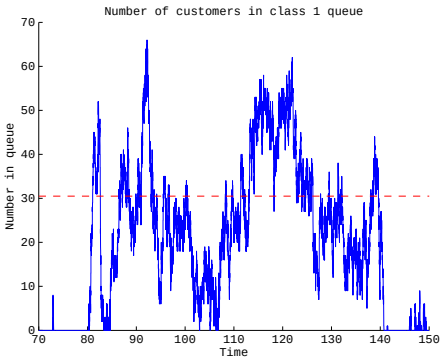


Figure 2.8: $Q_1(t)$ for FQR-T, with overload over $[80, 140]$, $n = 100$.

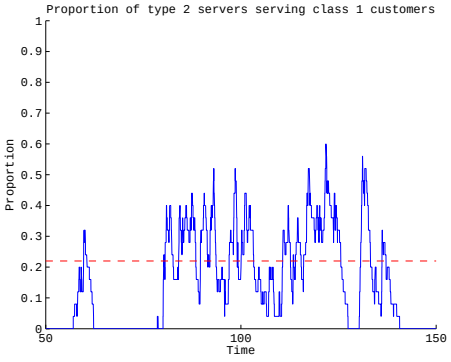


Figure 2.9: $Z_{1,2}(t)/25$ for FQR-T, with overload over $[80, 140]$, $n = 25$.

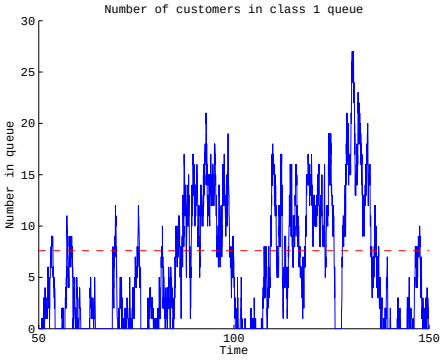


Figure 2.10: $Q_1(t)$ for FQR-T, with overload over $[80, 140]$, $n = 25$.

Mathematical Analysis We now provide further support. We first review mathematical analysis of the $M/M/n+M$ model; we next contrast with single-server models; afterwards we discuss implications for our X model

The M/M/n+M model. Consider the $M/M/n+M$ model with arrival rate λ, service rate μ and abandonment rate θ. First, it is useful to consider the special case in which $\theta = \mu$; then the number in system is distributed the same as in an $M/M/\infty$ system with service rate $\theta = \mu$. Thus the number in system at time t has a Poisson distribution for each fixed initial state. An explicit expression for the mean $m(t)$ at time t, starting empty, is given in (20) of [24]. More generally, the mean $m(t)$ satisfies an ordinary differential equation (ODE); see Corollary 4 of [24]. These results show that $m(t)$ and the entire distribution reaches steady state approximately at time c/μ, some constant c times the mean service time $1/\mu$. The constant c depends on our criterion; the critical time constant is $1/\mu$, a mean service time.

For the more general overloaded $M/M/n+M$ model (without assuming that $\theta = \mu$), it is helpful to consider the deterministic fluid approximation in [79]. Formula (2.17) there shows that the fluid approximation for the number in queue, $q(t)$, starting with all the servers busy, again evolves as the $M/M/\infty$ ODE, but with arrival rate $\lambda - n\mu$ and service rate θ. That implies that the fluid queue content (approximating the number in queue), starting from all servers busy but no queue, reaches steady state approximately at time c/θ, some constant c times the mean abandonment time $1/\theta$. That too will be approximately c/μ provided that θ is not too different from μ. The critical time constant here is $1/\theta$, a mean time to abandon.

To illustrate this mathematical analysis, we do a simulation of the $M/M/n+M$ model. We base our example here on Example 2.1.1 in §4.2. In that example, the service rates in both pools are $\mu_{i,i} = 1$, the abandonment rates are $\theta_i = 0.4$ and the number of agents in each pool is 100. In this example the arrival rates changed at some instant from $(\lambda_1, \lambda_2) = (90, 90)$ to $(\lambda_1, \lambda_2) = (130, 90)$. We show what happens if class 1 receives no help from service-pool 2. Then the class-1 queue behaves like an $M/M/100+M$ queue. Figure 2.11 depicts a simulated sample path of an $M/M/100+M$ queue, when the system is initialized

empty at time 0. The average steady-state queue length in the overload incident is about

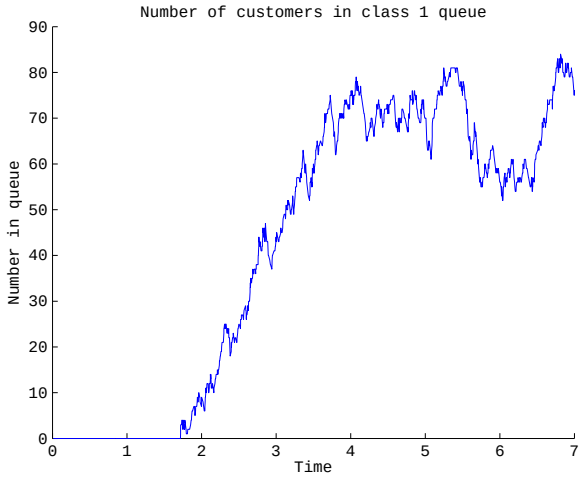


Figure 2.11: Time to reach steady state.

75, and it can be seen that this steady-state value is reached within about 4 time units when the system is initialized empty. (Time is measured in units of mean service times). If we assume, as in our example above, that the system was operating before the arrival rates changed, then most of the agents were probably busy, and the time to reach the new steady state is about 2 time units (two mean service times).

Single-Server models. In the introduction we stated that the number in system tends to approach steady state more quickly in many-server queues with abandonment than in single-server queues without abandonment. We should begin with a qualification: Slow approach to steady state occurs for single-server systems without abandonment when the system is heavily loaded. For single-server queues, we refer to Section III.7.3 of [20] on the relaxation time. Sections 4.6 and 5.1 of [76] gives conventional heavy-traffic approximations (when $\rho \uparrow 1$ with n fixed, where $\rho \equiv \lambda/n\mu$ is the traffic intensity) for the time required for the mean number in system to reach steady state in the general $G/G/n$ model with fixed n and without customer abandonment. The time required to reach steady state is

approximately $c/(1-\rho)^2$ mean service times, where c is a constant depending on the number of servers, n, the variability of the arrival and service processes (quantified explicitly) and again the criterion. Clearly the time to reach steady state can be quite long when ρ is high.

The X model. For our X model, there are two implications of the $M/M/n+M$ analysis above: First, when the overload incident begins, the queue length should be negligible, so that the fluid content in a newly overloaded queue will grow approximately linearly at rate $\lambda - n\mu$, because the opposite force $\theta q(t)$ will be small, since $q(t)$ is initially small. That means that the threshold will be quickly passed if there is a significant unbalanced overload.

For our more complicated X model with the QR-T control, after the threshold has been exceeded, the theoretical analysis for the $M/M/n + M$ model above provides a rough heuristic analysis indicating what should happen, but the actual evolution still depends on the state of the six-dimensional Markov chain $(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2)$. Thus we rely on simulation to confirm that the actual behavior is indeed similar to what occurs in these simple many-server models. We remark that the state-space collapse discussed in the next subsection indicates that $(Q_1(t), Q_2(t))$ should evolve approximately as a one-dimensional process, suggesting that the analysis above should not be too far off when the service rates $\mu_{i,j}$ do not differ greatly.

2.7.2 More on FQR

In this section we present additional background on FQR; for more, see [29, 30, 31, 32]. We first illustrate the state-space collapse (SSC). The conditions for SSC are satisfied if either the service rates only depend upon the customer class or the service rates only depend upon the agent pool. To illustrate, suppose that the service rates are independent of both class and pool, with $\mu_{1,1} = \mu_{1,2} = \mu_{2,1} = \mu_{2,2} = 1.0$. Figure 2.12 shows the plots of typical sample paths of the two queue-length processes when $\lambda_1 = \lambda_2 = m_1 = m_2 = 100$ and

$\theta_1 = \theta_2 = 0.2$. From Figure 2.12, we can clearly see the SSC.

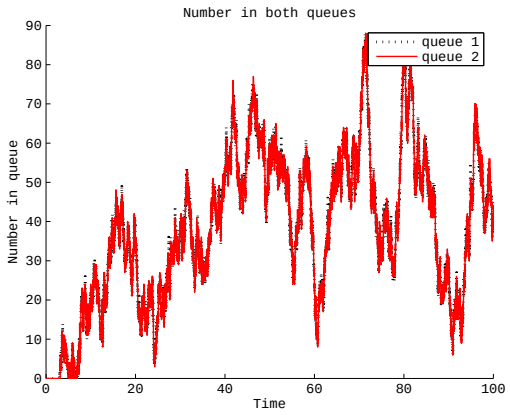


Figure 2.12: State-Space Collapse

We observed that, with FQR, it is possible to choose the ratio parameter r (or, equivalently, the queue proportions p_i) in order to determine the optimal level of staffing to achieve desired service-level differentiation. For example, under normal loading, our goal may be to choose staffing levels as small as possible subject to having 80% of class-1 customers wait less than 20 seconds, while 80% of class-2 customers wait less than 60 seconds. To see how this can be done with FQR, let T_i be the class-i delay target (e.g., $T_1 = 0.033$ and $T_2 = 0.100$ for 20 seconds and 60 seconds if the mean service times are 10 minutes); let W_i be the class-i waiting time before starting service; let p_i be the queue proportion determined by r. As explained in [29], the following string of approximations show how the individual class-i performance targets $P(W_i > T_i) \leq \alpha$, for both i, can be reduced into a single-class single-pool performance target $P(W > T) \leq \alpha$ for an appropriate choice of

the queue proportions p_i and the aggregate target T:

$$\begin{aligned} P(W_i > T_i) &\approx P(Q_i > \lambda_i T_i) \approx P\left(p_i Q_\Sigma > \lambda_i T_i\right) \approx P\left(Q_\Sigma > \sum_{k=1}^{2} \lambda_k T_k\right) \\ &\approx P\left(\lambda W > \sum_{k=1}^{2} \lambda_k T_k\right) \approx P(W > T) \leq \alpha \,, \end{aligned} \tag{2.7.1}$$

where we define $p_i \equiv \lambda_i T_i/(\lambda_1 T_1 + \lambda_2 T_2)$, $\lambda \equiv \lambda_1 + \lambda_2$ and $T \equiv (\lambda_1 T_1 + \lambda_2 T_2)/(\lambda_1 + \lambda_2)$. The first approximation in (5.5.37) follows by a heavy-traffic generalization of Little's law, establishing that the steady-state queue-length and waiting-time random variables are related approximately by $Q_i \approx \lambda_i W_i$. The second approximation in (5.5.37) is due to SSC: $Q_i \approx p_i Q_\Sigma$. The third approximation is obtained by choosing p_i as specified above. The fourth approximation in (5.5.37) follows from the heavy-traffic generalization of Little's law once again, for the entire system: $Q_\Sigma \approx \lambda W$ for λ as defined above, where W is the waiting time for an arbitrary customer. The fifth and final approximation follows by the appropriate definition of the aggregate target T, as defined above. With this reduction, we can determine the overall staffing by using elementary established methods for the single-class single-pool model. That is, we choose the total number of agents, m, so that $P(W > T) \leq \alpha$ in the $M/M/m + M$ model. We then let $m_i = p_i m$. From (5.5.37) and the fact that $r = p_1/(1 - p_1)$, we see that the required ratio is

$$r = \frac{p_1}{1 - p_1} = \frac{\lambda_1 T_1}{\lambda_2 T_2}. \tag{2.7.2}$$

For the X model (and more generally), Theorem 4.1 of Gurvich and Whitt [30] shows that, if the service rates only depend on the service pool or the class (but not both), then FQR is asymptotically optimal to minimize linear staffing costs subject to service-level constraints, as above, in the QED many-server heavy-traffic regime.

As was shown in §2.2.1, with inefficient sharing FQR without the thresholds we add

in FQR-T can cause the queues in the general X-model system to explode when there is no abandonment, because of the inefficient sharing. We now show that there is also serious performance degradation when we include customer abandonment. We use the same example as in §2.2.1, only adding abandonments with rates $\theta_1 = \theta_2 = 0.2$. As before, there are 100 agents in each pool. The arrival rates are $\lambda_1 = \lambda_2 = 99$ and the service rates are $\mu_{1,1} = \mu_{2,2} = 1$ and $\mu_{1,2} = \mu_{2,1} = 0.8$. To describe the performance degradation, we compare the performance to the no-sharing case. When there is no sharing, 2% of the customers abandon, the mean queue length is 10 and the mean conditional waiting time given that the customer is served is 0.10. On the other hand, for FQR with $r = 1$, again about 39% of the agents are busy serving customers from the other class. That reduces the effective service rate for each class from 100 to 92.2. As a consequence, about 7% of the customers abandon, the mean queue length is 34 and the average conditional waiting time given that the customer is served is 0.35.

Figures 2.13 and 2.14 show the sample paths of the number of agents in pool 1 helping class-2 customers, and the class-1 queue, respectively. Due to the symmetry of the system in our example, the $Z_{2,1}$ and Q_2 figures are very similar, and the fluid approximations for both queues and $Z_{i,j}$'s are equal.

In contrast, to illustrate how FQR-T performs, we consider the same example: Example 2.2.1 with abandonments at rate $\theta_i = 0.2$. We let $r_{1,2} = r_{2,1} = 1$, so that there is no change from FQR above, and we let the thresholds be $k_{1,2} = k_{2,1} = 10$. The results of a simulation experiment are shown in Figures 2.15 and 2.16. Numerical values were given in §2.2.2. The performance is greatly improved with FQR-T.

2.7.3 Optimal Solution for the Fluid Model

In this section we provide additional material supplementing §2.3.

Figure 2.13: $Z_{2,1}(t)/100$ for FQR with $r = 1$.

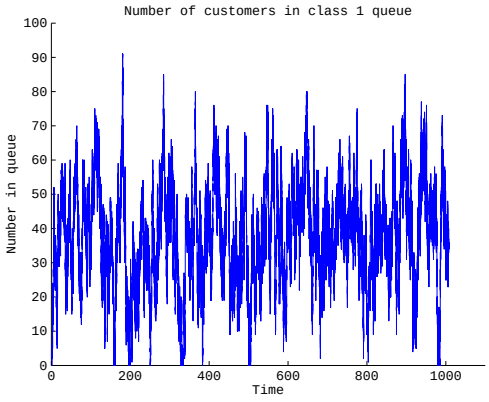


Figure 2.14: $Q_1(t)$ for FQR with $r = 1$.

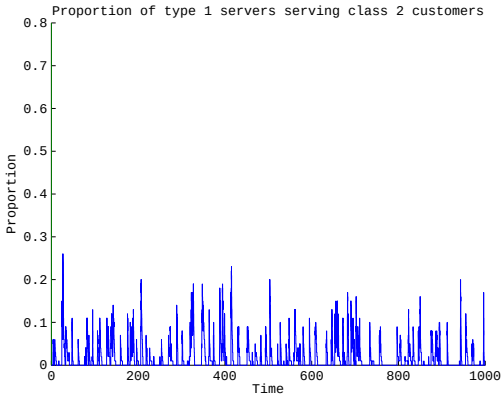


Figure 2.15: $Z_{2,1}(t)/100$ with FQR-T, $r = 1$.

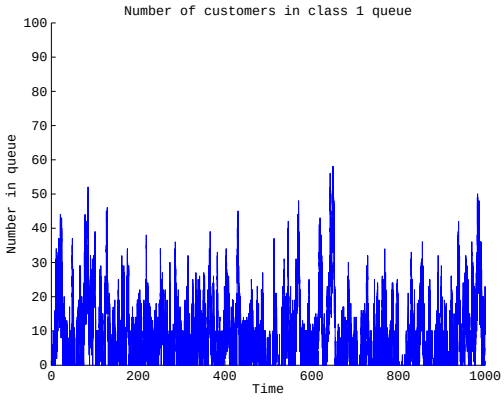


Figure 2.16: $Q_1(t)$ with FQR-T, $r = 1$.

Optimal Values Beyond the Boundaries It is natural to have the cost function C be smooth, in which case the optimal solution can be found by simple calculus. The following result concludes that, if the optimal solution found by calculus falls outside the feasible set, then the actual optimum value is obtained at the nearest boundary point. Let $a \wedge b \equiv \min\{a, b\}$ and $a \vee b \equiv \max\{a, b\}$. We omit the proof, which is a standard convexity result.

Proposition 2.7.1. (optimal values beyond the boundaries) *Let $\bar{Z}_{1,2}$ and $\bar{Z}_{2,1}$ be the values of $Z_{1,2}$ and $Z_{2,1}$ yielding minimum values of $C_{1,2}$ and $C_{2,1}$ in* (2.3.5)*, and let $\hat{Z}_{1,2}$ and $\hat{Z}_{2,1}$*

be the corresponding values yielding the minima ignoring the constraints in Proposition 2.3.3. Then $\bar{Z}_{1,2} = \hat{Z}_{1,2} \vee 0 \wedge m_2$, $\bar{Z}_{2,1} = \hat{Z}_{2,1} \vee 0 \wedge m_1$ *and* $(Z^*_{1,2}, Z^*_{2,1})$ *can assume only two possible values:* $(\bar{Z}_{1,2}, 0)$ *or* $(0, \bar{Z}_{2,1})$.

2.7.4 The Relation between r and Z

In §2.3.2 we observed that there is a one-to-one correspondence between the queue ratio $r \equiv Q_1/Q_2$ and the real variable $Z_{1,2} - Z_{2,1}$ used to specify the optimization problem in Proposition 2.3.3. That implies that there is a one-to-one correspondence between the fixed-agent-allocation optimization problem (choosing $Z_{1,2}$ and $Z_{2,1}$) and the queue-ratio control problem (choosing a state-dependent queue-ratio r) in the fluid-model context.

Proposition 2.7.2. (relating r and $Z_{1,2} - Z_{2,1}$) *For any given arrival-rate vector* (λ_1, λ_2) *or initial state* (q_1, s_1, q_2, s_2) *(without sharing), the queue ratio* $r \equiv Q_1/Q_2$ *is a strictly decreasing differentiable function of* $Z_{1,2} - Z_{2,1}$*, denoted by* ϕ*, as* $Z_{1,2} - Z_{2,1}$ *varies over its allowed domain in Proposition 2.3.3. Thus, the function* ϕ *has a unique inverse* ϕ^{-1} *and there exists a unique optimal* $r^* \equiv r^*(q_1, s_1, q_2, s_2)$*, which is characterized by*

$$r^* = \phi^{-1}(Z^*_{1,2} - Z^*_{2,1}), \tag{2.7.3}$$

where both r^* *and* $Z^*_{1,2} - Z^*_{2,1}$ *are understood to be functions of the initial state* (q_1, s_1, q_2, s_2)*. Moreover, there are two thresholds* $\eta_{1,2} > \eta_{2,1}$ *such that we want one-way sharing with pool* 2 *helping class* 1 *if* $r > \eta_{1,2}$*, in which case we let* $r_{1,2} = r^*$*; we want one-way sharing with pool* 1 *helping class* 2 *if* $r < \eta_{2,1}$*, in which case we let* $r_{2,1} = r^*$*; and we want no sharing at all if* $\eta_{2,1} \le r \le \eta_{1,2}$*. The thresholds are obtained from the thresholds* $\zeta_{1,2}$ *and* $\zeta_{2,1}$ *in Corollary* 2.3.1 *by* $\eta_{1,2} = \phi^{-1}(\zeta_{1,2})$ *and* $\eta_{2,1} = \phi^{-1}(\zeta_{2,1})$.

Proof: By (2.3.5), when pool 2 helps class 1, Q_1 is a strictly decreasing differentiable function of $Z_{1,2}$ and while Q_2 is a strictly increasing differentiable function of $Z_{1,2}$. On the

other hand, when pool 1 helps class 2, Q_1 is a strictly increasing differentiable function of $Z_{2,1}$ and while Q_2 is a strictly decreasing differentiable function of $Z_{2,1}$. Thus $r \equiv Q_1/Q_2$ is a strictly decreasing differentiable function of $Z_{1,2} - Z_{2,1}$ over its domain. ∎

2.7.5 Constant Weighted Queue Length

We now complete Proposition 2.3.4 by exhibiting the result for pool 1 helping class 2.

Proposition 2.7.3. (constant weighted queue lengths with pool 1 helping class 2) *Let*

$$a_{2,1} \equiv \frac{\mu_{2,1}\theta_1}{\mu_{1,1}\theta_2} \quad \textit{and} \quad \tilde{a}_{2,1} \equiv \frac{\mu_{2,1}}{\mu_{1,1}}. \tag{2.7.4}$$

Consider any initial state (λ_1, λ_2)*, or equivalently* (q_1, s_1, q_2, s_2)*, with* $s_2 = 0$*. Let*

$$w_{2,1} \equiv a_{2,1}\left(\frac{\lambda_1 - m_1\mu_{1,1}}{\theta_1}\right) + \left(\frac{\lambda_2 - m_2\mu_{2,2}}{\theta_2}\right) = a_{2,1}\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1}\right) + q_2. \tag{2.7.5}$$

Then

$$a_{2,1}\left(Q_1(Z_{1,2}) - \frac{S_1(Z_{2,1})\mu_{1,1}}{\theta_1}\right) + Q_2(Z_{2,1}) = w_{2,1} \tag{2.7.6}$$

for all $Z_{2,1}$ *with* $0 \le Z_{2,1} \le m_1$.

Just as with Proposition 2.3.4, Proposition 2.7.3 implies that the locus of all nonnegative queue-length vectors $(Q_1, Q_2) \equiv (Q_1(Z_{2,1}), Q_2(Z_{2,1}))$ associated with initial state (λ_1, λ_2), or equivalently (q_1, s_1, q_2, s_2), with $s_2 = 0$, is on the line $\{(Q_1, Q_2) : a_{2,1}Q_1 + Q_2 = w_{2,1}\}$ in the nonnegative quadrant. Thus, for any nonnegative constant $w_{2,1}$, the optimal queue-length vector (Q_1^*, Q_2^*) and the optimal queue-ratio $r_{2,1}^* \equiv Q_1^*/Q_2^*$ restricted to one-way sharing $(Z_{1,2} = 0)$ are the same for all initial states (q_1, s_1, q_2, s_2) with $s_2 = 0$ satisfying *(2.3.8)* and $q_2 \ge Q_2^*$. Moreover, $a_{2,1}Q_1^* + Q_2^* = w_{2,1}$. That same optimal queue-length vector and optimal queue ratio holds for all arrival pairs (λ_1, λ_2) where $s_2 = 0$, $Z_{1,2} = 0$

and

$$\lambda_1 + \tilde{a}_{2,1}\lambda_2 = \tilde{w}_{2,1} \equiv \frac{\theta_1\theta_2 w_{2,1} + a_{2,1}\theta_2 m_1\mu_{1,1} + \theta_1 m_2\mu_{2,2}}{a_{2,1}\theta_2}. \tag{2.7.7}$$

2.7.6 Structured Separable Cost Functions

At the end of §2.3.3, we observed that we can obtain explicit analytical expressions for the optimal ratio control if we impose additional structure on our cost function. We give the main results in this section and provide supporting details in the next section.

Main Results We first assume that C is separable, i.e., $C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2)$, where each component cost function C_i is strictly convex, strictly increasing and twice differentiable. We start by assuming that the derivatives C_i' are strictly increasing, so that their inverses exist. Let $\Psi(Q_1) \equiv C_1'(Q_1)$ and let Ψ^{-1} be its inverse. Then one of the following relations between the queue lengths should hold, when we choose the one that minimizes the cost:

$$Q_1 = \Psi^{-1}(a_{1,2}C_2'(Q_2)) \quad \text{or} \quad Q_1 = \Psi^{-1}(a_{2,1}C_2'(Q_2)), \tag{2.7.8}$$

for $a_{1,2}$ defined in (2.3.7) and $a_{2,1}$ defined in §2.7.5. If C_1' is not strictly increasing, then we work with the left-continuous inverse of Ψ defined by $\Psi^{\leftarrow} \equiv \{x : \Psi(x) \geq y\}$.

Power functions. If the cost functions C_i are simple power functions, i.e., $C_i(Q_i) \equiv c_i Q_i^{n_i}$ for $i = 1, 2$, then we have that either $Q_1^* = r_{1,2}^* Q_2^{*(n_2-1)/(n_1-1)}$ or $Q_1^* = r_{2,1}^* Q_2^{*(n_1-1)/(n_2-1)}$, where

$$r_{1,2}^* \equiv \sqrt[n_1-1]{a_{1,2}c_2 n_2/c_1 n_1} \qquad \text{and} \qquad r_{2,1}^* \equiv \sqrt[n_1-1]{a_{2,1}c_2 n_2/c_1 n_1}. \tag{2.7.9}$$

When $n_1 = n_2$, Q_1^*/Q_2^* is a fixed queue ratio, either $r_{1,2}^*$ or $r_{2,1}^*$ for $r_{i,j}^*$ as in (2.7.9). Thus, we need only to decide which way we should share, and then use FQR-T with the appropriate

$r^*_{i,j}$; i.e., we are in the setting of Figure 2.3 with constant ratios for which we have explicit expressions.

Quadratic functions. In practice it may be difficult to actually specify an appropriate cost function. Thus, for practical application we suggest quadratic functions: $C_i(Q_i) \equiv c_i Q_i^2 + b_i Q_i + a_i$ for $i = 1, 2$. These functions might be obtained by performing an approximation (e.g., via Taylor series approximation to an analytical expression or least squares fit to data). In this case, we have either

$$Q^*_1 - r^*_{1,2} Q^*_2 = k^*_{1,2}, \qquad \text{or} \qquad Q^*_1 - r^*_{2,1} Q^*_2 = k^*_{2,1}, \tag{2.7.10}$$

for

$$r^*_{1,2} \equiv \frac{a_{1,2} c_2}{c_1} = \frac{c_2 \mu_{2,2} \theta_1}{c_1 \mu_{1,2} \theta_2}, \qquad r^*_{2,1} \equiv \frac{a_{2,1} c_2}{c_1} = \frac{c_2 \mu_{2,1} \theta_1}{c_1 \mu_{1,1} \theta_2}. \tag{2.7.11}$$

and

$$k^*_{1,2} \equiv \frac{a_{1,2} b_2 - b_1}{2c_1}, \qquad k^*_{2,1} \equiv \frac{a_{2,1} b_2 - b_1}{2c_1}. \tag{2.7.12}$$

In other words, we keep a fixed-queue ratio centered about a constant $k^*_{i,j}$ instead of zero. That is, we employ new thresholds after sharing has been activated. (The current thresholds $k^*_{i,j}$ are not to be confused with the thresholds $k_{i,j}$ used with the queue-difference processes in (2.2.1) to test for the occurrence of overloads. We use $k^*_{1,2}$ only after sharing with pool 2 helping class 1.) From the two formulas in (2.7.11), we directly see how these ratio parameters and thresholds should depend on the model parameters. In particular, each ratio is either directly proportional or inversely proportional to each of six model parameters.

Quadratic and linear power functions. A natural simple cost function is the quadratic power function, which is is special case of the general power function with $n_1 = n_2 = 2$ and a special case of the general quadratic function with $b_1 = b_2 = a_1 = a_2 = 0$. The optimal control then is precisely FQR-T ($k^*_1 = k^*_2 = 0$), as indicated in Proposition 2.3.5.

In §2.7.7 we also discuss the special cases $n_1 = 2, n_2 = 1$ and $n_1 = 1, n_2 = 1$.

FQR-T without cost functions. Clearly, FQR-T could be employed directly without specifying any cost function, using engineering judgment to set the parameters. Even if that is the case, the queue-ratio formulas in (2.7.11) and possibly the centering formulas in (4.2.5) provide important insight into how the control parameters should depend on the model parameters.

2.7.7 Supporting Details About Structured Separable Cost Functions

We now supplement §2.7.6 by providing more details about the fluid model with a separable cost function. As before, we assume that C is separable, and that each component cost-function C_i is strictly convex, strictly increasing and twice differentiable. We then relax the strictly-increasing assumption, and consider linear functions.

Let $C(Q_1, Q_2)$ be a separable function. We can write C as a function of one variable $Z_{1,2}$ or $Z_{2,1}$, depending on which way the sharing is done.

$$C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2) \equiv C_{1,(i,j)}(Z_{i,j}) + C_{2,(i,j)}(Z_{i,j}) \equiv C(Z_{i,j}), \qquad (2.7.13)$$

where

$$C_{1,(1,2)}(Z_{1,2}) \equiv C_1\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} - \frac{Z_{1,2}\mu_{1,2}}{\theta_1}\right),$$
$$C_{2,(1,2)}(Z_{1,2}) \equiv C_2\left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} + \frac{Z_{1,2}\mu_{2,2}}{\theta_2}\right).$$

and

$$C_{1,(2,1)}(Z_{2,1}) \equiv C_1\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} + \frac{Z_{2,1}\mu_{1,1}}{\theta_1}\right),$$
$$C_{2,(2,1)}(Z_{2,1}) \equiv C_2\left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} - \frac{Z_{2,1}\mu_{2,1}}{\theta_2}\right).$$

Hence, the optimal $Z_{1,2}$ is achieved when

$$C'(Z_{1,2}) = -C'_{1,(1,2)}(Z_{1,2})\left(\frac{\mu_{1,2}}{\theta_1}\right) + C'_{2,(1,2)}(Z_{1,2})\left(\frac{\mu_{2,2}}{\theta_2}\right) = 0,$$

or equivalently,

$$C'_1(Q_1) = \frac{\mu_{2,2}\theta_1}{\mu_{1,2}\theta_2} C'_2(Q_2) \equiv a_{1,2}C'_2(Q_2).$$

Similarly, the optimal $Z_{2,1}$ is achieved when

$$C'_1(Q_1) = \frac{\mu_{2,1}\theta_1}{\mu_{1,1}\theta_2} C'_2(Q_2) \equiv a_{2,1}C'_2(Q_2).$$

The fact that C_i is strictly convex implies that $C''_i \geq 0$. If $C''_i > 0$ then C'_i is strictly increasing, and its inverse function exists. Let $\Psi(Q_1) \equiv C'_1(Q_1)$ and let Ψ^{-1} be its inverse. Then one of the following relations between the queues should hold:

$$\text{either} \quad Q_1 = \Psi^{-1}(a_{1,2}C'_2(Q_2)) \quad \text{or} \quad Q_1 = \Psi^{-1}(a_{2,1}C'_2(Q_2)), \tag{2.7.14}$$

where we choose the relation that minimizes the cost-function $C(Q_1, Q_2)$.

If the inverse of Ψ does not exist, then we can work with the left-continuous inverse of Ψ defined by $\Psi^{\leftarrow}(y) \equiv \{x : \Psi(x) \geq y\}$.

We now consider separable cost functions of the form:

$$C(Q_1, Q_2) = c_1 Q_1^{n_1} + c_2 Q_2^{n_2}, \qquad n_1, n_2 \in \mathbb{N}, \tag{2.7.15}$$

where each component is a power function. The optimal solution is given in the main chapter. We observe that the compatible-ratio condition $r_{1,2} \geq r_{2,1}$ in S 2.2.2 holds, because

$$\frac{r^*_{1,2}}{r^*_{2,1}} = \sqrt[n_1-1]{\frac{a_{1,2}}{a_{2,1}}} = \sqrt[n_1-1]{\frac{\mu_{1,1}\mu_{2,2}}{\mu_{1,2}\mu_{2,1}}} \geq 1,$$

under the inefficient-sharing condition (2.1.2). When $n_1 = n_2$ we get

$$\frac{Q_1^*}{Q_2^*} = r_{1,2}^* \quad \text{or} \quad \frac{Q_1^*}{Q_2^*} = r_{2,1}^*;$$

i.e., it is optimal to keep a fixed-queue ratio. Thus, we need only to decide on which way we should share, and then use FQR-T with the appropriate fixed queue ratio $r_{i,j}^*$.

These results explain why the optimal ratios in our numerical example with the cost function in (2.3.10) in §2.3.3 are almost constant. In the numerical example there are other terms, but the dominating ones are the quadratic terms. As the queues get larger, the influence of the smaller-power terms decreases, and the optimal ratios converge to fixed numbers. If the function is separable (as would be the case if our example had not had the Q_1Q_2 term), then the convergence is to the same ratios as if the only terms are $c_1Q_1^n + c_2Q_2^n$. The mixed terms of power n change these numbers. For the cost function in (2.3.10), the Q_1Q_2 terms is also of power 2, and hence the optimal ratios converge to different numbers than (2.7.9). But for that example, clearly the optimal ratios are nearly constant.

In §2.7.6 we introduced the general separable quadratic cost function to provide a tractable approximation for a broad range of possible cost functions. We observed that the optimal queue-ratio function becomes a shifted version of FQR-T, which is just FQR-T centered at points $k_{1,2}^*$ and $k_{2,1}^*$ instead of centered at 0. We now illustrate the resulting control for a candidate cost function. In order to make the linear components have approximately equal weight to the quadratic components when the queue lengths are about 50, we divide the coefficients c_i for the quadratic terms by 10. We also omit the mixed term Q_1Q_2, which violated the separability property. Instead of the cost function in (2.3.10), we now consider the cost function

$$C(Q_1, Q_2) \equiv 0.3Q_1^2 + 10Q_1 + 0.2Q_2^2 + 5Q_2. \tag{2.7.16}$$

The centering is depicted by the y-intercepts on the two lines in Figure 2.17.

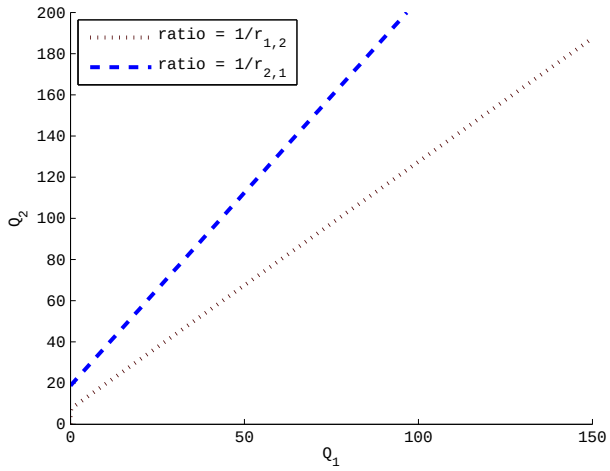


Figure 2.17: The optimal queue ratios (shifted FQR).

We now consider the two linear cases. When one or more of the component cost functions C_i is linear, we are led to modify our control. We indicate how our fluid-model analysis can be applied to generate alternative controls in these cases, but we do not examine their performance here.

$\mathbf{n_1 = 2, n_2 = 1}$. The cost-function $C(Q_1, Q_2) = c_1 Q_1^2 + c_2 Q_2$ has one quadratic term and one linear term. The special structure of this function (C_2 not strictly convex) changes the control. Now, there is no longer dependence on the two queues, since Q_2 no longer comes into play. By (2.7.14),

$$Q_1^* = \frac{a_{1,2}c_2}{2c_1} \equiv k_{1,2}^* \qquad \text{or} \qquad Q_1^* = \frac{a_{2,1}c_2}{2c_1} \equiv k_{2,1}^*.$$

Thus, we are no longer trying to keep a relation between the two queues, but instead we keep Q_1 not bigger than $k_{1,2}^*$ or $k_{2,1}^*$, depending which is optimal to use. To keep Q_1 at its optimal target, we modify our control: If class 1 is overloaded such that $q_1 > k_{1,2}^*$, then whenever $D_{1,2} \geq \max\{k_{1,2}, k_{1,2}^*\}$ every newly available agent takes his next customer from

the head of queue 1. Otherwise, every agent takes his next customer from the head of its own class queue.

If class 2 is overloaded, then we can have $Z_{2,1} > 0$ as long as we keep $Q_1 < k_{2,1}$. Hence, if $D_{2,1} < k_{2,1}$ and $Q_1 < k_{2,1}$, then every newly available agent takes his next customer from the head of Q_2. Otherwise, he will take his next customer from the head of his own class queue.

$\mathbf{n_1 = 1, n_2 = 1}$. The purely-linear cost function $C(Q_1, Q_2) = c_1 Q_1 + c_2 Q_2$ is even more different than the functions we considered so far. However, it is well known that a linear function attains its minima on the boundaries of its domain. In our setting, this means that we either try to keep the queue that needs help at zero, or that we do not help it at all. When $Z_{1,2} > 0$, we have

$$\begin{aligned} C(Q_1, Q_2) = C(Z_{1,2}) &= c_1 \left(q_1 - \frac{s_1 \mu_{1,1}}{\theta_1} - \frac{\mu_{1,2}}{\theta_1} Z_{1,2} \right) + c_2 \left(q_2 - \frac{s_2 \mu_{2,2}}{\theta_2} + \frac{\mu_{2,2}}{\theta_2} Z_{1,2} \right) \\ &= c_1 \left(q_1 - \frac{s_1 \mu_{1,1}}{\theta_1} \right) + c_2 \left(q_2 - \frac{s_2 \mu_{2,2}}{\theta_2} \right) + \left(\frac{c_2 \mu_{2,2}}{\theta_2} - \frac{c_1 \mu_{1,2}}{\theta_1} \right) Z_{1,2}. \end{aligned}$$

Thus, if

$$\frac{c_1 \mu_{1,2}}{\theta_1} \leq \frac{c_2 \mu_{2,2}}{\theta_2} \tag{2.7.17}$$

the function $C(Z_{1,2})$ is increasing, and its minima is attained when $Z_{1,2} = 0$. Otherwise, the function is decreasing, and it is optimal to take $Z_{1,2}$ as large as needed to ensure $Q_1 = 0$ (the simple calculus gives us that $Z^*_{1,2} = m_2$, but of course class 1 may not have enough arrivals to fill both service pools). This means that we either share completely, or not share at all.

Similarly, for

$$C(Q_1,Q_2)=C(Z_{2,1})=c_1\left(q_1-\frac{s_1\mu_{1,1}}{\theta_1}-\frac{\mu_{1,1}}{\theta_1}Z_{2,1}\right)+c_2\left(q_2-\frac{s_2\mu_{2,2}}{\theta_2}+\frac{\mu_{2,1}}{\theta_2}Z_{2,1}\right)$$
$$=c_1\left(q_1-\frac{s_1\mu_{1,1}}{\theta_1}\right)+c_2\left(q_2-\frac{s_2\mu_{2,2}}{\theta_2}\right)+\left(\frac{c_1\mu_{1,1}}{\theta_1}-\frac{c_2\mu_{2,1}}{\theta_2}\right)Z_{2,1},$$

if

$$\frac{c_2\mu_{2,1}}{\theta_2}\leq\frac{c_1\mu_{1,1}}{\theta_1} \tag{2.7.18}$$

then $C(Z_{2,1})$ is increasing, and its minima is attained at $Z_{2,1}=0$. Otherwise, $C(Z_{2,1})$ is decreasing, and it is optimal to take $Z_{2,1}$ as large as needed (and possible) to make sure that $Q_2=0$.

Rewriting the inefficient-sharing condition (2.1.2), we get

$$\frac{\mu_{1,1}}{\mu_{2,1}}\geq\frac{\mu_{1,2}}{\mu_{2,2}}.$$

If the two inequalities (3.5.7) and (3.5.8) hold together, then

$$\frac{\theta_1c_2}{\theta_2c_1}\leq\frac{\mu_{1,2}}{\mu_{2,2}}\quad\text{and}\quad\frac{\mu_{1,1}}{\mu_{2,1}}\leq\frac{\theta_1c_2}{\theta_2c_1},$$

but this contradicts the inefficient-sharing condition above, unless all the inequalities hold as equalities. Thus, we can have at most one of the inequalities, (3.5.7) or (3.5.8), hold under (2.1.2).

At first glance, it may seem from the discussion above that, when the holding cost is linear, we should not consider the system as an X model, but rather as an N model (sharing can be done in only one direction), if either (3.5.7) or (3.5.8) hold, or two independent I systems (no sharing at all), if none of these two inequalities hold. But that is not so. If there is spare capacity in one class, while the other class is overloaded, then it is always

optimal to use the extra agents to help the overloaded class. Since we do not know what the overload incident will produce, we cannot restrict the model to an N model in advance.

Let us summarize what we have found: The cost analysis leads us to give priority to either queue 1 or queue 2. Suppose that it is optimal to give priority to queue 1. That leads us to set the threshold for pool 2 helping class 1 at $k_{1,2} = 0$. In the fluid model that will either produce the desired result $Q_1 = 0$ or $Q_1 > 0$ and $Z_{1,2} = m_2$, with pool 2 devoting all its effort to class 1. There remains another case: when pool 1 has spare spare capacity. In that case, within the fluid model, if pool 2 is overloaded, then pool 1 should devote all the required spare capacity to serving class 2. We should have $Z_{2,1} = s_1 \wedge q_2$. If $s_1 > q_2$, then the help pool 1 provides to class 2 makes both queues empty, and there is remaining spare capacity for pool 1. On the other hand, if $s_1 \leq q_2$, then we have exactly $Z_{2,1} = s_1$. Overall, there are three possible end results in the fluid model: (i) $Q_1 > 0$ and $Z_{1,2} = m_2$, (ii) $Q_1 = Q_2 = 0$, (iii) $Q_2 > 0$, $Q_1 = 0$ and $Z_{2,1} = s_1$.

We now must consider how to implement that control in the actual system. As indicated above, to give priority to queue 1 at all times, we can set $k_{1,2} = 0$, and we always allow pool 2 to help class 1, even if $Z_{2,1} > 0$. The only difficulty is detecting whether or not pool 1 has spare capacity, so that we can have pool 1 helping class 2. For this purpose, we propose using a positive queue threshold for queue 2: We let an available agent in pool 1 help class 2 if, and only if, $Q_2 > k_{2,1}$, $Q_1 = 0$ and $Z_{1,2} = 0$.

Since we allow pool 2 to serve class 1 all the time, we could possibly have simultaneous two-way sharing (both $Z_{1,2} > 0$ and $Z_{2,1} > 0$), but there should be only minimal simultaneous two-way sharing. It remains to further investigate this case.

2.8 Additional Simulation Results

In this section we present additional simulation results.

2.8.1 Comparing the Two Controls

We now supplement the comparison of the two controls (the fixed staffing levels versus QR-T) in §2.5.2 by presenting detailed simulation results. These are given in Table 2.2, including the half-width of 95% confidence intervals and a comparison of the simulation to the fluid approximation.

As stated before, for each case, we conducted 5 independent simulation runs using QR-T, and 5 independent simulation runs with a fixed $Z_{1,2}$, each run with $300,000$ arrivals. The independent replications make it possible to reliably estimate confidence intervals using the t statistic with 4 degrees of freedom. The large number of arrivals ensures that the transient behavior in the beginning of the simulation, before reaching steady state, does not affect the final simulation estimates.

We now provide additional observations about our simulation results for this example. Another important observation is that FQR-T is doing a better job in keeping the ratio between the two queues close to the desired ratio. The accuracy becomes even better when the system is larger (see the "ratio" row in Table 2.1 in the $n = 400$ columns). We have also included a column showing the simulated standard deviations of the ratios. Note how small the standard deviations are when using FQR-T, in comparison to the standard deviations when using the fixed-$Z_{1,2}$ control. Since FQR-T is working towards keeping the ratio between the two queues fixed throughout, the ratio between the two queues at any time point is approximately $r_{1,2}$. It also makes the two queues strongly positively correlated, which reduces the overall variance. In contrast, under the fixed-$Z_{1,2}$ control, the two queues are independent with zero correlation.

The simulated ratio was calculated as a long-run average of the ratio between the two queues throughout the simulation time. We can compare it to Q_1/Q_2 from Table 2.1 which appears in §2.5.1 (in the $n = 100$ columns) which is also approximately 0.9. These agree

		Cost (in thousands)		actual ratio			actual $Z_{1,2}$	
policy		Approx.	Sim.	Approx.	Sim.	std.	Approx.	Sim.
FQR-T	$r = 1.20$	19.65	20.51 ± 0.64	1.20	1.07 ± 0.00	0.16 ± 0.01	15.0	15.9 ± 0.4
	$r = 1$	19.35	20.28 ± 0.81	1.00	0.90 ± 0.00	0.13 ± 0.01	16.7	17.7 ± 0.3
	$r = 0.83$	19.25	19.73 ± 0.64	0.83	0.76 ± 0.00	0.11 ± 0.00	18.4	18.9 0.5
	$r = 0.60$	19.56	21.16 ± 0.77	0.60	0.56 ± 0.00	0.08 ± 0.00	21.4	22.1 ± 0.3
	$r = 0.40$	20.75	22.31 ± 0.92	0.40	0.37 ± 0.00	0.06 ± 0.00	25.0	25.3 ± 0.3
fixed $Z_{1,2}$	$Z_{1,2} = 15$	19.65	21.47 ± 0.57	1.20	1.52 ± 0.08	1.93 ± 0.29	15.0	15.0 ± 0.0
	$Z_{1,2} = 17$	19.32	21.35 ± 0.46	0.96	1.13 ± 0.11	1.17 ± 0.45	17.0	17.0 ± 0.0
	$Z_{1,2} = 19$	19.26	20.86 ± 0.37	0.78	0.87 ± 0.07	0.75 ± 0.57	19.0	19.0 ± 0.0
	$Z_{1,2} = 22$	19.69	21.42 ± 0.60	0.56	0.61 ± 0.05	0.38 ± 0.04	22.0	22.0 ± 0.0
	$Z_{1,2} = 25$	20.75	22.63 ± 0.86	0.40	0.42 ± 0.01	0.33 ± 0.14	25.0	25.0 ± 0.0

Table 2.2: Full simulation results of Figure (2.4). The 'approx' columns show the anticipated results according to the fluid model, and the 'sim.' columns show the simulation results, together with half-width confidence interval.

closely because of state-space collapse, as in Figure 2.12 discussed in §2.7.2.

Finally, Table 2.2 shows that the fluid approximation tends to underestimate the actual average cost in the stochastic model. That is understandable, because the fluid model ignores stochastic fluctuations, which will tend to increase the average costs with a convex cost function. However, note that the fluid approximation does do an excellent job in describing the relative costs. In particular, the fluid model succeeds in locating the correct minima for both controls.

2.8.2 Performance of FQR-T Under Normal Loading

In §2.2.1 we saw that FQR-T performs well when $k_{1,2} = k_{2,1} = 10$ for Example 2.2.1 with $n = 100$ servers and $\lambda_i = 99$, showing that FQR without thresholds and one-way sharing can perform poorly. To supplement those simulation results and the simulation results in §2.5, in this section we present additional simulation results. As in Table 2.1, we consider three values of n: $n = 25$, $n = 100$ and $n = 400$. We let the arrival rates in both queues be $\lambda_i^{(n)} = 0.98n$. The service-rate and abandonment-rate parameters are fixed at $\mu_{i,i} = 1.0$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\theta_i = 0.2$. We let the thresholds be $k_{1,2}^{(n)} = k_{2,1}^{(n)} = 0.1n$, rounded up to 3.0 for $n = 25$. We compare the results of FQR-T to the $M/M/n + M$ model, which would prevail if there were absolutely no sharing at all. As before, we see that the mean queue lengths are actually slightly smaller with FQR-T. That shows that the little sharing that takes place with FQR-T is not so bad.

