State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems and Applications

A Thesis
Presented to
The Academic Faculty

by

Tolga Tezcan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
August 2006

State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems and Applications

Approved by:

J.G. Dai
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, Committee Chair

Amy Ward
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Advisor

Mor Armony
Leonard N. Stern School of Business
New York University

Ron Billings
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Anton Kleywegt H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: June 30, 2006

ACKNOWLEDGEMENTS

This thesis was a product of many years of work that has been influenced by many who have been a part of my academic life during my undergraduate studies at Bilkent University, and then during my graduate studies at Colorado State University-Pueblo and at Georgia Tech. I am grateful to all those teachers who have helped me to gain skills I now possess.

Dr. Jim Dai deserves the most credit for making this quest a success. He provided me not only with academic advisement and guidance in my research and classes but has also been a true mentor, friend and inspiration. He has played an important role in my personal and academic development for the last four years. He is always going to be the reference point as to what I should achieve as an academician and an advisor. I am grateful for the opportunity to work with him and for all the time he devoted towards my development. Each meeting with him was a great joy and source of excitement.

Dr. Amy Ward has also served as a co-advisor for almost a year. I would like to thank her for guiding me while exploring new research directions. It was a pleasure to have her as a co–advisor.

I would like to thank Dr. Ron Billings, Dr. Anton Kleywegt, and Dr. Mor Armony for serving on my dissertation committee. I also thank Dr. Leon McGinnis for his guidance and for providing the financial support during the first two years of my PhD studies. It was him who made it possible for me to realize my dream at Georgia Tech.

I thank Dr. Abhijit Gosavi, my master's thesis advisor, for encouraging me to take on stochastic analysis as my research direction and for his guidance in every stage of my graduate studies. I acknowledge Dr. Huseyin Sarper for his guidance and help during my stay at Colorado State University-Pueblo. I thank Dr. İhsan Sabuncuoğlu for his patience with me when I was still a novice researcher as an undergraduate.

I would like to thank the following professors who have influenced my academic formation and helped me during my Ph.D. studies at Georgia Tech; Dr. Hayriye Ayhan, Dr. Spyros

Reveliotis, and Dr. Anton Kleywegt. I also thank Dr. Gary Parker for helping me in every stage of my Ph.D. and for his availability whenever I needed help or advice.

I thank Dr. Janet Barnett for being an excellent teacher whose influence can still be seen in my proofs. She always will be my role model as a teacher. I also thank Dr. Christian Houdre for making probability a challenging but a fun course. Their enthusiasm and dedication to teach have helped me learn basics of analysis in great depth.

I would like to thank Dr. Costis Maglaras, Dr. Avi Mandelbaum, Dr. Assaf Zeevi and Dr. Mor Armony for providing feedback on earlier versions of the results presented in this theses.

I would like to thank Jim Luedtke, Yetkin Ileri, Rahul Rauniyar, and Deniz Dogan for their friendship, help and supportive presence during hard times. Jim and Yetkin have also helped edit some parts of this thesis.

I would like to thank my parents for all the sacrifices they have made in the past for me to realize my dream. They are the ones who taught me the value and joy of analytical thinking.

Finally, my wife Yıldız deserves as much credit as I do for completing this work. I am grateful for her being a part of my life and being by my side during hard times. I really appreciate her attempts to adjust her life around my always busy schedule. Her unconditional love, constant support and encouragement made this work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

We consider a class of queueing systems that consist of server pools in parallel and multiple customer classes. Customer service times are assumed to be exponentially distributed. We study the asymptotic behavior of these queueing systems in a heavy traffic regime that is known as the Halfin and Whitt many-server asymptotic regime. Halfin and Whitt [32] consider an $M/M/N$ system and fix the steady state probability that all servers are busy so that the probability an arriving customer must wait for service exceeds 0 and is strictly greater than one, while letting the arrival rate and the number of servers grow to infinity.

Our main contribution is a general framework for establishing state space collapse results in the Halfin and Whitt many-server asymptotic regime for parallel server systems having multiple customer classes. In our work, state space collapse refers to a decrease in the dimension of the processes tracking the number of customers in each class waiting for service and the number of customers in each class being served by various server pools. We define and introduce a "state space collapse" function, which governs the exact details of the state space collapse. Our notion of state space collapse contrasts with that in Harrison and Van Mieghem [36], which establishes a deterministic relationship between a lower-dimensional workload process and the queue length processes. Our methodology is similar in spirit to that in Bramson [13]; however, Bramson studies an asymptotic regime in which the number of servers is fixed and Bramson does not require a "state space collapse" function.

We illustrate the applications of our results in three different parallel server systems. The first system is a distributed parallel server system under the minimum-expected-delay faster-server-first (MED-FSF) or minimum-expected-delay load-balancing (MED-LB) policies. We prove that the MED-FSF policy minimizes the stationary distribution of total number of customers in the system. However, under the MED-FSF policy all the servers in our distributed system except those with the lowest service rate experience 100% utilization

but under the MED-LB policy, on the other hand, the utilizations of all the server pools are equal. We also show that under both policies the system performs as well as a corresponding single queue system. The second system we consider is known as the N-model. We show that when the service times only depend on the server pool providing service a static priority rule is asymptotically optimal. The optimality is in terms of stochastically minimizing linear holding costs over any finite time interval. Finally, we study two results conjectured in the literature for V-systems. First, we prove a state space collapse result conjectured in Armony and Maglaras [3]. Then, we propose a policy whose asymptotic performance is arbitrarily close to the conjectured performance of the policy proposed by Milner and Olsen [53] and prove a state space collapse result under this policy. We show for all of these systems that the conditions on the hydrodynamic limits can easily be checked using the standard tools that have been developed in the literature to analyze fluid models.

CHAPTER I

INTRODUCTION

Multi-class queueing networks have been extensively used to model queueing systems arising in manufacturing and service industries [34]. A special class of these networks, parallel server systems that are commonly used to model service systems with many servers; see [28], [47], [48], [49], and [57] for different applications, are of current interest. In a parallel server system, customers are handled by a set of server pools and leave the system after service. We also restrict our attention to systems with exponential service times. Similar to multiclass queueing networks, exact analysis of parallel server systems is limited. Even when available, results from classical queueing theory provide limited insight on the general properties of the performances of these systems and rarely can be used for optimization purposes.

An alternative tool for analyzing multiclass queueing networks is diffusion approximations. For heavily loaded networks, conventional heavy traffic approximations have been shown to provide accurate approximations for the system performance; see [15] and [65]. Under a conventional heavy traffic analysis almost all the customers are delayed in queue before their service starts. This is not the case in many systems that are especially seen in the service sector. For instance, a typical call center consists of many agents catering to a high volume of customers. A common performance target for call center managers is to have only a small percentage of customers delayed in queue before their service starts while keeping server utilizations high. It is empirically observed that this can be achieved when the number of servers in the system is large; see Section 4.1.2 in Gans et. al [28]. These systems are said to be operating under the quality and efficiency driven (QED) regime [28]. Starting with the seminal paper by Halfin and Whitt [32], these systems are analyzed under the Halfin and Whitt many-server limiting regime, which is also known as the QED regime since high utilization of the servers is achieved together with a high quality of service. It

has been shown that the Halfin and Whitt regime is the right asymptotic regime to study the service systems with many servers that are QED, that is, this regime captures the gain from economies of scale which cannot be captured by conventional heavy traffic analysis; see Garnett et al. [29] and Gans et al. [28]. Staffing rules such as the square-root safety rule has been established using these analyses which have not been observed from the classical queueing theory results such as the Erlang-C formula.

General procedures to establish the validity of conventional heavy traffic and many-server diffusion limits are different. To establish a conventional heavy traffic limit a sequence of systems having mean service times and inter-arrival times that become close for one or more stations in the network is considered. The number of servers are taken to be fixed for each system. Then, a heavy traffic limit theorem is established by studying the convergence of the stochastic processes associated with each system. On the other hand, a many-server diffusion limit is established by considering a sequence of systems with all servers having fixed service rates. The number of servers and the arrival rates grow to infinity in a way that the traffic intensity of one or more server pools converge to one.

In two companion papers, Bramson [13] and Williams [69], sufficient conditions are given under which a conventional heavy traffic limit theorem holds for a general class of multiclass queueing networks. The framework in Bramson [13] is also of independent interest; it enables one to show a state collapse (SSC) result by checking whether a similar SSC result holds for the hydrodynamic limits. His framework has been used to show SSC results in conventional heavy traffic diffusion limits for different multiclass queueing networks; see Bramson and Dai [14], Stolyar [61], Mandelbaum and Stolyar [52], Dai and Lin [21]. In these papers, the SSC results enabled the authors to establish the diffusion limits of the systems under consideration.