Due to the symmetry of the system under the parameters we chose, there is no difference between the steady-state values of both queues and service pools. Thus, we display only Q_1 and $Z_{1,2}$. We can see that as the system becomes larger, the sharing decreases, and the queue size gets closer to the queue length in an $M/M/n + M$ model.

2.8.3 Sensitivity Analysis For the Thresholds

We now consider different values for the thresholds with and without one-way sharing. Our objective is to perform a sensitivity analysis for the thresholds for a finite system (in this case having 100 agents in each service pool), as a complimentary to the asymptotic line of reasoning in §2.4. The simulation results, displayed in Table 2.4, are for systems having $\lambda_i = 98$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and $\theta_i = 0.2$, $i = 1, 2$. We vary the thresholds between $\kappa_{i,j} = 1$ and $\kappa_{i,j} = 30$, $i, j = 1, 2$. (Note that with $\kappa_{i,j} = 1$ FQR-T reduces to FQR.) For ease of exposition we take $r_{1,2} = r_{2,1} = 1$. The symmetry allows us to present

	n=25		n=100		n=400	
perf. meas.	I model	sim.	I model.	sim.	I model	sim.
$E[Q_1]$	5.1	4.8 ± 0.3	8.4	7.3 ± 1.0	11.3	10.5 ± 2.6
$E[Q_1/n]$	0.20	0.19 ± 0.01	0.08	0.07 ± 0.01	0.03	0.03 ± 0.01
$E[Z_{1,2}]$	$-$	1.3 ± 0.1	$-$	1.9 ± 0.2	$-$	1.3 ± 0.4
$E[Z_{1,2}/n]$	$-$	0.05 ± 0.01	$-$	0.02 ± 0.00	$-$	0.00 0.00

Table 2.3: A comparison of the exact I-model queues with simulation results for the steady-state performance measures of the X model in normal loading under FQR-T. The arrival rates are $\lambda_1^{(n)} = \lambda_2^{(n)} = 0.98n$ and the thresholds are $\kappa_{1,2}^{(n)} = \kappa_{2,1}^{(n)} = 0.1n$. Service rates are $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and the abandonment rates are $\theta_1 = \theta_2 = 0.2$

the results for $E[Q_1]$ and $E[Z_{1,2}]$ only, and consider $k_{1,2} = k_{2,1}$. (If $r_{1,2} \neq r_{2,1}$ then the sensitivity analysis should be performed for each of the two thresholds separately.)

Table 2.4 clearly shows the benefits of using one-way sharing, since even with $\kappa_{i,j} = 1$ the performance is almost as good as when we add thresholds of size 15. However, recall that the thresholds play a vital role in our control: In addition to helping prevent unwanted sharing, they act as "overload detectors": When $D_{i,j}(t)$ first crosses the threshold $\kappa_{i,j}$, we consider class-i queue to be overloaded, and sharing is activated with pool j helping queue i.

As discussed in §2.4, we do not want to have the thresholds too large, as they will fail to detect small overloads. Moreover, we see that it can actually be beneficial to share a little, even when the system is not overloaded; Observe that the average queue length in the case $\kappa_{i,j} = 30$ is larger than when $\kappa_{i,j}$ is $10, 15$ or 20. (See also the last paragraph in §2.2.)

Thus, in choosing the thresholds we need to make sure that under normal loadings they will not be crossed too often, but even small overloads will be detected. Here we see that any value in $\{10, \ldots, 20\}$ is reasonable, both with one-way sharing and without. To ensure

that even small overloads will be detected by the thresholds, it is probably best to take $10 \leq \kappa_{i,j} \leq 15$.

Insight From the Asymptotic Analysis of the Thresholds. The asymptotic analysis in §2.4 helps to find good candidates for the thresholds for larger systems. In our example, we can heuristically think of 30 as being of order $O(n)$, while 10 and 15 are of a smaller order, say $O(n^{0.6})$. Then $30 = 0.3n$, $15 \approx n^{0.6}$ and $10 \approx 2/3n^{0.6}$.

This line of reasoning hints at what the thresholds should be (approximately) for a larger system having the same service and abandonment parameters. For example, if $n = 1000$ then $\kappa_{i,j} = 0.3n = 300$ is too large, but $64 \approx n^{0.6} \leq \kappa_{i,j} \leq 2/3n^{0.6} \approx 42$ are good candidates for the thresholds. The threshold values can be determined using simulations, just as in Table 2.4.

	With One-Way Sharing		Without One-Way Sharing	
perf. meas.	$E[Q_1]$	$E[Z_{1,2}]$	$E[Q_1]$	$E[Z_{1,2}]$
$\kappa_{i,j} = 1$	8.4 ± 0.4	2.8 ± 0.3	29.9 ± 1.7	38.2 ± 0.6
$\kappa_{i,j} = 5$	8.3 ± 0.9	2.4 ± 0.3	8.6 ± 0.6	8.1 ± 0.1
$\kappa_{i,j} = 10$	7.5 ± 0.4	1.9 ± 0.2	7.4 ± 0.6	3.6 ± 0.2
$\kappa_{i,j} = 15$	7.2 ± 0.6	1.5 ± 0.2	7.1 ± 0.6	2.1 ± 0.1
$\kappa_{i,j} = 20$	7.5 ± 0.7	1.1 ± 0.2	7.3 ± 0.7	1.3 ± 0.2
$\kappa_{i,j} = 30$	8.2 ± 0.9	0.5 ± 0.1	8.2 ± 0.7	0.5 ± 0.1

Table 2.4: Sensitivity analysis of the effect of the thresholds in a system with 100 agents in each pool. The arrival rates are $\lambda_1 = \lambda_2 = 98$. Service rates are $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and the abandonment rates are $\theta_1 = \theta_2 = 0.2$. All the results are derived from five independent simulation runs.

Chapter 3

Transient and Stability Analysis

This chapter is devoted to the study of a dynamical system, represented by a three-dimensional *ordinary differential equations* (ODE). In Chapter 4 this ODE will be shown to arise as the *many-server heavy-traffic* (MS-HT) fluid limit of the overloaded Markovian X service-system model operating under FQR-T, as in Chapter 2. However, in this chapter we will derive the ODE heuristically, by considering the behavior of the stochastic X model when the number of servers in each service pool becomes large. In particular, we will apply a heavy-traffic *averaging principle* (AP) as an engineering principle, in order to justify the ODE considered here. As mentioned above, a rigorous justification is given in Chapters 4 and 5.

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 achieves a time-dependent steady state instantaneously in the MS-HT limit. In Chapter 4 we will show that convergence of the fluid-scaled sequence of overloaded X-model systems to this ODE holds, provided that the driving process is replaced by its long-run average behavior at each instant of time.

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 a limiting 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.

Since this chapter appears before the fluid limit is derived, we briefly explain how the fluid limit is derived. We also review the main points of Chapter 2 which will be needed for our analysis here.

3.1 Preliminaries

We now briefly summarize the essential conclusions of Chapter 2.

3.1.1 The Approximating Deterministic Fluid Model in Steady State

Given the model in §2.2.1, we want to determine an effective control and analyze its performance. In order to (approximately) minimize the expected cost over the overload incident, 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 [79]. Second, as demonstrated in §2.7.1, many-server systems approach steady state relatively quickly. (In this chapter, 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 Chapter 2 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.

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 $t \to \infty$, 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 §3.5 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

$$q^*_1 = \frac{\lambda_1 - \mu_{1,1} m_1 - \mu_{1,2} z^*_{1,2}}{\theta_1} \quad \text{and} \quad q^*_2 = \frac{\lambda_2 - \mu_{2,2}(m_2 - z^*_{1,2})}{\theta_2}. \tag{3.1.1}$$

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 (3.1.1) 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 $0 \le z^*_{1,2} \le m_2$.

It follow immediately from (3.1.1) that q^*_1 is decreasing with $z^*_{1,2}$, while q^*_2 is increasing with $z^*_{1,2}$. Consequently, for given arrival rates λ_1 and λ_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 Chapter 2.

However, as explained in Chapter 2, 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 be difficult to specify, it is natural to consider simple parametric spacial 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.

3.1.2 The FQR-T Control for the Original Queueing Model

Having found the optimal steady-state fluid levels for the fluid model, we suggested employing the FQR-T control (or its variants), which is described in §2.2. The purpose of the control is to automatically detect overloads immediately when they occur, and maintain the optimal ratio between the two queues when the system is overloaded.

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 $t \geq 0$.

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 Chapter 2, 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 $q^*_1 = r^*_{1,2}q^*_2 + \kappa_{1,2}$ for some $\kappa_{1,2} \in \mathbb{R}$ (assuming that pool 2 needs to help class 1), 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 $\kappa_{1,2}$ 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 $D_{i,j}(t) > \kappa_{1,2}$. 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 $\kappa_{1,2}$. We can think of FQR-T as the special case of shifted FQR-T with $\kappa_{1,2} = 0$.

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 [59] SSC is substantiated via simulation; in Chapter 4 it will be shown to hold asymptotically

in the MS-HT limit, which we describe in the next section.

In [59] 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.

3.2 The Many-Server Heavy-Traffic Fluid Limit

We first describe the convergence of the sequence of stochastic systems to the fluid limit, as was conjectured in [59] and will be established in Chapter 4. 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.)

3.2.1 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.,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \to \lambda_i \quad \text{and} \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \to m_i \quad \text{as} \quad n \to \infty, \tag{3.2.1}$$

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

$$\bar{Q}_i^n(t) \equiv \frac{Q_i^n(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^n(t) \equiv \frac{Z_{i,j}^n(t)}{n}, \quad i,j = 1,2, \quad t \geq 0. \tag{3.2.2}$$

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

$$\frac{k_{i,j}^n}{n} \to 0 \quad \text{and} \quad \frac{k_{i,j}^n}{\sqrt{n}} \to \infty \quad \text{as} \quad n \to \infty, \quad i,j = 1,2. \tag{3.2.3}$$

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 shold 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 §3.1.2, 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 $\kappa_{1,2}^n \to \infty$, then

$\tilde{D}^n_{1,2} \to \infty$ as $n \to \infty$. Hence, we redefine the difference process. Let

$$D^n(t) \equiv (Q^n_1(t) - \kappa^n) - rQ^n_2(t), \quad t \geq 0, \tag{3.2.4}$$

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

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

We now apply FQR using the process D^n in (3.2.4): if $D^n(t) > 0$, then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If $D^n(t) \leq 0$, then every newly available agent takes his new customer from the head of his own queue.

3.2.2 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 [57], we represent all the processes of interest in terms of mutually independent random-time-changed rate-1 Poisson processes: Let N^a_i, $N^s_{i,2}$ and N^u_i 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 $Z^n_{2,1}(t) = 0$, $Z^n_{1,1}(t) = m^n_1$ and $Z^n_{2,2}(t) = m^n_2 - Z^n_{1,2}(t)$, for all $t \geq 0$, so that we need only consider $Z^n_{1,2}$.

We thus obtain the following representation for the three processes Q^n_1, Q^n_2 and $Z^n_{1,2}$ in

terms of the queue-difference process D^n in (3.2.4):

$$
\begin{aligned}
Z^n_{1,2}(t) &\equiv Z^n_{1,2}(0) + N^s_{2,2}\left(\mu_{2,2}\int_0^t 1_{\{D^n(s)\geq 0\}}(m^n_2 - Z^n_{1,2}(s))\,ds\right) \\
&\quad - N^s_{1,2}\left(\mu_{1,2}\int_0^t 1_{\{D^n(s)<0\}}Z^n_{1,2}(s)\,ds\right), \quad t \geq 0. \\
Q^n_1(t) &\equiv Q^n_1(0) + N^a_1(\lambda^n_1 t) - N^s_{1,1}(m^n_1\mu_{1,1}t) - N^s_{1,2}\left(\mu_{1,2}\int_0^t 1_{\{D^n(s)\geq 0\}}Z^n_{1,2}(s))\,ds\right) \\
&\quad - N^s_{2,2}\left(\mu_{2,2}\int_0^t 1_{\{D^n(s)\geq 0\}}(m^n_2 - Z^n_{1,2}(s))\,ds\right) - N^u_1\left(\theta_1\int_0^t Q^n_1(s))\,ds\right), \quad t \geq 0. \\
Q^n_2(t) &\equiv Q^n_2(0) + N^a_2(\lambda^n_2 t) - N^s_{2,2}\left(\mu_{2,2}\int_0^t 1_{\{D^n(s)<0\}}(m^n_2 - Z^n_{1,2}(s))\,ds\right) \\
&\quad - N^s_{1,2}\left(\mu_{1,2}\int_0^t 1_{\{D^n(s)<0\}}Z^n_{1,2}(s)\,ds\right) - N^u_2\left(\theta_2\int_0^t Q^n_2(s))\,ds\right), \quad t \geq 0.
\end{aligned}
\tag{3.2.5}
$$

We then construct the usual martingale processes by subtracting the stochastic intensities, letting $M^{n,a}_i(t) \equiv N^a_i(\lambda^n_i t) - \lambda^n_i t$, $M^{n,u}_i(t) \equiv N^u_i\left(\theta_i\int_0^t Q^n_i(s))\,ds\right) - \theta_i\int_0^t Q^n_i(s)\,ds$ and $M^{n,s}_{i,2}(t) \equiv N^s_{i,2}(I^n_{i,2}(t)) - I^n_{i,2}(t)$, where $I^n_{i,2}(t)$ is the stochastic intensity used with the Poisson-process $N^s_{i,2}(t)$, e.g., $I^n_{1,2}(t) \equiv \mu_{1,2}\int_0^t 1_{\{D^n(s)<0\}}Z^n_{1,2}(s)\,ds$.

The fluid limit is a FWLLN. To express it, let $\mathcal{D}$ 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 [25, 78].

We next rewrite the equations in (3.2.5) by subtracting and adding the stochastic intensities, and then dividing each equation by n. It can be shown that $M^{n,a}_i/n \Rightarrow 0$, $M^{n,u}_i/n \Rightarrow 0$ and $M^{n,s}_{i,2}/n \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$, (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 $\mathcal{D}$ satisfies $Y^n = o_p(1)$ if $Y^n \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$. We thus have the associated

representation for the fluid-scaled queueing processes:

$$\begin{aligned}
\bar{Z}^n_{1,2}(t) &\equiv \bar{Z}^n_{1,2}(0) + \mu_{2,2}\int_0^t 1_{\{D^n(s)\geq 0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t 1_{\{D^n(s)<0\}}\bar{Z}^n_{1,2}(s)\,ds + o_p(1), \quad t \geq 0, \\
\bar{Q}^n_1(t) &\equiv \bar{Q}^n_1(0) + \bar{\lambda}^n_1 t - \bar{m}^n_1 t - \mu_{1,2}\int_0^t 1_{\{D^n(s)\geq 0\}}\bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{2,2}\int_0^t 1_{\{D^n(s)\geq 0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds - \theta_1\int_0^t \bar{Q}^n_1(s)\,ds + o_p(1), \quad t \geq 0, \\
\bar{Q}^n_2(t) &\equiv \bar{Q}^n_2(0) + \bar{\lambda}^n_2 t - \mu_{2,2}\int_0^t 1_{\{D^n(s)<0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t 1_{\{D^n(s)<0\}}Z^n_{1,2}(s)\,ds - \theta_2\int_0^t Q^n_2(s))\,ds + o_p(1), \quad t \geq 0.
\end{aligned} \tag{3.2.6}$$

The ODE we study is an approximation for the three-dimensional fluid-scaled process $\bar{X}^n \equiv (\bar{Q}^n_1, \bar{Q}^n_2, \bar{Z}^n_{1,2})$ with components defined in (3.2.6).

3.2.3 A Heuristic View of the AP

In fact, the ODE we study is the limit of the three-dimensional fluid-scaled process $\bar{X}^n \equiv (\bar{Q}^n_1, \bar{Q}^n_2, \bar{Z}^n_{1,2})$ with components defined in (3.2.6); i.e., in Chapter 4 we show that $\bar{X}^n \Rightarrow x$ in $\mathcal{D}_3$ as $n \to \infty$, 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 (3.2.6), 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 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 (3.2.4) is an order $O(1)$ process, without any spatial scaling, i.e., for each t, the sequence of unscaled random variables $\{D^n(t) : n \geq 1\}$ turns out to be stochastically bounded (or tight) in $\mathbb{R}$. That implies that D^n operates in a time scale that is different from the other processes Q_i^n and $Z_{1,2}^n$, which are scaled by dividing by n in (3.2.2) and (3.2.6). A heuristic explanation is that, with the MS-HT scaling in (3.2.1), in order for the two queues to change significantly (in a relative sense), which is captured by the scaling in (3.2.2), there needs to be $O(n)$ arrivals and departures from the queues. In contrast, the difference process D^n can never go far from 0, because it has drift pointing towards 0 from both above and below. Thus, the difference process oscillates more and more rapidly about 0 as n increases. 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

$$D_t^n \equiv D^n(X^n(t), s) \equiv \{D^n(t + s/n) : s \geq 0\}. \tag{3.2.7}$$

Dividing s by n in (3.2.7) 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 Chapter

4 is that, for each $t \geq 0$,

$$D_t^n \equiv \{D^n(X^n(t), s) : s \geq 0\} \Rightarrow D_t \equiv \{D(x(t), s) : s \geq 0\} \quad \text{in } \mathcal{D} \text{ as } n \to \infty, \quad (3.2.8)$$

where the limit $D_t \equiv \{D(x(t), s) : s \geq 0\}$ 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 (3.2.7), 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.,

$$\pi_{1,2}(x(t)) \equiv \lim_{s\to\infty} P(D(x(t), s) \geq 0) = \lim_{u\to\infty} \frac{1}{u} \int_0^u 1_{\{D(x(t),s)\geq 0\}}\, ds, \quad (3.2.9)$$

which depends on x because the distribution of $\{D(x(t), s) : s \geq 0\}$ depends on the value of $x(t) \in \mathbb{R}_3$.

To actually establish convergence for $\bar{X}^n$ in (3.2.6), we go further in Chapter 4 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 $n \geq n_0$,

$$\left| \frac{1}{\eta} \int_t^{t+\eta} 1_{\{D^n(X^n(t),s)\geq 0\}}\, ds - \pi_{1,2}(x(t)) \right| < \epsilon. \quad (3.2.10)$$

The local uniform convergence allows us to replace the indicator functions in the integrals in (3.2.6) with the $\pi_{1,2}$ functions in the fluid limit.

3.2.4 The Fluid-Limit ODE

The discussion in §§3.2.2 and 3.2.3 above is an outline of the convergence result in Chapter 4. A different approach appeared §4.2 of [59], where the ODE was developed directly,

assuming that the fluid limit exists, and is differentiable. The convergence $\bar{X}^n \Rightarrow x$, established in Chapter 4 based on the representation (3.2.6) together with the AP in (3.2.7)-(3.2.10), lead to the same ODE as in [59]. We now specify the ODE, which is the main subject of this chapter.

The general form of an ODE is $\dot{x}(t) = \Psi(x(t), t)$ for a function Ψ, 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

$$\dot{x}(t) \equiv (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t)) = \Psi(x(t)) \equiv \Psi(q_1(t), q_2(t), z_{1,2}(t)), \quad t \geq 0, \tag{3.2.11}$$

where $\Psi(x) : [0, \infty)^2 \times [0, m_2] \to \mathbb{R}_3$ can be displayed via

$$\begin{aligned}
\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))\left[z_{1,2}(t)\mu_{1,2} + z_{2,2}(t)\mu_{2,2}\right] - \theta_1 q_1(t) \\
\dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))\left[z_{2,2}(t)\mu_{2,2} + z_{1,2}(t)\mu_{1,2}\right] - \theta_2 q_2(t) \\
\dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{3.2.12}$$

with $\pi_{1,2} : [0, \infty)^2 \times [0, m_2] \to [0, 1]$ defined in (3.2.9).

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

$$\dot{x}(t) = \Psi(x(t)), \quad x(0) = w_0 \tag{3.2.13}$$

for $\Psi(x)$ in (3.2.11) - (3.2.12).

We remark that specifying the IVP in (3.2.11)-(3.2.13) does not fully characterize the limit of $\bar{X}^n$, given convergence of the initial conditions $\bar{X}^n(0) \to w_0$ w.p. 1, where $w_0 \geq 0$

is deterministic, as required in Chapter 4. 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 (3.2.13).

The proof of existence and uniqueness of a solution to (3.2.13), is tied to the characterization of $\pi_{1,2}$ in (3.2.12) and (3.2.9), 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.

3.3 The Fast-Time-Scale Process

Recall that the FTSP D_t is the limit of D_t^n without any scaling (see (3.2.8)), where D_t^n is the time-incremented difference process defined in (3.2.7) in terms of the queue-difference stochastic process $D^n \equiv (Q_1^n - \kappa^n) - rQ_2^n$ in (3.2.4). Since there is no scaling of space, the state space for the FTSP D_t is the countable lattice $\{\pm j \pm kr : j, k \in \mathbb{Z}\}$ in $\mathbb{R}$. To see this, first observe from (3.2.4) that D^n has state space $\{\pm j \pm kr - \kappa^n : j, k \in \mathbb{Z}\}$. Next, because of the subtraction in (3.2.7), D_t^n has state space $\{\pm j \pm kr : j, k \in \mathbb{Z}\}$. Finally, because of the convergence in (3.2.8), the FTSP D_t has this same state space.

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

Let $\lambda_+^{(j)}(m, x(t))$, $\lambda_+^{(k)}(m, x(t))$, $\mu_+^{(j)}(m, x(t))$ and $\mu_+^{(k)}(m, x(t))$ be the transition rates of the FTSMC D_t for transitions of $+j$, $+k$, $-j$ and $-k$, respectively, when $D(x(t), s) = m > 0$. Similarly, we define the transitions when $D(x(t), s) = m \leq 0$: $\lambda_-^{(j)}(m, x(t))$, $\lambda_-^{(k)}(m, x(t))$, $\mu_-^{(j)}(m, x(t))$ and $\mu_-^{(k)}(m, x(t))$. These rates are the limits of the rates of D_t^n as $n \to \infty$ with $\bar{X}^n(t) \Rightarrow x(t)$; convergence will be proved in Chapter 4.

First, for $D(x(t), s) = m \in (-\infty, 0]$, the upward rates are

$$\lambda_-^{(k)}(m, x(t)) = \lambda_1, \quad \text{and} \quad \lambda_-^{(j)}(m, x(t)) = \mu_{1,2} z_{1,2}(t) + \mu_{2,2} z_{2,2}(t) + \theta_2 q_2(t), \quad (3.3.1)$$

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

$$\mu_-^{(k)}(m, x(t)) = \mu_{1,1} z_{1,1}(t) + \theta_1 q_1(t) \quad \text{and} \quad \mu_-^{(j)}(m, x(t)) = \lambda_2, \quad (3.3.2)$$

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

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

$$\lambda_+^{(k)}(m, x(t)) = \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, x(t)) = \theta_2 q_2(t), \quad (3.3.3)$$

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

$$\begin{aligned}\mu_+^{(k)}(m, x(t)) &= \mu_{1,1}z_{1,1}(t) + \mu_{1,2}z_{1,2}(t) + \mu_{2,2}z_{2,2}(t) + \theta_1 q_1(t) \quad \text{and} \\ \mu_+^{(j)}(m, x(t)) &= \lambda_2,\end{aligned} \tag{3.3.4}$$

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

3.3.2 Representing the FTSMC D_t as a QBD

Further analysis is simplified by exploiting matrix geometric methods, as in [52]. In particular, we represent the integer-valued FTSMC $D_t \equiv \{D(x(t), s) : s \geq 0\}$ just constructed as a homogeneous continuous-time QBD, as in Definition 1.3.1 and §6.4 of [52]. 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 [52] (imitating the shape of a generator matrix of a birth-and-death process). In particular, we write

$$Q \equiv \begin{pmatrix} B & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{3.3.5}$$

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

written in block-triangular form composed of four $m \times m$ submatrices, i.e.,

$$B \equiv \begin{pmatrix} A_1^+ & B_\mu \\ B_\lambda & A_1^- \end{pmatrix} \quad \text{and} \quad A_i \equiv \begin{pmatrix} A_i^+ & 0 \\ 0 & A_i^- \end{pmatrix} \tag{3.3.6}$$

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, $n = 0, 1, 2, \ldots$ We assign original states $\phi(n)$ to positive integers n according to the mapping:

$$\phi(2nm + i) \equiv nm + i \quad \text{and} \quad \phi((2n+1)m + i) \equiv -nm - i + 1, \quad 1 \leq i \leq m. \tag{3.3.7}$$

Then we order the states in levels as follows

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

With this representation, the generator-matrix Q can be written in the form (3.3.5) above, where A_1 groups all the transitions within a level, A_0 groups the transitions from level $L(n)$ to level $L(n+1)$ and A_2 groups all transitions from level $L(n)$ to level $L(n-1)$. Matrix B groups the transitions within the boundary level $L(0)$, and is thus different than A_1.

To illustrate, consider an example with $r = 0.4$, so that we can choose $j = 2$ and $k = 5$,

yielding $m = 5$. The states are ordered in levels as follows

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

Then the submatrices B_μ, B_λ, A_i^+ and A_i^-, which form the block matrices B and A_i, $i = 0, 1, 2$, have the form in (3.3.9) with

$$\sigma_+ = \lambda_+^{(5)} + \lambda_+^{(2)} + \mu_+^{(5)} + \mu_+^{(2)} \quad \text{and} \quad \sigma_- = \lambda_-^{(5)} + \lambda_-^{(2)} + \mu_-^{(5)} + \mu_-^{(2)}. \tag{3.3.8}$$

(We solve a full numerical example with these matrices in §3.8.3.)

Henceforth in this chapter, we refer to D_t as the QBD, because this is the only QBD under consideration. However, we will consider other QBD's in Chapter 4. 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).

3.3.3 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 [52], 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 ν^+ and ν^- be the

$$B_\mu = \begin{pmatrix} 0 & \mu_+^{(2)} & 0 & 0 & \mu_+^{(5)} \\ \mu_+^{(2)} & 0 & 0 & \mu_+^{(5)} & 0 \\ 0 & 0 & \mu_+^{(5)} & 0 & 0 \\ 0 & \mu_+^{(5)} & 0 & 0 & 0 \\ \mu_+^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix} \qquad B_\lambda = \begin{pmatrix} 0 & \lambda_-^{(2)} & 0 & 0 & \lambda_-^{(5)} \\ \lambda_-^{(2)} & 0 & 0 & \lambda_-^{(5)} & 0 \\ 0 & 0 & \lambda_-^{(5)} & 0 & 0 \\ 0 & \lambda_-^{(5)} & 0 & 0 & 0 \\ \lambda_-^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_0^+ = \begin{pmatrix} \lambda_+^{(5)} & 0 & 0 & 0 & 0 \\ 0 & \lambda_+^{(5)} & 0 & 0 & 0 \\ 0 & 0 & \lambda_+^{(5)} & 0 & 0 \\ \lambda_+^{(2)} & 0 & 0 & \lambda_+^{(5)} & 0 \\ 0 & \lambda_+^{(2)} & 0 & 0 & \lambda_+^{(5)} \end{pmatrix} \qquad A_0^- = \begin{pmatrix} \mu_-^{(5)} & 0 & 0 & 0 & 0 \\ 0 & \mu_-^{(5)} & 0 & 0 & 0 \\ 0 & 0 & \mu_-^{(5)} & 0 & 0 \\ \mu_-^{(2)} & 0 & 0 & \mu_-^{(5)} & 0 \\ 0 & \mu_-^{(2)} & 0 & 0 & \mu_-^{(5)} \end{pmatrix}$$

$$A_1^+ = \begin{pmatrix} -\sigma_+ & 0 & \lambda_+^{(2)} & 0 & 0 \\ 0 & -\sigma_+ & 0 & \lambda_+^{(2)} & 0 \\ \mu_+^{(2)} & 0 & -\sigma_+ & 0 & \lambda_+^{(2)} \\ 0 & \mu_+^{(2)} & 0 & -\sigma_+ & 0 \\ 0 & 0 & \mu_+^{(2)} & 0 & -\sigma_+ \end{pmatrix} \qquad A_1^- = \begin{pmatrix} -\sigma_- & 0 & \mu_-^{(2)} & 0 & 0 \\ 0 & -\sigma_- & 0 & \mu_-^{(2)} & 0 \\ \lambda_-^{(2)} & 0 & -\sigma_- & 0 & \mu_-^{(2)} \\ 0 & \lambda_-^{(2)} & 0 & -\sigma_- & 0 \\ 0 & 0 & \lambda_-^{(2)} & 0 & -\sigma_- \end{pmatrix}$$

$$A_2^+ = \begin{pmatrix} \mu_+^{(5)} & 0 & 0 & \mu_+^{(2)} & 0 \\ 0 & \mu_+^{(5)} & 0 & 0 & \mu_+^{(2)} \\ 0 & 0 & \mu_+^{(5)} & 0 & 0 \\ 0 & 0 & 0 & \mu_+^{(5)} & 0 \\ 0 & 0 & 0 & 0 & \mu_+^{(5)} \end{pmatrix} \qquad A_2^- = \begin{pmatrix} \lambda_-^{(5)} & 0 & 0 & \lambda_-^{(2)} & 0 \\ 0 & \lambda_-^{(5)} & 0 & 0 & \lambda_-^{(2)} \\ 0 & 0 & \lambda_-^{(5)} & 0 & 0 \\ 0 & 0 & 0 & \lambda_-^{(5)} & 0 \\ 0 & 0 & 0 & 0 & \lambda_-^{(5)} \end{pmatrix}$$

(3.3.9)

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

$$\nu^+ A_0^+ \mathbf{1} < \nu^+ A_2^+ \mathbf{1} \quad \text{and} \quad \nu^- A_0^- \mathbf{1} < \nu^- A_2^- \mathbf{1}. \tag{3.3.10}$$

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

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

$$\begin{aligned} \delta_+(x(t)) &\equiv j\left(\lambda_+^{(j)}(x(t)) - \mu_+^{(j)}(x(t))\right) + k\left(\lambda_+^{(k)}(x(t)) - \mu_+^{(k)}(x(t))\right) \\ \delta_-(x(t)) &\equiv j\left(\lambda_-^{(j)}(x(t)) - \mu_-^{(j)}(x(t))\right) + k\left(\lambda_-^{(k)}(x(t)) - \mu_-^{(k)}(x(t))\right). \end{aligned} \tag{3.3.11}$$

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 (3.3.10):

Theorem 3.3.1. *The QBD D_t is positive recurrent if and only if*

$$\delta_-(x(t)) > 0 > \delta_+(x(t)). \tag{3.3.12}$$

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 §3.3.4 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, $\pi_{1,2}(x(t)) = 0$.

3.3.4 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 α_n^+ and α_n^- both

being $1 \times m$ vectors. Then the desired probability $\pi_{1,2}$ can be expressed as

$$\pi_{1,2} = \sum_{n=0}^{\infty} \sum_{j=1}^{m} \alpha_{n,j}^{+} = \sum_{n=0}^{\infty} \alpha_n^{+} \mathbf{1} = \sum_{n=0}^{\infty} \alpha_n \mathbf{1}_{+}, \tag{3.3.13}$$

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 [52], the steady-state distribution has the matrix-geometric form

$$\alpha_n = \alpha_0 R^n, \tag{3.3.14}$$

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 [52]), we have

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

Also, by Lemma 6.3.1 of [52], the boundary probability vector α_0 is the unique solution to the system

$$\alpha_0(B + RA_2) = 0 \quad \text{and} \quad \alpha \mathbf{1} = \alpha_0 (I - R)^{-1} \mathbf{1} = 1. \tag{3.3.15}$$

Finally, given the above, and using (3.3.13), we see that the desired quantity $\pi_{1,2}$ can be represented as

$$\pi_{1,2} = \alpha_0 (I - R)^{-1} \mathbf{1}_{+}, \tag{3.3.16}$$

where R is the $2m \times 2m$ *rate matrix* and α_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

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

and can be found efficiently by existing algorithms, as in [52] (see §3.8). 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 (3.3.12)) and then numerically calculate $\pi_{1,2}$. This allows us to numerically solve the ODE (3.2.11), as in §3.8. Moreover, we will use the representation (3.3.16), and results about the rate matrix R, to conclude topological properties of $\pi_{1,2}$.

3.4 Existence and Uniqueness of Solutions

We now start to analyze the ODE and IVP introduced in §3.2.4. In this section we show that a unique solution exists to the IVP (3.2.13) 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 Ψ in (3.2.12) to be (locally) Lipschitz continuous. However, Ψ 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 Ψ is indeed locally Lipschitz continuous in each of these regions.

3.4.1 Properties of Ψ

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 (3.2.4) 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 κ, equal to κ or below κ. 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:

$$\mathbb{S}^{\mathbf{b}} \equiv \{q_1 - rq_2 = \kappa\}, \quad \mathbb{S}^+ \equiv \{q_1 - rq_2 > \kappa\}, \quad \mathbb{S}^- \equiv \{q_1 - rq_2 < \kappa\}, \tag{3.4.1}$$

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

The boundary subset $\mathbb{S}^b$ is a hyperplane in the state space $\mathbb{S}$, and is therefore a closed subset. It is the subset of $\mathbb{S}$ 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 Ψ in (3.2.12) is Lipschitz continuous on $\mathbb{S}^+$ and $\mathbb{S}^-$, because $\pi_{1,2}(x) = 1$ when $x \in \mathbb{S}^+$, and $\pi_{1,2}(x) = 0$ when $x \in \mathbb{S}^-$, so that Ψ is linear in each region. However, Ψ is not linear on $\mathbb{S}^b$, so we must work harder there.

To analyze Ψ on $\mathbb{S}^b$, we exploit properties of the QBD introduced in §3.3. 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 (3.3.11). Let $\mathbb{A}$ be the set of all $x \in \mathbb{S}^b$ for which the QBD is positive recurrent, as given in (3.3.12); i.e., let

$$\mathbb{A} \equiv \{x \in \mathbb{S}^b \mid \delta_-(x) > 0 > \delta_+(x)\}. \tag{3.4.2}$$

From the continuity of the QBD drift-rates in (3.3.11), if follows that $\mathbb{A}$ is an open and connected subset of $\mathbb{S}^b$. Hence, $\mathbb{A}$ can be regarded as an open connected subset of $\mathbb{R}_2^+$ (since $\mathbb{S}^b$ is homoeomorphic to $\mathbb{R}^+ \times [0, m_2]$).

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 3.4.1. (local Lipschitz continuity) *A function $f : \mathbb{R}_n \to \mathbb{R}_m$ 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 3.4.1. *The function Ψ in* (3.2.12) *is locally Lipschitz continuous on $\mathbb{A}$.*

Proof: The key component of the function Ψ 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 (3.3.6) (see also the example in §3.3.2) and the definitions of the rates in (3.3.1)-(3.3.4), 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 [34] that the rate matrix R in (3.3.14), 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 (3.3.15) and the differentiability of R, that α_0 is also differentiable. Hence, from (3.3.16), 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 (3.3.15), we have

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

so that α_0' is continuous. The continuity of R' and α_0' implies that the derivative $\pi_{1,2}'$ is continuous on $\mathbb{A}$, which in turn implies that the derivative Ψ' is continuous on $\mathbb{A}$. That in turn implies that Ψ 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 [45]. Finally, since a continuous function on a compact set is bounded, Ψ satisfies this property. Hence Ψ is indeed locally Lipschitz continuous. ■

3.4.2 Solution to the ODE

The local Lipschitz continuity of Ψ 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 (3.2.13); e.g., see Theorem 2.2 of Teschl [68].

Theorem 3.4.2. (local existence and uniqueness) *If $w_0 \in \mathbb{A}$, then there exists a unique solution $x : [0, \delta) \to \mathbb{A}$ to the* IVP (3.2.13) *for some $\delta > 0$.*

Proof: By the classical Picard-Lindelöf theorem, Theorem 2.2 of Teschl [68] or Theorem 3.1 in [45], and Theorem 3.4.1, 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 δ_- and δ_+ are continuous functions, there necessarily exists δ with $0 < \delta \leq \delta_1$ such that $x(t) \in \mathbb{A}$ for all $t \in [0, \delta)$.

■

We now give sufficient conditions for the existence of a unique solution to the IVP (3.2.13) 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 (3.2.13) 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 3.4.3. (boundedness) *Every solution to the* IVP (3.2.13) *is bounded. In particular, the following upper bounds for the fluid queues hold:*

$$q_i(t) \leq q_i(0) \vee \lambda_i/\theta_i \quad t \geq 0, \quad i = 1, 2. \tag{3.4.3}$$

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 (3.4.3). 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

$$u_i(t) = \frac{\lambda_i}{\theta_i} + \left(u_i(0) - \frac{\lambda_i}{\theta_i}\right) e^{-\theta_1 t}, t \geq 0.$$

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 (3.4.3) 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

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

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

$$q_i(t) > u_i(t) \quad \text{for all} \quad t_1 < t < t_2. \tag{3.4.4}$$

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

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

Thus,

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

so that $q_1(t_3) < u_1(t_3)$, contradicting (3.4.4). A similar argument holds for q_2. ∎

Theorem 3.4.4. (global existence and uniqueness) *Let x be the unique solution to the* IVP (3.2.13) *on an interval $[0, \delta)$, established by Theorem* 3.4.2. *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 (3.2.13) *on $[0, \infty)$.*

In the proof of Theorem 3.4.4 we make use of the next lemma. For its proof see Theorem 3.3 in [45].

Lemma 3.4.1. *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 3.4.4: By Theorem 3.4.1, Ψ is locally Lipschitz continuous, and by Theorem 3.4.3, a solution to the IVP (3.2.13) is bounded. It follows from Lemma 3.4.1 that there exists a unique solution to (3.2.13) for all $t \geq 0$. ∎

In Section §3.6 we give sufficient conditions for the solution of the IVP (3.2.13) to lie entirely in $\mathbb{A}$, which by Theorem 3.4.4 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 §3.6.2, 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 3.4.2 to conclude that there exists a unique solution to the IVP (3.2.13) for all $t \geq 0$.

Remark 3.4.1. Theorems 3.4.1-3.4.4 also hold for solutions to the IVP (3.2.13) in $\mathbb{S}^-$ and $\mathbb{S}^+$. Indeed, they are elementary, because Ψ is Lipschitz continuous, since $\pi_{1,2}$ is constant in these regions. The boundedness used in the proof of Theorem 3.4.4, and proved in Theorem 3.4.3, applies in these two regions as well.

3.5 Fluid Stationarity

Our initial analysis of the overloaded X model in Chapter 2 was using a steady state (or stationary) fluid analysis. That is, we assumed that there exists a unique stationary point x^*

and that $x(t) \to x^*$ as $t \to \infty$ 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 (3.2.12). We then give conditions under which the fluid solution $x \equiv \{x(t) : t \geq 0\}$ converges to stationarity as $t \to \infty$. In §3.6, we show that it does so exponentially fast.

Definition 3.5.1. (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 Ψ in* (3.2.11) *and* (3.2.12)*. 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 λ_i, service-rate $\mu_{i,i}$ and abandonment-rate θ_i; see Theorem 2.3 in [79], especially equation (2.19). In particular,

$$q_i^a \equiv \frac{(\lambda_i - \mu_{i,i} m_i)^+}{\theta_i} \quad \text{and} \quad s_i^a \equiv \left(m_i - \frac{\lambda_i}{\mu_{i,i}} \right)^+, \quad i = 1, 2, \tag{3.5.1}$$

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 (3.2.12) 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 κ can be equal to zero). We thus require that all the class-1 fluid, *that is not served in pool* 1, minus the minimal amount of class-1 fluid that abandons, is larger than $\mu_{1,2}s_2^a$. With this requirement, pool 2 is assured to be full, assuming that it is initialized full. (If pool 2 is not initialized full, then it will fill up after some finite time period; see §3.7.)

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 in this chapter.**

Assumption A. ($\mathit{system\ overload,\ with\ class}\ 1\ \mathit{more\ overloaded}$)
Exactly one of the following must hold:

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

(II) $q_1^a \leq \kappa$ *and* $s_1^a = s_2^a = 0$.

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 [79].

It is significant that Assumption A involves only the parameters of the system, and requires no knowledge on the specific solution to the IVP (3.2.13). We will show that when this assumption holds, there exists a unique stationary point in $\mathbb{S}$ for every solution to (3.2.12).

We will use a different version of this assumption in Chapter 4, where we consider only limits in $\mathbb{A}$. Since we will want the system to be genuinely overloaded, Condition (I) will be slightly strengthen by assuming the the inequality is strong. See Assumption 1 in §4.3.

3.5.1 Uniqueness of the Stationary Point

By definition, a stationary point $x^* \in \mathbb{S}$ is such that $\Psi(x^*) = 0$. From (3.2.12), we see that this gives a system of three equations with three unknowns, namely, q_1^*, q_2^* and $z_{1,2}^*$. The apparent fourth variable $\pi_{1,2}^* \equiv \pi_{1,2}(x^*)$ is a function of the other three variables and its value is determined by x^*. In principle, the three equations in $\Psi(x) = 0$ can be solved directly to find all the roots of Ψ. However, $\pi_{1,2}^*$ is a complicated function of x^* having the complicated closed-form expression in (3.3.13) and (3.3.16).

Theorem 3.5.1 below states that *if there exists a stationary point for the fluid ODE* (3.2.12), then this point is unique, and must have the specified form. The uniqueness of x^* is proved by treating $\pi_{1,2}^*$ as a fourth variable, and adding a fourth equation to the three equations $\Psi(x) = 0$. However, it does not prove that a stationary point exists. In general, the solution $\pi_{1,2}^*$ we get from the system of four equations may not equal to $\pi_{1,2}(x^*)$, for the function $\pi_{1,2}$ defined in (3.2.9). The existence of a stationary point is more involved, and is proved later; See Corollary 3.5.6.

The proof of existence is immediate from the proof of uniqueness when $\pi_{1,2}(x^*)$ is

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

We now focus on uniqueness. Although $\pi^*_{1,2}$ is treated as a variable, we still impose conditions on it so that it can be a legitimate solution to (3.2.9). In particular, if $q^*_1 - rq^*_2 > \kappa$ then we let $\pi^*_{1,2} = 1$; if $q^*_1 - rq^*_2 < \kappa$, then we let $\pi^*_{1,2} = 0$. Equation (3.5.4) below shows that $0 \leq \pi^*_{1,2} \leq 1$ whenever $q^*_1 - rq^*_2 = \kappa$, i.e., whenever $x^* \in \mathbb{S}^b$.

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

$$z \equiv \frac{\theta_2(\lambda_1 - m_1\mu_{1,1}) - r\theta_1(\lambda_2 - m_2\mu_{2,2}) - \theta_1\theta_2\kappa}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}. \tag{3.5.2}$$

Theorem 3.5.1. (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* (3.2.13)*, which for z in* (3.5.2) *must take the form*

$$z^*_{1,2} = 0 \vee z \wedge m_2, \quad q^*_1 = \frac{\lambda_1 - m_1\mu_{1,1} - \mu_{1,2}z^*_{1,2}}{\theta_1}, \quad q^*_2 = \frac{\lambda_2 - \mu_{2,2}(m_2 - z^*_{1,2})}{\theta_2}. \tag{3.5.3}$$

Moreover,

$$\pi^*_{1,2} = \frac{\mu_{1,2}z^*_{1,2}}{\mu_{1,2}z^*_{1,2} + (m_2 - z^*_{1,2})\mu_{2,2}}. \tag{3.5.4}$$

Proof: We start with (3.5.4). This expression is easily derived from the third equation in (3.2.12), by equating $\dot{z}_{1,2}(t)$ to zero. Observe that if $z^*_{1,2} = m_2$ then $\pi^*_{1,2}$ in (3.5.4) 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 (3.2.12) we get the expressions of q^*_1 and q^*_2 in (3.5.3). 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.)

$$q_1(z) = \frac{\lambda_1 - \mu_{1,1}m_1 - \mu_{1,2}z}{\theta_1}, \quad q_2(z) = \frac{\lambda_2 - \mu_{2,2}(m_2 - z)}{\theta_2}, \quad \kappa = q_1(z) - rq_2(z). \tag{3.5.5}$$

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 (3.5.2). We can recover x^* from the solution to (3.5.5), 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 (3.5.5). 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 $q_1(m_2) - rq_2(m_2) > \kappa$ (and, trivially, $q_1(m_2) > \kappa$, $q_2(m_2) > 0$). This implies that $x^* \equiv (q_1(m_2), q_2(m_2), m_2) \in \mathbb{S}^+$ and, if it is indeed a solution to $\Psi(x) = 0$, then x^* is the unique stationary point for the ODE.

Now assume that the unique solution to (3.5.5) has $z < 0$. By replacing z with 0 we have $q_1(0) < q_1(z)$ and $q_2(0) > q_2(z)$, which imply that $q_1(0) - rq_2(0) < \kappa$. 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 (3.5.5) 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 (4.2.8). We start by rewriting $q_1(z)$ and $q_2(z)$ in (3.5.5) as

$$q_1(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1} z, \quad q_2(z) = \frac{\mu_{2,2}}{\theta_2}(z - S_2^a). \tag{3.5.6}$$

Now, it follows from Assumption A that

$$\kappa \leq q_1^a - \frac{\mu_{1,2}}{\theta_1} s_2^a \leq q_1^a - \frac{\mu_{1,2}}{\theta_1} S_2^a, \tag{3.5.7}$$

where the second inequality follows trivially, since $S_2^a \leq s_2^a$. From the third equation of (3.5.5), $\kappa = q_1(z) - r q_2(z)$. Combining this with (3.5.6), we see that

$$\kappa = q_1(z) - r q_2(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1} z - r \frac{\mu_{2,2}}{\theta_2}(z - S_2^a). \tag{3.5.8}$$

Combining (3.5.7) and (3.5.8), 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 (3.5.6) that $q_2(z) \geq 0$ and, by using the third equation in (3.5.5) again, $q_1(z) = r q_2(z) + \kappa \geq \kappa \geq 0$. ∎

An immediate consequence of the proof of Theorem 3.5.1 is that, in order to find the candidate stationary point x^*, one has to solve the three equations in (3.5.5). 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 3.5.2. *Let $x^* = (q_1^*, q_2^*, z_{1,2}^*)$ be the point defined in Theorem 3.5.1.*

1. *If $x^* \in \mathbb{S}^b$, then, for z defined in* (3.5.2),

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

2. *If $x^* = \mathbb{S}^+$, then*

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

3. *If $x^* \in \mathbb{S}^-$, then*

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

Proof: If $x^* \in \mathbb{S}^b$, then the solution to (3.5.5) 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 $q_1^* - rq_2^* > \kappa$, so that $\pi^*_{1,2} = 1$. Plugging $\pi^*_{1,2} = 1$ in the ODE for $z_{1,2}(t)$ in (3.2.12), 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 $q_1^* - rq_2^* < \kappa$, then $\pi_{1,2}^* = 0$, so that, by plugging this value of $\pi_{1,2}^*$ in the ODE for $z_{1,2}(t)$ in (3.2.12), 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). ∎

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

As we remarked at the beginning of this subsection, from the proofs of Theorem 3.5.1 and Corollary 3.5.2, and from the expression of π^* in (3.5.4), it is clear that x^* is a stationary point for the ODE (3.2.12) 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 3.5.3. *Let x^* be as in* (3.5.3)*. Then*

1. *$x^* \in \mathbb{S}^b$ if and only if*

$$\frac{\mu_{1,2} s_2^a}{\theta_1} \vee r q_2^a \;\le\; q_1^a - \kappa \;\le\; \frac{r\lambda_2}{\theta_2} + \frac{\mu_{1,2} m_2}{\theta_1}; \tag{3.5.9}$$

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

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

3. $x^* \in \mathbb{S}^-$ *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 3.5.2 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 (3.5.9).