A general theory to prove many-server limit theorems have not been established yet. Due to the differences in the procedures to obtain diffusion limits under the conventional heavy traffic and many-server asymptotic regimes, the results of Williams and Bramson cannot be readily extended to the many-server asymptotic analysis. In this study, we

extend the framework in Bramson [13] to show that multiplicative SSC results in many-server diffusion limits can be established by checking that the associated hydrodynamic limits satisfy certain conditions. Hydrodynamic limits are generalizations of fluid limits and their structure is determined using a functional weak law of large numbers. Our framework presents a relatively easy and a general method to establish SSC results in many-server diffusion limits, as Bramson's framework does in conventional heavy traffic diffusion limits.

Before we summarize our framework, we give the details of the Halfin and Whitt limiting regime. In [32], they consider a sequence of $GI/M/n$ systems. Let λ_n denote the arrival rate to the system with n servers, μ be the service rate of each server and $\rho_n = \lambda_n/(n\mu)$ denote the traffic intensity. They show that the steady state probability that there is at least one customer in the queue waiting for service converges to a limit $0 < \alpha < 1$ if and only if $\sqrt{n}(1-\rho_n) \to \theta$, for some $\theta > 0$, as $n \to \infty$. Let $X^n(t)$ denote the total number of customers, including those in service, in the nth system at time t. They show that the properly scaled number of customers in the system process, $\hat{X}^n(\cdot) = (X^n(\cdot)-n)/\sqrt{n}$, converges to a diffusion process if $\hat{X}^n(0)$ converges weakly. They also show that the steady state distribution of the queue length converges to the steady state distribution of the limiting diffusion process. This fact makes the use of diffusion limits attractive to approximate the steady state performance measures of these systems.

The current literature focuses on generalizing the analysis in Halfin and Whitt to other parallel server systems. However, this line of research is still in its early stages and only for a few systems with special topological properties similar diffusion limits have been established. Recent papers; see, for example, Armony [1], Armony and Maglaras [3, 2], Gurvich et al. [31], Maglaras and Zeevi [47, 48, 49], establish diffusion limits of certain classes of parallel server systems. As stated in Section 4 of Gans et al. [28], many-server asymptotic analysis is one of the most promising research directions in the analysis of parallel server systems with many servers, more specifically for the analysis and optimal control of call centers.

In this study, we begin with presenting a unifying framework for the analysis of parallel server systems. First, we determine the queueing equations that must be satisfied by a parallel server system. Then, we formulate a static planning problem which is similar to

that in Harrison [35]. Using this formulation we characterize a general many-server heavy traffic condition. The solution of this problem reveals the long-run proportion of servers that must be allocated to each customer class. Also, we obtain conditions under which the queue lengths will not grow without bound and restrict our attention to systems that satisfy these conditions. This is achieved by considering control policies that achieve the optimal long-run allocation proportions that is necessary for stability. We present a general fluid model framework that can be used to check if a control policy achieves these long-run allocations. It will be shown that for a policy that satisfy these conditions it is enough to check that a set of deterministic equations satisfy certain conditions in order to prove a SSC result.

Once we formulate the many-server heavy traffic condition we focus on establishing a framework to prove SSC results. As discussed above, SSC results play an important role in establishing conventional heavy traffic approximations as shown by Williams [69] and it is apparent from the current literature; see [1], [2], [3], and [31], that they are also crucial for establishing many-server diffusion limits.

The SSC framework established in Bramson hinges upon the observation that by slowing down the clock of the diffusion scaled processes, achieved by working with the hydrodynamic scaling, the events that happen instantaneously in the diffusion limits can be observed in more detail. The motivation behind using the hydrodynamic limits to study an SSC result in the current setting is based on this observation. We use the hydrodynamic scaling which is obtained by slowing down the time in the diffusion scaling used by Halfin and Whitt. Using this scaling we obtain the structure of the hydrodynamic limits. By utilizing the connection between the hydrodynamic limits and the diffusion limits, we show that certain conditions on hydrodynamic limits imply an SSC result in the diffusion limit.

In short, we show that in order for a state space collapse result to hold in the diffusion limit it must hold eventually for the hydrodynamic limits. The general structure and definition of hydrodynamic limits are complicated. It is not clear how one can check the required condition on hydrodynamic limits by directly using the definition. We overcome this hurdle by showing that the hydrodynamic limits of a general parallel server system must satisfy

a set of deterministic equations which we call the hydrodynamic model equations. These equations possess some of the nice properties of the fluid model equations but they are different. We illustrate how fluid model tools can be used to show that the SSC results for the hydrodynamic limits of three systems, to be explained below, hold.

Our results differ from Bramson's in several aspects. As described above, we focus on systems under the Halfin–Whitt many-server regime whereas Bramson focused on systems under conventional heavy-traffic. The main technical difficulty in extending Bramson's result to the many-server asymptotic regime stems from the number of servers going to infinity because it is assumed to be fixed at a finite value in the conventional heavy traffic analysis. Also, the class of SSC results we consider here is different than those considered by Bramson. Bramson [13] focused on establishing a relationship between the workload and queue length processes of a system. Although this relationship plays an important role in the conventional heavy traffic analysis, studying the workload process does not seem to help in the many-server asymptotic analysis. Loosely speaking, we use the term "state space collapse" to refer to a decrease in the dimension in the limit of the queueing processes associated with the system studied. To mathematically characterize such a result we introduce the notion of SSC functions which has not been used in the literature. Therefore, some of the results presented here does do not have corresponding counterparts in Bramson's framework. In addition, the hydrodynamic limits established in this study for the many-server setting are new.

We present the applications of our main result in three different parallel server systems. The first system we study is a distributed parallel server (DPS) system. A DPS system consists of a single customer class and multiple server pools. Each customer must be routed to a server pool or a queue at his arrival time following a routing policy. We focus on two control policies; the minimum-expected-delay–faster-server-first (MED–FSF) policy and the minimum-expected-delay–load-balancing (MED–LB) policy. Under both policies, if all servers are busy when a customer arrives at the system, the customer is routed to the queue that has the minimum expected delay. If there is an idle server at his arrival time, then under the MED–FSF policy the customer is routed to the fastest available pool and

under the MED–LB to the least utilized available pool. We show that the MED-FSF policy achieves complete resource pooling in the diffusion limit, that is, the many-server diffusion limit and its stationary distribution of a DPS system under the MED-FSF policy is equal in distribution to the corresponding system that has a single queue where all customers wait for service. However, under the MED–FSF policy all the servers in our DSP system except those with the lowest service rate experience 100% utilization. A common goal in call center management is to have all the servers to be utilized "fairly". The MED–LB policy achieves this objective. We show that the MED–LB policy asymptotically balances the load of the servers.

The second system we study is an N-system. An N-model consists of two customer classes and two server pools. The servers in the first pool can only serve class 1 customers whereas the second pool can serve either class. We assume that the second server's service rate is the same for both classes. We show that a static priority rule is asymptotically optimal for such N-models with many servers. The optimality is in terms of stochastically minimizing linear holding costs.

Finally, we illustrate the strength of our framework by solving two open problems from the literature. Both of these results are for a a parallel server system with a V-design, V-system for short. A V-system consists of a single server pool and multiple customer classes. We will focus on the case with two customer classes and assume that service rates of the customer classes are equal. The first policy we consider is a threshold policy proposed in Armony and Maglaras [3]. We prove that under their policy, the SSC result they conjecture holds. The second policy we consider is a more complicated threshold policy proposed by Milner and Olsen [53]. We first discuss why the SSC result they conjectured cannot be shown in standard function spaces used for the asymptotic analysis of queueing systems. Then, we propose a policy whose asymptotic performance is arbitrarily close to the conjectured performance of the policy proposed by Milner and Olsen. We prove a SSC result that is similar to that conjectured in Milner and Olsen under this policy.

The rest of this proposal is organized as follows. In the following section we review the related literature. We introduce the notation used in this thesis in Section 1.2. In Chapter 2

we introduce the parallel server systems and present the set of queueing equations that must be satisfied by these systems. In Chapter 3 we define the hydrodynamic limits and present our main results. Applications of our main results are presented in Chapter 4. The reader who wants to see the applications before the abstract theory can skip Chapter 3 and read Chapter 4 first. We conclude with some remarks and future research directions in Chapter 5.

1.1 Previous Work

In this chapter we review the literature. Standard references on conventional heavy traffic analysis include Harrison [34], Chen and Yao [15], and Whitt [65]. Results from classical queueing theory for the analysis of parallel server systems can be found on several textbooks; see, for example, Ross [59] and Gross and Harris [30].