Now Assume that (3.5.9) holds. It follows from the right-hand-side (RHS) inequality and the expression of z in (3.5.2) 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 (3.5.9), we see that, if $s_2^a = 0$ (and necessarily $q_2^a \geq 0 = s_2^a$), then

$$z \geq \frac{\theta_1\theta_2 rq_2^a - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} = 0.$$

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 (3.5.9) holds, then $s_2^a \leq z \leq m_2$. This was shown to to imply that $x^* \in \mathbb{S}^b$ in the proof of Theorem 3.5.1. (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 (3.5.9) 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 §3.6.2; see (3.6.5). It then follows that (3.3.12) holds at x^* if and only if $0 < z^* < m_2$.

Alternatively, in Corollary (3.5.6) we show that $\pi^*_{1,2}$ in (3.5.4) is indeed the value of (3.2.9) at the point x^*. It is easy to see that $0 < \pi^*_{1,2} < 1$ in (3.5.4) if and only if $0 < z^* < m_2$. ■

Remark 3.5.1. It follows from Corollary 3.5.3 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 $\mathbb{S}^+$. Consider the following example: There are 100 servers in each pool, serving their own class at rates $\mu_{1,1} = \mu_{2,2} = 1$. Type-2 servers serve class-1 customers at rate $\mu_{1,2} = 0.8$. Also, $\theta_1 = \theta_2 = 0.3$, $r = 0.8$ and $\kappa = 0$. Suppose that class 2 is not overloaded with $\lambda_2 = 90$. Then, for the stationary point to be in $\mathbb{S}^+$, we need to have $\lambda_1 > \mu_{1,1}m_1 + \mu_{1,2}m_2 + \theta_1 r\lambda_2/\theta_2 = 252$, i.e., the class-1 arrival rate is 252% larger than the total service rate of pool 1. If $\lambda_2 > 90$, especially if pool 2 is also overloaded, then λ_1 needs to be even larger than that.

3.5.2 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 $t \to \infty$. It does

not even guarantee that a solution to the IVP (3.2.13) is asymptotically stable in the sense that, if $\|x(0) - x^*\| < \epsilon$, then $x(t) \to x^*$ as $t \to \infty$, no matter how small ϵ is. In fact, there is not even a guarantee that $x(t)$ will remain in the ϵ-neighborhood of x^* for all $t \geq 0$. We will establish all of these properties in Theorem 3.5.4 below by showing that x^* in §3.5.1 is globally asymptotically stable, as defined below:

Definition 3.5.2. (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*

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

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 (3.5.3) 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 (3.2.12); i.e., x^* is a stationary point and $x(t) \to x^*$ as $t \to \infty$. (So far, We are unable to establish global asymptotic stability for x^* in the boundary region $\mathbb{S}^b - \to$.)

Theorem 3.5.4. (global asymptotic stability of x^*) *If the solution to* (3.2.13) *lies entirely in one of the regions $\mathbb{S}^+$, $\mathbb{S}^-$ or $\mathbb{A}$, then x^* in Theorem* 3.5.1 *is globally asymptotically stable.*

The proof of Theorem 3.5.4 relies on results from nonlinear-control theory for deterministic dynamical systems, specifically, Lyapunov stability theory; for background, see Chapter 4 of Khalil [45]. 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 3.5.3. (Lie derivative) *For a continuously differentiable function $V : E \to \mathbb{R}$, and a function $\Psi : E \to \mathbb{R}^n$, the Lie derivative of V along Ψ is defined by*

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

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

Definition 3.5.4. (Lyapunov-function candidate) *A continuously differentiable function $V : E \to \mathbb{R}$ is a Lyapunov-function candidate if:*

1. $V(0) = 0$
2. $V(x) > 0$ *for all x in $E - \{0\}$*

In proving Theorem 3.5.4 we use the following theorem, which is Theorem 4.2 pg. 124 in [45]:

Theorem 3.5.5. (global asymptotic stability for nonlinear ODE) *Let $x = 0$ be a stationary point of $\dot{x} = \Psi(x)$, $\Psi : E \to \mathbb{R}^n$, and let $V : \mathbb{R}^n_+ \to \mathbb{R}$ be a Lyapunov-function candidate. If*

1. $V(x) \to \infty$ *as $||x|| \to \infty$ and*
2. $\dot{V}(x) < 0$ *for all $x \neq 0$,*

then $x = 0$ is globally asymptotically stable as in Definition 3.5.2.

Notice that, under the conditions of Theorem 3.5.5, 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 §3.6.2. We say that $\beta_V(\alpha)$ is the α V-ball with center at x^* and radius α if

$$\beta_V(\alpha) \equiv \{x \in \mathbb{R}_n : \|V(x) - V(x^*)\| \leq \alpha\}. \tag{3.5.10}$$

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 $x(t) \to x^*$ as $t \to \infty$ for all initial values $x(0)$. We also establish this stronger "V-monotonicity."

proof of Theorem 3.5.4: Let $x \equiv \{x(t) : t \geq 0\}$ be the unique solution to (3.2.13), 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 (3.2.11), 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 (3.5.3). 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 (3.2.12), 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 (3.5.3)

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

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

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

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

Hence,

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

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

By Theorem 3.5.5, $y^* = 0$ is globally asymptotically stable for the modified system $g(y)$. Hence, x^* is globally asymptotically stable for the original system $\Psi(x)$. That is, for every initial value $x(0)$ we have that $x(t) \to x^*$, provided that x is in the same region ($\mathbb{S}^+$, $\mathbb{S}^-$ or $\mathbb{A}$) for all $t \geq 0$. ∎

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

Corollary 3.5.6. (existence and uniqueness of a stationary point) *Under Assumption A, there exists a unique stationary point x^* in $\mathbb{S}$ for the ODE in* (3.2.11) *and* (3.2.12)*, with x^* defined in* (3.5.3)*. As a consequence, we have $\pi_{1,2}(x^*) = \pi^*_{1,2}$ for $\pi_{1,2}$ in* (3.2.9) *and $\pi^*_{1,2}$ in* (3.5.4)*.*

Proof: Uniqueness of a stationary point for the ODE in (3.2.12) was fully treated in §3.5.1, so it suffices to consider only existence. We already observed after the proof of Corollary 3.5.2 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 3.5.4 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. ∎

3.6 Conditions for State-Space Collapse

Both our result establishing global existence and uniqueness of a solution x to the IVP (3.2.13) (Theorem 3.4.4) and our result establishing global asymptotic stability of the stationary point x^* to the ODE (3.2.11) (Theorem 3.5.4) 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 §3.7 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 (3.2.12) up to some finite time T (which is shown to be not very large).

3.6.1 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 3.6.1. (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*

$$q_2(0) \leq \lambda_2/\theta_2 \quad \textit{and} \quad q_1(0) \leq (\lambda_1 - m_1\mu_{1,1})/\theta_1. \tag{3.6.1}$$

If, in addition, the following inequalities are satisfied, then the solution to the IVP (3.2.13) *is in* $\mathbb{A}$ *for all* t*:*

$$\begin{aligned} (i) \quad \lambda_1 &< \nu m_2 + m_1\mu_{1,1} \quad \textit{and} \\ (ii) \quad \lambda_2 &< \nu m_2 \end{aligned} \tag{3.6.2}$$

Remark 3.6.1. The rate conditions in (3.6.2) 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 λ_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 3.6.1: We start by showing, under Condition (i), that $\delta_+(x(t))$ in (3.3.11) is strictly negative for each t. For a fixed t

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

if and only if

$$(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) - m_2\mu_{2,2} < -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t). \quad (3.6.3)$$

If $\mu_{2,2} > \mu_{1,2}$, then the left-hand side (LHS) of (3.6.3) is maximized at $z_{1,2}(t) = m_2$, and is equal to $-\mu_{1,2}m_2$. If $\mu_{2,2} < \mu_{1,2}$, the the LHS is maximized at $z_{1,2}(t) = 0$, and is equal to $-\mu_{2,2}m_2$. When $\mu_{2,2} = \mu_{1,2}$ the LHS is equal to $-\mu_{2,2}m_2 = -\mu_{1,2}m_2$. Overall, the LHS of (3.6.3) is smaller than or equal to $-\nu m_2$.

Since $q_2(0) \leq \lambda_2/\theta_2$, we conclude, using the bound in (3.4.3), that $\theta_2 q_2(t) \leq \lambda_2$ for all $t \geq 0$. This, together with the fact that $q_1(t) \geq 0$ for all t, implies that the RHS of (3.6.3) is larger than or equal to $-(\lambda_1 - m_1\mu_{1,1})$, so that

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

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

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

note that, for $\delta_-(x(t))$ in (3.3.11), 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

$$r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\mu_{2,2}m_2 > -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t). \quad (3.6.4)$$

It is easy to see that the LHS of (3.6.4) has a minimum value of $r(\mu_{1,2} \wedge \mu_{2,2})m_2 \equiv r\nu m_2$. By essentially the same arguments as in Theorem 3.4.3 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 (3.6.4) is smaller than or equal to $r\lambda_2$. Overall, we have

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

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

Since (3.3.12) 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 (3.2.13) must lie entirely in $\mathbb{A}$. ∎

Combining Theorems 3.4.4, 3.5.4 and 3.6.1, we have the following corollary providing sufficient conditions for all good results discussed so far:

Corollary 3.6.2. *If* (3.5.9) *holds with strict inequalities,* $x(0) \in \mathbb{A}$ *and the four inequalities in Theorem* 3.6.1 *hold, then* (i) *there exists a unique solution* x *to the IVP* (3.2.13) *which lies entirely in* $\mathbb{A}$ *and* (ii) *there exists a unique stationary point* x^* *to the ODE* (3.2.12) *which is globally asymptotically stable. That stationary point* x^* *is given in Corollary* 3.5.3.

3.6.2 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 $t \to \infty$. 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 α V-ball with center at x^* and radius α in (3.5.10). 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 $\beta_V(\alpha) \subseteq \to$. 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 α, we introduce the drift rates at stationarity, $\delta_+^* \equiv \delta_+(x^*)$ and $\delta_-^* \equiv \delta_-(x^*)$. It follows from the expressions in (3.3.11) that

$$\delta_+^* \equiv \delta_+(x^*) = -\mu_{2,2}(r+1)(m_2 - z_{1,2}^*) \quad \text{and} \quad \delta_-^* \equiv \delta_-(x^*) = \mu_{1,2}(r+1)z_{1,2}^*. \tag{3.6.5}$$

Thus, if $0 < z_{1,2}^* < m_2$, then the positive recurrence condition (3.3.12) holds at the stationary point x^*. (This agrees with (3.5.4) which has $0 < \pi_{1,2}^* < 1$ if and only if $0 < z_{1,2}^* < m_2$.)

In the next theorem we give explicit expressions for α. Observe that for reasonable rates, such as will hold in applications, α 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 §3.8.3 we show that, typically in applications, α is so large, that we can infer that x lies entirely in $\mathbb{A}$ without even solving the IVP! That is, the initial condition is already in the V-ball $\beta_V(\alpha)$.

Theorem 3.6.3. *Suppose that $x^* \in \mathbb{A}$ and let $\xi \equiv \min\{|\delta_+^*|, \delta_-^*\}$.*

1. *When $\mu_{2,2} \geq \mu_{1,2}$, let $\alpha = \xi / r\theta_2$*

2. *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 3.5.2 is a globally asymptotically stable stationary point.

Proof: To find a proper α for the V-ball $\beta_V(\alpha)$, we once again use the conditions (3.6.3) and (3.6.4). We first show how to find α for the case $\mu_{2,2} = B\mu_{1,2}$ for some $B \geq 1$, i.e., when $\mu_{1,2} \leq \mu_{2,2}$. Recall (proof of Theorem 3.5.4) 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

$$\begin{aligned} q_1^* - \frac{\epsilon_1}{C} < q_1(t) < q_1 + \frac{\epsilon_1}{C}, \qquad q_2^* - \epsilon_1 < q_2(t) < q_2^* + \epsilon_1 \quad \text{and} \\ z_{1,2}^* - \frac{\epsilon_1}{C-1} < z_{1,2}(t) < z_{1,2}^* + \frac{\epsilon_1}{C-1}, \quad t \geq T. \end{aligned} \tag{3.6.6}$$

To make sure $\delta_+(x(t)) < 0$, we use (3.6.3), reorganizing the terms. We need to have

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

By (3.6.6), the above inequality holds if

$$(\mu_{2,2}-\mu_{1,2})\left(z^*_{1,2}+\frac{\epsilon_1}{C-1}\right)+r\theta_2(q^*_2+\epsilon_1)-\theta_1\left(q^*_1-\frac{\epsilon_1}{C}\right)<-(\lambda_1-\mu_{1,1}m_1)+r\lambda_2+\mu_{2,2}m_2.$$

Plugging in the expressions for q^*_1, q^*_2 and $z^*_{1,2}$, we see that we need to find an $\epsilon_1>0$ such that

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

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

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

Similarly, to make sure that $\delta_-(x(t))>0$, we use (3.6.4), 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 (3.6.6) again (with a different ϵ_2), we see that it suffices to show that

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

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

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

Hence, we can take α as in (i).

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

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

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

3.6.3 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 (3.5.10) for an α identified in Theorem 3.6.3. It is easy to see that if $z^*_{1,2}$ is not too close to 0 or m_2, then α is relatively large, so that numerical issues do not rise. However, we want to know that the time T at which the solution enters this α-neighborhood of x^* should not be too large.

Definition 3.6.1. (exponential stability) *A stationary point x^* is said to be (globally) exponentially stable if there exist two real constants ϑ, $\beta > 0$ such that*

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

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

To show that x^* in (3.5.3) is exponentially stable, we use Theorem 3.4 on p. 82 of Marquez [50], which we state here for completeness.

Theorem 3.6.4. (exponential stability of the origin) *Suppose that all the conditions of Theorem 3.5.5 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

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

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 3.6.5. (exponential stability of x^*) *If the entire trajectory of the solution to the IVP* (3.2.13) *is in $\mathbb{A}$, then x^* in* (3.5.3) *is exponentially stable, and the following hold:*

1. *If $\mu_{1,2} > \mu_{2,2}$, then*

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

where

$$K_3 \equiv \max\{\theta_1, \theta_2, \mu_{1,2} - \mu_{2,2}\}. \tag{3.6.7}$$

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

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

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 3.5.5 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 (3.6.7), and the result follows from Theorem 3.6.4.

(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 3.5.5 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 $\dot{V}_2(x) \leq -K_4\|x\|$. ∎

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 3.7.1 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.

3.7 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

$$q_1(0) = q_2(0) = z_{1,2}(0) = z_{2,1}(0) = 0. \tag{3.7.1}$$

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 (3.2.11)-(3.2.12) 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 (3.4.1) in §3.4, but the assumption that the service pools are both full was introduced at the beginning of §3.2.2.) 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 (3.2.11) 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 (3.7.1), we should certainly hit $\mathbb{S}^b$. Sharing will begin only at a time T such that $q_1(T) - rq_2(T) = \kappa$. In this section we show that, if indeed $x^* \in \mathbb{A} \cup \mathbb{S}^+$, then $T < \infty$, where

$$T \equiv \inf\{t \geq 0 : x(t) \in \mathbb{S}^b\}. \tag{3.7.2}$$

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 (3.2.12). 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 (3.7.1), are given in the proof of Theorem 3.7.1 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 3.7.1.

Theorem 3.7.1. *If* $x^* \in \mathbb{A} \cup \mathbb{S}^+$, *if* (3.7.1) *holds and if* $z_{2,1}(t) \equiv 0$ *for all* $t \geq 0$, *then* $T < \infty$, *for* T *in* (3.7.2).

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 λ_i and service rate $\mu_{i,i}$, $i = 1, 2$; e.g., see [57] (where it assumed that $\lambda = \mu$, so that $\lambda/\mu = 1$). Therefore, the system evolution is described by the pair of ODE's

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

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

$$t_1 \equiv \min_{i=1,2} \inf\{t \geq 0 : z_{i,i}(t) = m_i\}. \tag{3.7.3}$$

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

$$t_2 \equiv \inf\{t \geq t_1 : q_1(t) = \kappa\}. \tag{3.7.4}$$

Recall that κ 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 [79]. The ODE describing the evolution of q_1 is

$$\dot{q}_1(t) = \lambda_1 - \mu_{1,1} m_1 - \theta_1 q_1(t), \quad t_1 \leq t < t_2, \quad \text{with} \quad q_1(t_1) = 0, \tag{3.7.5}$$

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

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

whose unique solution is

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

Hence, $t_3 \equiv \inf\{t \ge t_2 : z_{1,2}(t) + z_{2,2}(t) = m_2\}$, 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 (3.7.2).

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 $t_1 = \inf\{t \ge 0 : z_{2,2} = m_2\}$, for t_1 in (3.7.3). In that case q_2 begins to increase at time t_1, evolving according to the ODE of the overloaded Erlang-A model

$$\dot{q}_2(t) = \lambda_2 - \mu_{2,2}m_2 - \theta_2 q_2(t).$$

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 $t_1' \ge t_1$ at which $z_{1,1} = m_1$. Then q_1 increases, satisfying (3.7.5) with $q_1(t_1') = 0$. By the assumption on x^*, and following Corollary 3.5.3, there exists a time $T < \infty$ such that $q_1(T) - rq_2(T) = \kappa$. This is because $rq_2(t) \le rq_2^a < q_1^a - \kappa$ for all $t \le T$. On the other hand, it follows trivially from the solution to (3.7.5), that q_1^a is the globally asymptotically stable point of (3.7.5). Hence, for every $\epsilon > 0$, there exists t_ϵ such that $q_1(t) > q_1^a - \epsilon$ for all $t \ge t_\epsilon$. (This is because, by the initial conditions, $q_1(t) \le q_1^a$ for all t). Thus, we can find $\epsilon > 0$ such that

$$rq_2^a < q_1^a - \epsilon - \kappa < q_1(t) - \kappa \text{ for all } t \ge t_\epsilon. \tag{3.7.6}$$

The second scenario of the second case has pool 1 filled first at time t_1, so that q_1 starts increasing according to (3.7.5). If q_1 reaches κ before q_2 starts increasing, then we have the same behavior as when $s_2^a > 0$. However, if at time t_2 in (3.7.4) $q_2 > 0$, then the two queues will continue increasing independently until time T. Once again, (3.7.6) can be shown to hold, so that $T < \infty$. ∎

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 (3.3.12) holds at T, so that $x(T) \in \mathbb{A}$.

Remark 3.7.1. (*sharing in the wrong direction*) In Theorem 3.7.1 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

$$z_{2,1}(t) = z_{2,1}(t_5) e^{-\mu_{2,1}(t-t_5)}, \quad t \geq t_5. \tag{3.7.7}$$

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 3.7.1, 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 (3.7.5), 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 3.7.1 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 (3.2.11), 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.

3.8 A Numerical Algorithm to Solve the IVP

In this section we provide a numerical algorithm for solving the IVP (3.2.13). 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 §3.6.2 (when the sufficient conditions for global stability in §3.6.1 do not hold) depends on calculating the solution up to a finite time T, where we an observe that the solution is close enough to the stationary point x^*, for which an explicit expression is given in §3.5. 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 [59].

3.8.1 Computing $\pi_{1,2}(x)$ at a point x

In §3.3.2 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 (3.3.16) 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 α_0 are functions of x.)

We exploit the well-developed theory for QBD processes in Latouche and Ramaswami [52]. By Proposition 6.4.2 of [52], the matrix R is related to two other matrices, G and U, via

$$G = (-U)^{-1}A_2, \quad U = A_1 + A_0G \quad \text{and} \quad R = A_0(-U)^{-1}. \tag{3.8.1}$$

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

$$A_2 + A_1 G + A_0 G^2 = 0 \quad \text{and} \quad A_0 + R A_1 + R^2 A_2 = 0. \tag{3.8.2}$$

Hence, if can compute the matrix G, then the rate matrix R can be found via (3.8.1). Once R is known, we use (3.3.15) to compute α_0. With α_0 and R in hand, $\pi_{1,2}(x)$ is easily computed via (3.3.16).

It remains to compute the matrix G. In §8 of [52], 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 [52]. 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 (3.2.13). From our experience with this algorithm, it takes fewer than ten iterations to achieve a 10^{-6} precision (when calculating G).

3.8.2 Computing the Solution x

To compute the solution $\{x(t) : 0 \le t \le T\}$, 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 $\bar{D}(0) \equiv (q_1(0) - \kappa) - r q_2(0) > 0$ 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 (3.3.12) 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 3.4.4 and Remark 3.4.1. 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 $x(2h)$, $x(3h)$, $\ldots x(nh)$,

$$x((k+1)h) = x(kh) + h\Psi(x(kh)), \quad 0 \leq k \leq n, \tag{3.8.3}$$

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

If $z_{1,1}(0) < m_1$ or $z_{1,2}(0) + z_{2,2}(0) < m_2$, so that $x(0) \notin \mathbb{S}$, we use the appropriate fluid model for the alternative region, as specified in §3.7, 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 (3.8.3) by the different method. At any iteration, $\pi_{1,2}$ is computed as in §3.8.1. 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 §3.6.3), which is very smooth and stable. Hence, we have no problem with numerical stability with the Euler method.

In the numerical example in §3.8 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 10×10. It took less than 10 seconds for the algorithm to terminate (using a relatively slow, 1 GB memory, laptop). The same example, but with $r = 20/25$, so that the matrices are now 50×50, 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 3.2 in the following example).

The validity of the solution can be verified by comparing it to simulation results. See the example below. See also Chapter 2 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 (3.5.3). 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 3.1.

3.8.3 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 §3.3.2 with component rate matrices in (3.3.9). 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 [59] 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 (3.8.1), 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 (3.3.9).

Figures 3.1-3.4 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 3.1 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 3.2 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 §3.6.3. (Observe that if $\pi_{1,2}(x(t))$ is replaced by a constant in the ode (3.2.12), 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 (3.5.3)) we have $x^* = (366.7, 459.5, 237.5)$. Plugging $z_{1,2}^*$ in (3.5.4), 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.

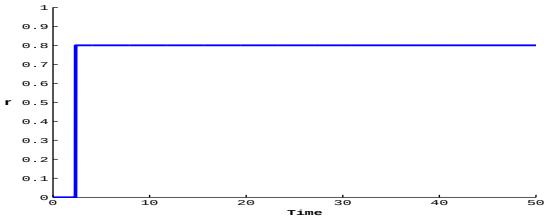


Figure 3.1: ratio between the queues.

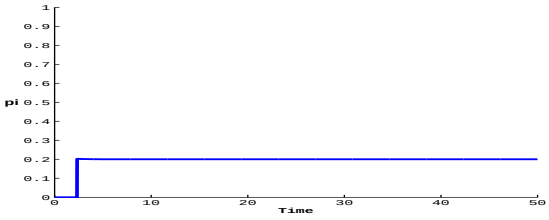


Figure 3.2: $\pi_{1,2}$ calculated at each iteration.

Note that in this example, the sufficient conditions for strong SSC in §3.6.1 do not hold. Specifically, condition (ii) in Theorem 3.6.1 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 3.6.1.

However, this example shows how useful the results of §3.6.2 are. By Theorem 3.6.3 we have $\alpha = \xi / r\theta_2$, where $\xi \equiv |\delta_+^*| \wedge \delta_-^*$. With the value of $z_{1,2}^*$ computed above, it follows

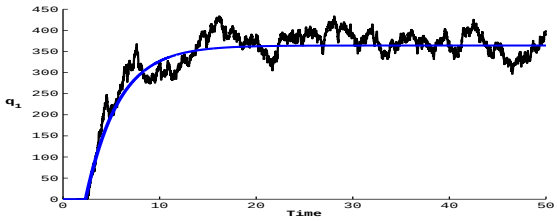


Figure 3.3: trajectory of q_1 together with a simulated sample path of Q_1.

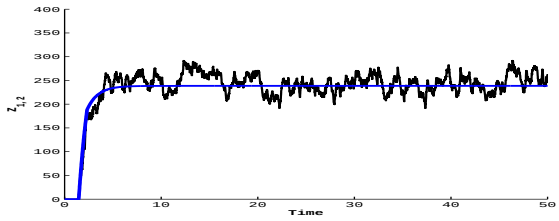


Figure 3.4: trajectory of $z_{1,2}$ together with a simulated sample path of $Z_{1,2}$.

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 entirely in $\mathbb{A}$ without even solving the algorithm. That is because $x(T) = (0, 0, 100) \in \beta_V(\alpha)$, and $\beta_V(\alpha) \subset\!\!\rightarrow$. (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 $x(T) \in \mathbb{A}$.)

3.9 Conclusions and Further Research

In this chapter we analyzed the deterministic ODE (3.2.11)-(3.2.12), 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 chapter is a vital link between Chapter 2 and the convergence proofs in Chapter 4.

We showed that the existence of a unique solution to the IVP (3.2.13) depends heavily on the characterization of the function Ψ in (3.2.11) and its topological properties. These properties, in turn, depend on the state space of Ψ, and the regions of the state space in which Ψ 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 chapter prepares the way for Chapter 4.

The connection to Chapter 2 is clear: First, we prove that the stationary point x^*, which was developed heuristically in Chapter 2 using flow-balance arguments, and was claimed to be the stationary point of (3.2.12) in [59], using reasonings similar to those in §3.5, 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 Chapter 2.

To fully connect to the model considered in Chapter 2, in §3.7 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 §3.5. 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 3.7.1. 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 (3.2.13), based on the matrix geometric method. This algorithm solves the different fluid models described in §3.7, and combines these solutions with the solution to (3.2.12) once the set $\mathbb{A}$, where the AP takes place, is hit.

Our main results in this chapter 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 Ψ appearing in (3.2.11) 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 3.7.1, and include the case in which $z_{2,1}$ becomes positive during the first transient period. We do make the following conjecture:

Conjecture 3.9.1. *Make Assumption A as usual and introduce lower thresholds as in Remark* 3.7.1. *If the appropriate ODE is defined for each relevant region, as in the proof of Theorem* 3.7.1, *then* $x(t) \to x^*$ *as* $t \to \infty$, *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.

3.10 Miscellany

3.10.1 More on the Algorithm

In this section we elaborate further on the algorithm introduced in §3.8. Let $\{t_m : m = 0, 1, 2, \ldots, n\}$ 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 (3.3.12) 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 §3.7. If the fluid state is not in $\mathbb{S}$, as in the initial period of the example in §3.8 and the example below, then we use the appropriate fluid model, as given in the proof of Theorem 3.7.1.

3.10.2 An Example with $\mathbf{x}^* \in \mathbb{S}^+$

We now consider the same example as in §3.8.3, 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 3.7.1, 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 (3.7.7). This is due to the numerical errors, as described in §3.7.

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 3.5.2 (ii).

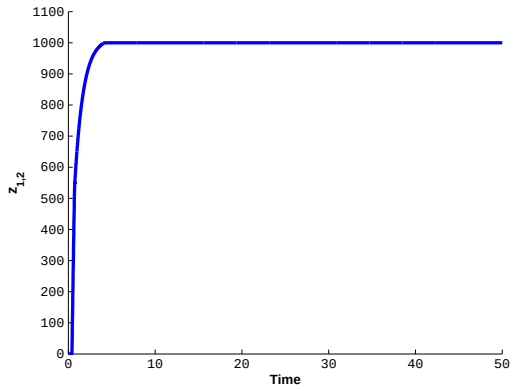


Figure 3.5: $z_{1,2}$ when λ_1 exceeds the system's capacity.

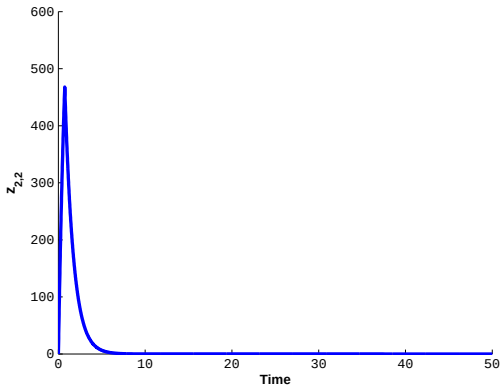


Figure 3.6: $z_{2,2}$ when λ_1 exceeds the system's capacity.

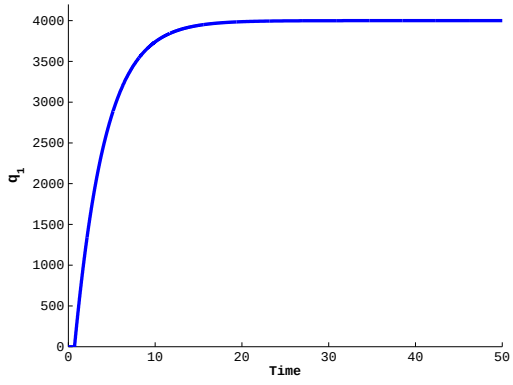


Figure 3.7: q_2 when λ_1 exceeds the system's capacity.

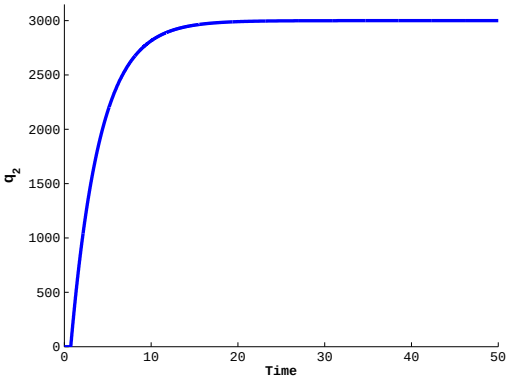


Figure 3.8: q_1 when λ_1 exceeds the system's capacity.

Chapter 4

Convergence to the Fluid Limit via the Averaging Principle

4.1 Overview

In this chapter we will prove that the solution to the ODE in Chapter 3 is indeed the MS-HT fluid limit of the overloaded X model; see Theorem 4.6.1; and see §4.3 for the key assumptions. In doing so, we will also prove the *averaging principle* (AP) which in turn will provide a strong version of *state-space collapse* (SSC) for the two-dimensional queue process and the server-assignment processes; for the SSC results, see Theorems 4.4.1, 4.4.2, 4.5.6 and 4.7.1. To streamline the reading, some of the more technical proofs appear separately in the next chapter.

We now consider the X model *during* the overload incident only, once sharing has begun; that will be captured by our main Assumptions 1 and 3 in §4.3. As a consequence, the model is stationary but the evolution is transient. Because of customer abandonment, the stochastic models will all be stable, approaching proper steady-state distributions. We will be proving a MS-HT limit for the system processes.

Convergence to the fluid limit will be established in roughly three steps: (i) representing the sequence of systems (§4.4), (ii) proving that the sequence considered is $\mathcal{C}$-tight (§4.8.1), and (iii) uniquely characterizing the limit (Chapter 3 and much of the rest of §4.3-§4.8, and Chapter 5).

The first representation step in §4.4 starts out in the usual way, involving rate-1 Poisson processes and martingales, as reviewed in [57]. However, the SSC in Theorem 4.4.1 requires a delicate analysis of the unscaled sequence; see §4.7, especially Lemma 4.7.4.

The second tightness step in §4.8.1 is routine, but the final characterization step is challenging. These last two steps are part of the standard compactness approach to proving stochastic-process limits; see [13], [25], [57] and §11.6 in [78]. As reviewed in [25] and [57], uniquely characterizing the limit is usually the most challenging part of the proof, but it is especially so here. Characterizing the limit is difficult because the FQR-T control is driven by a queue-difference process which is not being scaled and hence does not converge to a deterministic quantity with spatial scaling. However, the driving process operates in a different time scale than the fluid-scaled processes, asymptotically achieving a (time-dependent) steady state at each instant of time, yielding the AP.

As was shown in Chapter 3, the AP and the FTSP also complicate the analysis of the limiting ODE. First, it requires that the steady state of a continuous-time Markov chain (CTMC), whose distribution depends on the solution to the ODE, be computed at every instant of time. (As explained in Chapter 3, this argument may seem circular at first, since the distribution of the FTSP is determined by the solution to the ODE, while the evolution of the solution to the ODE is determined by the behavior of the FTSP. However, the separation of time scales explains why this construction is consistent.) The second complication is that the AP produces a singularity region in the state space, causing the ODE to be discontinuous in its full state space. Hence, both the convergence to the MS-HT fluid limit, and the analysis of the solution to the ODE depend heavily on the state space of the

ODE, which is characterized in terms of the FTSP. For that reason, many of the results in Chapter 3 are needed for proving convergence, and we summarize the essential results in §4.5 below.

There is now a substantial literature on fluid limits for queueing models, some of which is reviewed in [78]. For recent work on many-server queues, see [40, 44]. Because of the separation of time scales here, our work is in the spirit of fluid limits for networks of many-server queues in [8, 9], but again the specifics are quite different. Their separation of time scales justifies using a pointwise stationary approximation asymptotically, as in [51, 77].

4.2 Preliminaries

We briefly specify some of the notation we will be using.

4.2.1 Many-Server Heavy-Traffic (MS-HT) Scaling

We now add the subscript 6 to the process X, describing the X system, to emphasize that the original stochastic system under FQR-T is a six-dimensional *continuous time Markov chain* (CTMC), i.e.,

$$X_6(t) \equiv (Q_i(t), Z_{i,j(t)}; i, j = 1, 2), \quad t \geq 0 \tag{4.2.1}$$

To develop the fluid limit, we consider a sequence of X systems, $\{X_6^n : n \geq 1\}$ defined as in (4.2.1), indexed by n (denoted by superscript), with arrival rates and number of servers growing proportionally to n, i.e.,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \to \lambda_i \quad \text{and} \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \to m_i \quad \text{as} \quad n \to \infty, \tag{4.2.2}$$

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

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

$$\bar{X}_6^n(t) \equiv (\bar{Q}_i^n(t), \bar{Z}_{i,j}^n(t) : i,j = 1,2), \quad t \geq 0. \tag{4.2.3}$$

In this framework, with additional regularity conditions, we will prove that $\bar{X}_6^n \Rightarrow x_6$ in an appropriate framework (see §4.2.2), where x_6 is a deterministic continuous function.

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

$$\frac{k_{i,j}^n}{n} \to 0 \quad \text{and} \quad \frac{k_{i,j}^n}{\sqrt{n}} \to \infty \quad \text{as} \quad n \to \infty, \quad i,j = 1,2. \tag{4.2.4}$$

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.

We will also consider shifting thresholds $\kappa_{i,j}^n$, satisfying

$$\frac{\kappa_{i,j}^n}{n} \to \kappa_{i,j} \geq 0 \quad \text{as } n \to \infty, \quad i,j = 1,2. \tag{4.2.5}$$

These shifting thresholds can be of order n, i.e., $\kappa_{i,j} > 0$, if a version of FQR-T, the *shifted FQR-T* control, is employed. Shifted FQR-T is designed to keep the relation between the queues at $Q_1 \approx r_{1,2}Q_2 + \kappa_{1,2}$, or $Q_1 \approx r_{2,1}Q_2 + \kappa_{2,1}$, which is the optimal relation in

the stationary fluid model for the important class of separable quadratic cost functions; See §2.7. These shifting constants can also stand for the thresholds $k^n_{i,j}$, $i,j = 1,2$, if we choose not to drop them once sharing is initialized (for the reasons described in §3.1.2). In that case, the scale of $\kappa^n_{i,j}$ is as in (4.2.4). If the thresholds are dropped and the relation between the queues is a fixed ratio, then $\kappa^n_{i,j} = 0$ for all $n \geq 1$, $i,j = 1,2$. To summarize, we consider $\kappa^n_{i,j} = O(n)$, but without specifying their exact scale.

As before, let

$$D^n_{1,2}(t) \equiv (Q^n_1(t) - \kappa^n_{1,2}) - r_{1,2}Q^n_2(t), \quad t \geq 0, \tag{4.2.6}$$

and recall that FQR using the process $D^n_{1,2}$ in (4.2.6): if $D^n_{1,2}(t) > 0$, then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If $D^n_{1,2}(t) \leq 0$, then every newly available agent takes his new customer from the head of his own queue.

Let

$$\rho^n_i \equiv \frac{\lambda^n_i}{\mu_{i,i}m^n_i}, \quad \text{and} \quad \rho_i \equiv \lim_{n\to\infty} \rho^n_i = \frac{\lambda_i}{\mu_{i,i}m_i}, \quad i = 1,2. \tag{4.2.7}$$

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

Our results depend on the system being overloaded, where, without loss of generality, we assume that class 1 is more overloaded than class 2. However, in our case, a system can be overloaded even if one of the classes 2 is not overloaded by itself. We have the following quantities:

$$q^a_i \equiv \frac{(\lambda_i - \mu_{i,i}m_i)^+}{\theta_i} \quad \text{and} \quad s^a_i \equiv \left(m_i - \frac{\lambda_i}{\mu_{i,i}}\right)^+, \quad i = 1,2, \tag{4.2.8}$$

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

4.2.2 Conventions About Notation

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

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

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

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

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

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

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

stochastically bounded (SB), i.e., if

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

We say that Y is $o_P(f(t))$ if $\|Y\|_t/f(t)$ converges in probability (and thus, in distribution) to 0, i.e., if

$$\frac{\|Y\|_t}{f(t)} \Rightarrow 0 \quad \text{as} \quad t \to \infty.$$

If $f(t) \equiv 1$, then $Y = O_P(1)$ if it is SB, and $Y = o_P(1)$ if $\|Y\|_t \Rightarrow 0$. We define $O_P(f(n))$ and $o_P(f(n))$ in a similar way, but with the domain of f being $\mathbb{N}$, i.e., $f : \mathbb{N} \to [0, \infty)$.

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

4.3 The Main Assumptions

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

First, we have the two assumptions already made, (4.2.2) and (4.2.5). Our first new assumption is on the asymptotic behavior of the rates; it specifies the essential form of the overload. For the statement, recall the definitions in (4.2.2), (4.2.5) and (4.2.8), which describe the asymptotic behavior of the rates.

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

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

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

(2) $q_1^a - \kappa > rq_2^a$.

Condition (1) in Assumption 1 ensures that class 1 is asymptotically overloaded, even after receiving help from pool 2. To see why, first observe that, since $s_2^a \geq 0$, $q_1^a > \kappa \geq 0$, so that $\lambda_1 > \mu_{1,1}m_1$ and $\rho_1 > 1$. Hence, class 1 is overloaded. Next observe that $\mu_{1,2}s_2^a = \mu_{1,2}(1-\rho_2)^+$, and that $(1-\rho_2)^+$ is the amount of (steady-state fluid) extra service capacity in pool 2, if it were to serve only class-2 customers. Thus, Condition (1) in Assumption 1 implies that enough class-1 customers are routed to pool 2 to ensure that pool 2 is overloaded when sharing is taking place. This conclusion will be demonstrated in §4.7. Note that Condition (1) in Assumption 1 is slightly stronger than Condition (I) of Assumption A in Chapter 3. because here there is a strong inequality instead of a weak inequality.

Condition (2) in Assumption 1 ensures that class 1 is more overloaded than class 2 if class 2 is also overloaded. This condition helps ensure that there is no incentive for pool 1 to help pool 2, so that we can assume that $Z_{2,1}^n$ remains at 0.

We now expand upon the centering constants.

Assumption 2. $($*centering constants*$)$

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

(1) $\kappa^n \geq 0$ *for all* n *and* $\kappa^n/n \to \kappa$, *where* $0 \leq \kappa < \infty$.

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

In Assumption 2, the first condition is the standard scaling for the centering constants. If $\kappa = 0$, then we have FQR after sharing has been activated by passing the thresholds; if

$\kappa \neq 0$, then we have shifted FQR after sharing has been activated by passing the thresholds. From the perspective of the centering constants alone, it would suffice to consider $\kappa^n = n\kappa$. However, we have imposed additional conditions for the case $\kappa = 0$. We did this so that we could consider the FQR-T control with the original thresholds retained. As discussed in §4.2.1, we want those thresholds to be $o(n)$ but large compared to $O(\sqrt{n})$; e.g., we might have $\kappa^n = n^p$ for $1/2 < p < 1$. The regularity conditions involving scaling by $\log n$ is for results in §4.7 showing that the idleness is at most $O(\log n)$.

Our third assumption is about the initial conditions. We require that a fluid-scale limit exists at time 0, where the limit $x(0)$ satisfies the initial conditions required for the existence of a unique solution to the ODE, established in Chapter 3. The ODE and the FSTP will be reviewed here in §4.5. Specifically, Assumption 3 refers to the set $\mathbb{A}$ defined in (4.5.16) and expressed in (4.5.22). We will be explaining Assumption 3 in the next two sections. For the statement, recall the definition of the six-dimensional fluid-scaled process $\bar{X}_6^n$ in (4.2.3) and let $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ be the associated three-dimensional process. (In §4.4 we show that it suffices to consider $\bar{X}^n$.) We also need to separately specify initial conditions for the queue-difference processes in (4.2.6). We assume the queue-difference processes start in some fixed state.

Assumption 3. (*initial conditions*)

For each $n \geq 1$, $Z_{1,1}^n(0) = m_1^n$, $Z_{2,1}^n = 0$, $Q_1^n(0) \geq \kappa^n$, $Z_{1,2}^n(0) + Z_{2,2}^n(0) = m_2^n$ *and* $D_{1,2}^n(0) \equiv Q_1^n(0) - r_{1,2}Q_2^n(0) = j$ *for some fixed* j. *In addition,*

$$\bar{X}^n(0) \Rightarrow x(0) \in \mathbb{A} \quad \textit{as} \quad n \to \infty,$$

with $x(0)$ *being a deterministic element of* $\mathbb{R}_3$.

If the system is initialized not in $\mathbb{A}$, then other fluid models hold during the initial period

before $\mathbb{A}$ is hit; See §3.7. In this chapter we want to concentrate on time intervals on which the averaging principle is operating. The condition $x(0) \in \mathbb{A}$ implies that $q_1(0) \geq \kappa$, and that sharing is taking place (or at least initializing) at time 0, and that both service pools are full.

4.4 Representation of X_6^n

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

4.4.1 Starting with Rate-1 Poisson Processes

Let $A_i^n(t)$ count the number of class-i customer arrivals, let $S_{i,j}^n(t)$ count the number of serivce completions of class-i customers by agents in pool j, and let $U_i^n(t)$ count the number of class-i customers to abandon from queue, all in model n during the time interval $[0,t]$. Following common practice, as reviewed in §2 of [57], we represent these processes in terms of mutually independent rate-1 Poisson processes. We represent the counting processes A_i^n, $S_{i,j}^n$ and U_i^n as

$$\begin{aligned} A_i^n(t) &\equiv N_i^a(\lambda_i^n t), \\ S_{i,j}^n(t) &\equiv N_{i,j}^s\left(\mu_{i,j}\int_0^t Z_{i,j}^n(s)\,ds\right), \\ U_i^n(t) &\equiv N_i^u\left(\theta_i\int_0^t Q_i^n(s)\,ds\right), \quad t \geq 0, \end{aligned} \tag{4.4.1}$$

where N_i^a, $N_{i,j}^s$ and N_i^u for $i = 1,2; j = 1,2$ are eight mutually independent rate-1 Poisson processes.

The evolution of X_6^n in (4.2.3) is somewhat complicated because, at each service-completion epoch t, we need to know whether $D_{1,2}^n(t)$ is strictly positive or not, and whether there are any class-i customers in service-pool j, $i = 1, 2$, $i \neq j$. For example, fix n and consider a time $t > 0$ in which a type-2 server becomes available and that $D_{1,2}^n(t) > 0$. Then the newly available server should take a customer from the head of queue 1. However, if at the same time $Z_{2,1}^n(t) > 0$ then, according to the one-way sharing rule, he cannot take customers from queue 1. Hence, we need to be able to know at each time $t \geq 0$ whether $Z_{i,j}^n(t) > 0$. In addition, some customers may arrive to find idleness in their class service pool, so that they go immediately into service.

4.4.2 Simplification via SSC

However, since the system is assumed to be overloaded, it is reasonable to expect that the idleness process in the two service pools is asymptotically negligible in diffusion (and thus in fluid) scale. That means that $Z_{1,1}^n(t) + Z_{2,1}^n(t) \approx m_2^n$ and $Z_{2,2}^n(t) + Z_{1,2}^n(t) \approx m_2^n$ for all $t > 0$, provided that those approximations hold at $t = 0$. Also, since we assume that class 1 is more overloaded than class 2, it is reasonable to expect that $Z_{1,2}^n$ becomes positive before the threshold $k_{2,1}^n$ is crossed (for large n), so that $Z_{2,1}^n(t) = 0$, at least on some initial interval $[0, \tau]$, $\tau > 0$. If that is true, then $Z_{1,1}^n(t) \approx m_1^n$ and $Z_{2,2}^n(t) \approx m_2^n - Z_{1,2}^n(t)$, $t \in [0, \tau]$. The approximation signs will be replaced with equality with both diffusion and fluid scaling, producing a SSC result. Specifically, the dimension of the service process reduces from four to one in the limit with diffusion scaling. That will be proved in Theorem 4.7.1 below.

We now state a result which will allow us to represent the system in a relatively simple form, building on the SSC for the service process just explained (and which will be proved in §4.7). Recall that X_6^n has been defined in §4.2.1, the assumptions in §4.3 are in force, d_{J_1} denotes the standard Skorohod J_1 metric and $\breve{Y}^n \equiv Y^n/\sqrt{n}$ for any $Y^n \in \mathcal{D}_k$.