Early many-server diffusion approximations had appeared in Borovkov [12], Iglehart [42], and Whitt [64], with the limiting traffic intensity of the system converging to a values less than one, before Halfin and Whitt [32] studied the regime explained above. We restrict the rest of our review to the literature on the Halfin-Whitt many-server asymptotic analysis.

The analysis of Halfin and Whitt has been extended in several directions. Garnett et al. [29] studied the asymptotic analysis of an $M/M/n$ system with impatient customers and they have established similar results to those in Halfin and Whitt. Puhalski and Reiman [56] established the diffusion limit of a G/PH/n system, where PH stands for a phase type service time distribution. They have also established the many-server diffusion limits of a V-model parallel server system under a static priority policy. To the best of our knowledge, the first SSC result in the Halfin and Whitt regime appeared in Puhalski and Reiman [56]. Whitt [67] studies the many-server diffusion limit of a $G/H_2^*/n/m$ system, where H_2^* indicates that the service time distribution is an extremal distribution among the class of hyperexponential distributions. He later uses this analysis in [66] to approximate $G/GI/n/m$ systems.

Armony and Maglaras studied an $M/M/n$ system with two customer classes in [3, 2]. The hydrodynamic scaling we introduce in this study is similar to the scaling that is used in that paper. This scaling was also used in previous studies by Maglaras [50] and Fleming et al. [25] in different settings. Yet, the hydrodynamic scaling in the many-server

setting arises naturally from that in the conventional heavy traffic setting that is introduced by Bramson [13].

Gurvich et al. [31] studies a V-parallel server system with impatient customers. They show that a static buffer priority policy with a threshold policy is asymptotically optimal. Armony [1] studies an inverted-V-parallel server system and shows that the faster-server-first (FSF) policy is asymptotically optimal. The SSC results established in Gurvich et al. [31] and Armony [1] can be easily proved by using our main results.

Another approach to construct effective control policies is to formulate a Brownian control problem by mimicking the analysis in the conventional heavy traffic; see Harrison [33] for an introduction and Harrison [40] and Harrison and Williams [37] for current research in this field.

Harrison and Zeevi [38] and Atar et al. [7] study a V-parallel server system with impatient customers in the QED regime and find asymptotically optimal control policies. Atar [6, 5] follows a similar approach to that in [7] to find asymptotically optimal policies for tree-like systems. In a related recent paper by Atar et al. [8] discusses the null controllability of parallel server systems.

Although an analog of the framework in Williams [69] and Bramson [13] has not been built for the many-server regime, diffusion limits of some general Markovian queueing systems have been established. Mandelbaum et al. [51] and Pats [54] provided the diffusion limits for Markovian queueing network systems with exponential interarrival and service times in different settings. However, our approach differs from theirs in several aspects. We base our results on a multiclass queueing network setting and consider the control policies to be a part of the queueing system formulation. They do not consider control policies explicitly but assume that it can be incorporated in their "Markovian" service network formulation. Since the control policies are not explicitly studied they do not explicitly define the class of control policies that satisfy this assumption, but they illustrate the applications of their results in several multiclass parallel server systems. Not considering the control policies explicitly allows them to consider service and arrival processes whose rates can be time and state dependent.

1.2 Notation

The set of non-negative integers is denoted by $\mathbb{N}$. For an integer $d \geq 1$, the d-dimensional Euclidean space is denoted by $\mathbb{R}^d$ and $\mathbb{R}_+$ denotes $[0, \infty)$. Let $|x|$ denote the max norm on $\mathbb{R}^d$ given by $|x| = \max_{i=1,2...,d}\{|x_i|\}$. We also use $|x|$ to denote the max norm on $\mathbb{R}^{d_1 \times d_2}$, for $d_1 > 0$ and $d_2 > 0$. We use $\{x_n\}$ to denote a sequence whose nth term is x_n. For a function $f : \mathbb{R} \to \mathbb{R}$, we say that t is a regular point of f if f is differentiable at t and use $\dot{f}(t)$ to denote its derivative at t.