Theorem 4.4.1. (*Representation via SSC*) *As* $n \to \infty$, $d_{J_1}(\breve{X}_6^n, \breve{X}^{n,*}) \Rightarrow 0$ *in* $\mathcal{D}_6$, *where* $X^{n,*} \equiv X_6^n \equiv (Q_1^n, Q_2^n, Z_{1,1}^n, Z_{2,1}^n, Z_{1,2}^n, Z_{2,2}^n)$ *under the extra condition that* $Z_{1,1}^n = m_1^n$, $Z_{2,1}^n = 0$ *and* $Z_{1,2}^n + Z_{2,2}^n = m_2^n$, *with* $X^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$ *being represented via*

$$\begin{aligned} Z_{1,2}^n(t) &\equiv Z_{1,2}^n(0) + \int_0^t 1_{\{D_{1,2}^n(s)>0\}}\, dS_{2,2}^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,2}^n(t) \\ &= Z_{1,2}^n(0) + N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n)(s)\, ds\right) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right), \quad t \ge 0, \end{aligned} \tag{4.4.2}$$

$$\begin{aligned} Q_1^n(t) &\equiv Q_1^n(0) + A_1^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)>0\}}\, dS^n(t) \\ &\quad - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,1}^n(t) - U_1^n(t) \\ &= Q_1^n(0) + N_1^a(\lambda_1^n t) - N_{1,1}^s(\mu_{1,1} Z_{1,1}^n t) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}} Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_1^u\left(\theta_1\int_0^t Q_1^n(s)\, ds\right), \quad t \ge 0, \end{aligned} \tag{4.4.3}$$

$$\begin{aligned} Q_2^n(t) &\equiv Q_2^n(0) + A_2^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{2,2}^n(t) \\ &\quad - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,2}^n(t) - U_2^n(t) \quad t \ge 0 \\ &= Q_2^n(0) + N_2^a(\lambda_2^n t) \\ &\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right) \\ &\quad - N_2^u\left(\theta_2\int_0^t Q_2^n(s)\, ds\right), \quad t \ge 0. \end{aligned} \tag{4.4.4}$$

With a slight abuse of notation, henceforth we use $X^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$ to refer to both its direct representation in $\mathcal{D}_3$ and (by virtue of Theorem 4.4.1) the essentially three-dimensional process $X^{n,*}$ in $\mathcal{D}_6$.

Theorem 4.4.1 is achieved as a corollary of Theorem 4.7.1, which will be stated and proved in §4.7. Without it, we could not write the representation (4.4.2)-(4.4.4). In fact, if we do not know that $Z_{2,1}^n$ is asymptotically negligible, then the evolution of X_6^n becomes intractable. Specifically, the system may oscillate between different directions of sharing, with $Z_{1,2}^n$ being positive at some instances, and $Z_{2,1}^n$ being positive at other instances. The system may also get "stuck" with $Z_{2,1}^n(t) > 0$ and $Z_{1,2}^n(t) = 0$ for all $t > t_0$, for some $t_0 > 0$, even though we want to have sharing in the other direction. (See Lemma 4.7.2 below. If at some $t_0 \geq 0$ we have that $z_{2,1}(t_0) > 0$ then $z_{2,1}(t) > 0$ for all $t > t_0$, where $z_{2,1}$ is the fluid limit of $\bar{Z}_{2,1}$.) These situations are ruled out by Theorem 4.7.1 and Theorem 4.4.1.

4.4.3 Simplification via Martingales

We now obtain further simplification using the familiar martingale representation, again see [57]. Consider the representation of X^n in (4.4.2) - (4.4.4) above, and let

$$\begin{aligned} M_i^{n,a}(t) &\equiv N_i^a(\lambda_i^n t) - \lambda_i^n t, \\ M_i^{n,u}(t) &\equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s))\, ds\right) - \theta_i \int_0^t Q_i^n(s)\, ds, \\ M_{i,2}^{n,s}(t) &\equiv N_{i,2}^s(J_{i,2}^n(t)) - J_{i,2}^n(t), \end{aligned} \tag{4.4.5}$$

where $J_{i,2}^n(t)$ are the compensators of the Poisson-processes $N_{i,2}^s(t)$ in (4.4.2)-(4.4.4), $i = 1, 2$, e.g.,

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

The quantities in (4.4.5) can be shown to be martingales (with respect to an appropriate filtration); See [57]. However, we will not use any martingale property, and call those terms martingales for convenience.

The following lemma follows easily from the FSLLN for Poisson processes and the $\mathcal{C}$-tightness to be established in Theorem 4.8.1:

Lemma 4.4.1. (*fluid limit for the martingale terms*) *As* $n \to \infty$,

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

Proof: By Lemma 4.8.1, the sequence $\{\bar{X}_6^n : n \geq 1\}$ is tight in $\mathcal{D}$. Thus any subsequence has a convergent subsequence. By the proof of Lemma 4.8.1, the sequences $\{J_{i,j}^n/n\}$ are also $\mathcal{C}$-tight, so that $\{J_{i,j}^n/n\}$, $i = 1, 2$, all converge along a converging subsequence as well. Consider a converging subsequence $\{X^n\}$ and its limit $\bar{X}$, which is continuous by Lemma 4.8.1. Then the claim of the lemma follows for the converging subsequence from the FSLLN for Poisson processes and the continuity of the composition map at continuous limits, e.g., Theorem 13.2.1 in [78]. In this case, the limit of each fluid-scaled martingale is the zero function $0e \in \mathcal{D}$, regardless of the converging subsequence we consider, and is thus unique. Hence we have completed the proof. ∎

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

$$\bar{M}^n \equiv \bar{X}^n - \bar{C}^n, \tag{4.4.6}$$

where $\bar{X}^n$ is defined in (4.4.2)-(4.4.4) and, with an abuse of notation, $\bar{C}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$

for

$$\begin{aligned}
\bar{Z}^n_{1,2}(t) &\equiv \bar{Z}^n_{1,2}(0) + \mu_{2,2}\int_0^t 1_{\{D^n_{1,2}(s)>0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t 1_{\{D^n_{1,2}(s)\le 0\}}\bar{Z}^n_{1,2}(s)\,ds, \\
\bar{Q}^n_1(t) &\equiv \bar{Q}^n_1(0) + \bar{\lambda}^n_1 t - \bar{m}^n_1 t - \mu_{1,2}\int_0^t 1_{\{D^n_{1,2}(s)>0\}}\bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{2,2}\int_0^t 1_{\{D^n_{1,2}(s)>0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds - \theta_1\int_0^t \bar{Q}^n_1(s)\,ds, \\
\bar{Q}^n_2(t) &\equiv \bar{Q}^n_2(0) + \bar{\lambda}^n_2 t - \mu_{2,2}\int_0^t 1_{\{D^n_{1,2}(s)\le 0\}}(\bar{m}^n_2 - \bar{Z}^n_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t 1_{\{D^n_{1,2}(s)\le 0\}}\bar{Z}^n_{1,2}(s)\,ds - \theta_2\int_0^t \bar{Q}^n_2(s))\,ds.
\end{aligned} \tag{4.4.7}$$

(We have used the same notation $(\bar{Q}^n_1, \bar{Q}^n_2, \bar{Z}^n_{1,2})$ in the definition of the different procsses $\bar{X}^n$ in (4.4.2)-(4.4.4) and $\bar{C}^n$ in (4.4.7) above. The following result shows that this anomaly causes no problem. Recall that d_{J_1} denotes the standard J_1 metric.

Theorem 4.4.2. *As $n \to \infty$, $\bar{M}^n \Rightarrow 0$, so that $d_{J_1}(\bar{X}^n, \bar{C}^n) \Rightarrow 0$ in $\mathcal{D}_3$ as $n \to \infty$, where $\bar{X}^n$ is defined in* (4.4.2)-(4.4.4) *and $\bar{C}^n$ is defined in* (4.4.7)*.*

Proof: Since the weak limit of the centered fluid-scaled Poisson processes in (4.4.5) is the (continuous) 0 function, the sum of any two or more of those processes also converges to $0 \equiv 0e$ in $\mathcal{D}$, by the continuity of addition at continuous limits, and is therefore $o_P(1)$. Hence we get $\bar{M}^n \Rightarrow 0$ as $n \to \infty$ directly from Lemma 4.4.1, from which the remaining convergence follows directly. ∎

As a consequence of Theorem 4.4.2, henceforth we can focus on $\bar{C}^n$ in (4.4.7) instead of $\bar{X}^n$ in (4.4.2)-(4.4.4). We will do so, but redefining $\bar{X}^n$: We let $\bar{X}^n \equiv \bar{C}^n$; i.e., henceforth we let $\bar{X}^n \equiv (\bar{Q}^n_1, \bar{Q}^n_2, \bar{Z}^n_{1,2})$ in (4.4.7).

Theorem 4.4.2 reduces the expression of $\bar{X}^n$ to the random rates of the Poisson processes, and reveals the basic structure of the limiting ODE in (4.5.13). Due to Theorem

4.4.1, the representation in (4.4.7) is equivalent to the representation of the six-dimensional process $\bar{X}_6^n$, for X_6^n in (4.2.3). Hence, proving that $\bar{X}^n$ converges to a unique deterministic limit, will imply the convergence of $\bar{X}_6^n$ to a limit in a three-dimensional hyperplane of $\mathcal{D}_6$, which is homeomorphic to $\mathcal{D}_3$. It is thus enough to work with the three-dimensional process in (4.4.7). Given Theorems 4.4.1 and 4.4.2, we will show that

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

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

4.5 The FTSP and the ODE

Even though Theorems 4.4.1 and 4.4.2 allow us to consider only the three-dimensional process X^n in (4.4.7), we still must cope with the indicator functions in the integrands in (4.4.7), which appear because of the FQR routing. Thus, the key to a successful analysis of X^n is understanding the behavior of the stochastic queue-difference process $D_{1,2}^n \equiv (Q_1^n - \kappa^n) - r_{1,2}Q_2^n$ in (4.2.6) when some, but not all, type-2 servers are helping class-1 customers, and the system is overloaded in the sense of Assumption 1.

In Chapter 3 we presented and analyzed a three dimensional ODE (which we refer to simply as "the ODE" since it is the only ODE under consideration). This ODE was conjectured to arise as the limit of the fluid-scaled version of X^n in (4.4.2)-(4.4.4). In this chapter we will prove that conjecture. Specifically, we will show that $\bar{X}^n$ indeed converges weakly to the solution of that three-dimensional ODE, so that the fluid limit of $\bar{X}^n$ and the solution to the ODE coincide. However, *the ODE is well defined and its solution exists as an element of* $\mathcal{C}_3$, regardless of any convergence results.

Since an understanding of the ODE, its state space and its solution is required in order

to characterize the fluid limit, we begin by defining the ODE (motivated by the sequence $\bar{X}^n$). In doing so, we will be reviewing Chapter 3; see Chapter 3 for a complete analysis of the ODE. Recall that the ODE is driven by a stochastic process, whose local steady-state distributions govern the evolution of the solution to the ODE. We thus start by defining the driving process, which we call the FTSP. To understand the FTSP, we need to better understand the queue-difference process.

4.5.1 The Drift Rates of the Queue-Difference Processes

In this subsection we specify the transition rates of the queue-difference process $\{D^n_{1,2}(t) : t \geq 0\}$ in (4.2.6) at any time t_0 conditional on $X^n(t_0) = \Gamma^n$, where sharing is taking place; i.e., we consider the transition rates of the process

$$D^n \equiv D^n(\Gamma^n) \equiv \{D^n(\Gamma^n, t) : t \geq t_0\} \equiv \{D^n_{1,2}(X^n(t_0), t) : t \geq t_0\} \tag{4.5.1}$$

at time t_0 conditional on $X^n(t_0) = \Gamma^n$, where Γ^n is a deterministic state, under the assumption that sharing is taking place. (We will explain when sharing will be taking place in the following subsections.) The initial difference at time t_0 is $D^n_{1,2}(X^n(t_0), t_0) = Q^n_1(t_0) - r_{1,2}Q^n_2(t_0)$, where $(Q^n_1(t_0), Q^n_2(t_0))$ is part of $X^n_6(t_0)$. To be well defined, the state Γ^n should be for the full CTMC X^n_6. The transition rates are independent of time t_0 for any given process state Γ^n. However, because of §4.4, it suffices to focus on the three-dimensional process $\bar{X}^n$. In other words, we can think of Γ^n as a state of X^n, i.e., a vector in $\mathbb{N}^2 \times [0, m^n_2]$. Thus the transition rates in (4.5.2)-(4.5.5) below, under this simplifying assumption, are asymptotically correct with $o(n)$ terms as $n \to \infty$ (which we omit).

To simplify analysis, we will work with an integer state space. Thus we assume that the shifting thresholds $\kappa^n_{1,2}$ in (4.2.6) are integers and that $r_{1,2}$ is rational, in particular, $r_{1,2} = j/k$ for positive integers j and k. We then look at queue differences measured in

units of $1/k$. Hence, we have transitions of $\pm j$ and $\pm k$ instead of the original values of ± 1 and $\pm r$.

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

First, for $D^n(\Gamma^n, t_0) = m \leq 0$ with $\Gamma^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$, the upward rates are

$$\begin{aligned} &\lambda_-^{(k)}(n, m, \Gamma^n) \equiv \lambda_1^n, \quad \text{and} \\ &\lambda_-^{(j)}(n, m, \Gamma^n) \equiv \mu_{1,2} Z_{1,2}^n + \mu_{2,2}(m_2^n - Z_{1,2}^n) + \theta_2 Q_2^n, \end{aligned} \tag{4.5.2}$$

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

$$\mu_-^{(k)}(n, m, \Gamma^n) \equiv \mu_{1,1} m_1^n + \theta_1 Q_1^n \quad \text{and} \quad \mu_-^{(j)}(n, m, \Gamma^n) \equiv \lambda_2^n, \tag{4.5.3}$$

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

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

$$\lambda_+^{(k)}(n, m, \Gamma^n) \equiv \lambda_1^n \quad \text{and} \quad \lambda_+^{(j)}(n, m, \Gamma^n) \equiv \theta_2 Q_2^n, \tag{4.5.4}$$

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

$$\begin{aligned}\mu_+^{(k)}(n,m,\Gamma^n) &\equiv \mu_{1,1}m_1^n + \mu_{1,2}Z_{1,2}^n + \mu_{2,2}(m_2^n - Z_{1,2}^n) + \theta_1 Q_1^n \quad \text{and} \\ \mu_+^{(j)}(n,m,\Gamma^n) &\equiv \lambda_2^n,\end{aligned} \tag{4.5.5}$$

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

Using these transition rates, we can define the *drift rates* for $D^n(X^n(t),t) \equiv D^n(\Gamma^n,t)$, conditional upon $X^n(t) = \Gamma^n$. Let these drift rates in the regions $(0,\infty)$ and $(-\infty,0]$ be denoted by $\delta_+^n(X^n(t))$ and $\delta_-^n(X^n(t))$, respectively, Combining (4.5.20) and (4.5.2)-(4.5.5), we obtain

$$\begin{aligned}\delta_+^n(X^n(t)) &\equiv j[\lambda_1^n - \mu_{1,1}m_1^n + (\mu_{2,2}-\mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n(t) - \theta_1 Q_1^n(t)] \\ &\qquad - k[\lambda_2^n - \theta_2 Q_2^n(t)], \\ \delta_-^n(X^n(t)) &\equiv j[\lambda_1^n - \mu_{1,1}m_1^n - \theta_1 Q_1^n(t)] \\ &\qquad - k[\lambda_2^n + (\mu_{2,2}-\mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n - \theta_2 Q_2^n(t)].\end{aligned} \tag{4.5.6}$$

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

4.5.2 The FSTP

The FTSP can perhaps be best understood as being the limit of a family of *time-expanded queue-difference processes*, defined for each $n \geq 1$ by

$$D_e^n(\Gamma^n, s) \equiv D_{1,2}^n(t_0 + s/n), \quad s \geq 0. \tag{4.5.7}$$

where we condition on $X^n(t_0) = \Gamma^n$ for some deterministic vector Γ^n assuming possible values of $X^n(t_0) \equiv (Q_1^n(t_0), Q_2^n(t_0), Z_{1,2}^n(t_0))$. (The time t_0 is an arbitrary initial time.) We choose Γ^n so that sharing will occur (or will occur eventually for n large enough). Since we divide s in (4.5.7) by n, we are effectively dividing the rates by n. We are applying a "microscope" to "expand time" and look at the behavior after the initial time more closely. That is in contrast to the usual time contraction with conventional HT limits. See [75] for a previous limit using time expansion. We will explain the limit in detail in §4.5.6 below.

With that in mind, we see that the FTSP should have the same state space as $D_{1,2}^n$. When we relate the FTSP to the expanded queue-difference process in §4.5.6 below, we will also relate the initial differences, which so far are unspecified here. Since we already converted to an integer state space, the FTSP will be a continuous-time Markov chain (CTMC) on $\mathbb{Z}$. With that convention, the FTSP $\{D(\gamma, s) : s \geq 0\}$ has transition rates among the integers determined at any time s (in the newly introduced "infinitesimal" time scale) by both its current state $D(\gamma, s) \equiv m$ and the vector γ. The vector γ is a possible state of the fluid model $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ at some time t, where averaging may take place. Thus $\gamma \in [0, \infty)^2 \times [0, m_2]$. Specifically, γ can be any vector in the subset $\mathbb{A}$ defined in (4.5.16) below.

Given the current state m, we let the rates of the FTSP D as a function of γ be the limit of the rates of $D^n(\Gamma^n, \cdot)$ divided by n, where the rates of $D^n(\Gamma^n, \cdot)$ are themselves a function of the current state $D^n(\Gamma^n, 0) = m$ with $\Gamma^n/n \to \gamma$ as $n \to \infty$. Since $\Gamma^n/n \to \gamma$ as $n \to \infty$, there will be sharing in all systems for all n sufficiently large. (For the corresponding rates of the queue-difference process $D^n(\Gamma^n, \cdot)$ itself, see (4.5.2)-(4.5.5).)

Since the drift rates of $D^n(\Gamma^n, t)$ in (4.5.6) are linear functions of the state $X^n(t)$, we have

$$\delta_+^n(X^n(t)) \Rightarrow \delta_+(\bar{X}(t)) \quad \text{and} \quad \delta_-^n(X^n(t)) \Rightarrow \delta_-(\bar{X}(t)) \tag{4.5.8}$$

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

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

Paralleling the definitions in (4.5.2)-(4.5.5), we define the transition rates for $D(\gamma)$ as follows: First, for $D(\gamma, s) = m \in (-\infty, 0]$ with $\gamma \equiv (q_1, q_2, z_{1,2})$, the upward rates are

$$\begin{aligned} &\lambda_-^{(k)}(m, \gamma) \equiv \lambda_1, \quad \text{and} \\ &\lambda_-^{(j)}(m, \gamma) \equiv \mu_{1,2} z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_2 q_2. \end{aligned} \tag{4.5.9}$$

Similarly, the downward rates are

$$\mu_-^{(k)}(m, \gamma) \equiv \mu_{1,1} m_1 + \theta_1 q_1 \quad \text{and} \quad \mu_-^{(j)}(m, \gamma) \equiv \lambda_2 \tag{4.5.10}$$

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

$$\lambda_+^{(k)}(m, \gamma) \equiv \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, \gamma) \equiv \theta_2 q_2. \tag{4.5.11}$$

The downward rates are

$$\begin{aligned} &\mu_+^{(k)}(m, \gamma) \equiv \mu_{1,1} m_1 + \mu_{1,2} z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_1 q_1 \quad \text{and} \\ &\mu_+^{(j)}(m, \gamma) \equiv \lambda_2. \end{aligned} \tag{4.5.12}$$

4.5.3 The ODE

We can now present the three-dimensional ODE in terms of the FTSP D. Let $\dot{x} \equiv (\dot{q}_1, \dot{q}_2, \dot{z}_{1,2})$, where $\dot{x}(t)$ is the derivative evaluated at time t, and

$$\begin{aligned}
\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))\left[z_{1,2}(t)\mu_{1,2} + z_{2,2}(t)\mu_{2,2}\right] - \theta_1 q_1(t) \\
\dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))\left[z_{2,2}(t)\mu_{2,2} + z_{1,2}(t)\mu_{1,2}\right] - \theta_2 q_2(t) \\
\dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{4.5.13}$$

with $\pi_{1,2}(x(t)) \equiv P(D(x(t), \infty) > 0)$ for each $t \geq 0$, where $D(x(t), \infty)$ has the limiting steady-state distribution as $s \to \infty$ of the FTSP $D(\gamma, s)$ for $\gamma = x(t)$.

Equivalently, we have the following integral representation of the ODE in (4.5.13):

$$\begin{aligned}
z_{1,2}(t) &\equiv z_{1,2}(0) + \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t (1 - \pi_{1,2}(x(s)))z_{1,2}(s)\,ds, \\
q_1(t) &\equiv q_1(0) + \lambda_1 t - m_1 t - \mu_{1,2}\int_0^t \pi_{1,2}(x(s))z_{1,2}(s))\,ds \\
&\quad - \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\,ds - \theta_1\int_0^t q_1(s)\,ds, \\
q_2(t) &\equiv q_2(0) + \lambda_2 t - \mu_{2,2}\int_0^t (1 - \pi_{1,2}(x(s)))(m_2 - z_{1,2}(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t (1 - \pi_{1,2}(x(s)))z_{1,2}(s)\,ds - \theta_2\int_0^t q_2(s))\,ds.
\end{aligned} \tag{4.5.14}$$

The integral representation is closely related to the integral representation of $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ in (4.4.7); $\bar{X}^n$ has been replaced by x and the indicators $1_{D_{1,2}^n(s)>0}$ have been replaced by $\pi_{1,2}(x(s))$.

Since $\gamma = x(t)$, the relevant FTSP at time t depends on the solution of the ODE at time t, $x(t)$. Since the right side of the ODE has $\pi_{1,2}(x(t))$, the evolution of the ODE beyond t

in turn depends on the FTSP, specifically, upon the steady-state distribution of that FTSP. Given $x(t)$ for some $t > 0$, we can determine the FTSP $\{D(x(t), s) : s \geq 0\}$. Given that FTSP, we can determine the steady-state quantity $\pi_{1,2}(x(t))$. Then $\pi_{1,2}(x(t))$ appears on the right side of the ODE in (4.5.13), determining the future of the ODE. We provided an efficient algorithm to solve this ODE coupled with the FTSP in Chapter 3. The efficiency is based on the QBD structure discussed in §4.5.5.

4.5.4 The State Space of the ODE

Since the ODE in (4.5.13) is driven by the family of FTSP $D(\gamma, \cdot)$ (just as the stochastic systems are driven by the process $D^n_{1,2}$), we divide the state space of the fluid limit according to the relation that holds between q_1 and q_2, and the behavior of the FTSP in the different regions.

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

$$\mathbb{S}^{\mathbf{b}} \equiv \{q_1 - rq_2 = \kappa\}, \quad \mathbb{S}^+ \equiv \{q_1 - rq_2 > \kappa\}, \quad \mathbb{S}^- \equiv \{q_1 - rq_2 < \kappa\}, \tag{4.5.15}$$

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

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

The boundary subset $\mathbb{S}^b$ is a hyperplane in the state space $\mathbb{S}$, and is therefore a closed subset. It is the subset of $\mathbb{S}$ in which the AP is taking place, and the function $\pi_{1,2}$ can assume its full range of values, $0 \leq \pi_{1,2}(\gamma) \leq 1$, $\gamma \in \mathbb{S}^b$. Let $\mathbb{A} \subset \mathbb{S}^b$ be the set in which $D(x, \cdot)$ is positive recurrent. We have $0 < \pi_{1,2}(\gamma) < 1$ if and only if $\gamma \in \mathbb{A}$. Thus, for each $\gamma \in \mathbb{S}^b$, we define

$$\mathbb{A} \equiv \{\gamma \in \mathbb{S}^b : 0 < \pi_{1,2}(\gamma) < 1\}. \tag{4.5.16}$$

4.5.5 The Fundamental QBD structure

Characterizing the set $\mathbb{A}$ in (4.5.16) is essential to our analysis. Our analysis is simplified by exploiting matrix geometric methods, as in [52]. In particular, we represent the integer-valued FTSP $D \equiv \{D(\gamma, s) : s \geq 0\}$ constructed above as a homogeneous continuous-time QBD, as in Definition 1.3.1 and §6.4 of [52]. To do so, we 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 [52] (imitating the shape of a generator matrix of a birth-and-death process). In particular, for each three-dimensional state γ, we write

$$Q \equiv Q(\gamma) \equiv \begin{pmatrix} B & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{4.5.17}$$

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

$$B \equiv \begin{pmatrix} A_1^+ & B_\mu \\ B_\lambda & A_1^- \end{pmatrix} \quad \text{and} \quad A_i \equiv \begin{pmatrix} A_i^+ & 0 \\ 0 & A_i^- \end{pmatrix} \tag{4.5.18}$$

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

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

the mapping:

$$\begin{aligned} \phi(2nm+i) &\equiv nm+i \quad \text{and} \\ \phi((2n+1)m+i) &\equiv -nm-i+1, \quad 1 \le i \le m. \end{aligned} \tag{4.5.19}$$

Then we order the states in levels as follows

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

With this representation, the generator-matrix Q can be written in the form (4.5.17) above, where A_1 groups all the transitions within a level, A_0 groups the transitions from level $L(n)$ to level $L(n+1)$ and A_2 groups all transitions from level $L(n)$ to level $L(n-1)$. Matrix B groups the transitions within the boundary level $L(0)$, and is thus different than A_1. An example is given in §4.5.5.

QBD's having a generator matrix Q of the form (4.5.17)-(4.5.18) will be repeatedly constructed in our proofs. We thus refer to the QBD structure, represented by the generator matrix Q as specified by (4.5.18) as the *fundamental QBD*.

To determine when the AP holds, we need to be able to determine when the FTSP D is positive recurrent. Fortunately, QBD theory allows us to determine that easily for each γ, as explained in Chapter 3 and summarized below.

Let δ_+ and δ_- be the drift of the QBD in the positive and negative region, respectively (see §3.3.3. See [52] for the general theory); i.e., let

$$\begin{aligned} \delta_+(\gamma) &\equiv j\left(\lambda_+^{(j)}(\gamma)-\mu_+^{(j)}(\gamma)\right)+k\left(\lambda_+^{(k)}(\gamma)-\mu_+^{(k)}(\gamma)\right), \\ \delta_-(\gamma) &\equiv j\left(\lambda_-^{(j)}(\gamma)-\mu_-^{(j)}(\gamma)\right)+k\left(\lambda_-^{(k)}(\gamma)-\mu_-^{(k)}(\gamma)\right). \end{aligned} \tag{4.5.20}$$

By our construction of the rates above, it holds that $\delta_-(\gamma) > \delta_+(\gamma)$ for every $\gamma \in \mathbb{S}$. Below we repeat Theorem 3.3.1 with the modified notation:

Theorem 4.5.1. *The QBD representing the FTSP* $\{D(\gamma, s) : s \geq 0\}$ *is positive recurrent if and only if*

$$\delta_-(\gamma) > 0 > \delta_+(\gamma). \tag{4.5.21}$$

For every $\gamma \in \mathbb{R}_3$, the set $\mathbb{A}$ in (4.5.16) where the AP is operating, is the same set in which (4.5.21) holds, i.e.,

$$\mathbb{A} \equiv \{\gamma \in \mathbb{S} : \delta_-(\gamma) > 0 > \delta_+(\gamma)\}. \tag{4.5.22}$$

From the continuity of the QBD drift-rates in (4.5.20), if follows that $\mathbb{A}$ is an open and connected subset of $\mathbb{S}^b$. Hence, $\mathbb{A}$ can be regarded as an open connected subset of $\mathbb{R}^+_2$ (since $\mathbb{S}^b$ is homoeomorphic to $\mathbb{R}^+ \times [0, m_2]$). Our proofs (here and in Chapter 3) rely on the fact that if $x(s) \in \mathbb{A}$, then for some $h > 0$, $x(u) \in \mathbb{A}$, $0 < u < h$. In particular, if $x(0) \in \mathbb{A}$, then there exists a $\delta > 0$ such that $\{x(t) : 0 \leq t < \delta\} \subset \mathbb{A}$, as stated in Theorem 3.4.2, which we repeat below:

Theorem 4.5.2. *If* $x(0) \in \mathbb{A}$, *then there exists a unique solution* $x \in \mathcal{C}_3([0, \delta))$ *to the fluid ODE* (4.5.13) *for some* $\delta > 0$.

We will initially work on an interval $[0, \delta)$, $\delta > 0$, over which we can guarantee that the AP and Theorem 4.5.2 hold. After the convergence is established, this interval can be extended, typically all the way to ∞; see §3.6. However, the extension of the initial interval $[0, \delta)$ depends only on the solution to the ODE. Thus, it suffices to prove the convergence over $[0, \delta)$ no matter how small δ is. We will characterize a $\delta > 0$ in §4.8.3. For the rest of the discussion, assume that $\delta > 0$ is known.

4.5.6 The FTSP Arising as a Limit

We now present some results in which the FTSP $D \equiv \{D(\gamma, s) : s \geq 0\}$ arises as a limit. These results connect the queue difference process $D^n \equiv \{D^n_{1,2}(t) : t \geq 0\}$ defined in (4.2.6) and (4.5.1) and the time-expanded queue-difference processes D^n_e in (4.5.7) to the FTSP defined above. These results help explain the main theorem.

We first formalize the separation of time scales using the time-expanded queue-difference processes D^n_e defined in (4.5.7). The following result "explains" the AP, but does not complete the proof of the FWLLN. We prove this theorem in §5.1.

Theorem 4.5.3. *If $\Gamma^n/n \to \gamma$ and $D^n(\Gamma^n, 0) \Rightarrow D(\gamma, 0)$ in $\mathbb{R}$ as $n \to \infty$, where $\gamma \in \mathbb{A}$, then*

$$\{D^n_e(\Gamma^n, s) : s \geq 0\} \Rightarrow \{D(\gamma, s) : s \geq 0\} \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty, \tag{4.5.23}$$

where D^n_e is the time-expanded queue-difference process in (4.5.7) *and D is the FTSP; i.e., we have convergence of the sequence of time-inhomogeneous CTMC's to a limiting time-homogeneous CTMC.*

Of course, we are actually interested in the queue-difference processes. We will obtain the following result in Corollary 4.8.5. Recall the definition of stochastic boundedness (SB) in §4.2.2.

Theorem 4.5.4. *Over an appropriate interval, $[0, \delta)$, the sequence of stochastic processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \delta\} : n \geq 1\}$ is SB in $\mathcal{D}$, so that the sequence of random variables $\{D^n_{1,2}(t) : n \geq 1\}$ is SB in $\mathbb{R}$ for each t, $0 \leq t \leq \delta$.*

Nevertheless, one implication of Theorem 4.5.3 is that, as n increases, $D^n_{1,2}$ fluctuates "too much" in the neighborhood of every point $t \in [0, \delta)$ for the sequence of stochastic processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \delta\} : n \geq 1\}$ to be $\mathcal{D}$-tight. If the sequence were tight, then

it would have a convergent subsequence. If $D^n_{1,2}$ were to converge on $[0, \delta)$ to a process in $\mathcal{D}$ along that subsequence, then the limiting process must have at most finitely many discontinuities exceeding any constant $\epsilon > 0$, see e.g., Lemma 1 on p. 122 of [13]. However, for every neighborhood of any $t \in [0, \delta]$, there would necessarily be infinitely many jumps of size 1 in the limit as $n \to \infty$. Moreover, every t would have to be a discontinuity point of the limit, but there can be only countably many discontinuities. Hence, the limit process could not be an element of $\mathcal{D}$. Hence tightness does not hold.

However, we do obtain a proper limit for the sequence of random variables $\{D^n_{1,2}(t) : n \geq 1\}$ in $\mathbb{R}$ for each fixed t by exploiting the AP. After we prove Theorem 4.6.1, we will establish the following pointwise AP result, which helps explain the AP. See [77] for a similar result. We prove this theorem in §5.1 after proving Theorem 4.5.3.

Theorem 4.5.5. (*pointwise AP*) *Fix $t \in [0, \delta)$. As $n \to \infty$, $D^n_{1,2}(t) \Rightarrow D(x(t), \infty)$ in $\mathbb{R}$ as $n \to \infty$, where $x(t)$ is the solution to the ODE at time t and $D(x(t), \infty)$ has the limiting steady-state distribution of the FTSP $D(\gamma, s)$ for $\gamma = x(t)$.*

Remark 4.5.1. Even though the limit of $\bar{X}^n$ turns out to be deterministic, Theorems 4.5.3 and 4.5.5 imply that the process $D^n_{1,2}$ does not become deterministic as $n \to \infty$. Given Theorems 4.5.3 and 4.5.5, we see that indeed *the deterministic ODE is driven by a stochastic process*. More precisely, the evolution of the (deterministic) solution to the ODE over $[0, \delta)$ is governed by a stochastic process, although the ODE describing that evolution is itself deterministic, depending on the time-dependent steady-state distribution of the FTSP's.

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

As an immediate consequence of Theorem 4.5.4, we obtain the following SSC result.

Corollary 4.5.1. (*SSC of the queue process*) *As* $n \to \infty$,

$$c_n^{-1}((Q_1^n - \kappa^n) - r_{1,2}Q_2^n) \Rightarrow 0 \quad \textit{in} \quad \mathcal{D}([0,\delta))$$

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

Corollary 4.5.1 shows that the two-dimensional scaled queue process is effectively a one-dimensional process as $n \to \infty$. Combining Theorem 4.4.1 and Corollary 4.5.1 gives the following SSC result, which reduces the dimension of the process from the original six dimension, to only two when we consider the fluid-scaled or diffusion-scaled versions of the process X_6^n in (4.2.3). In particular, asymptotically, the six-dimensional process $\bar{X}_6^n \in \mathcal{D}_6$ actually exists in a two-dimensional hyperplane of $\mathcal{D}_6$, which is homeomorphic to $\mathcal{D}_2$ over the interval $[0,\delta)$. For $\mathcal{D}_3 \equiv \{(a_1,a_2,a_3) : a_1,a_2,a_3 \in \mathcal{D}\}$, $\bar{X}_3^n$ is asymptotically an element of the two-dimensional hyperplane $\{(a_1, r_{1,2}a_1 + \kappa, a_3) : a_1, a_3 \in \mathcal{D}\}$ of $\mathcal{D}_3$.

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

Theorem 4.5.6. (*Complete SSC*) *As* $n \to \infty$, $d_{J_1}(\breve{X}_6^n, \breve{X}_2^n) \Rightarrow 0$ *in* $\mathcal{D}_6([0,\delta))$ *as* $n \to \infty$, *where* $X_2^n \equiv (Q_1^n, r_{1,2}Q_1^n + \kappa^n, Z_{1,2}^n)$.

Remark 4.5.2. The SSC result in Theorem 4.4.1 is stated for $\mathcal{D}_6 \equiv \mathcal{D}_6([0,\infty))$, while the SSC in Corollary 4.5.1, and thus also Theorem 4.5.6, holds on $\mathcal{D}_6([0,\delta))$. However, the SSC result in Corollary 4.5.1 and Theorem 4.5.6 can be extended as long as the solution to the ODE is in $\mathbb{A}$, since the SSC of the queue process is implied by the AP. (This will become clear in the proofs.) As we mentioned above, the solution to the ODE is typically in $\mathbb{A}$ for all $t \geq 0$.

4.6 The FWLLN

In this chapter we will establish a FWLLN for scaled versions of the vector stochastic process (X_6^n, Y_8^n), where

$$X_6^n \equiv (Q_i^n, Z_{i,j}^n) \in \mathcal{D}_6 \quad \text{and} \quad Y_8^n \equiv (A_i^n, S_{i,j}^n, U_i^n) \in \mathcal{D}_8, \quad i,j = 1,2, \tag{4.6.1}$$

For the FWLLN, we focus on the scaled vector process

$$(\bar{X}_6^n, \bar{Y}_8^n) \equiv n^{-1}(X_6^n, Y_8^n), \tag{4.6.2}$$

for X_6^n and Y_8^n in (4.6.1). Recall that Assumptions 1-3 are in force.

Theorem 4.6.1. (FWLLN) *There exists $\delta > 0$ such that*

$$(\bar{X}_6^n, \bar{Y}_8^n) \Rightarrow (x, y) \quad \textit{in} \quad \mathcal{D}_{14}([0,\delta)) \quad \textit{as} \quad n \to \infty, \tag{4.6.3}$$

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

$$x \equiv (q_i, z_{i,j}) \quad \textit{and} \quad y \equiv (a_i, s_{i,j}, u_i), \quad i = 1,2; j = 1,2; \tag{4.6.4}$$

$z_{2,1} = s_{2,1} = m_1 - z_{1,1} = m_2 - z_{2,2} - z_{1,2} = 0e$ *and $(q_1, q_2, z_{1,2})$ being the unique solution to the three-dimensional ODE in* (4.5.13)*. The remaining limit function y is defined in terms of x:*

$$\begin{aligned} a_i(t) &\equiv \lambda_i t, \quad s_{i,j}(t) \equiv \mu_{i,j} \int_0^t z_{i,j}(s)\, ds, \\ u_i(t) &\equiv \theta_i \int_0^t q_i(s)\, ds \quad \textit{for} \quad t \geq 0, \quad i = 1,2; \quad j = 1,2. \end{aligned} \tag{4.6.5}$$

The time interval $[0, \delta)$ *can be expanded to the largest interval* (*typically* $[0, \infty)$) *such that there exists a unique solution to the ODE in* (4.5.13).

Theorem 4.6.1 established convergence over some interval $[0, \delta)$. Theorem 4.6.1 concludes by stating that the interval can be extended whenever the solution to the ODE can be extended. Ensuring convergence over $[0, \delta)$ will usually imply convergence over an interval $[0, T)$, for some $T \gg \delta$, often even $T = \infty$. First, the convergence over $[0, \delta)$ implies that the SSC results in the next section, §4.7, hold globally - see the explanation right above Lemma 4.7.3. Second, once the convergence is established, and the unique solution to the ODE (4.5.13) is known to exist (Theorem 4.5.2), we can use the results in §3.6, to infer whether we can extend the convergence to the whole halfline $[0, \infty)$ *by analyzing the limiting ODE itself, and not the stochastic sequence* X^n. In particular, the solution to the ODE (4.5.13) can be extended to the entire halfline $[0, \infty)$ by showing that $x(t) \in \mathbb{A}$ for all $t \geq 0$. Often, this can be done without even solving the ODE; see Theorem 3.4.4 and §3.6.

By Theorem 4.5.6, it is enough to present the fluid limit of $(\bar{Q}_1^n, \bar{Z}_{1,2}^n)$, since each queue determines the other in the limit. Nevertheless, in Theorem 4.6.1 we presented the fluid limit for both queues. We did so, because the three-dimensional framework applies in other regions. For example, in Chapter 3 we analyzed that same ODE in all three regions. More importantly, even in our settings, when Assumption 3 holds and the solution is in $\mathbb{A}$ over $[0, \delta)$, it is helpful to solve the fluid equations without explicitly forcing the SSC relation between the queues. Having the solution satisfying $q_1(t) - r_{1,2}q_2(t) = \kappa$ strongly indicates that the numerical solution to the fluid ODE is correct; See the last paragraph in §2.9.2.

The rest of this chapter is devoted to the proof of Theorem 4.6.1. Most proofs of supporting theorems and lemmas appear in Chapter 5 (in order of appearance in this chapter).

4.7 SSC for the Service Process

In this section we establish state-space collapse (SSC) for the service process

$$Z^n \equiv (Z^n_{1,1}, Z^n_{1,2}, Z^n_{2,1}, Z^n_{2,2});$$

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

Unlike our main convergence result - Theorem 4.6.1 - which is proved on an initial interval, the SSC of the service process holds globally on $[0, \infty)$ for FQR-T, given Assumptions 1-3. However, here we do not yet show that a limit of $\bar{Z}^n_{1,2}$ as $n \to \infty$ exists. We only show that, when analyzing Z^n, it is sufficient to consider $Z^n_{1,2}$, prove that its fluid-scaled and diffusion-scaled versions converge and then characterize the limits. That will be done for the fluid-scaled case in the next section (and Chapter 5).

Here is the SSC result to be established in this section. Note that it directly implies Theorem 4.4.1.

Theorem 4.7.1. (*global SSC of the service process*) *As* $n \to \infty$,

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

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

$$\bar{I}^n_j \equiv I^n_j/n \quad \text{and} \quad \hat{I}^n_j \equiv I^n_j/\sqrt{n}, \quad j = 1, 2. \tag{4.7.1}$$

Theorem 4.7.1 will be proved in two steps. First, we show that $Z^n_{2,1} \Rightarrow 0$; second, we show that $\hat{I}^n_1$ and $\hat{I}^n_2$ are asymptotically negligible. By the first step, $I^n_1 = m^n_1 - Z^n_{1,1} + o_P(1)$, so that we can disregard the $o_P(1)$ term in the second step.

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

4.7.1 Extreme-Value Limits for QBD Processes

We are unaware of any established extreme-value limits for QBD processes, so we establish the following result here. Recall that a QBD has states (i, j), where i is the level and j is the phase. If we only consider the level we get the level process; it is an elementary function of a QBD. The proof of this theorem, like most others. appears in Chapter 5.

Theorem 4.7.2. $($*extreme value for QBD*$)$ *If $\mathcal{L}$ is the level process of a positive recurrent* $($*homogeneous*$)$ *QBD process* $($*with a finite number of phases*$)$*, then there exists $c > 0$ such that*

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

Both the statement and the proof of Theorem 4.7.2 are complicated by the discreteness of the integer-valued process $\mathcal{L}$. The proof is also somewhat complicated by the continuity of time. It is well known that the stationary distribution of the QBD level is asymptotically geometric, e.g., see §9.1 in [52]. Hence, we are unambiguously in the light-tailed case,

but we do not have the conventional convergence in law to the Gumbel distribution if we subtract by $c \log t$ instead of divide. Indeed, there do not exist scaling functions $a(t)$ and $b(t)$ such that $a(t)(\|\mathcal{L}\|_t - b(t))$ converges in law to a proper limit as $t \to \infty$; see Sections 1.5 and 1.7 of [53]. Even though the conventional extreme-value limit cannot hold, Theorem 4.7.2 evidently is not in best possible form. First, we should have $\|\mathcal{L}\|_t / \log t \Rightarrow c$ for a specific constant c (which is easy to identify); second, we should have tightness of the family $\{\|\mathcal{L}\|_t - c \log t : t \geq 1\}$ for that same constant c; e.g., see Example C.2.6 of [3] and Problem 4.2 of [7], but our weaker implication of such results suffices for the application here and has a relatively simple proof; see §5.2.

4.7.2 Basic Stochastic-Order Bounds

As we mentioned before, the proofs will involve stochastic-order bounds, using sample-path stochastic order, involving coupling; see [74], Ch. 4 of [54] and §2.6 of [56]. We briefly discuss those bounds for a sequence of stochastic processes $\{Y^n : n \in \mathbb{N}\}$. We will bound the process Y^n, for each $n \geq 1$, by a process Y_b^n; i.e., for each n, we will establish conditions under which it is possible to construct stochastic processes $\tilde{Y}_b^n$ and $\tilde{Y}^n$ on a common probability space, with $\tilde{Y}_b^n$ having the same distribution as Y_b^n, $\tilde{Y}^n$ having the same distribution as Y^n, and every sample path of $\tilde{Y}_b^n$ lies below (or above) the corresponding sample path of $\tilde{Y}^n$. We will then write $Y_b^n \leq_{st} (\geq_{st}) Y^n$. However, we will not introduce this "tilde" notation; Instead, we will use the original notation Y^n and Y_b^n. As a first step, we will directly give both processes, Y^n and Y_b^n identical arrival processes, the Poisson arrival processes specified for Y^n. We will then show that the remaining construction is possible by increasing (decreasing) the departure rates so that, whenever $Y^n = Y_b^n$, any departure in Y^n also leads to a departure in Y_b^n. That is justified by having the conditional departure rates, given the full histories of the systems up to time t, be ordered.

The stochastic-order bounds will be of the form

$$Y^n(t) = Y^n(0) + \sum_{i=1}^{k} N_i\left(\int_0^t J_i^n(s)\,ds\right), \quad t \geq 0, \tag{4.7.2}$$

where N_i, $i = 1, 2, \ldots, k$, denote independent rate-1 Poisson processes, and J_i^n is a stochastic process that serves as a random time change of the Poisson process N_i. If we are concerned with the fluid limit of Y^n, then we next divide both sides of (4.7.2) by n, subtract and then add back J_i^n to get the representation

$$\begin{aligned}\bar{Y}^n(t) \equiv Y^n(t)/n = \bar{Y}^n(0) + n^{-1}\int_0^t J_i^n(s)\,ds \\ + n^{-1}\sum_{i=1}^{k}\left[N_i\left(\int_0^t J_i^n(s)\,ds\right) - \int_0^t J_i^n(s)\,ds\right].\end{aligned} \tag{4.7.3}$$

The third step is to apply a version of the continuous mapping theorem to (4.7.3) (The purpose of the bounds is to be able to use the continuous mapping theorem, which can not be used on X^n.) However, to avoid unnecessary repetitions, we will not write the second step (4.7.3) and write only the representation as in (4.7.2), with the understanding that the continuous mapping theorem is actually applied to the version of $\bar{Y}^n$ in (4.7.3).

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

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

represented via

$$\begin{aligned}
Q^n_{1,a}(t) &= Q^n_{1,a}(y) + N^a_1(\lambda^n_1 t) - N^s_{1,1}(\mu_{1,1} m^n_1 t) \\
&\quad - N^u_1\left(\theta_1 \int_0^t (Q^n_{1,a}(s) \vee 0)\, ds\right), \\
Q^n_{2,a}(t) &= Q^n_{2,a}(y) + N^a_2(\lambda^n_2 t) - N^s_{1,2}\left(\mu_{1,2} \int_0^t Z^n_a(s)\, ds\right) \\
&\quad - N^s_{2,2}\left(\mu_{2,2} \int_0^t (m^n_2 - Z^n_a(s))\, ds\right) \\
&\quad - N^u_2\left(\theta_2 \int_0^t (Q^n_{2,a}(s) \vee 0)\, ds\right), \\
Z^n_a(t) &= Z^n_a(y) - N^s_{1,2}\left(\mu_{1,2} \int_0^t Z^n_a(s)\, ds\right).
\end{aligned} \tag{4.7.4}$$

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

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

via

$$\begin{aligned} Q^n_{1,b}(t) &= Q^n_{1,b}(y) + N^a_1(\lambda^n_1 t) - N^s_{1,1}(\mu_{1,1} m^n_1 t) - N^s_{1,2}\left(\mu_{1,2}\int_0^t Z^n_b(s)\right) \\ &\quad - N^u_1\left(\theta_1 \int_0^t (Q^n_{1,b}(s) \vee 0)\, ds\right), \\ Q^n_{2,b} &= Q^n_{2,b}(y) + N^a_2(\lambda^n_2 t) - N^u_2\left(\theta_2 \int_0^t Q^n_{2,b}(s)\, ds\right), \\ Z^n_b(t) &= Z^n_b(y) + N^s_{2,2}\left(\mu_{2,2}\int_0^t (m^n_2 - Z^n_b(s))\, ds\right), \end{aligned} \tag{4.7.5}$$

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

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

Lemma 4.7.1. *For $y \geq 0$ consider the processes $\{X^n_a(t) : t \geq y\}$ in* (4.7.4) *and $\{X^n_b(t) : t \geq y\}$ in* (4.7.5)*, for which the following holds for all $n \geq 1$:*

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

Also assume that $\bar{X}^n_a(y) \equiv X^n_a(y)/n \Rightarrow X_a(y)$ and $\bar{X}^n_b(y) \Rightarrow X_b(y)$ in $\mathbb{R}$ as $n \to \infty$. Then $\{\bar{X}^n_a(t) : t \geq y\} \Rightarrow \{X_a(t) : t \geq y\}$ and $\{\bar{X}^n_b(t) : t \geq y\} \Rightarrow \{X_b(t) : t \geq y\}$ in $\mathcal{D}_3$ as $n \to \infty$, where X_a and X_b are defined as follows: For $t \geq y$, $X_a(t) \equiv (Q_{1,a}(t), Q_{2,a}(t), Z_a(t))$

satisfies the following integral equation

$$\begin{aligned} Q_{1,a}(t) &= Q_{1,a}(y) + \lambda_1 t - \mu_{1,1} m_1 t - \theta_1 \int_0^t (Q_{1,a}(s) \vee 0)\, ds, \\ Q_{2,a}(t) &= Q_{2,a}(y) + \lambda_2 t - \mu_{1,2} \int_0^t Z_a(s)\, ds - \mu_{2,2} \int_0^t (m_2 - Z_a(s))\, ds \\ &\quad - \theta_2 \int_0^t (Q_{2,a}(s) \vee 0)\, ds, \\ Z_a(t) &= Z_a(y) + \mu_{2,2} m_2 t - \mu_{2,2} \int_0^t Z_a(s)\, ds, \end{aligned} \tag{4.7.6}$$

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

$$\begin{aligned} Q_{1,b}(t) &= Q_{1,b}(y) + \lambda_1 t - \mu_{1,1} m_1 t - \mu_{1,2} \int_0^t Z_b(s)\, ds \\ &\quad - \theta_1 \int_0^t (Q_{1,b}(s) \vee 0)\, ds, \\ Q_{2,b}(t) &= Q_{2,b}(y) + \lambda_2 t - \theta_2 \int_0^t Q_{2,b}(s)\, ds, \\ Z_b(t) &= Z_b(y) + \mu_{2,2} m_2 t - \mu_{2,2} \int_0^t Z_b(s)\, ds, \end{aligned} \tag{4.7.7}$$

Proof: By the continuous mapping theorem, applied to the integral representation, Theorem 4.1 in [57], $\bar{Z}_a^n \equiv Z_a^n/n$ and $\bar{Z}_b^n \equiv Z_b^n/n$ converge to the processes Z_a and Z_b with continuous sample paths. We can then apply Theorem 4.1 in [57] again, to conclude that the fluid-scaled queue lengths, $\bar{Q}_{i,a}^n \equiv Q_{i,a}^n/n$ and $\bar{Q}_{i,b}^n \equiv Q_{i,b}^n/n$, $i = 1, 2$, converge as well. (Note that $h(s) \equiv \theta(s \vee 0)$ is Lipschitz continuous, as required for the integral representation to be continuous.) ∎

Note that the condition $\bar{X}_a^n(y) \Rightarrow X_a(y)$ and $\bar{X}_b^n(y) \Rightarrow X_b(y)$ in $\mathbb{R}$ as $n \to \infty$ holds for $y = 0$ with $X_a(0) = X_b(0) = x(0)$, where $x(0)$ is deterministic, by Assumption 3 and our construction (since we take $X_a^n(0) = X_b^n(0) = X^n(0)$). In that case, and whenever $X_a(y)$ and $X_b(y)$ are deterministic, the limits X_a and X_b are deterministic functions. Indeed, we

anticipate that the limits X_a and X_b will be deterministic, but we use the more general form in our proof of Lemma 4.8.11, exploiting convergence along subsequences, where we do not yet know that the limit is deterministic.

4.7.3 The $Z^n_{2,1}$ Process

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

Theorem 4.7.3. *(global one-way sharing)* $Z^n_{2,1} \Rightarrow 0$ *in* $\mathcal{D}$ *as* $n \to \infty$.

The proof of Theorem 4.7.3 relies on three lemmas, which we state now. The proofs of these lemmas and Theorem 4.7.3 appear in 5.2. The first lemma proves a special case which implies Theorem 4.7.3. The other two lemmas prove a local version of the theorem, i.e., that $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ as $n \to \infty$ for some $\tau > 0$. In the proof of Theorem 4.7.3 we extend the local result to the full halfline $[0, \infty)$.

Our first lemma treats the simplest case.

Lemma 4.7.2. *If* $z_{1,2}(0) > 0$*, then, for all* $T > 0$, $P(\inf_{0 \leq t \leq T} \bar{Z}^n_{1,2}(t) > 0) \to 1$ *as* $n \to \infty$. *As a consequence,* $Z^n_{2,1} \Rightarrow 0$ *as* $n \to \infty$.

Given Lemma 4.7.2, it remains to consider only the case $z_{1,2}(0) = 0$. Hence, we assume that $z_{1,2}(0) = 0$ for the rest of this section. Here is the outline of the proof: The SSC statement for $Z^n_{2,1}$ will first be proved locally on an interval $[0, \tau]$, for some $\tau > 0$. Then, we can use later results, proving that $\bar{Z}^n_{1,2}(t) \Rightarrow z_{1,2}(t)$ as $n \to \infty$ on $[0, \delta]$ for some $\delta \leq \tau$, to extended the local SSC statement to a global one. That is, our proof follows three steps: (1) We first prove that $\|Z^n_{2,1}\|_\tau \Rightarrow 0$, for some $\tau > 0$. (2) For some δ satisfying $0 < \delta \leq \tau$, we can use the local result established in the first step, to prove Theorem 4.6.1, and deduce that the *deterministic* fluid limit $z_{1,2}(t)$ of $\bar{Z}^n_{1,2}(t)$ exists over $[0, \delta]$. (3) Finally,

we show that $z_{1,2}(t_0) > 0$ for some t_0, $0 < t_0 < \delta \leq \tau$, so that Lemma 4.7.2 can be applied to extend the local statement in step (1) to a global one. *We emphasize at the outset that the extension to a global statement is not circular, since the convergence of the $\bar{Z}^n_{1,2}$ process over $[0, \delta]$ (established in Theorem 4.6.1) uses only the local SSC result (since we take $\delta \leq \tau$).*

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

Lemma 4.7.3. *If either (i) $\kappa > 0$ or (ii) $r_{1,2} > r_{2,1}$ and $q_1(0) > 0$, then there exists τ, $0 < \tau \leq \infty$, such that*

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

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

The proof of Lemma 4.7.3 relies on a fluid argument. That fluid reasoning fails when $\kappa = 0$ and $r_{2,1} = r_{1,2} \equiv r$ or when $\kappa = 0$ and $q_1(0) = 0$, since then $q_1(0) - r_{1,2}q_2(0) = q_1(0) - r_{2,1}q_2(0)$. In these cases we will rely on the threshold $k^n_{2,1}$, and construct a finer sample-path stochastic-order bound for the stochastic system.

When we consider the stochastic sequence $\{X^n\}$, we need to have $rQ^n_2(t) - Q^n_1(t) > k^n_{2,1}$ in order to initialize sharing, with pool 1 helping class 2. It is thus clear that we need to consider the stochastic fluctuations of the weighted queue-length processes $D^n_{2,1}$, and show that the probability of the threshold $k^n_{2,1}$ being crossed over an initial interval $[0, \tau]$ converges to 0 as $n \to \infty$. Arguments relying solely on the fluid-scaled processes (which are of order $O_P(n)$) are too crude, and cannot reveal whether $k^n_{2,1}$ is exceeded on an interval, since $k^n_{2,1}$ is taken to be $o(n)$. We treat that case in the next lemma by appealing to the extreme-value result established in Theorem 4.7.2.

Remark 4.7.1. Recall that the two initial thresholds $k^n_{1,2}$ and $k^n_{2,1}$ are designed to prevent sharing when the two classes are not overloaded, and are thus chosen to satisfy $k^n_{i,j}/\sqrt{n} \to$

∞ as $n \to \infty$. Once sharing starts, with pool 2 helping queue 1, $k^n_{1,2}$ may be dropped (unless shifted-FQR is employed, in which case $k^n_{1,2} = \kappa^n = O(n)$), but $k^n_{2,1}$ is kept, in order to prevent sharing in the other direction. In the proof of the next lemma, Lemma 4.7.4, we will see that when sharing is taking place, it is enough to have $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$. This suggests that, once sharing starts, we can replace the original threshold $k^n_{2,1}$, with a new and smaller threshold, which satisfies $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$.

In the next lemma we treat the cases not treated in Lemma 4.7.3. In addition to $z_{1,2}(0) = 0$, we assume that $\kappa = 0$ and that $q_1(0) - r_{2,1}q_2(0) = 0$. This latter assumption implies that either $q_1(0) = 0$ (so that $q_2(0) = 0$ as well), or, if $q_1(0) > 0$, then necessarily $r_{1,2} = r_{2,1}$.

Lemma 4.7.4. *Assume that $\kappa = 0$ and that $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$. Also assume that $q_1(0) - r_{2,1}q_2(0) = 0$ (where $r_{2,1}$ is a rational number). Then there exists τ, $0 < \tau \leq \infty$, such that*

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

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

Lemmas 4.7.3 and 4.7.4 prove that, for some $\tau > 0$, $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ as $n \to \infty$. We will use this local result in the proof of Theorem 4.6.1, which shows that, for some $0 < \delta \leq \tau$, $\{\bar{X}^n(t) : 0 \leq t \leq \delta\} \Rightarrow \{x(t) : 0 \leq t \leq \delta\}$, where x is deterministic. In particular, $\bar{Z}^n_{1,2}(t) \Rightarrow z_{1,2}(t)$ over $[0, \delta]$, where $z_{1,2}(t)$ is deterministic. Recall that Theorem 4.6.1 relies only on the local version of Theorem 4.7.3 established already.

Remark 4.7.2. The conclusion of Lemma 4.7.2 reveals a disadvantage of the one-way sharing rule for very large systems. The lemma concludes that, for large n, if for some $\epsilon > 0$ and $t_0 \geq 0$ $Z^n_{1,2}(t_0) > \epsilon n$, then $Z^n_{1,2}(t)$ is very likely not to reach 0 for a long time, thus preventing sharing in the opposite direction, even if that would prove beneficial to do so at a later time, e.g., because there is a new overload incident in the opposite direction.

In practice, we thus may want to relax the one-way sharing rule. One way of relaxing the one-way sharing rule is by dropping it entirely, and relying only on the thresholds $k^n_{1,2}$ and $k^n_{2,1}$ to prevent sharing in both directions simultaneously (at least until the arrival rates change again). Another modification is to introduce lower thresholds on the service processes, denoted by $s^n_{i,j}$, $i \neq j$, such that pool 2 is allowed to start helping class 1 at time t if $D^n_{2,1} > k^n_{2,1}$ and $Z^n_{1,2}(t) < s^n_{1,2}$, and similarly in the other direction.

We do not analyze either of these modified controls In this chapter. We observe that a global result stating that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$ will be much harder to show, because we cannot use the reasoning in Lemma 4.7.2. Specifically, showing that $Z^n_{1,2}$ becomes positive in fluid scale and never empties, does not imply that $Z^n_{2,1} \Rightarrow 0$, since sharing may be allowed at time t even if $Z^n_{1,2}(t) > 0$. Nevertheless, Lemmas 4.7.3 and 4.7.4 still hold, so that $Z^n_{2,1}(t) = 0$ for all $t \in [0, \tau)$ for some $\tau > 0$ and all n large enough. Since the convergence to the fluid limit in Theorem 4.6.1 is initially established for an interval $[0, \delta)$, we can decrease δ if necessary, so that $\delta \leq \tau$. Once convergence of the fluid limit to its stationary point is established (using the results in §2.7), we have that the fluid cannot leave $\mathbb{A}$, and $z_{2,1}$ is guaranteed to remain zero throughout.

4.7.4 The Idleness Processes

We next address the two idleness processes. We will use the standard concept of stochastic boundedness, extended to stochastic processes, which was defined in §4.2.2.

Theorem 4.7.4. *For $j = 1, 2$, $I^n_j / \log n$ is SB, which implies that $\hat{I}^n_j \Rightarrow 0$ as $n \to \infty$.*

Remark 4.7.3. The proof Theorem 4.7.4 uses the result in the previous subsection, namely that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$. Hence, the statement of the theorem should first be shown to hold on $[0, \tau]$, for τ in Lemmas 4.7.3 and 4.7.4. Once the local result is shown to hold, it is used to prove Theorem 4.6.1, so that the convergence of $\bar{X}^n$ to the deterministic fluid limit

x is established over an interval $[0, \delta]$, for some $0 < \delta \leq \tau$. In the proof of Theorem 4.7.3 this was shown to imply that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$ over the entire halfline $[0, \infty)$. We can thus extend the proof of Theorem 4.7.4 to the entire halfline as well. For that reason, the statement of the theorem refers to the global result and its proof also assumes that $Z^n_{2,1}$ is asymptotically null globally.

4.8 Proofs of the Main Theorem

We now come to the proof of Theorem 4.6.1. There are eight subsections here. In §4.8.1 we establish tightness. In §4.8.2 we establish explicit stochastic bounds on all the processes, which control the total rate of transitions. In §4.8.3 we identify an interval $[0, \delta)$ over which the frozen difference processes are positive recurrent, asymptotically. In §4.8.4 we state a continuity result for QBD's that we will apply. In §4.8.5 we establish stochastic-process bounds. In §4.8.6 we establish bounds for the integrals over small subintervals. In §4.8.7 we complete the proof of Theorem 4.6.1, exploiting the preparation in the previous subsections. The string of inequalities in (5.5.37) in 5.5.5 shows what is needed. Finally, in §5.1.2 we prove Theorem 4.5.5. Most of the proofs for this section appear in §5.5.

4.8.1 Tightness

We start by establishing tightness.

Lemma 4.8.1. *The sequence* $\{(\bar{X}^n_6, \bar{Y}^n_8) : n \geq 1\}$ *in* (4.6.2) *is* $\mathcal{C}$*-tight in* $\mathcal{D}_{14}$.

For background on tightness, see [13, 57, 78]. We recall a few key facts: Tightness of a sequence of k-dimensional stochastic processes in $\mathcal{D}_k$ is equivalent to tightness of all the one-dimensional component stochastic processes in $\mathcal{D}$. For a sequence of random elements of $\mathcal{D}_k$, $\mathcal{C}$-tightness implies $\mathcal{D}$-tightness and that the limits of all convergent subsequences

must be in $\mathcal{C}_k$; see Theorem 15.5 of the first 1968 edition of [13]. Thus it suffices to verify conditions (6.3) and (6.4) of Theorem 11.6.3 of [78]. Hence, it suffices to prove SB of the sequence of stochastic processes evaluated at time 0 and appropriately control the oscillations, using the modulus of continuity on $\mathcal{C}$. We obtain the stochastic boundedness at time 0 immediately from Assumption 3 in §4.3. We show that we can control the oscillations in our proof of Lemma 4.8.1. The resulting tightness implies that the sequence of stochastic processes is SB. We give an alternative proof of SB in §4.8.2, which yields explicit bounds on the limit processes.

Since the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (4.6.2) is $\mathcal{C}$-tight by Lemma 4.8.1, every subsequence has a further subsequence which converges to a continuous limit. We conclude this section by applying the modulus-of-continuity inequalities established in the proof of Lemma 4.8.1 to deduce additional smoothness properties of the limits of all converging subsequence.

Lemma 4.8.2. *If $(\bar{X}_6, \bar{Y}_8)$ is the limit of a subsequence of $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in $\mathcal{D}_{14}$, then each component in $\mathcal{D}$, say $\bar{X}_i$, has bounded modulus of continuity; i.e., for each $T > 0$, there exists a constant $c > 0$ such that*

$$w(\bar{X}_i, \zeta, T) \leq c\zeta \quad w.p.1 \tag{4.8.1}$$

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

In closing this subsection, we remark that we cannot employ these bounds on the modulus of continuity to directly deduce that the limit $(\bar{X}_6, \bar{Y}_8)$ is either differentiable or deterministic. For example, a nonlinear piecewise-linear function with bounded slope is Lipschitz continuous without being differentiable, and the random function At, $t \geq 0$, where A is a bounded (non-deterministic) random variable satisfies (4.8.1) without itself being deterministic.

The $\mathcal{C}$-tightness result in Lemma 4.8.1 implies that every subsequence of the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (4.6.1) has a further converging subsequence in $\mathcal{D}_{14}$, whose limit is in the function space $\mathcal{C}_{14}$. However, by Theorem 4.7.1, it suffices to focus on $\bar{X}^n$ in $\mathcal{D}_3$, where the limits of the subsequences will be in $\mathcal{C}_3$. To establish the convergence of the sequence $\bar{X}^n$, we must show that every converging subsequence converges to the same (unique) limit. We thus need to characterize the limit of any converging subsequence, show that it is deterministic and that it satisfies the ODE (4.5.13) of Theorem 4.6.1. The existence and uniqueness of the solution to the ODE over an interval $[0, \delta)$, for some $\delta > 0$, is stated in Theorem 4.5.2. This δ can be increased as long as the solution x to the limiting ODE 4.5.13 remains in $\mathbb{A}$. In this section we will characterize an initial interval $[0, \delta]$ for which the solution is ensured to be in $\mathbb{A}$. Since we will be using the results of §4.7, we can decrease δ if necessary, so that $\delta \leq \tau$, for τ defined in Lemmas 4.7.3 and 4.7.4.

4.8.2 Explicit Stochastic Bounds

In this section we establish some explicit stochastic bounds on the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (4.6.1) and (4.6.2). These bounds complement the material in §4.8.1 and will be used to control the transition rates of the queue-difference stochastic processes $D_{1,2}^n$.

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

$$\begin{aligned} S_{i,j}^n(t) &\leq N_{i,j}^s\left(\mu_{i,j} m_j^n t\right), \\ Q_i^n(t) &\leq Q_i^n(0) + A_i^n(t), \\ U_i^n(t) &\leq N_i^u\left(\theta_i [Q_i^n(0)t + A_i^n(t)t]\right), \quad t \geq 0. \end{aligned} \tag{4.8.2}$$

We apply the FWLLN for the Poisson process with (4.8.2) and Assumption 3 to obtain the following lemma.

Lemma 4.8.3. $\bar{Y}_8^n \leq \bar{Y}_{bd}^n$, *where* $\bar{Y}_{bd}^n \Rightarrow y_{bd}$ *in* $\mathcal{D}_8$, *with*

$$\begin{aligned} y_{bd}(t) \equiv (&\lambda_1 t, \lambda_2 t, \mu_{1,1} m_1 t, 0, \mu_{1,2} m_2 t, \mu_{2,2} m_2 t, \\ &\theta_1[q_1(0)t + \lambda_1 t^2], \theta_2[q_2(0)t + \lambda_2 t^2]) \quad \textit{in} \quad \mathbb{R}_8. \end{aligned} \tag{4.8.3}$$

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

For the following comparison result we use the same sample-path stochastic-order construction as in §4.7.

Lemma 4.8.4. *If* $Q_\Sigma^n(0) \leq_{st} Q_{bd}^n(0)$ *in* $\mathbb{R}$, *then* $Q_\Sigma^n \leq_{st} Q_{bd}^n$ *in* $\mathcal{D}$.

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

Lemma 4.8.5. *If* $\bar{Q}_{bd}^n(0) \Rightarrow q_{bd}(0)$ *in* $\mathbb{R}$ *w.p.1, where* $q_{bd}(0)$ *is deterministic, then we have the FWLLN*

$$\bar{Q}_{bd}^n \Rightarrow q_{bd} \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty, \tag{4.8.4}$$

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

$$q_{bd}(t) \leq q_{bd}^* \equiv q_{bd}(0) \vee (\lambda/\theta) \quad \textit{for all} \quad t \geq 0. \tag{4.8.5}$$

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

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

Applying the continuous mapping theorem for the integral representation, Theorem 4.1 in [57], we have that $\bar{Q}_{bd}^n \Rightarrow q_{bd}$ in $\mathcal{D}$ as $n \to \infty$, where q_{bd} satisfies the ODE in the statement of the lemma. The solution to this ODE is easily seen to be $q(t) = \lambda/\theta + (q(0) - \lambda/\theta)e^{-\theta t}$, from which (4.8.5) follows. ∎

Lemma 4.8.5 implies that the sequence $\{\bar{Q}_{bd}^n : n \geq 1\}$ is C-tight in $\mathcal{D}$ whenever there is convergence of the initial conditions. Together with Lemma 4.8.4, that implies the following result.

Corollary 4.8.1. *The sequence $\{\bar{Q}_{\Sigma}^n : n \geq 1\}$ is SB in $\mathcal{D}$. For each $t > 0$, The limit of any converging subsequence of $\{\|\bar{Q}_{\Sigma}^n\|_t : n \geq 1\}$, where $n \to \infty$, is almost surely contained in the bounded interval $[0, (q_1(0) + q_2(0)) \vee (\lambda/\theta)]$.*

Proof: We use Assumption 3 to ensure that there is convergence of the initial conditions: $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}_6$ as $n \to \infty$, where $x(0)$ is deterministic. We can then let the initial conditions in Lemma 4.8.5 be $q_{bd}(0) \equiv q_1(0) + q_2(0)$. Hence, we get

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

That FWLLN for $\bar{Q}_{bd}^n$ implies that $\{\bar{Q}_{bd}^n\}$ is SB, which in turn implies that $\{\bar{Q}_{\Sigma}^n\}$ is SB. Moreover, we get the final conclusion of Corollary 4.8.1. ∎

We now have the following strengthening of the SB conclusion that can be deduced from Lemma 4.8.1.

Corollary 4.8.2. *The sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in* (4.6.1) *and* (4.6.2) *is SB in $\mathcal{D}_{14}$. For each $t > 0$, the limit of any convergent subsequence of the sequence $\{\|(\bar{X}_6^n, \bar{Y}_8^n)\|_t : n \geq 1\}$*

is contained in a compact subset of $\mathbb{R}_{14}$.

We also want to control the changes in the queue-length processes over intervals. For that purpose, let $T^n(t)$ be the total number of transitions of the process $(\bar{X}_6^n, \bar{Y}_8^n)$ in the time interval $(0, t]$.

Lemma 4.8.6. *For* $0 \le t < t + u$ *with* $u > 0$,

$$\sup_{t \le s \le t+u} \{|Q_1^n(s) - Q_1^n(t)| + |Q_2^n(s) - Q_2^n(t)|\} \le T^n(t+u) - T^n(t) \le_{st} T_b^n(u), \quad (4.8.6)$$

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

$$c \equiv \lambda_1 + \lambda_2 + \mu_{1,1} m_1 + (\mu_{1,2} \vee \mu_{2,2}) m_2 + (\theta_1 \vee \theta_2) \left((q_1(0) + q_2(0)) \vee \left(\frac{\lambda_1 + \lambda_2}{\theta_1 \vee \theta_2} \right) \right). \quad (4.8.7)$$

As a consequence, $n^{-1} T_b^n \Rightarrow T_b$ *in* D *as* $n \to \infty$*, where* $T_b(t) \equiv ct$*,* $t \ge 0$*, for* c *in* (4.8.7)*. Thus, for any* $(t, u, \tilde{c}, \epsilon)$ *with* $0 \le t < t + u$*,* $\tilde{c} > c$ *and* $\epsilon > 0$*, there exists* $n_0 \equiv n_0(t, u, \tilde{c}, \epsilon)$ *such that*

$$P(T^n(t+u) - T^n(t) > \tilde{c} n u) \le \epsilon \quad \textit{for all} \quad n \ge n_0. \quad (4.8.8)$$

Proof: Apply Lemma 4.8.3 to bound the rate of arrivals and service completions. Apply Corollary 4.8.2 to bound the total queue content, then multiply by $\theta_1 \vee \theta_2$ to bound the rate of abandonments. ■

4.8.3 Positive Recurrence of the Frozen Difference Process

We defined the transition rates of the queue-difference process in (4.5.1). We assumed that $X^n(t_0) = \Gamma^n$ where Γ^n is some fixed deterministic state where sharing is taking place, and specified the transition rates at time t_0. We now consider the *constant-rate QBD* with those transition rates. We also extend the definition by letting Γ^n be a random variable,

where it is understood that Γ^n only determines the constant transition rates, and does not otherwise affect the future evolution of the stochastic process. Let $D_f^n(\Gamma^n) \equiv \{D_f^n(\Gamma^n, t) : t \geq 0\}$ denote this process. (Since t_0 plays no role in (4.5.1), we take it to be 0.) We use the subscript f because we refer to this constant-rate QBD as the *frozen queue-difference process*, thinking of the constant transition rates being achieved because the state has been frozen at the state Γ^n. (As in §4.5.1 the now-constant transition rates in (4.5.2)-(4.5.5) are asymptotically correct as $n \to \infty$ with extra $o(n)$ terms, which we omit.)

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

It is important that this frozen difference process $D_f^n(\Gamma^n)$ can be directly identified with a version of the FSTP, because both are QBD's with the same structure. Indeed, the frozen-difference process can be defined as a version of the FTSP with special state and basic model parameters λ_i and m_j, and transformed time. In order to express the relationship, we indicate the dependence upon the arrival rates and number of servers. In particular,

$$\{D_f^n(\lambda_i^n, m_j^n, \Gamma^n, s) : s \geq 0\} \stackrel{\rm d}{=} \{D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, ns) : s \geq 0\}, \tag{4.8.9}$$

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

$$D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, 0) \equiv D_f^n(\lambda_i^n, m_j^n, \Gamma^n, 0) \equiv Q_1^n(0) - r_{1,2}Q_2^n(0), \tag{4.8.10}$$

where (Q_1^n, Q_2^n) is part of the state Γ^n. This can be checked by verifying that the constant transition rates are indeed identical for the two processes, referring to (4.5.2)-(4.5.5) and (4.5.9)-(4.5.12). Since $\lambda_i^n/n \to \lambda_i$, $i = 1, 2$ and $m_j^n/n \to m_j$, $j = 1, 2$, by virtue of the MS-HT scaling in (4.2.2), we will have the transition rates of $D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, \cdot)$ converge to those of $D(\gamma) \equiv D(\lambda_i, m_j, \gamma, \cdot)$ whenever $\Gamma_n/n \to \gamma$. Of course, (4.8.9)

should not be surprising, because we defined the FTSP in terms of the queue-difference process by a limit that asymptotically reverses (4.8.9): The transition rates of $D(\gamma)$ were defined to be the limit of the transition rates of $D^n(\Gamma_n)/n$ when $\Gamma/n \to \gamma$.

Since the process $D_f^n(X^n(t_0), t)$ has the same QBD structure as the FTSP D, a version of Theorem 4.5.1 holds, i.e., for a given fixed $X^n(t_0)$, the frozen difference process $\{D_f^n(X^n(t_0), t) : t \geq 0\}$ is positive recurrent if and only if

$$\delta_+^n(X^n(t_0)) < 0 < \delta_-^n(X^n(t_0)). \tag{4.8.11}$$

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

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

$$B_n(\xi, \eta) \equiv \{ \sup_{t \in [0,\xi]} \delta_+^n(X^n(t)) < -\eta \quad \text{and} \quad \inf_{t \in [0,\delta]} \delta_-^n(X^n(t)) > \eta \}. \tag{4.8.12}$$

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

Lemma 4.8.7. *There exist $\xi > 0$ and $\eta > 0$ such that $P(B_n(\xi, \eta)) \to 1$ as $n \to \infty$, where $B_n(\xi, \eta)$ is the subset in* (4.8.12)*, on which the process $\{D_f^n(X^n(t), s) : s \geq 0\}$ is positive recurrent for all $t \in [0, \xi]$.*

4.8.4 Continuity of the FTSP QBD

In the remaining proof, we will ultimately reduce everything down to the behavior of the FTSP QBD D. First, we intend to analyze the inhomogeneous queue-difference processes

$D^n(\Gamma^n)$ in terms of associated homogeneous (constant-rate) processes $D_f^n(\Gamma^n)$ introduced in §4.8.3, obtained by freezing the transition rates at the transition rates in the initial state Γ^n. In (4.8.9) above, we showed that the frozen-difference processes can be represented directly in terms of the FTSP, by transforming the model parameters (λ_i, m_j) and the fixed initial state γ and scaling time. In the following subsections, we will appropriately bound the queue-difference processes $D^n(\Gamma^n)$ above and below by associated frozen-queue difference processes, and then transform them into versions of the FTSP D. For the rest of the proof, we will exploit a continuity property possessed by this family of QBD processes. We will be applying this to the FTSP D.

To set the stage, we review basic properties of the QBD process. From the transition rates defined in (4.5.9)-(4.5.12), we see that there are only 8 different transition rates overall. The generator Q in (4.5.17) is based on the four basic $2m \times 2m$ matrices B, A_0, A_1, and A_2, involving the 8 transition rates. By Theorem 6.4.1 and Lemma 6.4.3 of [52], when the QBD is positive recurrent, the FTSP steady-state probability vector has the matrix-geometric form $\alpha_n = \alpha_0 R^n$, where α_n and α_0 are $1 \times 2m$ probability vectors and R is the $2m \times 2m$ rate matrix, which is the minimal nonnegative solutions to the quadratic matrix equation $A_0 + RA_1 + R^2A_2 = 0$, and can be found efficiently by existing algorithms, as in [52]; See Chapter 3 for applications in our settings. If the drift condition (4.5.21) holds, then the spectral radius of R is strictly less than 1 and the QBD is positive recurrent (Corollary 6.2.4 of [52]). As a consequence, we have $\sum_{n=0}^{\infty} R^n = (I - R)^{-1}$. Also, by Lemma 6.3.1 of [52], the boundary probability vector α_0 is the unique solution to the system $\alpha_0(B + RA_2) = 0$ and $\alpha \mathbf{1} = \alpha_0(I - R)^{-1}\mathbf{1} = 1$.

Like any irreducible positive recurrent CTMC, the positive recurrent QBD is regenerative, with successive visits to any state constituting an embedded renewal process. As usual for QBD's (see [52]), we can choose to analyze the system directly in continuous time or in discrete time by applying uniformization, where we generate all potential transitions from

a single Poisson process with a rate exceeding the total transition rate out of any state. In continuous time we focus on the interval between successive visits to the regenerative state; in discrete time we focus on the number of Poisson transitions between successive visits to the regenerative state.

Let τ be the return time and let N be the number of Poisson transitions (with specified Poisson rate). Because of the QBD structure, the return time τ has a moment generating function (mgf) $\phi_\tau(\theta) \equiv E[e^{\theta\tau}]$, for which there exists a critical value $\theta^* > 0$ such that $\phi_\tau(\theta) < \infty$ for $\theta < \theta^*$ and $\phi_\tau(\theta) = \infty$ for $\theta > \theta^*$, while the number of transitions, N, has the generating function (gf) $\psi_N(z) \equiv E[z^N]$, for which there exists a radius of convergence z^* with $0 < z^* < 1$ such that $\psi_N(z) < \infty$ for $z < z^*$ and $\psi_N(z) = \infty$ for $z > z^*$.

Moreover, the mgf $\phi_\tau(\theta)$ and gf $\psi_N(z)$ can be expressed directly in terms of the finite QBD defining matrices. It is easier to do so if we choose a regenerative state, say s^*, in the boundary region (corresponding to the matrix B in (4.5.17)). To illustrate, we discuss the gf. With s^* in the boundary level, in addition to the transitions within the boundary level and up to the next level from the boundary, we only need consider the number of transitions, plus starting and ending states, from any level above the boundary down one level. Because of the QBD structure, these key downward first passage times are the same for each level above the boundary, and are given by the probabilities $G_{i,j}[k]$ and the associated matrix generating function $G(z)$ on p. 148 of [52]. Given $G(z)$, it is not difficult to write an expression for the generating function $\psi_{N^n}(z)$, just as in the familiar BD case; e.g., see §4.3 of [52].

We will be interested in the *cumulative process*

$$C(t) \equiv \int_0^t (f(D(s)) - E[f(D(\infty))])\, ds \quad t \geq 0, \tag{4.8.13}$$

for the special function $f(x) \equiv 1_{\{x \geq 0\}}$. Cumulative processes associated with regenerative

processes obey CLT's and FCLT's, depending upon assumptions about the basic cycle random variables τ and $\int_0^\tau f(D(s))\,ds$, where we assume for this definition that $D(0) = s^*$; see §VI.3 of [7] and [28]. From [14], we have the following CLT with a Berry-Esseen bound on the rate of convergence (stated in continuous time, unlike [14]): For any bounded measurable function f, there exists t_0 such that

$$|E[f(C(t))/\sqrt{t}] - E[f(N(0,\sigma^2))]| \leq \frac{K}{\sqrt{t}} \quad \text{for all} \quad t > t_0, \tag{4.8.14}$$

where

$$\sigma^2 \equiv E\left[\left(\int_0^\tau f(D(s)) - E[f(D(\infty))]\,ds\right)^2\right], \tag{4.8.15}$$

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

There is significant simplification in our case, because the function f in (4.8.14) is an indicator function. Hence, we have the simple domination:

$$\int_0^\tau |f(D(s)|\,ds = \int_0^\tau f(D(s))\,ds \leq \tau \quad \text{w.p.1} \tag{4.8.16}$$

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

We will exploit the following continuity result for QBD's.

Lemma 4.8.8. $(\textit{continuity of QBD's})$ *Consider a sequence of irreducible, positive recurrent QBD's having the structure of the fundamental QBD in §4.5.5, with generator matrices* $\{Q_n : n \geq 1\}$ *of the form* (4.5.17)*. If* $Q_n \to Q$ *as* $n \to \infty$*, where the positive-recurrence drift condition* (4.5.21) *holds for* Q*, then there exists* n_0 *such that the positive-recurrence drift condition* (4.5.21) *holds for* Q_n *for* $n \geq n_0$*. For* $n \geq n_0$*, the quantities* $(R, \alpha_0, \alpha, \phi_\tau, \theta^*, \psi_N, z^*, \sigma^2, K)$ *indexed by* n *are well defined for* Q_n*, where* σ^2 *and* K *are given in* (4.8.14) *and* (4.8.15)*, and converge as* $n \to \infty$ *to the corresponding quantities associated with the QBD with generator matrix* Q*.*

Proof: First, continuity of R, α_0 and α follows from the stronger differentiability in an open neighborhood of any $\gamma \in \mathbb{A}$, which was shown to hold in the proof of Theorem 5.1 in Chapter 3, building on Theorem 2.3 in [34]. The continuity of σ^2 follows from the explicit representation in (4.8.15) above (which corresponds to the solution of Poisson's equation). We use the QBD structure to show that the basic cycle variables τ and $\int_0^\tau f(D(s))\,ds$ are continuous function of Q, in the sense of convergence in distributions (or convergence of mgf's and gf's) and then for convergence of all desired moments, exploiting (4.8.16) and the mgf of τ to get the required uniform integrability. Finally, we get the continuity of K from [14] and the continuity of the third absolute moments of the basic cycle variables, again exploiting the uniform integrability. We will have convergence of the characteristic functions used in [14]. However, we do not get an explicit expression for the constants K. ∎

We use the continuity of the steady-state distribution α in (5.5.33) in §5.5.5. In addition, we use the following corollary to Lemma 4.8.8 in (5.5.32) in §5.5.5.

Corollary 4.8.3. *If* $(\lambda_i^n, m_j^n, \gamma_n) \to (\lambda_i, m_j, \gamma)$ *for our FTSP QBD's, where* (4.5.21) *holds*

for (λ_i, m_j, γ)*, then for all* $\epsilon > 0$ *there exist* t_0 *and* n_0 *such that*

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

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

Proof: First apply Lemma 4.8.8 for the steady-state probability vector α, to find n_0 such that $|P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)| - P(D(\lambda_i, m_j, \gamma, \infty) > 0)| < \epsilon/2$ for all $n \geq n_0$. By the triangle inequality, henceforth it suffices to work with $P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)$ in place of $P(D(\lambda_i, m_j, \gamma, \infty) > 0)$ in the statement to be proved. By (4.8.14), for any M, there exists t_0 such that for all $t \geq t_0$,

$$\begin{aligned} P\left(|\frac{1}{t}\int_0^t 1_{\{D(\lambda_i^n, m_j^n, \gamma_n, s)>0\}}\, ds - P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)| > \frac{M}{\sqrt{t}}\right) \\ < P(|N(0, \sigma^2(\lambda_i^n, m_j^n, \gamma_n))| > M) + \frac{K(\lambda_i^n, m_j^n, \gamma_n)}{\sqrt{t}}. \end{aligned} \tag{4.8.17}$$

Next, choose M so that $P(|N(0, \sigma^2(\lambda_i, m_j, \gamma))| > M) < \epsilon/2$. Then, invoking Lemma 4.8.8, increase n_0 and t_0 if necessary so that $|\sigma^2(\lambda_i^n, m_j^n, \gamma_n)) - \sigma^2(\lambda_i, m_j, \gamma))|$ and $|K(\lambda_i^n, m_j^n, \gamma_n) - K(\lambda_i, m_j, \gamma)|$ are sufficiently small so that the right side of (4.8.17) is less than $\epsilon/2$ for all $n \geq n_0$ and $t \geq t_0$. If necessary, increase t_0 and n_0 so that $M/\sqrt{t_0} < \epsilon/2$. With those choices, the objective is achieved. ∎

4.8.5 Process Bounds

Our next step is to find a $\xi > 0$ for which we can uniformly bound the frozen difference processes $\{D_f^n(X^n(t), \cdot)\}$ and the queue-difference processes $\{D_{1,2}^n(t)\}$ for all $t \in [0, \xi]$, with two QBD's - one from above and the other from below. We thus translate the uniformity of the bounds on the drifts, established in Lemma 4.8.7, to a uniformity of bounds on the family of process $\{D_f^n(X^n(t), \cdot)\}$ for $t \in [0, \xi]$. Having two bounding QBD's will

eventually allow us to use a sandwiching argument. Now, instead of sample path stochastic order, we use rate order, denoted by $X_1 \leq_r X_2$, by which we mean that, from every integer state and for every possible state that can be reached from that state in a single transition, both (i) the transition rates up in CTMC X_1 are less than or equal to the corresponding transition rates up in CTMC X_2, and (ii) the transition rates down in CTMC X_1 are greater than or equal to the corresponding transition rates down in CTMC X_2.

Lemma 4.8.9. *There exist $\xi > 0$ and $\eta > 0$, random vectors X_M^n and X_m^n, and a sequence of sets $\{B_n(\xi,\eta) : n \geq 1\}$ in the underlying probability space with $P(B_n(\xi,\eta)) \to 1$ as $n \to \infty$, such that, for $0 \leq t \leq \xi$,*

$$\begin{aligned} D_f^n(X_m^n,\cdot) &\leq_r D_f^n(X^n(t),\cdot) \leq_r D_f^n(X_M^n,\cdot), \\ D_f^n(X_m^n,\cdot) &\leq_r D_{1,2}^n(t) \leq_r D_f^n(X_M^n,\cdot), \end{aligned} \tag{4.8.18}$$

where the bounding processes $D_f^n(X_M^n,\cdot)$ and $D_f^n(X_m^n,\cdot)$, and thus also the interior processes $D_f^n(X^n(t),\cdot)$, satisfy (4.8.12) *on $B_n(\xi,\eta)$, $n \geq 1$, and are thus positive recurrent.*

When $r_{1,2} = 1$, rate order directly implies the stronger sample path stochastic order, but not more generally, because the upper (lower) process can jump down below (up above) the lower (upper) process when the lower process is at state 0 or below, while the upper process is just above state 0. Nevertheless, we can obtain the following stochastic order bound, involving a finite gap. However, there is no gap when $r_{1,2} = 1$ because then $j = k = 1$.

Corollary 4.8.4. *Let $\zeta \equiv (j \vee k) - 1$. Under the conditions of Lemma* 4.8.9*, there exist $\xi > 0$ and $\eta > 0$, random vectors X_M^n and X_m^n, and a sequence of sets $\{B_n(\xi,\eta) : n \geq 1\}$ in the underlying probability space with $P(B_n(\xi,\eta)) \to 1$ as $n \to \infty$, such that, whenever*

$$\begin{aligned} D_f^n(X_m^n,0) - \zeta &\leq_{st} D_f^n(X^n(0),0) \leq_{st} D_f^n(X_M^n,0) + \zeta, \\ D_f^n(X_m^n,0) - \zeta &\leq_{st} D_{1,2}^n(0) \leq_{st} D_f^n(X_M^n,0) + \zeta, \end{aligned} \tag{4.8.19}$$

in $\mathbb{R}$*,*

$$\begin{aligned} D_f^n(X_m^n,\cdot)-\zeta &\leq_{st} D_f^n(X^n(t),\cdot)\leq_{st} D_f^n(X_M^n,\cdot)+\zeta, \\ D_f^n(X_m^n,t)-\zeta &\leq_{st} D_{1,2}^n(t)\leq_{st} D_f^n(X_M^n,t)+\zeta, \end{aligned} \tag{4.8.20}$$

in $\mathcal{D}([0,\xi]))$*, where the bounding processes* $D_f^n(X_M^n,\cdot)$ *and* $D_f^n(X_m^n,\cdot)$*, and thus also* $D_f^n(X^n(t),\cdot)$*, satisfy* (4.8.12) *on* $B_n(\xi,\eta)$*,* $n\geq 1$*, and are thus positive recurrent.*

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

As an immediate corollary to Corollary 4.8.4, we can deduce stochastic boundedness (SB) as $n\to\infty$. The following corollary implies Theorem 4.5.4.

Corollary 4.8.5. *For* $n\geq 1$*, let* $\mathcal{S}^n$ *be the set of all processes* $\{D_{1,2}^n(t):0\leq t\leq\xi\}$ *and* $\{D_f^n(X^n(t),s):0\leq s\leq\xi\}$ *for* $0\leq t\leq\xi$ *with* ξ *from Corollary* 4.8.4*. (The sets* $\mathcal{S}^n$ *form an uncountably infinite subset of the space* $\mathcal{D}([0,\xi]).$*) Suppose that condition* (4.8.19) *is satisfied. Then the sequence* $\{\mathcal{S}^n:n\geq 1\}$ *is SB. Consequently, the sequence of processes* $\{\{D_{1,2}^n(t):0\leq t\leq\xi\}:n\geq 1\}$ *is SB in* $\mathcal{D}([0,\xi])$*, so that the sequence* $\{D_{1,2}^n(t):n\geq 1\}$ *is SB in* $\mathbb{R}$ *for each* t *with* $0\leq t\leq\xi$*.*

Proof: By letting $n \to \infty$ in Corollary 4.8.5, we are able to exploit the stochastic order bound in (4.8.20), where the bounds are positive recurrent, satisfying (4.8.12). ∎

We will later show that the conclusions of Corollary 4.8.5 hold when ξ is replaced by δ, where $[0, \delta)$ is the interval over which there exists a unique solution to the ODE in $\mathbb{A}$. Together with Theorem 4.5.3, Corollary 4.8.5 proves that the sequence of processes $\{\{D^n_{1,2}(t) : 0 \le t \le \xi\} : n \ge 1\}$ is SB but not tight in $\mathcal{D}([0, \xi])$; the oscillations are too rapid.

4.8.6 Special Construction to Bound the Integrals

The comparisons in Lemma 4.8.9 and Corollary 4.8.4 are important, but they are not directly adequate for our purpose. The sample-path stochastic order bound works fine for the special case of $r_{1,2} = 1$, but not more generally, because of the gap ζ. However, we now show that an actual gap will only be present rarely, if we choose the interval length ξ small enough and n big enough. We use the construction in the previous section, exploiting the fact that we have rate order, where the bounding rates can be made arbitrarily close to each other by choosing the interval length ξ suitably small.

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

$$D^n_{1,2}(0) = D^n_f(X^n_m, 0) = D^n_f(X^n_M, 0) = D^n_f(X^n(t), 0) = j \tag{4.8.21}$$

for some fixed j, for all t, $0 \le t \le \xi$.

Lemma 4.8.10. *Assume that condition* (4.8.21) *holds. For any $\epsilon > 0$, there exist $\xi > 0$ and $\eta > 0$, random vectors X^n_M and X^n_m, associated QBD processes $\{D^n_f(X^n_M, s) : s \ge 0\}$ and $\{D^n_f(X^n_m, s) : s \ge 0\}$ (with constant transition rates), and a sequence of sets $\{B_n(\xi, \eta) : n \ge 1\}$ in the underlying probability space with $P(B_n(\xi, \eta)) \to 1$ as $n \to \infty$, such that,*

on the set $B_n(\xi,\eta)$*,*

$$\begin{aligned} \delta_+^n(X_m^n) &< -\eta \quad \textit{and} \quad \delta_-^n(X_m^n) > \eta, \\ \delta_+^n(X_M^n) &< -\eta \quad \textit{and} \quad \delta_-^n(X_M^n) > \eta \end{aligned} \tag{4.8.22}$$

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

$$\begin{aligned} \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_m^n,s)>0\}}\,ds - \epsilon \le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X^n(t),s)>0\}}\,ds \\ \le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_M^n,s)>0\}}\,ds + \epsilon \\ \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_m^n,s)>0\}}\,ds - \epsilon \le \frac{1}{\xi}\int_0^\xi 1_{\{D_{1,2}^n(s)>0\}}\,ds \\ \le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_M^n,s)>0\}}\,ds + \epsilon. \end{aligned} \tag{4.8.23}$$

4.8.7 Proof of Theorem 4.6.1

By the tightness established in Lemma 4.8.1, we know that every subsequence of $\{\bar{X}^n : n \in \mathbb{N}\}$ has a further subsequence converging weakly in $\mathcal{D}_3$. We will be considering a converging subsequence with limit $\bar{X}$, but without changing the indexing notation. (We understand that n runs through a subsequence.) It suffices to show that the limit $\bar{X}$ is deterministic and satisfies the ODE in (4.5.13) or, equivalently, the integral representation in (4.5.14).

By Theorems 4.4.1 and 4.4.2, which draws on §4.7, it suffices to focus on the integral representation for $\bar{X}^n$ in (4.4.7). Many of the terms converge directly to their counterparts in (4.5.14) because of the assumed MS-HT scaling in §4.2.1 and the convergence $\bar{X}^n \Rightarrow \bar{X}$ through the subsequence obtained from the tightness. Indeed, the only exceptions are the integral terms involving the indicator functions. However, these integral terms are easily

seen to be tight as well, as a conseuence of the tightness of the sequences $\{\bar{Z}^n_{i,j} : n \geq 1\}$ established in §4.8.1. Hence, we can consider a subsequence of our original converging subsequence in which all these integral terms converge to proper limits as well. Hence we have the integral representation in (4.4.7) converge to the system

$$\begin{aligned}
\bar{Z}_{1,2}(t) &= z_{1,2}(0) + \mu_{2,2}\bar{I}_{z,1}(t) - \mu_{1,2}\bar{I}_{z,2}(t) \\
\bar{Q}_1(t) &= q_1(0) + \bar{\lambda}_1 t - \bar{m}_1 t - \mu_{1,2}\bar{I}_{q,1,1}(t) \\
&\quad - \mu_{2,2}\bar{I}_{q,1,2}(t) - \theta_1 \int_0^t \bar{Q}_1(s)\, ds, \\
\bar{Q}_2(t) &= q_2(0) + \bar{\lambda}_2 t - \mu_{2,2}\bar{I}_{q,2,1}(t) \\
&\quad - \mu_{1,2}\bar{I}_{q,2,2}(t) - \theta_2 \int_0^t \bar{Q}_2(s))\, ds.
\end{aligned} \tag{4.8.24}$$

We have exploited the assumed convergence of the initial conditions in Assumption 3 to replace $\bar{X}(0)$ by $x(0)$ in (4.8.24). In more detail, for one integral term we have

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

through the final converging subsequence.