For each positive integer d, $\mathbb{D}^d[0, \infty)$ denotes the d-dimensional Skorohod path space; see [24]. For $x, y \in \mathbb{D}^d[0, \infty)$ and $T > 0$ we set

$$\|x(t) - y(t)\|_T = \sup_{0 \leq t \leq T} |x(t) - y(t)|.$$

The space $\mathbb{D}^d[0, \infty)$ is endowed with the J_1 topology and the weak convergence in this space is considered with respect to this topology. For a sequence of functions $\{x_n\} \in \mathbb{D}^d[0, \infty)$, the sequence is said to converge uniformly on compact sets to $x \in \mathbb{D}^d[0, \infty)$ as $n \to \infty$, denoted by $x_n \to x$ u.o.c., if for each $T > 0$

$$\|x_n(t) - x(t)\|_T \to 0 \text{ as } n \to \infty.$$

The term FSLLN stands for functional strong law of large numbers and FCLT stands for functional central limit theorem; see [15] for details.

We assume that all the random variables are defined in the same probability space $(\Omega, \mathcal{F}, P)$. A stochastic process can be viewed as a function from $\Omega \times [0, \infty)$ to $\mathbb{R}$. In several occasions, we will need to analyze the sample paths of stochastic processes. In those cases, we will explicitly express the dependence by writing $X(\cdot, \omega)$ for the sample path associated with $\omega \in \Omega$ of a stochastic process X. If the sample paths of a subset of Ω is analyzed, we omit ω from the notation.

CHAPTER II

PARALLEL SERVER SYSTEMS

We consider a system with parallel server pools and several customer classes. A server pool consists of several servers whose service capacities and service capabilities are the same. Customers arrive to the system exogenously and upon arrival they are routed to one of the buffers (or queues). Two customers that are routed to the same buffer are said to be in the same class. Each customer is only served by one of servers. Once the service of a customer is completed by one of the servers, he leaves the system. We refer to these systems as *parallel server systems*.

We use I to denote the number of arrival streams, J to denote the number of server pools, and K to denote the number of customer classes. For notational convenience, we define $\mathcal{I} = \{1, \ldots, I\}$ the set of arrival streams, $\mathcal{J} = \{1, \ldots, J\}$ the set of server pools, and $\mathcal{K} = \{1, \ldots, K\}$ the set of customer classes.

We denote the arrival rate for the ith stream by λ_i. The customers arriving in the ith stream are called type i customers and customers that are routed to buffer k are called the class k customers. We assume that the set of pools that can handle class k customers is fixed and denote it by $\mathcal{J}(k)$. Similarly, we assume that the set of queues that servers in pool j can handle is fixed and denote it by $\mathcal{K}(j)$.

Once a customer joins a queue he cannot swap to other queues nor can he renege. After the customer is routed to a queue, say queue k, he proceeds directly to service if there is an available server in one of the pools in $\mathcal{J}(k)$. We assume that the service time of a class k customer by a server in pool j is exponentially distributed with rate $\mu_{jk} > 0$ for all $k \in \mathcal{J}(k)$. We denote the number of servers in pool j by N_j, for $j \in \mathcal{J}$ and we set $N = (N_1, N_2, \ldots, N_J)$. We denote the total number of servers in the system by $|N|$.

In order to operate a multiclass parallel server system control policies must also be given. A control policy must specify a routing policy that can be used to route an arriving

customer to one of the buffers and a scheduling policy that can be used to dispatch a server to serve a customer. Such dispatching is needed in two circumstances. First, whenever a server completes the service of a customer and there exist multiple customers in different classes that the server can handle. Second, whenever a customer arrives and there exist one or more idle servers who can handle that customer class. One can imagine a countless number of policies to serve these purposes. We restrict our attention to control policies that are head-of-the-line and non-idling. A scheduling policy is said to be non-idling if a server never idles when there is a customer waiting in one of the queues that can be served by that server and head-of-the-line (HL) if each server can only serve one customer at any given time and the customers in the same queue are served on a first-in-first-out (FIFO) basis. We assume that the control policy is *non-preemptive*; once the service of a customer starts it can not be interrupted before it is finished. We do not make any assumptions about the routing policy. We call a control policy non-idling and HL if the associated scheduling policy is non-idling and HL.