At this point, it suffices to identify the limit of each integral term with the corresponding term in the integral representation in (4.5.14). That will uniquely characterize the limit over an initial interval $[0, \delta)$ because, by Theorem 4.5.2, there exists a unique solution to the ODE over an initial interval $[0, \delta)$. Since each of these integrals can be treated in essentially the same way, we henceforth focus only on the term $\bar{I}_{q,1,2}(t)$. Thus, it suffices to show that

$$\bar{I}_{q,1,2}(t) = \int_0^t \pi_{1,2}(\bar{X}(s))\bar{Z}_{1,2}(s))\, ds \tag{4.8.25}$$

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

to show that

$$\bar{I}_{q,1,2}(t+\xi) - \bar{I}_{q,1,2}(t) = \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t)\xi + o(\xi) \quad \text{as} \quad \xi \to 0. \tag{4.8.26}$$

We achieve that goal by applying Lemma 4.8.11 below. ■

Recall that $[0, \delta)$ is the interval where the ODE has a unique solution. It is initially reduced to satisfy the requirements of §4.7, but then can be increased once a smaller interval has been treated. However, here we reduce δ again if necessary, so that $\delta < \xi$ for ξ in Lemmas 4.8.7, 4.8.9 and 4.8.10. After Lemma 4.8.11 and Theorem 4.6.1 have been proved for this reduced δ, δ can be further increased to the point where the existence of a unique solution to the ODE has been determined. Below we will be introducing a new ξ less than this new δ.

Lemma 4.8.11. (convergence of the integral terms) *For any $\epsilon > 0$ and t with $0 \le t < \delta$, with δ specified above, there exists $\xi \equiv \xi(\epsilon, \delta, t)$ with $0 < \xi < \delta - t$ and n_0 such that*

$$P\left(|\frac{1}{\xi}\int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}}\bar{Z}^n_{1,2}(s)\, ds - \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t)| > \epsilon\right) < \epsilon \tag{4.8.27}$$

for all $n \ge n_0$.

Chapter 5

Remaining Proofs in Chapter 4

This chapter is dedicated the remaining proofs in Chapter 4, and consists of five sections. The material is presented in the order of the associated material in Chapter 4. Section 5.1 contains the proofs of Theorems 4.5.3 and 4.5.5 in §4.5. Section 5.2 contains the proofs for theorems and lemmas establishing SSC for the service processes in §4.7. Sections 5.3 and 5.4 contain supplementary material for §4.7. In particular, Section 5.3 displays the bounding QBD used in the proof of Lemma 4.7.4, while Section 5.4 provides more on the idleness processes, going beyond Theorem 4.7.4 in §4.7.

Section 5.5 contains the proofs for the theorems and lemmas completing the proof of Theorem 4.6.1 in §4.8. Section 5.5 has four subsections, corresponding to the subsections of §4.8 where the results are located.

5.1 Remaining Proofs in Section 4.5

In this section we provide the two remaining proof in §4.5: We prove Theorems 4.5.3 and 4.5.5.

5.1.1 Proof of Theorem 4.5.3

We first establish the claimed convergence of processes in (4.5.23). For any $\gamma \in \mathbb{A}$, the limiting FTSP $\{D(\gamma, s) : s \geq 0\}$ is a CTMC with bounded constant transition rates, as specified in §4.5.2. (In this section we view the FTSP as a CTMC rather than as a QBD process.) Hence, the FTSP can make only finitely many transitions in any bounded interval. Moreover, there are only four possible transitions from any state, and there are only two possible forms for these transitions, depending upon whether $D(\gamma, s) > 0$ or $D(\gamma, s) \leq 0$. Thus, the FTSP is a well-defined random element of $\mathcal{D}$. In this framework of integer-valued processes, convergence in $\mathcal{D}$ is equivalent to convergence of the finite-dimensional distributions (fidi's).

The converging processes $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ defined in (4.5.7) are more complicated, having time-dependent transition rates, but they have essentially the same structure. For each n and s, these processes also have only four possible transitions from any state, and there are only two possible forms for these transitions, depending upon whether $D_e^n(\Gamma^n, s) > 0$ or $D_e^n(\Gamma^n, s) \leq 0$. By assumption, the initial conditions converge. Since $\Gamma^n/n \to \gamma$ as $n \to \infty$, and because of the special time scaling in (4.5.7), we have uniform convergence of the time-varying transition rates of $D_e^n(\Gamma^n, s) > 0$ to the constant transition rates of the FTSP over the interval $[0, t]$. Hence, we have convergence of the fidi's, and thus convergence in $\mathcal{D}$.

We now elaborate on the way this last step can be formalized. That can be done cleanly using a uniformization framework, as in Theorem 3.1 of [51], in which all transitions of $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ are generated from a single Poisson process with constant rate. However, there is a complication, because in general the transition rates are not unbounded above. One approach to this problem is to use adaptive uniformization as in [55] and references cited therein. However, by Corollary 4.8.1, the scaled total queue content $n^{-1}Q_\Sigma^n$

is stochastically bounded above by a process $n^{-1}Q^n_{bd}$, which converges in law to the deterministic finite bound $q_{bd}(t) \leq q^*_{bd}$ given in (4.8.5). Hence, $D^n_e(\Gamma^n, \cdot)$ is asymptotically equivalent to a process with uniformly bounded transition rates. (For a direct stochastic bound on the number of transitions over a subinterval, see Lemma 4.8.6.) Hence, without loss of generality, we work with the asymptotically equivalent processes that do have uniformly bounded transition rates. However, we do not introduce new notation; instead we simply act as if $n^{-1}Q^n_\Sigma$ is bounded above and the transition rates of $\{D^n_e(\Gamma^n, s) : s \geq 0\}$ are bounded above. Hence, we just apply standard uniformization.

Given the Poisson process with a fixed rate, which exceeds the transition rate out of any state, all potential transitions are the transition epochs of the Poisson process. The actual transitions at the transition epochs of the Poisson process occur according to a discrete-time Markov chain (DTMC). However, in our nonstationary context, the DTMC is nonstationary as well. In particular, as in [51], we can express the time-dependent transition function as

$$\begin{aligned} P^{(n)}_{i,j}(t) &\equiv P(D^n_e(\Gamma^n, t) = j | D^n_e(\Gamma^n, 0) = i) \\ &= \sum_{k=0}^{\infty} \frac{e^{-\eta t}(\eta t)^k}{k!} \int \cdots \int_{0 \leq s_1 < s_2 < \cdots < s_k \leq t} (\prod_{l=1}^{k} P^{(n)}_\eta(s_l))_{i,j} \frac{k!}{t^k} \, ds_1 \cdots ds_k, \end{aligned} \tag{5.1.1}$$

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

Thus, for any given time interval $[0, t]$ and $\epsilon > 0$, we can find an integer ν such that the total number of transitions of all of the processes $\{D^n_e(\Gamma^n, s) : s \geq 0\}$ over $[0, t]$ is at most ν with probability $1 - \epsilon$. This will apply to all processes under discussion. Moreover, the occurrence of those ν transitions is distributed over $[0, t]$ according to ν i.i.d. uniformly random variables, using the classical property of the Poisson process. We can thus take the number ν and the locations of the transitions as fixed, independent of n. We

are then left with the product of ν DTMC transition matrices at time-varying locations, as shown in (5.1.1). These transition matrices here are infinite matrices, but each has at most 5 positive entries in each row. For any given ν and initial state, we can only reach a finite number of states. So, at this point, these transition matrices actually are equivalent to finite matrices. Moreover, these transition matrices converge to the common limiting transition matrix corresponding to the FTSP, uniformly. Hence, we can uniformly bound the difference between the product of these ν matrices and the corresponding product for the FTSP, independent of their time-varying locations. In that way, we can bound the total error by an arbitrarily small quantity by choosing first ν and then n to be suitably large. ■

5.1.2 Proof of Theorem 4.5.5

By Corollary 4.8.5, the sequence of random variables $\{D^n_{1,2}(t) : n \geq 1\}$ is SB. Since SB is equivalent to tightness in $\mathbb{R}$, every subsequence has a converging subsequence. We show that every such converging subsequence must converge to the random variable $D(x(t), \infty)$, which has the steady-state distribution of the FTSP D determined by the fluid state $x(t)$ at time t. That implies that the entire sequence must converge, so that completes the proof.

To characterize the limit of a convergent subsequence, we exploit the continuity of, first, $x(t)$ and, second, $D(x(t), \infty)$, exploiting Lemma 4.8.8. With these properties, we obtain the following lemma, which relates the FTSP at finite times to its steady-state distribution.

Lemma 5.1.1. *For any t_0 with $0 \leq t_0 < \delta$, where δ is chosen to ensure that the ODE has a unique solution x with $x(t) \in \mathbb{A}$ for all $t \in [0, \delta)$, and any $\epsilon > 0$, there exists s_0 and $\zeta > 0$ such that $t_0 + \zeta < \delta$ and*

$$D(x(t_0), \infty) - \epsilon \leq_{st} D(x(t), s) \leq_{st} D(x(t_0), \infty) + \epsilon \quad \textit{in} \quad \mathbb{R} \tag{5.1.2}$$

for all $s \geq s_0$ and all $t \in (t_0 - \zeta, t_0 + \zeta)$.

Proof: As stated above, Lemma 4.8.8 establishes continuity in t of the distributions of the steady-state variables $D(x(t), \infty)$ of the FTSP D. It also establishes continuity for the distribution of the return time to a fixed regeneration state. Thus, we can establish uniform (geometric) rate of convergence to the steady state distribution as $s \to \infty$ (uniform in t near t_0) by exploiting a coupling construction, as in Lemma VII.2.9 of [7]. The proof there provides explicit expressions to provide uniform bounds on the rate of convergence for t in a small neighborhood of any t_0. ∎

Next, by Theorem 4.6.1, $\bar{X}^n \Rightarrow x$ in $D([0, \delta))$ as $n \to \infty$, where x is a deterministic continuous function with $x(t) \in \mathbb{A}$ for all $t \in [0, \delta)$. (We do not apply Theorem 4.5.5 in the proof of Theorem 4.6.1.) Then we can apply Theorem 4.5.3, just proved above, to obtain

$$D^n_{1,2}(X^n(t), t + s_0/n) = D^n_e(X^n(t), s_0) \Rightarrow D(x(t), s_0) \quad \text{as} \quad n \to \infty. \tag{5.1.3}$$

From the proof of Theorem 4.5.3 we can conclude the convergence is uniform for t in a neighborhood of t_0. Hence we can apply Lemma 5.1.1 to conclude that there exists n_0 such that

$$D(x(t_0), \infty) - 2\epsilon \leq_{st} D^n_{1,2}(X^n(t), t + s_0/n) \leq_{st} D(x(t_0), \infty) + 2\epsilon \quad \text{in} \quad \mathbb{R} \tag{5.1.4}$$

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

In closing, we remark that a minor variant of Lemma 4.8.11 (proved in the same way) establishes the weaker limit for local averages:

$$\lim_{\xi \downarrow 0} \lim_{n \to \infty} \frac{1}{\xi} \int_t^{t+\xi} 1_{\{D^n_{1,2}(s) \leq k\}} ds \to P(D(x(t), \infty) \leq k) \quad \text{for all} \quad k, \tag{5.1.5}$$

but (5.1.5) and tightness alone are evidently insufficient to establish the desired result.

5.2 Remaining Proofs in Section 4.7

Proof of Theorem 4.7.2: Our proof is based on regenerative structure. The intervals between successive visits to the state $(0, j)$ constitute an embedded renewal process for the QBD. Since the QBD is positive recurrent, these cycles have finite mean. Given the regenerative structure, our proof is based on the observation that, if the process $\mathcal{L}$ were continuous real-valued with an exponential tail, instead of integer valued with a geometric tail, then we could establish the conventional convergence in law of $\|\mathcal{L}\|_t - c \log t$ to the Gumbel distribution, which implies our conclusion. Hence, we bound the process $\mathcal{L}$ above w.p.1 by another process $\mathcal{L}_b$ that is continuous real-valued with an exponential tail and which inherits the regenerative structure of $\mathcal{L}$.

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

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

where X is an exponential random variable distributed as X_i that is independent of $\mathcal{L}(t)$. We now consider the stationary version of $\mathcal{L}$, which makes $\mathcal{L}_b$ stationary as well. We let the desired constant c be the mean of the exponential random variables X_i. If we make c

sufficiently large, then we clearly have $P(\mathcal{L}_b(t) > x) \sim e^{-x/c}$ as $x \to \infty$, because the first and third terms become asymptotically negligible as $x \to \infty$. (We choose c to make $\mathcal{L}(t)$ asymptotically negligible compared to X.)

It now remains to establish the conventional extreme-value limit for the bounding process $\mathcal{L}_b$. For that, we exploit the exponential tail of the stationary distribution, just established, and regenerative structure. There are two approaches to extreme-value limits for regenerative processes, which are intimately related, as shown by Rootzén [63]. One is based on stationary processes, while the other is based on the cycle maxima, i.e., the maximum values achieved in successive regenerative cycles. First, if we consider the stationary version, then we can apply classical extreme-value limits for stationary processes as in [53]. The regenerative structure implies that the mixing condition in [53] is satisfied; see Section 4 of [63].

However, the classical theory in [53] and the analysis in [63] applies to sequences of random variables as opposed to continuous-time processes. In general, the established results for stationary sequences in [53] do not extend to stationary continuous-time processes. That is demonstrated by extreme-value limits for positive recurrent diffusion processes in [15, 23]. Proposition 3.1, Corollary 3.2 and Theorem 3.7 of [15] show that, in general, the extreme-value limit is not determined by the stationary distribution of the process.

However, continuous time presents no difficulty in our setting, because the QBD is constant between successive transitions, and the transitions occur in an asymptotically regular way. It suffices to look at the embedded discrete-time process at transition epochs. That is a standard discrete-time Markov chain associated with the continuous-time Markov chain represented as a QBD. Let $N(t)$ denote the number of transitions over the interval $[0, t]$. Then $\mathcal{L}_b(t) = \mathcal{L}_d(N(t))$, where $\mathcal{L}_d(n)$ is the embedded discrete-time process associated with $\mathcal{L}_b$. Since $N(t)/t \to c' > 0$ w.p.1 as $t \to \infty$ for some constant $c' > 0$, the results directly established for the discrete-time process D_d are inherited with minor modification

by $\mathcal{L}_b$. Indeed, the maximum over random indices already arises when relating extremes for regenerative sequences to extremes of i.i.d. sequences; see p. 372 and Theorem 3.1 of [63]. In fact, there is a substantial literature on extremes with a random index, e.g., see Proposition 4.20 and (4.53) of [61] and also [64]. Hence, for the QBD we can initially work in discrete time, to be consistent with [53, 63]. After doing so, we obtain extreme-value limits in both discrete and continuous time, which are essentially equivalent.

So far, we have established an extreme-value limit for the stationary version of $\mathcal{L}_b$, but our process $\mathcal{L}_b$ is actually not a stationary process. So it is natural to apply the second approach based on cycle maxima, which is given in [63, 6] and Section VI.4 of [7]. We would get the same extreme-value limit for the given version of $\mathcal{L}_b$ as the stationary version if the cycle maximum has an exponential tail. Moreover, this reasoning would apply directly to continuous time as well as discrete time. However, Rootzén [63] has connected the two approaches (see p. 380 of [63]), showing that all the versions of the regenerative process have the same extreme-value limit. Hence, the given version of the process $\mathcal{L}_b$ has the same extreme-value limit as the stationary version, already discussed. Moreover, as a consequence, the cycle maximum has an exponential tail if and only if the stationary distribution has an exponential tail. Hence, we do not need to consider the cycle maximum directly. ■

Remark 5.2.1. (an alternative proof) An alternative proof of Theorem 4.7.2 would be based on a direct demonstration that the cycle maximum of $\mathcal{L}$ has a geometric tail. That alternative reasoning has the advantage that it applies directly in continuous time; see [6] and Section VI.4 of [7]. However, we are unaware of such a result in the literature. Evidently, it can be derived from the known behavior of the first passage times between levels. By Theorem 8.2.2 of [52], the probability of moving from level 0 to level $k+1$ before returning to level 0 is asymptotically geometric as $k \to \infty$. However, the return to level 0 may not be in the same phase as the initial phase. Hence, we must consider the random evolution within

level 0 until we either hit the initial phase or leave level 0, and then the random number of those returns until we do return to level 0 in the same phase as the initial phase. Evidently that will not alter the geometric tail, but that remains to be shown.

In fact, if we show that the cycle maximum has a geometric tail, then we need not construct the bounding process $\mathcal{L}_b$. Instead, we can directly apply the extreme-value theorem for regenerative processes with geometric tail, Theorem 6 in [4] or Problem 4.2 on p. 185 of [7], from which our conclusion would follow. In particular, it is well known that the maximum queue length over a busy cycle in an $M/M/1$ is asymptotically geometric. We can thus use Theorem 6, and, more directly, the example on p. 112 in [4], for the extreme-value bound for the $M/M/1$ queue-length process, which we apply in the proof of Theorem 4.7.4.

Proof of Lemma 4.7.2: By Assumption 3, the condition $z_{1,2}(0) > 0$ implies that $P(Z^n_{1,2}(0) > 0) \to 1$ as $n \to \infty$. Clearly, for every $n \geq 1$, $Z^n_{1,2}$ is stochastically bounded from below, in sample-path stochastic order, by a process Z^n_b which has $Z^n_b(0) = Z^n_{1,2}(0)$, has only departures and no new arrivals, i.e., $Z^n_{1,2} \geq_{st} Z^n_b$ for all $n \geq 1$ and $t \geq 0$, where

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

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

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

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

It follows that $z_b(t) \geq z_b(0)e^{-\mu_{1,2}t}$, so that $z_b(t) > 0$ for all $t \geq 0$. Thus $P(\inf_{0\leq s\leq t} Z^n_b(s) > 0) \to 1$ as $n \to \infty$. The stochastic order bound implies that the same is true for $Z^n_{1,2}$, which

proves the first claim of the lemma. The second claim that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$ follows from the first together with the one-way sharing rule. ∎

Proof of Lemma 4.7.3: When either of the conditions (i) or (ii) holds, $d_{2,1}(0) < 0$, where $d_{2,1}(t) \equiv r_{2,1}q_2(t) - q_1(t)$, $t \geq 0$. Under condition (i), by Assumption 3, $-d_{2,1}(0) \geq d_{1,2}(0) \equiv q_1(0) - r_{1,2}q_2(0) = \kappa$. If $\kappa = 0$ and Condition (ii) holds, then $d_{2,1}(0) < r_{1,2}q_2(0) - q_1(0) = \kappa = 0$.

We will construct a sample-path stochastic-order bound from above for $D^n_{2,1}$, and show that this bounding process is asymptotically strictly negative on an interval $[0, \tau]$, for some $\tau > 0$. To stochastically bound $D^n_{2,1}$, we consider a sequence of systems $\{X^n_b : n \geq 1\}$ in (4.7.5) initialized at time 0 with $X^n_b(0) \equiv X^n(0)$, $n \geq 1$. Thus, $Q^n_{i,b}(0) = Q^n_i(0)$, and both service pools start full with only their own customers. (Recall that we are considering the case $Z^n_{1,2}(0) = 0$ for all n large enough.)

Let $D^n_b \equiv r_{2,1}Q^n_{2,b} - Q^n_{1,b}$ be the weighted difference process in X^n_b. By construction, $Q^n_{1,b} \leq_{st} Q^n_1$ and $Q^n_{2,b} \geq_{st} Q^n_2$, so that $D^n_b \geq_{st} D^n_{2,1}$. Now, as was shown in §4.7.2, $\bar{X}^n_b \Rightarrow x_b$ as $n \to \infty$, for x_b in (4.7.7). Hence, $\bar{D}^n_b \equiv D^n_b/n \Rightarrow d_b \equiv r_{2,1}q_{2,b} - q_{1,b}$ as $n \to \infty$, with $d_b(0) < 0$.

The limit process $q_{1,b}(t)$ may eventually become negative as t increases, at which point it becomes meaningless as a stochastic-order bound for q_1. However, the continuity of $q_{1,b}$, together with the initial condition, $q_{1,b}(0) > 0$, implies that we can find a time $\tau_1 > 0$, such that $q_{1,b}(t) > 0$ for all $t \in [0, \tau_1]$. Similarly, the continuity of d_b implies that there exists $\tau_b > 0$, where $\tau_b \equiv \inf\{t \geq 0 : d_b(t) = 0\}$. Then, for $\tau^n_b \equiv \inf\{t \geq 0 : D^n_b(t) \geq 0\}$, by applying a version of Theorem 13.6.4 in [78], the continuous mapping theorem gives $\tau^n_b \Rightarrow \tau_b$. Now, for $\tau^n \equiv \inf\{t \geq 0 : D^n_{2,1}(t) \geq 0\}$ we have that $\tau^n \geq_{st} \tau^n_b$. Taking $\tau \equiv \tau_1 \wedge \tau_b$ gives the first claim of the statement.

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

Proof of Lemma 4.7.4: We will prove the lemma by constructing a QBD process that serves as a stochastic-order bound for the process $D^n_{2,1}$ over some interval $[0, \tau]$. The claims will then follow from an application of the extreme-value limit in Theorem 4.7.2. As a first step, we define the following processes:

For $s \geq 0$, let $X^n_*(s) \equiv (Q^n_{1,a}(s), Q^n_{2,a}(s), Z^n_b(s))$, where $Q^n_{i,a}$, $i = 1, 2$, are defined in (4.7.4) and Z^n_b is defined in (4.7.5). For a fixed $s > 0$ and a fixed $X^n_*(s)$, define the following processes:

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

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

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

conditional on $X^n_*(s)$, is a continuous-time Markov chain as a function of the time argument t. (That is because X^n_* is constructed independently of D^n_*.) The key observation here is that the conditional process D^n_* (given $X^n_*(s)$), can be analyzed as a QBD, just as in §2.4. In particular, if $r_{2,1} = j/k$, where j, k are positive integers with no common divisors, then the process $\tilde{D}^n_* \equiv j Q^n_{2,*} - k Q^n_{1,*}$ is a CTMC with state space in the nonnegative integers, and can be represented as a QBD; See §5.3. Moreover, the process $\tilde{D}^n_*$ is positive recurrent

if and only if D^n_* is.

Our next objective is to replace the family of processes $\{\tilde{D}^n_*(X^n_*(s),t) : t \geq 0\}$ (there is a different process for each $X^n_*(s)$) with one positive-recurrent QBD which will bound $D^n_{2,1}$ from above over an entire interval $[0,\tau]$, for some $\tau > 0$, and then translate the scaling by n in X^n_* to a scaling by n of the time argument t. More specifically, we continue the proof in two steps: in the first step we find a positive recurrent QBD $D^n_*(X^n_m,t)$, such that $D^n_*(X^n_m,\cdot) \geq_{st} D^n_*(X^n_*(s),\cdot)$ for all $s \in [0,\tau]$. In the second step, the bounding process $D^n_*(X^n_m,\cdot)$ is shown to be equal in distribution to a rate-1 QBD on the interval $[0,a_n\tau]$, for some $\{a_n\}$ such that $a_n/n \to 1$ as $n \to \infty$. The second step allows us to employ Theorem 4.7.2 and show that the probability that the threshold $k^n_{2,1}$ is crossed over $[0,\tau]$ converges to 0 as $n \to \infty$.

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

Given $X^n_*(s)$, the process $\{\tilde{D}^n_*(X^n_*(s),t) : t \geq 0\}$ has upward jumps of size j with rate $\hat{\lambda}_j(X^n_*(s)) \equiv \lambda^n_2$, and downward jumps of size j (away from the boundary) with rate $\hat{\mu}_j(X^n_*(s)) \equiv \mu_{2,2}(m^n_2 - Z^n_b(s)) + \theta_2 Q^n_{2,a}(s)$. It has upward jumps of size k with rate $\hat{\lambda}_k(X^n_*(s)) \equiv \mu_{1,1}m^n_1 + \mu_{1,2}Z^n_b(s) + \theta_1 Q^n_{1,a}(s)$, and downwards jumps of size k (away from the boundary) with rate $\hat{\mu}_k(X^n_*(s)) \equiv \lambda^n_1$. Now, by Theorem 7.2.3 in [52], for a given $X^n_*(s)$, $\tilde{D}^n_*(X^n_*(s),\cdot)$ is positive recurrent if and only if $\tilde{\delta}_*(X^n_*(s)) < 0$, where

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

Since $\bar{X}_*^n \equiv X_*^n/n \Rightarrow x_* \equiv (q_{1,a}, q_{2,a}, z_b)$, for z_b in (4.7.7) and $q_{i,a}$, $i = 1, 2$ in (4.7.6), we can define for every $s \geq 0$ the functions $\hat{\lambda}_j(x_*(s))$, $\hat{\mu}_j(x_*(s))$, $\hat{\lambda}_k(x_*(s))$ and $\hat{\mu}_k(x_*(s))$ to be the limits of $\hat{\lambda}_j(X_*^n(s))/n$, $\hat{\mu}_j(X_*^n(s))/n$, $\hat{\lambda}_k(X_*^n(s))/n$ and $\hat{\mu}_k(X_*^n(s))/n$, respectively, as $n \to \infty$.

By the linearity of $\tilde{\delta}_*$ and the continuity of the addition mapping when the limits are continuous, e.g. Theorem 12.7.1 in [78], we have that $\tilde{\delta}_*(X_*^n(s))/n \Rightarrow \tilde{\delta}_*(x_*(s))$, where

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

Note that, by our construction of X_*^n, $x(0) = x_*(0)$ (that is because $X_a^n(0) = X_b^n(0) = X^n(0)$ for all $n \geq 1$). It is easy to see that, if $r_{2,1} = r_{1,2}$ (recall also that $z_{1,2}(0) = 0$), then $\tilde{\delta}_*(x_*(0)) = -\delta_-(x(0))$ for $\delta_-(x(0))$ in (4.5.20). Since, by Assumption 3, $\delta_-(x(0)) > 0$, it holds that $\tilde{\delta}_*(x_*(0)) < 0$.

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

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

which implies that $\lambda_2 > \mu_{2,2}m_2$, since by Assumption 1, $q_1^a \equiv \lambda_1 - \mu_{1,1}m_1 > 0$. It follows from the latter inequality and the fact that $r_{2,1} < r_{1,2}$, that $\tilde{\delta}_*(x_*(0)) < 0$. To summarize, $\tilde{\delta}_*(x_*(0)) < 0$ in both cases considered in the statement of the lemma.

Since x_* and $\tilde{\delta}_*(x_*)$ are continuous functions, we can find $\tau > 0$ such that

$$\sup_{s\in[0,\tau]} \tilde{\delta}_*(x_*(s)) < 0.$$

Hence, there exists $\eta_1 > 0$ such that

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

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

We now construct a single bounding QBD process that bounds $\tilde{D}_*^n(X_*^n(s), \cdot)$ for all $s \in [0, \tau]$. For that purpose, let $X_m^n \equiv (Q_{1,m}^n, Q_{2,m}^n, Z_m^n)$, where

$$Q_{1,m}^n \equiv \|Q_{1,a}^n\|_\tau, \quad Q_{2,m}^n \equiv \inf_{0\leq t\leq\tau} Q_{2,a}^n(t) \quad \text{and} \quad Z_m^n \equiv \|Z_b^n\|_\tau.$$

Applying the continuous mapping theorem for the supremum function, e.g., Theorem 12.11.7 in [78], we have that $\bar{X}_m^n \equiv X_m^n/n \Rightarrow x_m \equiv (q_{1,m}, q_{2,m}, z_m)$, with $q_{1,m} \equiv \|q_{1,a}\|_\tau$, $q_{2,m} \equiv \inf_{0\leq t\leq\tau} q_{2,a}(t)$ and $z_m \equiv \|z_b\|_\tau$.

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

$$\begin{aligned} Q_{1,*}(t) &= q_{1,a}(0) + N_1^a(\lambda_1 t) - N_{1,1}^s(\mu_{1,1}m_1 t) - N_{1,2}^s(\mu_{1,2}z_m t) - N_1^u(\theta_1 q_{1,m} t), \\ Q_{2,*}(t) &= q_{2,a}(0) + N_2^a(\lambda_2 t) - N_{2,2}^s(\mu_{2,2}(m_2 - z_m)t) - N_2^u(\theta_2 q_{2,m} t). \end{aligned}$$

By our choice of x_m, the QBD D_* is positive recurrent. Observe that for every sequence of sample paths $\{X_m^n : n \in \mathbb{N}\}$, the scaling in $D_*^n(X_m^n, \cdot)$ is equivalent to scaling time by a factor of order $O(n)$ in D_*. That is, for every $T > 0$, and every sample path of X_m^n

contained in the sets B^n defined above $\{D^n_*(X^n_m, t) : 0 \le t \le T\} \stackrel{d}{=} \{D_*(a_n t) : 0 \le t \le T\}$, with $a_n/n \to 1$ as $n \to \infty$.

Let $M_*(t) \equiv \sup_{s\in[0,t]} D_*(s)$ denote the running maximum of the positive recurrent QBD D_*. It follows from Theorem 4.7.2 that there exists $c > 0$ such that

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

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

■

Proof of Theorem 4.7.3: By Lemma 4.7.2, we only need to consider the case $z_{1,2}(0) = 0$. By Lemmas 4.7.3 and 4.7.4, there exists $\tau > 0$ such that

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

Hence, the claim of the theorem will follow from Lemma 4.7.2 and Theorem 4.6.1 if we show that for some t_0 satisfying $0 < t_0 \le \delta \le \tau$ it holds that $z_{1,2}(t_0) > 0$, where $z_{1,2}$ is the (deterministic) fluid limit of $\bar{Z}^n_{1,2}$ as $n \to \infty$ (shown to exist in the proof of Theorem 4.6.1 on $[0, \delta]$). We will actually show a somewhat stronger result, namely, that for any $0 < \epsilon \le \delta$ there exists $t_0 < \epsilon$ such that $z_{1,2}(t_0) > 0$. We prove that by assuming the contradictory statement: for some $0 < \epsilon \le \delta$ and for all $t \in [0, \epsilon]$, $z_{1,2}(t) = 0$.

Since, by our contradictory assumption, $z_{1,2}(t) = 0$ over $[0, \epsilon]$, we have that $Z^n_{1,2} = o_P(n)$. Recall also that $Z^n_{2,1} = o_P(1)$ over $[0, \epsilon]$ (since $\epsilon \le \tau$, and τ is chosen according to Lemmas 4.7.3 and 4.7.4). Define the processes

$$L^n_1 \equiv Q^n_1 + Z^n_{1,1} + Z^n_{1,2} - m^n_1 \quad \text{and} \quad L^n_2 \equiv Q^n_2 + Z^n_{2,1} + Z^n_{2,2} - m^n_2, \tag{5.2.1}$$

representing the excess number in system for each class. Note that $(L_i^n)^+ = Q_i^n$, $i = 1, 2$. Then,

$$\begin{aligned} L_i^n(t) = L_i^n(0) + N_i^a(\lambda_i^n t) - N_{i,i}^s\left(\mu_{i,i}\int_0^t (L_i^n(s) \wedge 0)\, ds\right) \\ - N_i^u\left(\theta_i \int_0^t (L_i^n(s) \vee 0)\, ds\right) + o_P(n), \quad i = 1, 2 \end{aligned} \tag{5.2.2}$$

for $0 \le t \le \delta$ as $n \to \infty$, where N_i^a, $N_{i,i}^s$ and N_i^u are independent rate-1 Poisson processes. The $o_P(n)$ terms are replacing the (random-time changed) Poisson processes related to $Z_{1,2}^n$ and $Z_{2,1}^n$, which can be disregarded when we consider the fluid limits of (5.2.2).

Letting $\bar{L}_i^n \equiv L_i^n/n$, $i = 1, 2$, and applying the continuous mapping theorem for the integral representation function in (5.2.2), Theorem 4.1 in [57], (see also Theorem 7.1 and its proof in [57]), we have that $(\bar{L}_1^n, \bar{L}_2^n) \Rightarrow (\bar{L}_1, \bar{L}_2)$ as $n \to \infty$, where, for $i = 1, 2$,

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

so that

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

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

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

Recall that $\rho_1 > 1$, so that $\lambda_1 - \mu_{1,1}m_1 > 0$. Together with the initial condition,

$L_1(0) \geq 0$, we see that, in that case, $\bar{L}_1(t) \geq 0$ for all $t \geq 0$. First assume that $\rho_2 \geq 1$. Then, by similar arguments, $\bar{L}_2(t) \geq 0$ for all $t \geq 0$. In that case, we can replace $\bar{L}_i$ with q_i, $i = 1, 2$, and write

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

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

$$\begin{aligned} d_{1,2}(t) &= q_1^a + (q_1(0) - q_1^a)e^{-\theta_1 t} - r\left(q_2^a + (q_2(0) - q_2^a)e^{-\theta_2 t}\right) \\ &= (q_1^a - rq_2^a) + (q_1(0) - q_1^a)e^{-\theta_1 t} - r(q_2(0) - q_2^a)e^{-\theta_2 t}. \end{aligned} \tag{5.2.3}$$

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

It is easy to see that

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

Hence, $d'_{1,2}(0) = \lambda_1 - \mu_{1,1}m_1 - \theta_1 q_1(0) - r(\lambda_2 - \mu_{2,2}) + r\theta_2 q_2(0)$. If follows from (4.5.20) and the assumption $z_{1,2}(0) = 0$, that $d'_{1,2}(0) = \delta_-(x(0))$. By Assumption 3, $x(0) \in \mathbb{A}$, so that $d'_{1,2}(0) > 0$, and $d_{1,2}$ is strictly increasing at 0. Now, since $d_{1,2}(0) = \kappa$, we can find $t_1 \in (0, \epsilon]$, such that $d_{1,2}(t) > \kappa$ for all $0 < t < t_1$. This implies that $P(\inf_{0<t\leq t_1} D^n_{1,2}(t) > 0) \to 1$ as $n \to \infty$.

It follows from the representation of $Z^n_{1,2}$ in (4.4.2) that for any $t \in [0, t_1]$,

$$\bar{Z}^n_{1,2}(t) = \frac{N^s_{2,2}\left(\mu_{2,2}m_2^n t\right)}{n} + o_P(1). \tag{5.2.4}$$

The $o_P(1)$ term follows from our assumption that $\bar{Z}^n_{1,2}(t) \Rightarrow 0$ as $n \to \infty$. However,

by the FSLLN for Poisson processes, the fluid limit $z_{1,2}$ of $\bar{Z}^n$ in 5.2.4 satisfies $z_{1,2}(t) = \mu_{2,2}m_2 t > 0$ for every $0 < t \leq t_1$. We thus get a contradiction to our assumption that $z(t) = 0$ for all $t \in [0, \epsilon]$.

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

Proof of Theorem 4.7.4: We will start working with the processes L_1^n and L_2^n defined in (5.2.1) (but recall that, by Theorem 4.7.3 $Z_{2,1}^n \Rightarrow 0$, and in particular $\hat{Z}_{2,1}^n \Rightarrow 0$). For each $n \geq 1$, we will bound the two-dimensional process (L_1^n, L_2^n) below in sample-path stochastic order by another two-dimensional process $(L_{1,b}^n, L_{2,b}^n)$.

We construct the lower-bound process $(L_{1,b}^n, L_{2,b}^n)$ by increasing the departure rates in both processes L_1^n and L_2^n, making it so that each goes down at least as fast, regardless of the state of the other. First, we place reflecting upper barriers on the two queues. This is tantamount to making the death rate infinite in these states and all higher states. We place the reflecting upper barrier on L_1^n at κ^n, where $\kappa^n \geq 0$; we place the reflecting upper barrier on L_2^n at 0. With the upper barrier at κ^n, the departure rate of L_1^n is bounded above by $\mu_{1,1}m_1^n + \theta_1\kappa^n + \mu_{1,2}Z_{1,2}^n(t)$, based on assuming that pool 1 is fully busy serving class 1 (since $\mu_{2,1}Z_{2,1}^n(t) = o_p(1)$ we ignore it), that L_1^n is at its upper barrier, and that $Z_{1,2}^n(t)$ agents from pool 2 are currently busy serving class 1 in the original system. Second, with the upper barrier at 0, the departure rate of L_2^n is bounded above by $\mu_{2,2}m_2^n - \mu_{1,2}Z_{1,2}^n(t)$, based on assuming that pool 2 is fully busy with $Z_{1,2}^n(t)$ agents from pool 2 currently busy serving class 1, and that L_2^n is at its upper barrier 0. Thus, we give $L_{1,b}^n$ and $L_{2,b}^n$ these

bounding rates at all times.

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

$$U^n \equiv \mu_{2,2}(L^n_1 - \kappa^n) + \mu_{1,2}L^n_2 \quad \text{and} \quad U^n_b \equiv \mu_{2,2}(L^n_{1,b} - \kappa^n) + \mu_{1,2}L^n_{2,b}. \tag{5.2.5}$$

By the established sample-path stochastic order $(L^n_1, L^n_2) \geq_{st} (L^n_{1,b}, L^n_{2,b})$ and the monotonicity of the linear map in (5.2.5), we get the associated sample-path stochastic order $U^n \geq_{st} U^n_b$. Moreover, the stochastic process U^n_b is independent of the process $Z^n_{1,2}$, because of the particular linear combination we have chosen for the one-dimensional processes U^n and U^n_b in (5.2.5). We have chosen that linear combination so that the number of pool-2 agents working on class 1 does not matter.

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

$$\begin{aligned} \mu^n_b &\equiv \mu_{2,2}(\mu_{1,1}m^n_1 + \theta_1\kappa^n + \mu_{1,2}Z^n_{1,2}(t)) \\ &\quad + \mu_{1,2}(\mu_{2,2}m^n_2 - \mu_{2,2}Z^n_{1,2}(t))) \\ &= \mu_{2,2}(\mu_{1,1}m^n_1 + \theta_1\kappa^n) + \mu_{1,2}\mu_{2,2}m^n_2. \end{aligned} \tag{5.2.6}$$

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

$$\begin{aligned} \delta^n_b &\equiv \lambda^n_b - \mu^n_b = \mu_{2,2}(\lambda^n_1 - m^n_1\mu_{1,1} - \theta_1\kappa^n) \\ &\quad + \mu_{1,2}(\lambda^n_2 - m^n_2\mu_{2,2}). \end{aligned} \tag{5.2.7}$$

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

$$\delta_b \equiv \mu_{2,2}(\lambda_1 - m_1\mu_{1,1} - \theta_1\kappa) + \mu_{1,2}(\lambda_2 - m_2\mu_{2,2}) > 0, \tag{5.2.8}$$

with the inequality following from Assumption 1.

Now we observe that $-U_b^n$ is equivalent to the number in system in a stable $M/M/1$ queueing model with traffic intensity $\rho_*^n \to \rho_* < 1$. Let Q_* be the number-in-system process in an $M/M/1$ system having arrival rate equal to $\lambda_* \equiv \mu_{2,2}(m_1\mu_{1,1} + \theta_1\kappa) + \mu_{1,2}m_2\mu_{2,2}$, service rate $\mu_* \equiv \mu_{2,2}\lambda_1 + \mu_{1,2}\lambda_2$ and traffic intensity $\rho_* \equiv \lambda_*/\mu_* < 1$. Observe that the scaling in U_b^n is tantamount to accelerating time by a factor of order $O(n)$ in Q_*. That is, $\{-U_b^n(t) : t \geq 0\}$ can be represented as $\{Q_*(c_n t) : t \geq 0\}$, where $c_n/n \to 1$ as $n \to \infty$.

Let $M_*(t) \equiv \|Q_*\|_t$. We can now apply the extreme-value result in Theorem 4.7.2 for the $M/M/1$ queue above (since an $M/M/1$ is trivially a QBD) to conclude that $M_*(t) = O_P(\log(t))$. This implies that $U_b^n/\log(n)$ is SB.

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

5.3 The Bounding QBD in Lemma 4.7.4

In this section we add some more supporting detail to §4.7. In particular, we now describe how to present the process $\tilde{D}_*^n \equiv jQ_*^n - kQ_*^n$ in the proof of Lemma 4.7.4 as a QBD for each n. To that end, let $m \equiv j \vee k$. We divide the state space $\mathbb{N} \equiv \{0, 1, 2, \dots\}$ into level

of size m: Denoting level i by $L(i)$, we have

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

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

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

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

For example, if $j = 2$ and $k = 3$, then

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

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

where $\hat{\mu}_\Sigma^n \equiv \hat{\mu}_3^n + \hat{\mu}_2^n$ and $\sigma^n \equiv \hat{\mu}_\Sigma^n + \hat{\lambda}_2^n + \hat{\lambda}_3^n$.

Let $A^{(n)} \equiv A_0^{(n)} + A_1^{(n)} + A_2^{(n)}$. Then $A^{(n)}$ is an irreducible CTMC infinitesimal generator

matrix. It is easy to see that its unique stationary probability vector, $\nu^{(n)}$, is the uniform probability vector, attaching probability $1/m$ to each of the m states. Then by Theorem 7.2.3 in [52], the QBD is positive recurrent if and only if

$$\nu A_0^{(n)} \mathbf{1} < \nu A_2^{(n)} \mathbf{1},$$

where $\mathbf{1}$ is the vector of all 1's. This translates to the stability condition given in the proof of Lemma 4.7.4.

5.4 More on the Idleness Processes

In this section we present additional results about the idleness processes, going beyond Theorem 4.7.4. We treat pools 1 and 2 in the following subsections.

5.4.1 The Idleness Process in Pool 1

We now show how to analyze the idleness in pool 1 without paying attention to what happens in pool 2. This provides a more elementary derivation of the results for I_1^n in Theorem 4.7.4.

We start by showing that Q_1^n is never "too much" below κ^n if κ^n is large enough, where "large enough" in our setting is $\kappa^n/\log(n) \to \infty$ as $n \to \infty$. Since the thresholds in FQR-T are of order greater than $O(\sqrt{n})$, this includes the case in which the thresholds are kept throughout (i.e., they are not dropped once they are crossed, so that $\kappa^n = k_{1,2}^n$), and the case in which κ^n is the centering constant used in shifted FQR-T, where $\kappa^n/n \to \kappa > 0$.

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

$$\rho_* \equiv \frac{\mu_{1,1} m_1 + \theta_1 \kappa}{\lambda_1} < 1, \tag{5.4.1}$$

where the inequality follows from Assumption 1.

We define the difference-process

$$E_1^n \equiv \kappa^n - Q_1^n. \tag{5.4.2}$$

We will focus on the positive part: $(E_1^n)^+(t) \equiv \max\{E_1^n(t), 0\}$.

Lemma 5.4.1. *If $\kappa^n/\log(n) \to \infty$ as $n \to \infty$, then $(E_1^n)^+/\log(n)$ is SB.*

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

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

We can bound the process Q_1^n from below by Q_b^n in the sense of sample-path stochastic order; i.e., for each n, it is possible to construct stochastic processes $\tilde{Q}_b^n$ and $\tilde{Q}_1^n$ on a common probability space, with $\tilde{Q}_b^n$ having the same distribution as Q_b^n, $\tilde{Q}_1^n$ having the same distribution as Q_1^n, and every sample path of $\tilde{Q}_b^n$ lies below the corresponding sample path

of $\tilde{Q}_1^n$. The stochastic bound is constructed directly by generating the same arrival processes to both systems. We let departures from service coincide in both systems whenever $Z_b^n = Z_{1,1}^n$. Similarly, we let abandonments from Q_1^n coincide with abandonments from Q_b^n whenever both queues are equal. The argument follows the reasonings in Theorems 6 and 9 in [74].

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

$$E_b^n(t) \leq_{st} Q_*^n(t) = N_*^a\left((\mu_{1,1}m_1^n + \theta_1\kappa^n)t\right) - N_*^s\left(\lambda^n \int_0^t 1_{\{Q_*^n(s)>0\}}\, ds\right), \qquad (5.4.3)$$

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

Let Q_* be the number-in-system process in an $M/M/1$ system having arrival rate equal to $\mu_{1,1}m_1 + \theta_1\kappa$ and service rate λ_1, so that ρ_* in (5.4.1) is the traffic intensity to Q_*, and $\rho_* < 1$. Observe that the effect of increasing the size of the $M/M/m_1^n/\kappa^n + M$ system and its arrival rate (by increasing m_1^n, κ^n and λ_1^n) is tantamount to accelerating time by a factor of order $O(n)$ in Q_*. That is, $\{E_b^n(t) : t \geq 0\}$ is stochastically bounded from above (in sample-path stochastic order) by $\{Q_*(c_n t) : t \geq 0\}$, where $c_n/n \to 1$ as $n \to \infty$, for every $t \geq 0$. We can now apply extreme-value theory for the $M/M/1$ queue. In particular, if we let $M_*(t) \equiv \max\{Q_*(s) : 0 \leq s < t\}$, then $\|E_b^n\|_t$ is bounded from above, in the sample-path stochastic-order sense, by the process $M_*(c_n t)$.

Since the queue length is discrete, with a geometric stationary distribution, a standard extreme-value limit does not exist. Nevertheless, we can bound the $\limsup$ above; in

particular, it follows from Theorem 6 in [4] and the example following it, (see also Problem 4.2 pg. 185 of [7]), that, for $c = [(\mu_{1,1}m_1 - \theta_1\kappa)(1-\rho_*)]^{-1} > 0$,

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

where

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

and $b(t) \to -\log(1-\rho_*)/\log(\rho_*)$ as $t \to \infty$. The last inequality is the result in [4].

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

■

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

Corollary 5.4.1. *Under the conditions of Lemma* 5.4.1*, for any* $T > 0$*,*

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

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

Proposition 5.4.1. *If* $\kappa^n/\log(n) \to c$, *where* $0 \le c < \infty$, *then* $I_1^n/\log(n)$ *is SB.*

Proof: The proof is similar to the proof of Lemma 5.4.1. If we prove the result for any bounded sequence, then the result will follow trivially for any unbounded sequence. We thus assume that $0 \le \kappa^n \le M < \infty$. We use the same sample-path stochastic-order $M/M/1$-bound Q_*^n in (5.4.3) to bound I_1^n, only now we replace κ^n with M in the representation (5.4.3). Since M becomes negligible relative to the scaling by n as n increases, the traffic intensity for the process Q_*, defined in the proof of Lemma 5.4.1, is $\rho_* = \mu_{1,1}m_1/\lambda_1$, so that $\rho_* < 1$ by 1. Hence, the bound M_* in the proof of Lemma 5.4.1, applies to I_1^n. ∎

We can combine Corollary 5.4.1 and Proposition 5.4.1. To that end, we define the process

$$L_1^n \equiv Q_1^n + Z_{1,1}^n - m_1^n. \tag{5.4.4}$$

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

Corollary 5.4.2. *The sequence* $(\kappa^n - L_1^n)^+/\log(n)$ *is SB. Hence,* $I_1^n/\log(n)$ *is SB.*

5.4.2 The Idleness Process in Pool 2

We now turn to the pool-2 idleness process. We establish a stronger property away from the time origin.