We also focus on Markovian policies that use information on the queue-length and number of customers in service to allocate servers to customer classes at the time of an arrival to or a departure from the system. We define a strictly increasing sequence $\{\sigma_l\}_{l=0}^{\infty}$ that specifies the successive times at which an arrival occurs to, or a departure occurs from, some class in the network. These time points naturally depend on the policy and can be constructed as described below. We assume that a policy takes action only when the state of the system is changed via an arrival or a departure. Therefore, the server allocations remain constant between $[\sigma_n, \sigma_{n+1})$ for $n \geq 1$. The new allocations for the next interval $[\sigma_{n+1}, \sigma_{n+2})$ are assigned based on the state of the system during the previous interval $[\sigma_n, \sigma_{n+1})$ and the event happened at time σ_{n+2}. Let e_i denote the I dimensional ith unit vector and E_{jk} denote the $J \times K$ matrix with all of its entries equal to 0 except its jkth entry equal to 1. Also e_0 and E_0 denote the zero matrices, i.e., all of their entries are equal to zero, with dimensions I and $J \times K$, respectively.

Let $Q(t) = (Q_k(t); k \in \mathcal{K})$, where $Q_k(t)$ denotes the number of class i customers in queue at time t and $Z(t) = (Z_{jk}(t); i \in \mathcal{I}, j \in \mathcal{J}(i))$, where $Z_{jk}(t)$ denotes the number

of class k customers being served by a server in server pool j at time t. To specify the allocation scheme we assume that with each policy π there exists a measurable function $f_\pi : \mathbb{N}^I \times \mathbb{N}^{J\times K} \times \mathbb{N}^I \times \mathbb{N}^{J\times K} \to \mathbb{N}^I \times \mathbb{N}^{J\times K}$ such that

$$f_\pi\left(Q(\sigma_n), Z(\sigma_n), e, E\right) = \left(Q(\sigma_{n+1}), Z(\sigma_{n+1})\right) \tag{2.1}$$

gives the new allocations, where $e = e_i, E = E_0$ if the event at time σ_{n+1} is an arrival to class i, $e = e_0, E = E_{jk}$ if it is a class k departure from server pool j. We call f_π the transition function for policy π and we say that a policy is *admissible* if it is non-idling, HL, non-preemptive, and has the Markovian structure described above.

A few issues must be addressed in our policy description. First, more than one event may occur at the same time. In that case we order the occurrence times of the events arbitrarily and let policy π make successive allocations at a time point. Also, a policy must satisfy some physical constraints, e.g.; it cannot allocate more servers from a server pool than the number of servers available. These constraints are formulated in the next section within the system equations.

2.1 The dynamics of parallel server systems

In this section we describe the dynamics of a parallel server system. Actually, we will describe in detail the dynamics of a "perturbed" system. The perturbed system is closely related to the parallel server system, and it allows us to write down queueing network equations that are similar to the ones in the standard multiclass queueing networks. Specifically, the number of service completions in the perturbed system is more amenable to analysis than it is in the original system. The equivalence of these two systems, under the exponential service time assumption, will be discussed at the end of this section. Note that a control policy for routing and server scheduling is needed to operate the perturbed system. Like the parallel server system, we assume that each control policy for the perturbed system is non-preemptive, HL, and non-idling. We denote a generic non-idling and HL control policy by π.

The perturbed system is identical to the parallel server system except that its service mechanism is modified as follows. At any given time, when $n \geq 1$ servers in pool j serve

n class k customers, the n servers work on a single class k customer, where as in the original system each customer in service receives service from a server. The remaining $n-1$ customers are said to be *locked for service*; they do not receive any service even though they have left queue k. The single customer in service, called the *active customer*, can be chosen arbitrarily among the n customers. We assume that the service efforts from the n servers are *additive* in that service of the active customer is completed when the total time spent by all servers on the customer reaches the service requirement of the customer. When the service of the active customer is completed, the customer departs the system, and one of the servers working on that customer is freed. At this point, the remaining $n-1$ servers choose a new active customer, and the freed server is either assigned to a class, say k', or stays idle following a non-idling scheduling policy. In the former case, the server locks a class k' customer, with a given service requirement, for service. If there is an active customer that is currently being served by n' servers in pool j that are working on class k' customers, the new server joins the service efforts of these n' servers on the active customer. Otherwise, the locked class k' becomes an active customer, served by the new server.

The object of study in this paper is a stochastic process $\mathbb{X} = (A, A_q, A_s, Q, Z, T, Y, B, D)$, where $\mathbb{X}$ is defined via the perturbed system and each of its components is explained in the next few paragraphs. The notation used in this section is inspired by that used in Puhalskii and Reiman [56] and Armony [1].