Proposition 5.4.2. *For all* ϵ *and* T *satisfying* $0 < \epsilon < T < \infty$,

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

Proof: Much of the argument here repeats the proof of Theorem 4.7.4. For the first statement, we will create a stochastic lower bound and show that it satisfies the statement. We will exploit a linear combination of processes associated with the two queues. For that

purpose, we define the process

$$L_2^n \equiv Q_2^n + Z_{1,2}^n + Z_{2,2}^n - m_2^n, \tag{5.4.5}$$

representing the excess number in system for class 2. Then let U^n be the linear combination of the processes L_i^n, $i = 1, 2$, defined in (5.4.4) and (5.4.5):

$$U^n \equiv \mu_{2,2}(L_1^n - \kappa^n) + \mu_{1,2}L_2^n. \tag{5.4.6}$$

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

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

Just as in the proof of Lemma 5.4.1, we will bound the process U^n in (5.4.6) below in sample-path stochastic order by another process, U_b^n, a one-dimensional birth-and-death (BD) process. As a first step, we give U_b^n the same Poisson arrival processes as the original system has. Thus, U_b^n has constant birth rate $\lambda_b^n \equiv \mu_{2,2}\lambda_1^n + \mu_{1,2}\lambda_2^n$.

We next bound the pair of processes (L_1^n, L_2^n) below in sample-path stochastic order by another two-dimensional process $(L_{1,b}^n, L_{2,b}^n)$. We construct the lower-bound process$(L_{1,b}^n, L_{2,b}^n)$ by increasing the departure rates in both processes L_1^n and L_2^n, making it so that each goes down at least as fast, regardless of the state of the other. First, we place reflecting upper barriers on the two queues. This is tantamount to making the death rate infinite in these states and all higher states. We place the reflecting upper barrier on L_1^n at $\kappa^n + \epsilon_1 n$; we place the reflecting upper barrier on L_2^n at $\epsilon_1 n$. With the upper barrier at $\epsilon_1 n$, the departure rate of

L_1^n is bounded above by $\mu_{1,1}m_1^n + \theta_1\kappa^n + \theta_1\epsilon_1 n + \mu_{1,2}Z_{1,2}^n(t)$, based on assuming that pool 1 is fully busy serving class 1, that L_1^n is at its upper barrier, and that $Z_{1,2}^n(t)$ agents from pool 2 are currently busy serving class 1 in the original system. Second, with the upper barrier at $\epsilon_1 n$, the departure rate of L_2^n is bounded above by $\mu_{2,2}m_2^n + \theta_2\epsilon_1 n - \mu_{1,2}Z_{1,2}^n(t)$, based on assuming that pool 2 is fully busy with $Z_{1,2}^n(t)$ agents from pool 2 currently busy serving class 1, and that L_2^n is at its upper barrier $\epsilon_1 n$. Thus, we give $L_{1,b}^n$ and $L_{2,b}^n$ these bounding rates at all times

Of course, as constructed, the evolution of $(L_{1,b}^n, L_{2,b}^n)$ depends on the process $Z_{1,2}^n$ associated with the original system. However, we can avoid this difficulty by looking at the special linear combination in (5.2.5); i.e., we define the associated process

$$U_b^n \equiv \mu_{2,2}(L_{1,b}^n - \kappa^n) + \mu_{1,2}L_{2,b}^n. \tag{5.4.7}$$

By the sample-path stochastic order $(L_1^n, L_2^n) \geq_{st} (L_{1,b}^n, L_{2,b}^n)$, we get the associated sample-path stochastic order $U^n \geq_{st} U_b^n$. Moreover, the stochastic process U_b^n is independent of the process $Z_{1,2}^n$, because of the particular linear combination we have chosen for the one-dimensional processes U^n and U_b^n in (5.2.5) and (5.4.7). We have chosen that linear combination so that the number of pool-2 agents working on class 1 does not matter.

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

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

$$\begin{aligned}\mu_b^n &\equiv \mu_{2,2}(\mu_{1,1}m_1^n + \theta_1\kappa^n + \theta_1\epsilon_1 n + \mu_{1,2}Z_{1,2}^n(t)) \\ &\quad +\mu_{1,2}(\mu_{2,2}m_2^n + \theta_2\epsilon_1 n - \mu_{2,2}Z_{1,2}^n(t))) \\ &= \mu_{2,2}(\mu_{1,1}m_1^n + \theta_1\kappa^n) + \mu_{1,2}\mu_{2,2}m_2^n + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1 n. \qquad (5.4.8)\end{aligned}$$

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

$$\begin{aligned}\delta_b^n &\equiv \lambda_b^n - \mu_b^n = \mu_{2,2}(\lambda_1^n - m_1^n\mu_{1,1} - \theta_1\kappa^n) \\ &\quad +\mu_{1,2}(\lambda_2^n - m_2^n\mu_{2,2}) + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1 n. \qquad (5.4.9)\end{aligned}$$

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

$$\delta_b \equiv \mu_{2,2}(\lambda_1 - m_1\mu_{1,1} - \theta_1\kappa) + \mu_{1,2}(\lambda_2 - m_2\mu_{2,2}) + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1. \qquad (5.4.10)$$

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

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

$$\bar{U}_b^n \Rightarrow u_b \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \qquad (5.4.11)$$

where

$$\bar{U}_b^n(t) \equiv U_b^n(t)/n \quad \text{and} \quad u_b(t) \equiv \delta_b t \wedge \epsilon_1 \quad \text{for} \quad t \geq 0, \qquad (5.4.12)$$

with $u_b(0) = 0$.

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

$$P(\inf_{\epsilon \le t \le T} U^n(t) > 0) \to 1 \quad \text{as} \quad n \to \infty. \tag{5.4.13}$$

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

5.5 Remaining Proofs in Section 4.8

5.5.1 Remaining Proofs in §4.8.1

Proof of Lemma 4.8.1: For background on tightness, see [13, 57, 78]. We recall a few key facts: Tightness of a sequence of k-dimensional stochastic processes in $\mathcal{D}_k$ is equivalent to tightness of all the one-dimensional component stochastic processes in $\mathcal{D}$. For a sequence of random elements of $\mathcal{D}_k$, C-tightness implies $\mathcal{D}$-tightness and that the limits of all convergent subsequences must be in $\mathcal{C}_k$; see Theorem 15.5 of the first 1968 edition of [13]. Thus it suffices to verify conditions (6.3) and (6.4) of Theorem 11.6.3 of [78]. Hence, it suffices to prove SB of the sequence of stochastic processes evaluated at time 0 and appropriately control the oscillations, using the modulus of continuity on $\mathcal{C}$. We obtain the stochastic boundedness at time 0 immediately from Assumption 3 in §4.3. We show that we can control the oscillations below. The resulting tightness implies that the sequence of stochastic processes is SB.

We now show how to control the oscillations. For that purpose, let $w(x, \zeta, T)$ is the

modulus of continuity of the function $x \in \mathcal{D}$, i.e.,

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

Using the representations (4.4.1)-(4.4.4), for $t_2 > t_1 \ge 0$ we have

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

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

$$\begin{aligned}w(Q_1^n/n,\zeta,T) \;\le\; & w(A_1^n/n,\zeta,T) + w(S^n/n,\zeta,T) + w(S_{1,1}^n/n,\zeta,T) \\ & + w(U_1^n/n,\zeta,T). \end{aligned} \tag{5.5.1}$$

Then observe that we can bound the oscillations of the service processes $S_{i,j}^n$ by the oscillations in the scaled Poisson process $N_{i,j}^s(n\cdot)$. In particular, by (4.4.1),

$$w(S_{i,j}^n/n,\zeta,T) \le w(N_{i,j}^s(n\mu_{i,j}m_j\cdot)/n,\zeta,T) \le w(N_{i,j}^s(n\cdot)/n, c\zeta, T) \tag{5.5.2}$$

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

$$\begin{aligned} Q_i^n(t) &\le Q_i^n(0) + A_i^n(t), \\ |U_i^n(t_2) - U_i^n(t_1)| &= |N_i(\theta_i \int_{t_1}^{t_2} Q_i^n(s)\, ds| \\ &\le |N_i(n\theta(\bar{Q}_i^n(0) + \bar{A}_i^n(T)(t_2-t_1))|. \end{aligned} \tag{5.5.3}$$

Let $q_{bd} \equiv 2(q_i(0)+T)$, where $\bar{Q}_i^n(0) \Rightarrow q_i(0)$ by Assumption 3, and let B_n be the following

subset of the underlying probability space:

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

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

$$w(U_i^n/n, \zeta, T) \leq w(N_i^u(nq_{bd}\cdot)/n, \zeta, T) \leq w(N_i^u(n\cdot)/n, c\zeta, T) \tag{5.5.4}$$

for some constant $c > 0$.

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

$$\begin{aligned} w(Q_i^n/n, \zeta, T) \quad \leq \quad & w(N_i^a(n\cdot)/n, c\zeta, T) + 2\sum_{i=1}^{2}\sum_{j=1}^{2} w(N_{i,j}^s(n\cdot)/n, c\zeta, T) \\ & + w(N_i^u(n\cdot)/n, c\zeta, T). \end{aligned} \tag{5.5.5}$$

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

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

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

We now turn to the sequence $\{\bar{Z}_{1,2}^n\}$. Let $A_{1,2}^n(t)$ denote the total number of class-1 arrivals up to time t, who will eventually be served by type-2 servers in system n. Let $\bar{A}_{1,2}^n \equiv A_{1,2}^n/n$ and $\bar{S}_{1,2}^n(t) \equiv S_{1,2}^n(t)/n$, for $S_{1,2}^n(t)$ in (4.4.1). Since

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

we have

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

However, for A^n_1 in (4.4.1),

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

Since $\bar{A}^n_1 \Rightarrow \lambda_1 e$ in $\mathcal{D}$, the sequence $\{\bar{A}^n_1\}$ is tight. Together with (5.5.2), that implies that the sequence $\{\bar{Z}^n_{1,2}\}$ is tight as well. Finally, we observe that the tightness of $\{\bar{Y}^n_8\}$ follows from (5.5.2), (5.5.4) and the convergence of $\bar{A}^n_i$. ∎

Proof of Lemma 4.8.2: Apply the bounds on the modulus of continuity involving Poisson processes in the proof of Lemma 4.8.1 above. For a Poisson process N, let $\hat{N}^n \equiv \sqrt{n}(\bar{N}^n - e)$, where $\bar{N}^n(t) \equiv N(nt)/n$, $t \geq 0$. By the triangle inequality, for each n, ζ, and T,

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

Since, $w(x, \zeta, T)$ is a continuous function of x for each fixed ζ and T, we can apply this bound with the inequalities in the proof of Lemma 4.8.1 to deduce (4.8.1). ∎

5.5.2 Remaining Proof in §4.8.3

Proof of Lemma 4.8.7: Consider the drift rates of the QBD-version of D^n_f in (4.5.6), and observe that, by the linearity of the drift expressions and Assumption 3, $\delta^n_+(X^n(0))/n \Rightarrow \delta_+(x(0))$ and $\delta^n_-(X^n(0))/n \Rightarrow \delta_-(x(0))$ for δ_+ and δ_- in (4.5.20). Also by Assumption 3,

$x(0) \in \mathbb{A}$ so that (4.5.21) holds. This implies that there exists $\eta > 0$ such that

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

i.e., (4.8.11) holds at $t = 0$ with probability converging to 1 as $n \to \infty$.

To prove the lemma, we bound the drifts in (4.5.6). We do that by bounding the change in the components of $X^n(t)$ in a short interval after time 0. To do that, we use the stochastic-order bounds in (4.7.4)-(4.7.5). Recall the rather special ordering obtained there:

$$(-Q_{1,a}^n, Q_{2,a}^n, Z_a^n) \leq_{st} (-Q_1^n, Q_2^n, Z_{1,2}^n) \leq_{st} (-Q_{1,b}^n, Q_{2,b}^n, Z_b^n). \tag{5.5.6}$$

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

$$\delta_+^n(X^n(t)) \leq_{st} \delta_+^n(X_+^n(t)), \qquad \delta_-^n(X^n(t)) \geq_{st} \delta_-^n(X_-^n(t)) \tag{5.5.7}$$

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

$$\lim_{n\to\infty} P\left(\sup_{t\in[0,\xi]} \delta_+^n(X_+^n(t)) < -\eta \quad \text{and} \quad \inf_{t\in[0,\xi]} \delta_-^n(X_-^n(0)) > \eta\right) = 1. \tag{5.5.8}$$

To construct the processes X_+^n and X_-^n with these properties, we use the bounding processes X_a^n and X_b^n in (4.7.4) and (4.7.5) (appearing again in (5.5.6). Specifically, we let

$$X_+^n \equiv (Q_{1,a}^n, Q_{2,b}^n, Z_+^n) \quad \text{and} \quad X_-^n \equiv (Q_{1,b}^n, Q_{2,a}^n, Z_-^n), \tag{5.5.9}$$

respectively, where $Z_+^n = Z_b^n$ if $\mu_{2,2} \geq \mu_{1,2}$, and $Z_+^n = Z_a^n$ otherwise. $Z_-^n = Z_a^n$ if

$\mu_{2,2} \geq \mu_{1,2}$, and $Z^n_- = Z^n_b$ otherwise. As a consequence, for each $t \geq 0$, the drifts satisfy

$$\begin{aligned}\delta^n_+(X^n_+(t)) &\equiv j[\lambda^n_1 - \mu_{1,1}m^n_1 - (\mu_{1,2} - \mu_{2,2})Z^n_+(t) - \mu_{2,2}m^n_2 - \theta_1 Q^n_{1,b}(t)] \\ &\quad - k[\lambda^n_2 - \theta_2 Q^n_{2,a}(t)], \\ \delta^n_-(X^n_-(t)) &\equiv j[\lambda^n_1 - \mu_{1,1}m^n_1 - \theta_1 Q^n_{1,a}(t)] \\ &\quad - k[\lambda^n_2 - (\mu_{1,2} - \mu_{2,2})Z^n_-(t) - \mu_{2,2}m^n_2 - \theta_2 Q^n_{2,b}(t)].\end{aligned} \tag{5.5.10}$$

We have directly defined the processes in (5.5.9) to ensure that the inequalities in (5.5.7) are satisfied.

Assume that $X^n_+(0) = X^n_-(0) = X^n(0)$. By Assumption 3, $\bar{X}^n(0) \Rightarrow x(0)$ as $n \to \infty$, so that the condition in Lemma 4.7.1 holds at $t = 0$. Hence, by Lemma 4.7.1, $\bar{X}^n_+ \Rightarrow x_+ \equiv (q_{1,b}, q_{2,a}, z_+)$, where $z_+ = z_a$ if $\mu_{2,2} \geq \mu_{1,2}$ and $z_+ = z_b$ otherwise. Also, $\bar{X}^n_- \Rightarrow x_- \equiv (q_{1,a}, q_{2,b}, z_-)$, where $z_- = z_b$ if $\mu_{2,2} \geq \mu_{1,2}$ and $z_- = z_a$ otherwise. Hence, by the linearity of the functions δ^n_+ and δ^n_-,

$$\delta^n_+(X^n_+)/n \Rightarrow \delta_+(x_+) \quad \text{and} \quad \delta^n_-(X^n_-)/n \Rightarrow \delta_-(x_-) \quad \text{in } \mathcal{D} \text{ as } n \to \infty. \tag{5.5.11}$$

Since $x_+(0) = x_-(0) = x(0) \in \mathbb{A}$, and by the continuity of $\delta_+(\cdot)$ and $\delta_-(\cdot)$, we can find $\xi > 0$ and $\eta > 0$, such that $\delta_+(x_+(t)) < -\eta$ and $\delta_-(x_-(t)) > \eta$. for all $t \in [0, \xi]$. That implies that we have (5.5.8). Together with (5.5.7), that concludes the proof. ∎

5.5.3 Remaining Proof in §4.8.5

Proof of Lemma 4.8.9: We can apply essentially the same reasoning as in the proof of Lemma 4.8.7. We only need to change the order. Now we aim to achieve:

$$\begin{aligned} &\delta_+^n(X_m^n(t)) \le \delta_+^n(X^n(t)) \le \delta_+^n(X_M^n(t)), \quad \text{and} \\ &\delta_-^n(X_m^n(t)) \le \delta_-^n(X^n(t)) \le \delta_-^n(X_M^n(t)) \end{aligned} \tag{5.5.12}$$

instead of (5.5.7). Moreover, we will do so such that the two bounding QBD's are positive recurrent over some interval $[0, \xi]$ on the sets B_n where $P(B_n) \to 1$ as $n \to \infty$. In other words, we will use random vectors X_M^n and X_m^n instead of full processes.

We again use the stochastic-order bounds in (4.7.4)-(4.7.5), with the ordering in (5.5.6). To construct X_M^n, let

$$X_{M^+}^n \equiv (Q_{1,M}^n, Q_{2,M}^n, Z_{M^+}^n) \quad \text{and} \quad X_{M^-}^n \equiv (Q_{1,M}^n, Q_{2,M}^n, Z_{M^-}^n), \tag{5.5.13}$$

where

$$\begin{aligned} Q_{1,M}^n &\equiv \inf_{0\le t\le \xi} Q_{1,b}^n(t) \vee 0, \quad Q_{2,M}^n \equiv \|Q_{2,b}^n\|_\xi, \\ Z_{M^+}^n &\equiv \inf_{0\le t\le \xi} Z_+^n(t), \qquad\quad Z_{M^-}^n \equiv \|Z_-^n\|_\xi, \end{aligned} \tag{5.5.14}$$

with $Z_+^n(t) \equiv Z_b^n$ and $Z_-^n(t) \equiv Z_a^n$ if $\mu_{2,2} \ge \mu_{1,2}$, and $Z_+^n(t) \equiv Z_a^n$ and $Z_-^n(t) \equiv Z_b^n$ otherwise. Note that we can regard $\{X_{M^+}^n(\xi) : \xi \ge 0\}$ as a stochastic process as a function of ξ, but we work with the final value $X_{M^+}^n \equiv X_{M^+}^n(\xi)$, and similarly for $X_{M^-}^n$. Let $\{D_f^n(X_M^n, s) : s \ge 0\}$ have the rates determined by $X_{M^-}^n$ when $D_f^n(X_M^n, s) \le 0$, and the rates determined by $X_{M^+}^n$ when $D_f^n(X_M^n, s) > 0$.

We do a similar construction for X_m^n. Let

$$X_{m^+}^n \equiv (Q_{1,m}^n, Q_{2,m}^n, Z_{m^+}^n) \quad \text{and} \quad X_{m^-}^n \equiv (Q_{1,m}^n, Q_{2,m}^n, Z_{m^-}^n), \tag{5.5.15}$$

where

$$\begin{aligned} Q^n_{1,m} &\equiv \|Q^n_{1,a}\|_\xi, \quad Q^n_{2,m} \equiv \inf_{0\le t\le \xi} Q^n_{2,b}(t) \vee 0, \\ Z^n_{m^+} &\equiv \|Z^n_+\|_\xi, \quad\quad Z^n_{m^-} \equiv \inf_{0\le t\le \xi} Z^n_-(t). \end{aligned} \tag{5.5.16}$$

with $Z^n_+(t) \equiv Z^n_a$ and $Z^n_-(t) \equiv Z^n_b$ if $\mu_{2,2} \ge \mu_{1,2}$, and $Z^n_+(t) \equiv Z^n_b$ and $Z^n_-(t) \equiv Z^n_a$ otherwise (the reverse of what is done in (5.5.14)). Let $\{D^n_f(X^n_m, s) : s \ge 0\}$ have the rates from $X^n_{m^-}$ when $D^n_f(X^n_m, s) \le 0$, and the rates from $X^n_{m^+}$ when $D^n_f(X^n_m, s) > 0$. By this construction, we achieve the ordering in (4.8.18). We cover the rates of $D^n_{1,2}(t)$ too because we can make the identification: the rates of $D^n_{1,2}(t)$ given $X^n(t)$ coincide with the rates of $D^n_f(X^n(t), \cdot)$.

It remains to find a ξ such that both the processes $\{D^n_f(X^n_m, s) : s \ge 0\}$ and $\{D^n_f(X^n_M, s) : s \ge 0\}$ are positive recurrent. To do so, we will use a minor modification of the reasoning in the final step of the proof of Lemma 4.8.7. We use Lemma 4.7.1, which concludes that the bounding processes as functions of ξ have fluid limits. By Lemma 4.7.1, we can conclude that $\bar{X}^n_{m^+} \equiv n^{-1}X^n_{m^+} \Rightarrow x^+_m$, $\bar{X}^n_{m^-} \equiv n^{-1}X^n_{m^-} \Rightarrow x^-_m$, $\bar{X}^n_{M^+} \equiv n^{-1}X^n_{M^+} \Rightarrow x^+_M$ and $\bar{X}^n_{M^-} \equiv n^{-1}X^n_{M^-} \Rightarrow x^-_M$ in $\mathcal{D}$, where x_{m^+}, x^-_m, x^+_M and x^-_M are all continuous with $x^+_m(0) = x_m{-}(0) = x^+_M(0) = x^-_M = x(0) \in \mathbb{A}$. Hence, we can find ξ' such that $x_m(\xi) \in \mathbb{A}$ and $x_M(\xi) \in \mathbb{A}$ for all $\xi \in [0, \xi']$. Hence, we can choose ξ such that the constant vectors $x_m \equiv x_m(\xi)$ and $x_M \equiv x_M(\xi)$ both arbitrarily close to $x(0)$.

Finally, we use the linearity of the drift function to deduce the positive recurrence of the processes depending upon n. As in (5.5.11), we have

$$\begin{aligned} \delta^n_-(X^n_{m^-})/n &\Rightarrow \delta_-(x^-_m), \quad \delta^n_+(X^n_{m^+})/n \Rightarrow \delta_+(x^+_m), \\ \delta^n_-(X^n_{M^-})/n &\Rightarrow \delta_-(x^-_M), \quad \text{and} \quad \delta^n_+(X^n_{M^+})/n \Rightarrow \delta_+(x^+_M). \end{aligned} \tag{5.5.17}$$

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

5.5.4 Remaining Proof in §4.8.6

Proof of Lemma 4.8.10: We start with the processes $D_f^n(X_m^n, \cdot)$ and $D_f^n(X_M^n, \cdot)$ already constructed in §§4.8.5 and 5.5.3, with the understanding that the interval length ξ will in general need to be redefined, now depending on ϵ. Since the initial state has been frozen in $D_f^n(X_m^n, \cdot)$, $D_f^n(X_M^n, \cdot)$ and $D_f^n(X^n(t), \cdot)$, these three processes are stationary CTMC's (have stationary transition rates), but $D_{1,2}^n(t)$ is a nonstationary CTMC. In the following we construct modified versions of these processes, but so as not to alter their individual distributions.. For the following, we regard all the processes as CTMC's and use the natural order on the integer state space (instead of the special order in the QBD structure).

As in the proof of Theorem 4.5.3, we can apply uniformization. As explained there, without loss of generality, we can regard the transition rates in $D_{1,2}^n$ as being uniformly bounded. Thus, for for all n suiitably large, and for each process under consideration, we can generate all potential transitions from constant-rate Poisson processes. Because of the scaling by $O(n)$ in (4.2.2), the Poisson processes for model n can be given rate αn, $n \geq 1$, for some positive constant α. The constant α is chosen so that the rate αn exceeds the maximum total transition rate out of any state for any of the processes for each $n \geq 1$. Then the actual transitions of the process are governed by a DTMC. The Poisson process generates potential transitions. When there is not a real transition, that is captured in the DTMC by a transition from that state back to itself. By choosing the Poisson transition rate sufficiently large, for every state in the state space, there is positive probability of a one-step transition immediately back to that same state. Hence, the DTMC is aperiodic as well as ireducible and positive recurrent. Note that the Poisson process captures the scaling by n.

For the new construction, we use a regenerative approach, using the regenerative structure discussed in §4.8.4. Provided that the QBD's $D_f^n(X_M^n, \cdot)$ and $D_f^n(X_m^n, \cdot)$ are positive

recurrent, which will hold on $B_n(\xi, \eta)$ by virtue of the construction in §4.8.5, successive visits to any fixed state constitute regenerative cycles for these stationary CTMC's with constant transition rates. It is convenient to let the regenerative state, denoted by s^*, be contained in the boundary of the QBD.

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

$$D_f^n(X_m^n, 0) = D_f^n(X_M^n, 0) = D_f^n(X^n(t), 0) = D_{1,2}^n(0) = s^*, \tag{5.5.18}$$

but we will later show that this initial condition is not needed; e.g., it can be replaced by the SB condition imposed in Assumption 3. We then focus on successive visits to that fixed state for the upper bound process.

For the new construction, we couple all four processes; i.e., we start by constructing *all* the processes together, starting in their common initial state, based on the rate order established in (4.8.18). That means that we use a single Poisson process with rate αn to generate potential transitions for all the processes under consideration. We match the actual transitions as much as possible in order to keep the processes evolving together as much as possible. We will choose ξ to ensure that the transition probabilities differ by only a negligible amount, so the processes will only rarely have different transitions during a single regenerative cycle. Even though we cannot achieve full sample path stochastic order for the stochastic processes over the full time interval, we can keep all the processes together over each regenerative cycle, with high probability. (Recall that the number of transitions in each regenerative cycle is of order $O(1)$, but the transitions are occurring at rate $O(n)$, so we are *not* succeeding in keeping the process paths identical over positive time intervals, but that is not needed. Because we are concerned with the integrals in (4.8.23), it suffices to have the *proportion* of time that the paths are identical be large. Also recall that the inequalities in (4.8.23) need not hold w.p.1; we are only claiming that the probability

that they hold should converge to 1 as $n \to \infty$.)

Our general idea is to construct an alternating renewal process for each n, which involves a sequence $\{(U^n_{1,k}, U^n_{2,k}) : k \geq 1\}$ of i.i.d pairs of nonnegative random variables, $U^n_{1,k}$ and $U^n_{2,k}$. These variables measure times in the full process and so will be $O(1/n)$. The first random variable $U^n_{1,k}$ is the geometric random sum of the cycle lengths of all the regenerative cycles where the processes all coincide, while the second interval $U^n_{2,k}$ is a subsequent interval on which the processes do not necessarily coincide. The second interval ends when all processes are in the regenerative state together. We then repeat the construction. We will make the first interval $U^n_{1,k}$ much longer than the second interval $U^n_{2,k}$, ensuring that the proportion of time that the processes all agree is arbitrarily close to 1 (falling within the ϵ gaps in (4.8.23)). The cycles will have $O(1)$ transitions, but since the transitions occur according to the Poisson process at rate αn, the cycle lengths are asymptotically negligible, making the limiting proportions all that matters.

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

We do a careful analysis to ensure that the second random variable $U^n_{2,k}$ is appropriately

controlled, independent of ξ, and then we choose ξ suitably small to make the first interval relatively long, so that the long-run proportion of time that the process is in the second interval, which is

$$\frac{E[U_{2,k}^n]}{E[U_{1,k}^n] + E[U_{2,k}^n]}, \tag{5.5.19}$$

is as small as desired. In fact, our construction will make $E[U_{1,k}^n] \uparrow \infty$ as $\xi \uparrow \infty$, while $E[U_{2,k}^n] \downarrow 0$ as $\xi \uparrow \infty$. Since the Poisson rate αn produces a time scaling of order $O(n)$, the cycles are occurring more rapidly as $n \to \infty$. In that way we can achieve the inequalities in (4.8.23) with probabiity converging to 1 as $n \to \infty$. Since we are working with indicator functions in (4.8.23), in computing the bound we allow the worst case, in which the indicator functions differ by 1 throughout the second interval.

We now present the details. Let the random number of transitions in a regenerative cycle for the upper bound process $D_f^n(X_M^n, \cdot)$ be N^n. Since the events are occurring at rate of order $O(n)$, we can use a version of the time-expanded queue-difference process for $D_f^n(X_M^n, \cdot)$, as in (4.5.7). By Theorem 4.5.3, we have $N^n \Rightarrow N$ as $n \to \infty$, where N is the corresponding random number of transitions during a regenerative cycle for the FTSP $D(x_M, \cdot)$, using the same designated regenerative state, where $\bar{X}_M^n \Rightarrow x_M$ as $n \to \infty$, as in §4.8.5. Moreover, because of the special QBD structure we also have additional regularity properties.

Let p_n be the probability mass function of N^n, i.e., $p_n(k) \equiv P(N^n = k)$. As in §4.8.4, From the convergence $N^n \Rightarrow N$ and the QBD structure of all processes, we know that p_n has a proper generating function (gf) $\psi_{N^n}(z) \equiv E\left[z^{N^n}\right]$. Combining the QBD and gf structure, we can conclude that there is an integer k_0 such that we can bound the probabilities $p_n(k)$ above and below by

$$C_L \tilde{q}^k \le p_n(k) \le C_U q^k \quad \text{for all} \quad k \ge k_0, \tag{5.5.20}$$

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

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

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

With that modified construction in place, let $W_i^n = 1$ if the lower bound process $D_f^n(X_m^n, \cdot)$ makes a different transition from the upper bound process $D_f^n(X_M^n, \cdot)$ at the i^{th} transition of the Poisson process, given that has not happened so far. Given our revised construction above, we can assume that the sequence $\{W_i^n : i \geq 1\}$ is a sequence of i.i.d

random variables with $P(W_i^n = 1) = \phi_n$, where $\phi_n \to \phi$ as $n \to \infty$ and $\phi \downarrow 0$ as $\xi \downarrow 0$. To see why, recall that, by Lemma 4.7.1, $\bar{X}_M^n \Rightarrow x_M$ and $\bar{X}_m^n \Rightarrow x_m$ in $\mathcal{D}_6$ as $n \to \infty$, where $x_M(0) = x_m(0) = (x(0), x(0))$. Hence, by taking ξ small enough and n large enough, we can make $\bar{X}_M^n$ and $\bar{X}_m^n$ arbitrarily close for all $t \in [0, \xi]$. Consequently, the probability that any of the processes differ at step $k \geq 1$ during a regenerative cycle, depends on the number of transitions during a regenerative cycle being at least k. Hence,

$$\begin{aligned} P(A_n) \equiv P(\text{any processes differ}) &= \sum_{k=1}^{\infty} \phi_n (1-\phi_n)^{k-1} \sum_{j=k}^{\infty} p_n(j) \\ &\leq \sum_{k=1}^{k_0} \phi_n (1-\phi_n)^{k-1} + \sum_{k=k_0+1}^{\infty} \phi_n (1-\phi_n)^{k-1} \sum_{j=k}^{\infty} C_U q^j \\ &= \phi_n \left(\sum_{k=1}^{k_0} (1-\phi_n)^{k-1} + \frac{C_U q}{1-q} \sum_{k=k_0+1}^{\infty} [(1-\phi_n)q]^{k-1} \right) \\ &\leq C_1 \phi \end{aligned} \tag{5.5.21}$$

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

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

$$E[V_{1,k}^n] = \frac{E[N^n]}{P(A_n)} \geq \frac{C_2 E[N]}{\phi} \quad \text{for all suitably large} \quad n, \tag{5.5.22}$$

where $C_2 < 1/C_1$ for C_1 in (5.5.21). We obtain the lower bound in (5.5.22) by applying

the convergence of the means $E[N^n] \to E[N]$ as $n \to \infty$, indicated above. Thus,

$$E[U^n_{1,k}] \geq \frac{C_2 E[N]}{\phi n \alpha} \quad \text{for all suitably large} \quad n, \tag{5.5.23}$$

as well. The main point is that we can make these means in (5.5.22) and (5.5.23) large in the relevant scale by making ϕ suitably small, which we can achieve by the proper choice of ξ.

We now want to show that $V^n_{2,k}$, the number of transitions of the Poisson process with rate $n\alpha$ in the second interval $U^n_{2,k}$, can be suitably controlled. To go with (5.5.22), it suffices to show that $V^n_{2,k}$ is SB as $n \to \infty$. Equivalently, it suffices to show that $nU^n_{2,k}$ is SB as $n \to \infty$. We will consider the two parts of this second interval in turn.

First consider the exceptional cycle. Let N^n_e be the random number of transitions in an exceptional regenerative cycle for the upper bound process. First, N^n_e is not distributed the same as N^n, because longer cycles are more likely to become exceptional cycles than shorter ones, because they generate more opportunities for a difference. Nevertheless, we can bound $E[N^n_e]$ above. To do so, we need to bound $P(A_n)$ below, instead of above as in (5.5.21). We can do so by using the lower bound for the probabilities $p_n(k) \equiv P(N^n = k)$ in (5.5.20).

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

$$E[N^n_e] = E[N^n | A_n] = \frac{E[N^n; A_n]}{P(A_n)}. \tag{5.5.24}$$

We start with the numerator of (5.5.24):

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

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

Note that, by Abel's Lemma (Lemma 5.1 pg. 64 in [41]), $\psi_{N^n}(z_n)$ and, consequently, $\frac{d}{dz_n}\psi_{N^n}(z_n)$ are continuous from the left at $z_n = 1$. Also, $z_n \to 1$ (from the left) as $\phi_n \to 0$. Hence, the numerator of (5.5.24) converges to 0 as $\phi_n \to 0$. We next show that the rate of convergence to 0 is the same as that of the denominator of (5.5.24), so that (5.5.24) is bounded from above by a constant. By (5.5.21) and Fubini's theorem,

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

Applying L'Hôpital's rule and Abel's lemma, we see that the limit of (5.5.24) as $\phi_n \to 0$ (by taking n to infinity and then ξ to zero) is bounded from above by a constant. Specifically,

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

for some constant C_3. (Recall that $E[N^n] \to E[N]$ and $E[(N^n)^2] \to E[N^2]$ as $n \to \infty$ by (5.5.20).)

For the next step, we will also want to bound the tail probabilities of N_e^n. By a minor

variation of the argument in (5.5.24), we can show they are bounded by a random variable with a geometric tail. If $k_1 \geq k_0$, then

$$\begin{aligned} P(N_e^n \geq k_1) &= \frac{P(N^n \geq k_1; A_n]}{P(A_n)} = \frac{\sum_{k=k_1}^{\infty} \sum_{j=1}^{k} \phi_n (1-\phi_n)^{j-1} p_n(k)}{\sum_{k=1}^{\infty} \sum_{j=1}^{k} \phi_n (1-\phi_n)^{j-1} p_n(k)} \\ &\leq \frac{\sum_{k=k_1}^{\infty} [1-(1-\phi_n)^k] C_U q^k}{\sum_{k=k_0}^{\infty} [1-(1-\phi_n)^k] C_L \tilde{q}^k} \leq C_4 [(1-\phi)q]^{k_1} \end{aligned} \tag{5.5.25}$$

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

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

Now observe that the intervals between successive visits of all four processes to this

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

By (i) the convergence of $\bar{X}_M^n \Rightarrow x_M$, (ii) the convergence of the transition rates of $\{D_f^n(X_M^n, s) : s \geq 0\}$ defined in (4.5.2)-(4.5.5) to the transition rates of the FTSP $\{D(x_M, s) : s \geq 0\}$ defined in (4.5.9)-(4.5.12) as $n \to \infty$, which is justified by (4.8.9) and the following discussion, and (iii) Lemma 4.8.8, we deduce that $\pi^n(s^*) \to \pi(s^*)$ as $n \to \infty$, where $\pi(s^*)$ is the steady-state probability of the FTSP, i.e., $\pi(s^*) = P(D(x_M, \infty) = s^*)$. Hence, for this special initial condition, we have established the bound $E[N_c^n] \leq C_7/\pi(s^*)^4 < \infty$ for $C_7 > 1$ for all n suitably large (depending on our choice of C_7).

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

together as possible, for each of the three pairs of processes. We have already described how to analyze the probability of a difference occurring over successive transitions, which can be (and will be) made negligible.

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

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

We can bound this expected number of transitions until the four processes reach the regeneration state together if we can bound the first hitting time of s^*. That is so, because we can bound the expected number of transitions for all four independent processes to reach the regeneration state together, if at transition k all four processes have visited state

s^* at least once in the last k transitions. That makes the other three discrete-time processes distributed as index shifted versions of the upper-bound DTMC.

e now want to bound the first passage time to s^* for each of the processes not starting in s^*. he first passage time can be controlled provided the initial condition can be controlled. e thus control the separation between the processes that can occur during the rest of the exceptional cycle, after the first non-identical transition. After the first non-identical transition, we focus on the upper bound process. We say that the exceptional cycle ends when the upper bound process next hits the regenerative state. However, because of the non-identical transitions, the other processes typically will not hit the regenerative state at that same transition epoch. It is evident that, as long as the processes stay together on the same side of 0, the probability of a second different transition during the cycle will be negligible. However, we lose control when the processes are on different sides of state 0. Fortunately, it suffices to use a crude bound on the maximum possible separation of the processes during the exceptional cycle. We can suppose that the maximum possible separation is achieved at each transition over the entire cycle. The worst case would have the separation increase by $K \equiv 2(j \vee k)$ at every transition. (The two processes would have a transition at the same time going the maximum possible distance away from each other.) Hence, since the total number of transitions of the upper bound process in the exceptional cycle after the initial non-identical transition is N_e^n, then the other processes are in a state within KN_e^n states of the regenerative state, where the upper bound process $D_f^n(X_M^n, \cdot)$ will be at the end of the exceptional cycle. In (5.5.25) we have shown that this random bound on the initial difference has a geometric tail, so that the probability of large differences are controlled. Since the first passage time (number of transitions) from any fixed state to s^* has a generating function, the number of transitions until all the processes have hit s^* is SB. Consequently, N_c^n is SB.

We now specify what we do if there are differences within the period considered above.

If there were any differences (an event of small probability), then we repeat the construction for the second part of the second interval using four independent versions of $D_f^n(X_M^n, \cdot)$ until the four processes are again together in the regenerative state. This second try will produce a number of transitions $N_{c,2}^n$ different from $N_{c,1}^n \equiv N_c^n$ in the first try, but actually somewhat more favorable (tending to be smaller) because the initial conditions are more favorable, with three of the four processes likely to be starting in the regenerative state and the interior process differing at most by the gap ζ, by virtue of Corollary 4.8.4. (By the independence of the pairs, two or more differences will be asymptotically negligible compared to a single difference.) So, if the second try is needed, we will be able to control $N_{c,2}^n$ just as we can control $N_{c,1}^n$.

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

The final task is to show that the special initial conditions imposed in (5.5.18) are actually not needed. However, given the assumed condition (4.8.21), it suffices to assume that state j is the specified regenerative state. Alternatively, we could add an extra initial period at the beginning. During this initial period we generate all processes from a common Poisson process and proceed until the upper bound process hits the designated regenerative

state. If all processes stay together in the designated regenerative state, then we can proceed with the construction above. With high probability, all processes will move together throughout this initial period. If that does not occur, we can have a subsequent interval of the kind $U^n_{2,k}$ analyzed above, before we get to the regenerative state. A similar story will hold if we generalize the initial conditions in a controlled way. That completes the proof. ∎

5.5.5 Remaining Proof in §4.8.7

Proof of Lemma 4.8.11: First, let $\delta > 0$, $\epsilon > 0$ and t with $0 < t < \delta$ be given, where the δ is chosen so that $\delta < \xi$ for ξ in Lemmas 4.8.7, 4.8.9 and 4.8.10. Below we will be introducing a new ξ less than this δ.

We start by observing that versions of Lemmas 4.8.9 and 4.8.10 hold on an interval $[t, t + \xi]$, where $\xi \equiv \xi(t)$ satisfies $0 < \xi < \delta - t$. Before, we started with the convergence $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}^3$ at time 0 based on Assumption 3. Now, instead, we base the convergence $\bar{X}^n(t) \Rightarrow \bar{X}(t)$ at time t on the convergence we have along the converging subsequence. Since the processes are Markov processes, we can construct the processes after time t, given only the value of $X^n(t)$, independently of what happens on $[0, t]$. We apply Lemma 4.8.7 to deduce that $P(\bar{X}(t) \in \mathbb{A}) = 1$ (which is justified by our choice of δ).

We now indicate how the proofs of Lemmas 4.8.9 and 4.8.10 need to be modified, proceeding forward after time t. Let $X^{n,\xi}_M \equiv (X^{n,\xi}_{M^+}, X^{n,\xi}_{M^-})$ be defined similar to X^n_M in (5.5.13) and $X^{n,\xi}_m \equiv (X^{n,\xi}_{m^+}, X^{n,\xi}_{m^-})$ be defined similar to and X^n_m in (5.5.15), but with supremum and infimum taken over the interval $[t, t + \xi]$ (instead of over the interval $[0, \xi]$ as before (where the constants ξ need not be the same for each t; i.e., $\xi \equiv \xi(t)$). Recall that the associated bounding quantities are constructed from separate processes related to X^n only through their distributions. These too do not depend on the evolution of X^n after time t.

Reasoning as before, by virtue of Lemma 4.7.1, the limits $x_M^\xi \equiv (x_{M^+}^\xi, x_{M^-}^\xi)$ and $x_m^\xi \equiv (x_{m^+}^\xi, x_{m^-}^\xi)$ of $\bar{X}_M^{n,\xi}$ and $\bar{X}_m^{n,\xi}$ exist. (Since $\bar{X}(t)$ so far is a random variable, so are x_M^ξ and x_m^ξ. However, we can regard $\bar{X}(t)$ as a constant by conditioning upon it, without affecting the evolution after time t, because of the Markov property.) In particular, Applying the continuous mapping theorem for the supremum, Theorem 12.11.7 in [78], we have that $X_{M^+}^{n,\xi}/n \Rightarrow x_{M^+}^\xi \equiv (q_{1,M}^\xi, q_{2,M}^\xi, z_{M^+}^\xi)$ and $X_{M^-}^{n,\xi}/n \Rightarrow x_{M^-}^\xi \equiv (q_{1,M}^\xi, q_{2,M}^\xi, z_{M^-}^\xi)$ as $n \to \infty$, where

$$\begin{aligned}
q_{1,M}^\xi &\equiv \inf_{t\le s\le t+\xi} q_1^\xi(s) \vee 0, \\
q_{2,M}^\xi &\equiv \sup_{t\le s\le t+\xi} q_2(s), \\
z_{M^+} &\equiv \begin{cases} \inf_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \\ \sup_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \end{cases} \\
z_{M^-} &\equiv \begin{cases} \inf_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \\ \sup_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \end{cases}
\end{aligned} \tag{5.5.26}$$

Similarly, $X_{m^+}^{n,\xi}/n \Rightarrow x_{m^+}^\xi \equiv (q_{1,m}^\xi, q_{2,m}^\xi, z_{m^+}^\xi)$ and $X_{m^-}^{n,\xi}/n \Rightarrow x_{m^-}^\xi \equiv (q_{1,m}^\xi, q_{2,m}^\xi, z_{m^-}^\xi)$ as $n \to \infty$, with

$$\begin{aligned}
q_{1,m}^\xi &\equiv \sup_{t\le s\le t+\xi} q_1(s), \\
q_{2,m}^\xi &\equiv \inf_{t\le s\le t+\xi} q_2(s) \vee 0, \\
z_{m^+} &\equiv \begin{cases} \inf_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \\ \sup_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \end{cases} \\
z_{m^-} &\equiv \begin{cases} \inf_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \\ \sup_{t\le s\le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \end{cases}
\end{aligned} \tag{5.5.27}$$

The two bounding frozen difference processes are $\{D_f^n(X_M^{n,\xi}, s) : s \ge t\}$ and $\{D_f^n(X_m^{n,\xi}, s) : s \ge t\}$. As a consequence of this construction, we can conclude that there exists $\xi > 0$ and

an integer n_1 such that the drift rates of these bounding processes satisfy both the inequalities in (4.8.12) in order for them to be positive recurrent and the rate order in (4.8.18) with probability at least $1 - \epsilon/6$ for all $n \geq n_1$.

We next apply Lemma 4.8.10 to conclude that there exists a new ξ, taken no bigger than the one created so far, such that the following variants of the integral inequalities in (4.8.23) hold with probability at least $1 - \epsilon/6$ as well:

$$\begin{aligned} \frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(X_m^{n,\xi},s)\}}\,ds - \frac{\epsilon}{6m_2} &\leq \frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_{1,2}^n(s)>0\}}\,ds \\ &\leq \frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(X_M^{n,\xi},s)>0\}}\,ds + \frac{\epsilon}{6m_2}. \end{aligned} \tag{5.5.28}$$

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

We now represent the bounding frozen queue-difference processes directly in terms of the FTSP, using the relation (4.8.9):

$$\begin{aligned} \{D_f^n(\lambda_i^n, m_j^n, X_m^{n,\xi}, t+s) : s \geq 0\} &\stackrel{\mathrm{d}}{=} \{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, t+sn) : s \geq 0\} \\ \{D_f^n(\lambda_i^n, m_j^n, X_M^{n,\xi}, t+s) : s \geq 0\} &\stackrel{\mathrm{d}}{=} \{D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, t+sn) : s \geq 0\}. \end{aligned} \tag{5.5.29}$$

Upon making a change of variables, the bounding integrals in (5.5.28) become

$$\begin{aligned} \frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(\lambda_i^n,m_j^n,X_m^{n,\xi},s)>0\}}\,ds &\stackrel{\mathrm{d}}{=} \frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n,m_j^n/n,X_m^{n,\xi}/n,s>0\}}\,ds \\ \frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(\lambda_i^n,m_j^n,X_M^{n,\xi},s)>0\}}\,ds &\stackrel{\mathrm{d}}{=} \frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n,m_j^n/n,X_m^{n,\xi}/n,s)>0\}}\,ds. \end{aligned} \tag{5.5.30}$$

For each integer k, we have the iterated limits

$$\begin{aligned}
&\lim_{n\to\infty}\lim_{s\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s) = k) \\
&\quad = \lim_{s\to\infty}\lim_{n\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s) = k), \\
&\lim_{n\to\infty}\lim_{s\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s) = k) \\
&\quad = \lim_{s\to\infty}\lim_{n\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s) = k),
\end{aligned} \tag{5.5.31}$$

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

By Corollary 4.8.3, we also have the associated double limit for the averages over intervals of length $O(n)$ as $n \to \infty$

$$\begin{aligned}
&\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s)>0\}}\, ds \Rightarrow P(D(\lambda_i, m_j, x_m^\xi, \infty) > 0) \equiv \pi_{1,2}(x_m^\xi), \\
&\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s)>0\}}\, ds \Rightarrow P(D(\lambda_i, m_j, x_M^\xi, \infty) > 0) \equiv \pi_{1,2}(x_M^\xi).
\end{aligned} \tag{5.5.32}$$

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

Invoking Lemma 4.8.8, choose ξ less than or equal to the previous value of ξ such that

$$|\pi_{1,2}(x_m^\xi) - \pi_{1,2}(\bar{X}(t))| \le \frac{\epsilon}{6m_2}. \tag{5.5.33}$$

For that ξ, applying (5.5.32), choose $n_2 \geq n_1$ such that

$$\begin{aligned} & P\left(|\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n,m_j^n/n,X_m^{n,\xi}/n,s)>0\}}\,ds - \pi_{1,2}(x_m)| > \frac{\epsilon}{6m_2}\right) < \frac{\epsilon}{6} \\ \text{and} \quad & P\left(|\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n,m_j^n/n,X_M^{n,\xi}/n,s)>0\}}\,ds - \pi_{1,2}(x_M)| > \frac{\epsilon}{6m_2}\right) < \frac{\epsilon}{6} \end{aligned} \tag{5.5.34}$$

for all $n \geq n_2$.

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

$$P(\sup_{u:t\leq u\leq t+\xi}\{|\bar{X}^n(u) - \bar{X}(t)|\} > \epsilon/6) < \epsilon/6 \quad \text{for all} \quad n \geq n_3. \tag{5.5.35}$$

For the current proof, we will use the consequence

$$P(\sup_{u:t\leq u\leq t+\xi}\{|\bar{Z}^n_{1,2}(u) - \bar{Z}_{1,2}(t)|\} > \epsilon/6) < \epsilon/6 \quad \text{for all} \quad n \geq n_3. \tag{5.5.36}$$

We now show the consequences of the selections above. We will directly consider only the upper bound; the reasoning for the lower bound is essentially the same. Without loss of generality, we take $\epsilon \leq 1 \wedge m_2$. From above, we have the following relations (explained afterwards) holding with probability at least $1-\epsilon$ (counting $\epsilon/6$ once each for (5.5.26), (5.5.27), (5.5.28), (5.5.36) and twice for (5.5.34)):

$$
\begin{aligned}
(a)\quad & \int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}} \bar{Z}^n_{1,2}(s)\, ds \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}}\, ds \\
(b)\quad & \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_t^{t+\xi} 1_{\{D^n_f(\lambda^n_i, m^n_j, X^n_M, s)>0\}}\, ds + \frac{\epsilon\xi}{6m_2} \right) \\
(c)\quad & \stackrel{\mathrm{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_0^{\xi} 1_{\{D(\lambda^n_i/n, m^n_j/n, X^{n,\xi}_M/n, t+sn)>0\}}\, ds + \frac{\epsilon\xi}{6m_2} \right) \\
(d)\quad & \stackrel{\mathrm{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\frac{1}{n\xi} \int_0^{n\xi} 1_{\{D(\lambda^n_i/n, m^n_j/n, X^{n,\xi}_M/n, t+s)>0\}}\, ds + \frac{\epsilon}{6m_2} \right) \\
(e)\quad & \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\pi_{1,2}(x^{\xi}_M) + \frac{2\epsilon}{6m_2} \right) \\
(f)\quad & \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\pi_{1,2}(\bar{X}(t)) + \frac{3\epsilon}{6m_2} \right) \\
(g)\quad & \le \bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t))\xi + \frac{\pi_{1,2}(\bar{X}(t))}{6}\epsilon\xi + \frac{1}{2}\epsilon\xi + \frac{\xi\epsilon^2}{12m_2} \\
(h)\quad & \le \bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t))\xi + \frac{3}{4}\epsilon\xi \\
& \le (\bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t)) + \epsilon)\xi \quad \text{for all} \quad n \ge n_0 \equiv n_3.
\end{aligned}
\tag{5.5.37}
$$

We now explain the steps in (5.5.37): First, for (a) we replace $\bar{Z}^n_{1,2}(s)$ by $\bar{Z}_{1,2}(t)$ for $t \le s \le t+\xi$ by applying (5.5.36). For (b), we apply Lemma 4.8.10. For (c), we use the alternative representation in terms of the FTSP in (5.5.29). For (d), we use the change of variables in (5.5.30). For (e), we use (5.5.34), exploiting the convergence in (5.5.32). For (f), we use (5.5.33). Step (g) is simple algebra, exploiting $\bar{Z}_{1,2}(t) \le m_2$. Step (h) is more algebra, exploiting $\pi_{1,2}(\bar{X}(t)) \le 1$, and $\epsilon \le 1 \wedge m_2$. That completes the proof of the lemma. ∎

Chapter 6

Diffusion Refinements

In this chapter we use the fluid limit, together with the SSC result, to establish diffusion limits when the system is overloaded and the fluid limit is in $\mathbb{A}$. However, our results here depend on Conjecture 6.2.1, which we did not prove yet. Essentially, Conjecture 6.2.1 strengthens the AP result to diffusion scale. We intend to prove this result in the future.

6.1 The Diffusion Limit

Let $q_s(t)$ be the sum of the two fluid-limit queues: $q_s(t) \equiv q_1(t)+q_2(t)$. Similarly, let $Q_s^n(t)$ be the total queue-length process in system n. For simplicity of exposition, we assume that the thresholds are dropped once crossed, so that $k_{1,2}^n = \kappa = 0$ for all $n \geq 1$.

For $t \geq 0$ we define the diffusion-scaled processes:

$$\begin{aligned} \hat{Q}_s^n(t) &\equiv \frac{Q_s^n(t) - nq_s(t)}{\sqrt{n}}; \qquad \hat{Z}_{1,2}^n(t) \equiv \frac{Z_{1,2}^n(t) - nz_{1,2}(t)}{\sqrt{n}}; \\ \hat{Q}_1^n(t) &\equiv \frac{Q_1^n(t) - nq_1(t)}{\sqrt{n}}; \qquad \hat{Q}_2(t) \equiv \frac{Q_2^n(t) - nq_2(t)}{\sqrt{n}}. \end{aligned} \tag{6.1.1}$$

Let

$$p_1 \equiv \frac{r}{1+r}, \qquad p_2 \equiv 1 - p_1 = \frac{1}{1+r} \tag{6.1.2}$$

Theorem 6.1.1. *Let T be such that $x(t) \in \mathbb{A}$ over $[0,T)$. (Hence, $T \geq \delta$ for δ in Theorem 4.6.1, and possibly $T = \infty$.) Assume that*

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

and that Conjecture 6.2.1 holds. Then we have the joint convergence

$$\left(\hat{Q}_s^n, \hat{Q}_1^n, \hat{Q}_2^n, \hat{Z}_{1,2}^n\right) \Rightarrow \left(\hat{Q}_s, p_1\hat{Q}_s, p_2\hat{Q}_s, \hat{Z}_{1,2}\right) \quad \textit{in } \mathcal{D}_4([0,T)) \textit{ as } n \to \infty, \tag{6.1.3}$$

where $(\hat{Q}_s, \hat{Z}_{1,2})$ is the unique solution of the following two-dimensional stochastic integral equation:

$$\begin{aligned}
\hat{Q}_s(t) &= \hat{Q}_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}_{1,2}(s)\, ds - (p_1\theta_1 + p_2\theta_2) \int_0^t \hat{Q}_s(s)\, ds \\
&\quad + B_1\left(\gamma_1(t)\right), \\
\hat{Z}_{1,2}(t) &= \hat{Z}_{1,2}(0) - \int_0^t \left[(\mu_{2,2} - \mu_{1,2})\pi_{1,2}(x(s)) + \mu_{1,2}\right] \hat{Z}_{1,2}(s)\, ds + B_2(\gamma_2(t)),
\end{aligned} \tag{6.1.4}$$

where, for $i = 1, 2$, B_i are independent standard BM's, and γ_i are the following strictly-increasing time-scale functions:

$$\begin{aligned}
\gamma_1(t) &\equiv (\lambda_1 + \lambda_2 + m_1\mu_{1,1} + \mu_{2,2}m_2)t + (p_1\theta_1 + p_2\theta_2) \int_0^t q_s(u)\, du \\
&\quad + (\mu_{1,2} - \mu_{2,2}) \int_0^t z_{1,2}(u)\, du, \\
\gamma_2(t) &\equiv \int_0^t \left(\mu_{1,2} - (\mu_{2,2} + \mu_{1,2})\pi_{1,2}(x(u))\right) z_{1,2}(u)\, du + \mu_{2,2}m_2 \int_0^t \pi_{1,2}(x(u))\, du,
\end{aligned} \tag{6.1.5}$$

It is easy to see that γ_1 and γ_2 are indeed strictly increasing since their derivatives are positive;

$$\begin{aligned}\dot{\gamma}_1(t) &= \lambda_1 + \lambda_2 + m_1\mu_{1,1} + \mu_{2,2}m_2 + (p_1\theta_1 + p_2\theta_2)q_s(t) + (\mu_{1,2} - \mu_{2,2})z_{1,2}(t) \\ &\geq \lambda_1 + \lambda_2 + m_1\mu_{1,1} + \mu_{1,2}z_{1,2}(t) > 0,\end{aligned}$$

where the first inequality is due to the fact that $q_s(t) \geq 0$ and $0 \leq z_{1,2}(t) \leq m_2$ for all $t \geq 0$. Similarly, since $0 < \pi_{1,2}(x(t)) < 1$ and $z_{1,2}(t) \leq m_2$ for all $t \geq 0$ we have

$$\begin{aligned}\dot{\gamma}_2(t) &= \mu_{1,2}z_{1,2}(t) - (\mu_{2,2} + \mu_{1,2})\pi_{1,2}(x(t))z_{1,2}(t) + \mu_{2,2}m_2\pi_{1,2}(x(t)) \\ &= \mu_{1,2}(1 - \pi_{1,2}(x(t)))z_{1,2}(t) + \mu_{2,2}(m_2 - z_{1,2}(t))\pi_{1,2}(x(t)) > 0.\end{aligned}$$

The stochastic process $(\hat{Q}_s, \hat{Z}_{1,2})$ is evidently difficult to analyze; Apart from being a two-dimensional diffusion process, the time arguments of the Brownian-motion parts of $(\hat{Q}_s, \hat{Z}_{1,2})$ have no closed-form solutions. However, if we know that the fluid solution converges to stationarity, then it does so exponentially fast, according to Theorem 2.7.4. Since we are mainly interested in the steady state variance of the diffusion limits, it is reasonable to initialize "close" to this fluid stationary point in order to simplify the expressions in (6.1.4). We do this in the next corollary. We then further simplify the diffusion expressions. (see also [59] for more discussion on diffusion approximations for this model. In particular, for simple heuristics which are shown to approximate the diffusion limits exceptionally well.)

Corollary 6.1.1. *If, in addition to the conditions of Theorem 6.1.1,* $x(0) = x^*$ *for* x^* *in* (3.5.3) *(so that* x *is stationary, and hence* $T = \infty$ *in the statement of Theorem 6.1.1), then* $\gamma_i(t) = \xi_i t$, $i = 1, 2$, *for* $\gamma_i(t)$ *in* (6.1.5), *where*

$$\xi_1 = 2(\lambda_1 + \lambda_2) \quad \textit{and} \quad \xi_2 = \frac{2\mu_{1,2}\mu_{2,2}z^*_{1,2}(m_2 - z^*_{1,2})}{\mu_{1,2}z^*_{1,2} + (m_2 - z^*_{1,2})\mu_{2,2}}. \tag{6.1.6}$$

Then, $\hat{Q}_s$ and $\hat{Z}_{1,2}$ are the unique solutions to the following integral equation

$$\begin{aligned}\hat{Q}_s(t) &= \hat{Q}_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}_{1,2}(s)\, ds - (p_1\theta_1 + p_2\theta_2) \int_0^t \hat{Q}_s(s)\, ds \\ &\quad + \sqrt{\xi_1} B_1(t) \\ \hat{Z}_{1,2}(t) &= \hat{Z}^n_{1,2}(0) - \zeta \int_0^t \hat{Z}_{1,2}(s)\, ds + \sqrt{\xi_2} B_2(t),\end{aligned} \tag{6.1.7}$$

where B_1 and B_2 are independent standard BM's and

$$\zeta \equiv \frac{\mu_{1,2}\mu_{2,2} m_2 z^*_{1,2}}{\mu_{1,2} z^*_{1,2} + \mu_{2,2}(m_2 - z^*_{1,2})}. \tag{6.1.8}$$

Hence, $\hat{Z}_{1,2}$ can be expressed separately, without referring to $\hat{Q}_s$, as a one-dimensional Ornstein-Uhlenbeck (OU) process with steady-state distribution

$$\hat{Z}_{1,2}(\infty) \stackrel{d}{=} N\left(0,\ 1 - \frac{z^*_{1,2}}{m_2}\right).$$

Proof: By the definition of a stationary point, if $x(0) = x^*$ then $x(t) = x^*$ for all $t > 0$. Then $q_i(t) = q^*_i$ and $z_{1,2}(t) = z^*_{1,2}$ and $\pi_{1,2}(x(t)) = \pi^*_{1,2}$, for $\pi^*_{1,2}$ in (3.5.4). The expressions in (6.1.6) follow easily from the expressions in (6.1.5), by replacing the time-dependent fluid quantities by their stationary values, described in (3.5.3). Replacing $\pi_{1,2}(x(t))$ by the expression for $\pi^*_{1,2}$ in (3.5.4), and $z_{1,2}(t)$ by $z^*_{1,2}$, gives us the expression for $\hat{Z}_{1,2}$, which is well-known to be the equation of an OU process with the specified steady-state distribution (e.g., see pg. 218 of [42]). ∎

Corollary 6.1.2. *If, in addition to the assumptions of Theorem 6.1.1, $\mu_{2,2} = \mu_{1,2} \equiv \nu$, then the two diffusion-limit processes $\hat{Q}_s$ and $\hat{Z}_{1,2}$ are independent one-dimensional processes*

which satisfy the following integral equations

$$\begin{aligned}\hat{Q}_s(t) &= \hat{Q}_s(0) - \eta_2 \int_0^t \hat{Q}_s(s)\, ds + B_1\left(\tilde{\gamma}_1(t)\right), \\ \hat{Z}_{1,2}(t) &= \hat{Z}_{1,2}(0) - \nu z^*_{1,2} \int_0^t \hat{Z}_{1,2}(s)\, ds + B_2\left(\tilde{\gamma}_2(t)\right),\end{aligned} \tag{6.1.9}$$

where

$$\begin{aligned}\tilde{\gamma}_1(t) &\equiv 2(\lambda_1+\lambda_2)t + \left(\frac{\eta_1}{\eta_2} - q_s(0)\right) e^{-\eta_2 t} \\ \tilde{\gamma}_2(t) &\equiv \nu\left(m_2 \int_0^t \pi_{1,2}(x(u))\, du + \int_0^t z_{1,2}(u)\, du - 2\int_0^t \pi_{1,2}(x(u)) z_{1,2}(u)\, du\right).\end{aligned} \tag{6.1.10}$$

$$\eta_1 \equiv \lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\nu, \qquad \eta_2 \equiv p_1\theta_1 + p_2\theta_2, \tag{6.1.11}$$

and B_1 and B_2 are independent standard BM's.

Proof: It is immediate from the expressions of $\hat{Q}_s$ and $\hat{Z}_{1,2}$ in (6.1.4) that when $\mu_{1,2} = \mu_{2,2}$ the two diffusion processes are independent. Now, since $q_i = p_i q_s$ and $\mu_{1,2} = \mu_{2,2}$, it follows from (4.5.13) that $\dot{q}_s(t)$ satisfies the simple ordinary differential equation

$$\dot{q}_s(t) = (\lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\mu_{2,2}) - (p_1\theta_1 + p_2\theta_2) q_s(t) \equiv \eta_1 - \eta_2 q_s(t),$$

whose solution is

$$q_s(t) = \frac{\eta_1}{\eta_2} + \left(q(0) - \frac{\eta_1}{\eta_2}\right) e^{-\eta_2 t}$$

for η_1 and η_2 in (6.1.11). Plugging $q_s(t)$ in $\gamma_1(t)$ in (6.1.5) gives $\tilde{\gamma}_1(t)$.

The expressions of $\tilde{\gamma}_2(t)$ is immediate from $\gamma_2(t)$ in (6.1.5) when $\mu_{1,2} = \mu_{2,2} = \nu$. Also note that, in this case, $\zeta = \nu z^*_{1,2}$ for ζ in (6.1.8). ∎

The following corollary is immediate from the expressions of $\hat{Q}_s$ and $\hat{Z}_{1,2}$ in (6.1.9).

Note that when $\mu_{1,2} = \mu_{2,2} = \nu$ then $\pi^* = z^*_{1,2}/m_2$.

Corollary 6.1.3. *If, in addition to the conditions of Corollary 6.1.2, the initial conditions are such that $\hat{Q}_s(0) = q^*_s$ and $\hat{Z}_{1,2}(0) = z^*_{1,2}$ (so that $\pi_{1,2}(x(t)) = \pi^*_{1,2}$), then $\hat{Q}_s$ and $\hat{Z}_{1,2}$ are independent one-dimensional Ornstein-Uhlenbeck (OU) processes, i.e.,*

$$\hat{Q}_s = q^*_s - \eta_2 \int_0^t \hat{Q}_s(s)\, ds + \sqrt{2(\lambda_1 + \lambda_2)} B_1(t)$$
$$\hat{Z}_{1,2} = z^*_{1,2} - \nu z^*_{1,2} \int_0^t \hat{Z}_{1,2}(s)\, ds + \sqrt{2\nu z^*_{1,2}\left(1 - \frac{z^*_{1,2}}{m_2}\right)} B_2(t),$$

for η_2 in (6.1.11) *and $\pi^*_{1,2}$ in* (3.5.4)*, and where B_1 and B_2 are independent standard BM's. The two OU processes have the following steady-state distributions:*

$$\hat{Q}_s(\infty) \stackrel{d}{=} N\left(0\,,\, \frac{(1+r)(\lambda_1+\lambda_2)}{r\theta_1 + \theta_2}\right) \quad \textit{and} \quad \hat{Z}_{1,2}(\infty) \stackrel{d}{=} N\left(0\,,\, 1 - \frac{z^*_{1,2}}{m_2}\right).$$

Equivalently, $\hat{Z}_{1,2}(\infty) \stackrel{d}{=} N\left(0\,,\, 1 - \pi^*_{1,2}\right)$.

Remark 6.1.1. (Equivalence with the single-class model.) If, in addition to the conditions of Corollary 6.1.3, it also holds that $\theta_1 = \theta_2 \equiv \theta$, then the diffusion-limit process $\hat{Q}_s$ is the same as the limit obtained for the $M/M/n + M$ model in the Efficiency Driven (ED) regime, see [79]. That is, $\hat{Q}_s$ is an Ornstein-Uhlenbeck process with infinitesimal mean equal to θ and infinitesimal variance $2\lambda \equiv 2(\lambda_1 + \lambda_2)$. Thus, its steady-state distribution is normal with mean zero and variance λ/θ.

Theorem 6.1.1 and its corollaries illustrate the strength of the AP. A direct implication of the AP is SSC for both the fluid-scaled and the diffusion-scaled queue processes. But the AP implies more than just SSC; As we have seen, thanks to the AP, we can analyze the diffusion-scaled service-process $\hat{Z}_{1,2}(t)$ and its fluid counterpart $z_{1,2}(t)$. The AP also implies that $\hat{Z}_{1,2}(t)$ does not depend on the limiting diffusion queue processes. This may

seem surprising at first, since $Z^n_{1,2}(t)$ depends on the queues $Q^n_1(t)$ and $Q^n_2(t)$ for each n and t, and in the fluid limit, $z_{1,2}(t)$ depends on the fluid-queues $q_1(t)$ and $q_2(t)$, via $\pi_{1,2}(x(t))$.

In particular, when $\mu_{1,2} = \mu_{2,2}$ (service rates are pool dependent), the diffusion-limit queues are independent of the diffusion-limit service processes $\hat{Z}_{i,j}$, $i, j = 1, 2$. To see why this result is implied by the averaging principle, observe that the indictor functions in (6.2.6) below, which are functions of the two queues, are replaced by expressions involving $\pi_{1,2}(x(t))$, which do not depend on the queues. (This may be a little confusing, but $\pi_{1,2}(x(t))$ is a function of the deterministic fluid-limit queues, and does not depend on the actual queues.)

We can regard this result as a converse to SSC for the following reason: In our model, SSC of the queues implies that the two-dimensional process $(\hat{Q}_1, \hat{Q}_2)$, which is in general in $\mathcal{D}_2$, exists in the one-dimensional hyperplane $\hat{Q}_1 = r_{1,2}\hat{Q}_2$. That is, the two queues are strictly correlated, and behave as a one-dimensional process; The dimension of the state space collapses to one.

On the other hand, Corollaries 6.1.2 and 6.1.3 and Remark 6.1.1 imply that the two-dimensional process $(\hat{Q}_s, \hat{Z}_{1,2}) \in D^2$ can be decomposed into its two components. We get a "separation" of the state space $\mathcal{D}_2$ as each process exists in $\mathcal{D}$, independently of the other process. (This illustrates why the condition that the rates are pool dependent is sufficient to maintain stability in [29] and [31] for the X model, and more generally, for models whose routing graphs are cyclic.)

6.2 Proof of Theorem 6.1.1

We use the sample-path construction in §4.4 to construct martingale representations for the stochastic processes, as in [57]. The martingale representation is constructed without specifying any filtration, since we will not use any martingale property. We call this "the

martingale representation" for convenience. To achieve this martingale representation, we decompose the independent time-changed Poisson processes N_i^a, $N_{i,2}^s$ and N_i^u, $i = 1, 2$, in the following way:

$$\begin{aligned}
M_{1,1}^n(t) &\equiv N_{1,1}^s(m_1^n \mu_{1,1} t) - m_1^n \mu_{1,1} t \\
M_{i,2}^n(t) &\equiv N_{i,2}^s\left(\mu_{i,2} \int_0^t Z_{i,2}^n(s) ds\right) - \mu_{i,2} \int_0^t Z_{i,2}^n(s) ds, \quad i = 1, 2, \\
M_{a_i}^n(t) &\equiv N_i^a(\lambda_i^n t) - \lambda_i^n t, \quad i = 1, 2, \\
M_{u_i}^n(t) &\equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s) ds\right) - \theta_i \int_0^t Q_i^n(s) ds, \quad i = 1, 2.
\end{aligned} \tag{6.2.1}$$

The processes in (6.2.1) can be shown to be square-integrable martingales (with respect to an appropriate filtration), and we thus refer to them as "martingales".

Unlike in the fluid-limit proof, which was carried out using the compactness approach, the diffusion limits will be proved using the continuous mapping approach. It is significant that the the continuity of the integral representation below is due to the AP and SSC established before.

Lemma 6.2.1. (Continuity of the two-dimensional integral representation) *Consider the two-dimensional integral representation*

$$\begin{aligned}
x_1(t) &= b_1 + y_1(t) + \alpha_2 \int_0^t x_2(s)\, ds + \alpha_1 \int_0^t x_1(s)\, ds \\
x_2(t) &= b_2 + y_2(t) + \int_0^t g(s) x_2(s)\, ds
\end{aligned} \tag{6.2.2}$$

where $g : \mathbb{R} \to \mathbb{R}$ satisfies $g(0) = 0$ and is Lipschitz continuous with a Lipschitz constant c_g. The integral representation (6.2.2) *has a unique solution (x_1, x_2), so that the integral representation constitutes a function $f : \mathcal{D}_2 \times \mathbb{R}_2 \to \mathcal{D}_2$ mapping (x_1, x_2, b_1, b_2) into $(x_1, x_2) \equiv f(x_1, x_2, b_1, b_2)$. In addition, the function f is a continuous mapping from*

$\mathcal{D}_2 \times \mathbb{R}_2$ to $\mathcal{D}_2$. Moreover, if y_2 is continuous then x_2 is continuous. If both y_1 and y_2 are continuous, then x_1 is also continuous.

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

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

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

To show that f_2 is continuous we use Theorem 2.11 in [67] with $h(x_2(u), u) \equiv g(u)x_2(u)$. Clearly, condition (1) in that theorem holds, since $g(0) = 0$, and it remains to show that Condition (2) holds as well. For that purpose, choose $T > 0$ and let λ be a homeomorphism on $[0, T]$ with strictly positive derivative $\dot{\lambda}$. Then, for every $\varphi_1, \varphi_2 \in \mathcal{D}$

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

where $c_1 \equiv c_g T^2 \|\varphi_2\|_T$ and $c_2 \equiv \|g\|_T$.

For $x_1 = f_1(y_1, x_2, b_1)$ we can apply Theorem 4.1 in [57] with input $y \equiv y_1 + \alpha_2 \int_0^t x_2(u)\, du$. It follows from Theorem 2.11 in [67] that if y_2 is continuous then so is x_2. If, in addition, y_1 is continuous, then y is continuous and, by Theorem 4.1 in [57], so is

x_1. ■

Proof of Theorem 6.1.1: Following (4.4.3)–(4.4.4), we write the total queue-length process $Q_s^n(t) \equiv Q_1^n(t) + Q_2^n(t)$ using the martingale decomposition, as in [57]. Observe that the indicator functions in the representation (4.4.3) and (4.4.4) do not appear in the representation of $Q_s^n(t)$.

$$\begin{aligned} Q_s^n(t) &= Q_s^n(0) + N_1^a(\lambda_1^n t) + N_2^a(\lambda_2^n t) - N_{1,1}^s(m_1^n \mu_{1,1} t) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t Z_{1,2}^n(s))\,ds\right) - N_{2,2}^s\left(\mu_{2,2}\int_0^t Z_{2,2}^n(s)\,ds\right) \\ &\quad - N_1^u\left(\theta_1\int_0^t Q_1^n(s)\,ds\right) - N_2^u\left(\theta_2\int_0^t Q_2^n(s)\,ds\right), \quad t \geq 0 \\ &= Q_s^n(0) + (\lambda_1^n + \lambda_2^n)t - m_1^n\mu_{1,1}t - \mu_{1,2}\int_0^t Z_{1,2}^n(s)\,ds - \mu_{2,2}\int_0^t Z_{2,2}^n(s)\,ds \\ &\quad - \theta_1\int_0^t Q_1^n(s)\,ds - \theta_2\int_0^t Q_2^n(s)\,ds + M_s^n(t), \end{aligned}$$

where

$$M_s^n(t) \equiv \sum_{i=1}^2 M_{a_i}^n(t) - \sum_{i=1}^2 M_{u_i}^n(t) - \sum_{i=1}^2 M_{i,2}^n(t) - M_{1,1}^n(t). \tag{6.2.3}$$

From (4.5.13) it follows that $q_s \equiv q_1 + q_2$, the fluid counterpart of Q_s^n, evolves according to the integral equation:

$$\begin{aligned} q_s(t) &= q_s(0) + (\lambda_1 + \lambda_2)t - \mu_{1,1}m_1 t - \mu_{1,2}\int_0^t z_{1,2}(u)\,du - \mu_{2,2}\int_0^t z_{2,2}(u)\,du \\ &\quad - \theta_1\int_0^t q_1(u)\,du - \theta_2\int_0^t q_2(u)\,du, \end{aligned}$$

so that, substituting q_1 with $p_1 q_s(u)$ and $q_2(u)$ with $p_2 q_s(u)$, we get

$$\begin{aligned} q_s(t) = q_s(0) + (\lambda_1 + \lambda_2)t - \mu_{1,1} m_1 t - \mu_{2,2} m_2 t \\ + (\mu_{2,2} - \mu_{1,2}) \int_0^t z_{1,2}(u)\, du - (p_1\theta_1 + p_2\theta_2) \int_0^t q_s(u)\, du \end{aligned}$$

We get

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

Obviously, the second and third terms in the expression above converge to zero. Recall that, by Theorem 4.7.1, $n^{-1/2}\|Z_{2,2}^n - (m_2^n - Z_{1,2}^n)\| \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$ so that $z_{2,2} = m_2 - z_{1,2}$. Also, $(m_2^n/n - m_2) \to 0$ as $n \to \infty$ by assumption. Hence,

$$\begin{aligned} \hat{Q}_s^n = \hat{Q}_s^n(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}_{1,2}^n(s)\, ds \\ - \theta_1 \int_0^t \hat{Q}_1^n(s)\, ds - \theta_2 \int_0^t \hat{Q}_2^n(s)\, ds + \hat{M}_s^n(t). \end{aligned} \tag{6.2.4}$$

Define

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

By applying the continuous-mapping theorem and the SSC result in Theorem 4.5.6, we

have that $\|\hat{Q}^n_s - \hat{Y}^n_s\| \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$. Hence we can write

$$\begin{aligned}\hat{Q}^n_s(t) = \hat{Q}^n_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}^n_{1,2}(s)\, ds - (p_1\theta_1 + p_2\theta_2) \int_0^t \hat{Q}^n_s(s)\, ds \\ + \hat{M}^n_s(t) + o_P(1).\end{aligned} \tag{6.2.5}$$

The next Lemma identifies the limit of the martingale $\hat{M}^n_s(t)$. It's proof is given in the end of this section.

Lemma 6.2.2. *Under the conditions of Theorem 6.1.1, $\hat{M}^n_s(t) \Rightarrow B(\gamma_1(t))$ in $\mathcal{D}$ as $n \to \infty$, where $\{B(t) : t \geq 0\}$ is a standard brownian motion, and $\gamma_1(t)$ is defined in* (6.1.5)*.*

To finish the proof, we apply Lemma 6.2.1 to the integral representation of $\hat{Q}^n_s(t)$. Assuming that $\hat{Z}^n_{1,2} \Rightarrow \hat{Z}_{1,2}$ (as will be shown next, building on Conjecture 6.2.1), we have that (6.2.5) is a continuous mapping from $\mathcal{D}$ to itself, and the convergence of $Q^n_s(t)$ to the limit in (6.1.3) is implied by the limit of $\hat{M}^n_s(t)$ in Lemma 6.2.2.

We now turn to the $\hat{Z}_{1,2}(t)$ process. We start with the representation (4.4.2) of $Z^n_{1,2}(t)$.

$$\begin{aligned}Z^n_{1,2}(t) = Z^n_{1,2}(0) + \mu_{2,2} \int_0^t 1_{\{D^n_{1,2}(s) \geq 0\}} (m^n_2 - Z^n_{1,2}(s))\, ds \\ - \mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s) \leq 0\}} Z^n_{1,2}(s) + M^n_Z(t),\end{aligned} \tag{6.2.6}$$

where

$$\begin{aligned}M^n_{Z_{1,2}}(t) &\equiv N^s_{1,2}\left(\mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s) \leq 0\}} Z^n_{1,2}(s)\, ds\right) - \mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s) \leq 0\}} Z^n_{1,2}(s)\, ds, \\ M^n_{Z_{2,2}}(t) &\equiv N^s_{2,2}\left(\mu_{2,2} \int_0^t 1_{\{D^n_{1,2}(s) \geq 0\}} Z^n_{2,2}(s)\, ds\right) - \mu_{2,2} \int_0^t 1_{\{D^n_{1,2}(s) \geq 0\}} Z^n_{2,2}(s)\, ds\end{aligned}$$

and

$$M^n_Z(t) \equiv M^n_{Z_{2,2}}(t) - M^n_{Z_{1,2}}(t). \tag{6.2.7}$$

Let $\widetilde{Z}^n_{1,2}(t)$ have the same representation as $Z^n_{1,2}(t)$, but with the indicator functions $1_{\{D^n_{1,2}(s)\geq 0\}}$ and $1_{\{D^n_{1,2}(s)<0\}}$ replaced by $\pi_{1,2}(x(s))$ and $1-\pi_{1,2}((s))$, respectively. Similarly, $\widetilde{M}^n_Z(t)$ is the same as in (6.2.7), with the indicator functions replaced by the appropriate counterparts.

$$\begin{aligned}\widetilde{Z}^n_{1,2} \equiv \widetilde{Z}^n_{1,2}(0) &+ \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m^n_2 - \widetilde{Z}^n_{1,2}(s))\,ds \\ &- \mu_{1,2}\int_0^t (1-\pi_{1,2}(x(s)))\widetilde{Z}^n_{1,2}(s)\,ds + \widetilde{M}^n_Z(t).\end{aligned} \tag{6.2.8}$$

If Conjecture 6.2.1 below indeed holds, then for all $T>0$, $n^{-1/2}\|Z^n_{1,2} - \widetilde{Z}^n_{1,2}\|_T \Rightarrow 0$ in $\mathcal{D}$ as $n\to\infty$. Assuming Conjecture 6.2.1, we work with $\hat{Z}_{1,2}(t)$ and $\hat{M}^n_Z(t) \equiv \hat{M}^n/\sqrt{n}$, but with the indicator functions replaced by the $\pi_{1,2}(x(s))$ expressions.

From (4.5.13) we see that

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

Upon centering $\hat{Z}^n_{1,2}(t)$ about the fluid limit, and dividing by $\sqrt{n}$ as in (6.1.1),

$$\begin{aligned}\hat{Z}^n_{1,2}(t) &= \hat{Z}^n_{1,2}(0) - \frac{\int_0^t \left[(\mu_{2,2}-\mu_{1,2})\pi_{1,2}(x(s)) + \mu_{1,2}\right]\hat{Z}_{1,2}(s)\,ds}{\sqrt{n}} \\ &\quad + \frac{\mu_{2,2}(m^n_2 - nm_2)\int_0^t \pi_{1,2}(x(s))\,ds}{\sqrt{n}} + \frac{M^n_Z(t)}{\sqrt{n}} \\ &= \hat{Z}^n_{1,2}(0) - \int_0^t \left[(\mu_{2,2}-\mu_{1,2})\pi_{1,2}(x(s)) + \mu_{1,2}\right]\hat{Z}^n_{1,2}(s)\,ds + \hat{M}^n_Z(t),\end{aligned} \tag{6.2.9}$$

where the last inequality follows from the fact that $(m^n_2 - nm_2)/\sqrt{n} \to 0$, as $n\to\infty$.

Now, $\pi_{1,2}(x(s))$ is locally Lipschitz continuous in $\mathbb{A}$ as a function of $x(s)$ by Theorem 2.5.1, and is thus Lipschitz continuous over compact sets. Moreover, $x(s)$ is Lipschitz continuous, as a function of the time argument s by Lemma 4.8.1. It follows that $\pi_{1,2}(x(s))$ is Lipschitz continuous as a function of the time argument s as well. We can thus apply

Lemma 6.2.1 and conclude that the representation in (6.2.9) is a continuous mapping from $\mathcal{D}$ to itself. The convergence to the limit-process $\hat{Z}_{1,2}(t)$ is implied by the next lemma, whose proof is similar to that of Lemma 6.2.2 and is thus omitted.

Lemma 6.2.3. *Under the conditions of Theorem 6.1.1, $\hat{M}_Z^n(t) \Rightarrow B(\gamma_2(t))$ in $\mathcal{D}$ as $n \to \infty$, where $\{B(t) : t \geq 0\}$ is a standard Brownian motion, and $\gamma_2(t)$ is defined in* (6.1.5)*.*

Lemma 6.2.3 completes the proof. ■

Proof of Lemma 6.2.2 Let

$$\begin{aligned} \hat{M}_S^n(t) &= \left(\hat{M}_{1,1}^n(t), \hat{M}_{1,2}^n(t), \hat{M}_{2,2}^n(t)\right) \quad \hat{M}_A^n(t) = \left(\hat{M}_{a_1}(t), \hat{M}_{a_2}(t)\right), \quad \text{and} \\ \hat{M}_u^n(t) &= \left(\hat{M}_{u_1}^n(t), \hat{M}_{u_2}^n(t)\right). \end{aligned}$$

To compress the notation, for $x \in \mathcal{D}_n$ and $t \in [0,\infty)^n$, we define $x(t) \equiv (x_1(t_1), x_2(t_2), \ldots, x_n(t_n))$. We start by proving that

$$\left(\hat{M}_A^n(t), \hat{M}_S^n(t), \hat{M}_u^n(t)\right) \Rightarrow \left(B_A(\lambda t), B_S\left(\mu \int_0^t z(s)\, ds\right), B_u\left(\theta \int_0^t q(s)\, ds\right)\right), \tag{6.2.10}$$

in $\mathcal{D}_7$, as $n \to \infty$. Here $B_A(t)$, $B_S(t)$ and $B_u(t)$ are, respectively, 2-, 3- and 2-dimensional independent Brownian motions. Using our compressed notation we have $\lambda t \equiv (\lambda_1 t, \lambda_2 t)$, $\mu z(s) \equiv (\mu_{1,1} z_{1,1}(s), \mu_{1,2} z_{1,2}(s), \mu_{2,2} z_{2,2}(s))$, $\theta q(s) \equiv (\theta_1 q_1(s), \theta_2 q_2(s))$. For example, $B_A(t) = (B_{A_1}(\lambda_1 t), B_{A_2}(\lambda_2 t))$, and similarly for $B_S(\cdot)$ and $B_u(\cdot)$.

The result of the lemma then follows from the definition of $\hat{M}_s^n(t)$ in (6.2.3), and the continuity of addition under continuous limits, e.g., Corollary 12.7.1 in [78].

For the Poisson processes defined in (4.4.1), let

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

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

$$\left(\tilde{M}^n_A(t), \tilde{M}^n_S(t), \tilde{M}^u(t)\right) \Rightarrow \left(\tilde{B}_A(t), \tilde{B}_S(t), \tilde{B}_u(t)\right), \quad \text{in } \mathcal{D}_7, \text{ as } n \to \infty,$$

where $\tilde{B}_A$, $\tilde{B}_S$ and $\tilde{B}_u$ are, respectively, 2-dimensional, 3-dimensional and 2-dimensional independent Brownian motions. See Theorem 4.2 and §9.1 in [57].

Let

$$\Phi^n_{A_i}(t) \equiv \frac{\lambda^n_i t}{n}, \qquad \Phi^n_{S_{i,j}}(t) \equiv \frac{\mu_{i,j}\int_0^t Z^n_{i,j}(s)\,ds}{n} \quad \text{and}$$
$$\Phi^n_{u_i}(t) \equiv \frac{\theta_i \int_0^t Q^n_i(s)\,ds}{n}, \quad i,j = 1,2.$$

Then, by the condition on the arrival rates, $\Phi^n_{A_i} \Rightarrow \lambda_i t$, $i = 1,2$. From the initial conditions in the statement of Theorem 6.1.1, the fluid limit and the continuity of the integral mapping, it follows that $\Phi^n_{S_{i,j}} \Rightarrow \mu_{i,j}\int_0^t z_{i,j}(s)\,ds$ and $\Phi^n_{u_i} \Rightarrow \theta_i \int_0^t q_i(s)\,ds$, $i,j = 1,2$ in $\mathcal{D}$ as $n \to \infty$.

Let $\Phi^n_A(t)$, $\Phi^n_{S_{i,j}}(t)$ and $\Phi^n_{u_i}(t)$ be the corresponding vector-valued processes. Then

$$\left(\Phi^n_A(t), \Phi^n_{S_{i,j}}(t), \Phi^n_{u_i}(t)\right) \Rightarrow \left(\lambda t,\, \mu \int_0^t z(s)\,ds,\, \theta \int_0^t q(s)\,ds\right),$$

in $\mathcal{D}_7$, as $n \to \infty$. By definition,

$$\left(\hat{M}_A^n(t), \hat{M}_S^n(t), \hat{M}_u^n(t)\right) = \left(\tilde{M}_A^n\left(\Phi_A^n(t)\right), \tilde{M}^n\left(\Phi_S^n(t)\right), \tilde{M}_u^n\left(\Phi_u^n(t)\right)\right),$$

and the result follows from the continuity of the composition mapping at continuous limits, Theorem 13.2.1 in [78]. ∎

As we stated above and as was made clear by the proof of Theorem 6.1.1, the convergence of the processes in (6.1.1) to the diffusion limits depend on the following conjecture, which we intend to prove in the future.

Conjecture 6.2.1. *Consider* $\hat{Z}_{1,2}^n$ *in* (6.1.1) *and* $\widetilde{Z}_{1,2}^n$ *in* (6.2.8). *Then* $\|\hat{Z}_{1,2}^n - \widetilde{Z}_{1,2}^n\| \Rightarrow 0$ *in* $\mathcal{D}$ *as* $n \to \infty$.

Bibliography

[1] Abate, J. and Whitt, W. 1988. Simple spectral representations for the $M/M/1$ queue. *Queueing Systems* **3**, 321–346.

[2] Aksin, Z., Armony, M. and Mehrotra, V. 2007. The modern call center: a multi-disciplinary perspective on operations management research. *Production Oper. Management* **16** (6) 665–688.

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

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

[5] Armony, M. and Gurvich, I. 2006. When promotions meet operations: cross-selling and its effect on call-center performance. *Manufacturing Service Oper. Management*, forthcoming.

[6] Asmussen, S. 1998. Extreme value Theory for Queues Via Cycle Maxima. *Extremes* 1:2, 137–168.

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

[8] Bassamboo, A., J. M. Harrison, A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. *Queueing Systems* **51** (3-4) 249–285.

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

[10] Bassamboo, A. and Zeevi, A. 2009. Near optimal data-driven staffing for large call centers. *Oper. Res.*, forthcoming.

[11] Bell, S. L. and Williams, R. J. 2005. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a thrshold policy. *Electronic J. Prob.* **10** 1044–1115.

[12] Bhandari, A., Scheller-Wolf, A. and Harchol-Balter, M. 2008. An exact and efficient algorithm for the constrained dynamic operatyor staffing problem for call centers. *Management Sci.* **54** (2) 339–353.

[13] Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. Wiley, New York.

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

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

[16] Bramson, M. 1998. State space collapse with applications to heavy traffic limits to multiclass queueing networks. *Queueing Systems* **30** 89–148.

[17] Brown, L. Gans, N., Mandelbaum A., Sakov, A., shen, H., Zeltyn, S. and Zhao, L. 2005. Statistical analysis of a call center: A queueing-science perspective. *J. Amer. Statist. Assoc.*, **100**, 36–50.

[18] Coddington, E. A. and Levinson, N. 1955. *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.

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

[20] Cohen, J. W. 1982. *The Single Server Queue,* second ed., North-Holland, Amsterdam.

[21] Dai, J.G. and Tezcan, T. 2009. State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.* Forthcoming.

[22] Dai, J.G., He, S. and Tezcan, T. 2009. Many-Server diffusion limits for $G/Ph/n+GI$ queues. *Preprint*.

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

[24] Eick, S. G., Massey, W. A. and Whitt, W. 1993. The physics of The $M_t/G/\infty$ queue. *Oper. Res.* **41** (4) 731–742.

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

[26] Gans, N., Koole, G. and Mandelbaum, A. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** (2) 79–141.

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

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

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

[30] Gurvich, I. and Whitt, W. 2009b. Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems. *Manufacturing Service Oper. Management*, **11** (2) 237–253.

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

[32] Gurvich, I. and Whitt, W. 2010b. Asymptotic optimality of queue-ratio routing for many-server service systems. working paper.

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

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

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

[36] Iglehart, D.L. 1968. Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* **2**, 429–441.

[37] Iglehart, D.L. and Whitt, W. 1970. Multiple channel queues in heavy traffic, I. *Adv. Appl. Prob.* **2** 150–177.

[38] Iglehart, D.L. and Whitt, W. 1970. Multiple channel queues in heavy traffic, II. sequences, networks and batches. *Adv. Appl. Prob.* **2** 355–369.

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

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

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

[42] Karlin, S. and Taylor, H. M. 1981. *A Second Course in Stochastic Processes.* Academic Press, New York.

[43] Karr, A.F. 1975. Weak Convergence of a Sequence of Markov Chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **33** 41–48.

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

[45] Khalil K. Hassan, 2002. *Nonlinear Systems.* Third Edition. Prentice Hall, New Jersey.

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

[47] Kingman, J.F.C. 1961. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* **57** 902–904.

[48] Kingman, J.F.C. 1962. On queues in heavy traffic. *J. of the Royal Statustucal Society* **24** 383–392.

[49] Kurtz, T.G., 1992. Averaging for Martingale Problems and Stochastic Approximations. *Applied Stochastic Analysis, Proc.* US-French Workshop, Lecture Notes in Control and Information Sciences Vol. 177 Springer, Berlin, 186–209.

[50] Marquez J. Horacio, 2003. *Nonlinear Control Systems.* Wiley, New Jersey.

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

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

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

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

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

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

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

[58] Pang, G. and Whitt, W. 2009. Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems*, **61**, 167–202. Available at: http://www.columbia.edu/∼ww2040/recent.html

[59] Perry, O. and Whitt, W. 2009. A fluid approximation for service systems responding to unexpected overloads. Submitted to *Oper. Res.*

[60] Prohorov, Y. V. 1956. Convergence of random processes and limit theorems in probability. Theor. Probability Appl. **1**, 157–214

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

[62] Rockafellar, T. R. 1970. *Convex Analysis*, Princeton University Press, Princeton, N.J.

[63] Rootzén, H. 1988. Maxima and Exceedances of Stationary Markov Chains. *Advances in Applied Probability.* **20** (2), 371–390.

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

[65] Smith, D.R. and Whitt, W. 1981. Resource sharing of efficiency in traffic systems. *Bell Systems Technical Journal*, **60** (13), 39-55.

[66] Stone, C. 1963. Limit theorems for random walks, birth and death processes and diffusion processes. *Illinois J. Math.* **4** 638–660.

[67] Talreja, R. and Whitt, W. 2008. Heavy-traffic limits for waiting times in many-server queues with abandonments. Available at: http://www.columbia.edu/∼ww2040/recent.html

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

[69] Tezcan, T. and Dai, J. G. 2006. Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Georgia Institute of Technology.

[70] Uchitelle, L. 2002. "Answering '800' Calls, extra income but no security", *The New York Times*, March 27, Section A, pg. 1, Column 5.

[71] Ward, A. R. and Glynn P. W. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43** 103–128.

[72] Whitt, W. 1971. Weak Convergence Theorems for Priority Queues: Preemptive-Resume Discipline. *Journal of Applied Probability*, **8**, No. 1, 74–94.

[73] Whitt, W. 1980. Continuity of generalized semi-Markov processes. *Math. Oper. Res.*, **5**, No. 4, pp. 494-501.

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

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

[76] Whitt, W. 1989. Planning queueing simulations. *Management Sci.* **35** (11) 1341–1366.

[77] Whitt, W. 1991. The Pointwise Stationary Approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci*, vol. 37, No. 3, 1991, pp. 307-314

[78] Whitt, W. 2002. *Stochastic-Process Limits*, New York: Springer.

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

[80] Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Science* 51 (2) 221–235.

[81] Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Operations Research* 54 (1) 37–54.

[82] Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* **15** (1) 88–102.

[83] Whitt, W. 2007. Proofs of the martingale FCLT. *Probability Surveys.* **4**, 268–302.

[84] Williams, R. J. 1998. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* **30**(1 - 2) 27–88.

[85] Zeltyn S., 2004 *Call centers with impatient customers: exact analysis and many-server asymptotics of the* $M/M/N + G$ *queue.* PhD thesis. Available at: http://iew3.technion.ac.il/serveng/References/MMNG_thesis.pdf