The first component is $A = (A_i : i \in \mathcal{I})$, where $A_i(t)$ denotes the total number of arrivals by time t for type i customers. We give more details about the structure of the arrival process in the next section. Here, we just mention that it is a delayed renewal process (see Ross [59]). The second component is $A_q = \{A_{ik}; i \in \mathcal{I}, k \in \mathcal{K}\}$, where $A_{ik}(t)$ denotes the total number of type i customers who are routed to queue k at the time of their arrival and who had to wait in the queue prior to receiving service and arrived at the system before time t. The third component is $A_s = (A_{ijk}; i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J}(k))$, where $A_{ijk}(t)$ denotes the total number of type i customers who have been routed to queue k and started their service immediately after their arrival at server pool j before time t. The component B is $(B_{jk} : j \in \mathcal{J}, k \in \mathcal{J}(k))$, where $B_{jk}(t)$ denotes the total number of class k customers who

are delayed in the queue and whose service started in pool j before time t.

The components Z and Q are $(Z_{jk} : j \in \mathcal{J}, k \in \mathcal{K}(j))$ and $Q = (Q_k : k \in \mathcal{K})$, respectively, where we use $Z_{jk}(t)$ to denote the total number of servers in pool j that serve class k customers and $Q_k(t)$ to denote the total number of customers in queue k at time t. The components T, D and Y are $(T_{jk} : j \in \mathcal{I}, k \in \mathcal{J}(j))$, $(D_{jk} : j \in \mathcal{J}, k \in \mathcal{K}(j))$, and $(Y_j : j \in \mathcal{J})$, respectively, where $T_{jk}(t)$ denotes the total time spent serving class k customers by all N_j servers of pool j, $D_{jk}(t)$ denotes the total number of class k customers whose service is completed by a server in pool j by time t, and $Y_j(t)$ denotes the total idle time experienced by servers of pool j up to time t. Note that $T_{jk} \leq N_j t$.

Let $\{S_{jk},\ j = 1, \ldots, I, k = 1, 2, \ldots, J\}$ be a set of independent Poisson processes with each process S_{jk} having rate $\mu_{jk} > 0$. We set $S = (S_{jk})$ and $\mu = (\mu_{jk})$. For the perturbed system, we model the total number of class k customers whose service is completed by servers in pool j via

$$D_{jk}(t) = S_{jk}(T_{jk}(t)), \quad t \geq 0. \tag{2.2}$$

The process $\mathbb{X}$ depends the control policy used in the perturbed system. In order to emphasize the dependence on the control policy π used, we use $\mathbb{X}_\pi$ to denote the process. Clearly, each component of A, A_q, A_s, B, T, D, and Y is a nondecreasing process, and each component of Q and Z is nonnegative. Furthermore, the process $\mathbb{X}_\pi$ satisfies the following

equations for all $t \geq 0$.

$$A_i(t) = \sum_{k \in \mathcal{K}} A_{ik}(t) + \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}(k)} A_{ijk}(t), \text{ for all } i \in \mathcal{I}, \tag{2.3}$$

$$Q_k(t) = Q_k(0) + \sum_{i \in \mathcal{I}} A_{ik}(t) - \sum_{j \in \mathcal{J}(k)} B_{jk}(t), \text{ for all } k \in \mathcal{K}, \tag{2.4}$$

$$Z_{jk}(t) = Z_{jk}(0) + \sum_{i \in \mathcal{I}} A_{ijk}(t) + B_{jk}(t) - D_{jk}(t), \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{2.5}$$

$$\sum_{k \in \mathcal{K}(j)} Z_{jk}(t) \leq N_j, \tag{2.6}$$

$$D_{jk}(t) = S_{jk}(T_{jk}(t)), \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{2.7}$$

$$T_{jk}(t) = \int_0^t Z_{jk}(s)\mathrm{d}s, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{2.8}$$

$$Y_j(t) = N_j t - \sum_k T_{jk}(t), \tag{2.9}$$

$$Q_k(t) \left(\sum_{j \in \mathcal{J}(k)} \left(N_j - \sum_{l \in \mathcal{K}(j)} Z_{jl}(t) \right) \right) = 0, \text{ for all } k \in \mathcal{K}, \tag{2.10}$$

$$\int_0^t \sum_{j \in \mathcal{J}(k)} \left(N_j - \sum_{l \in \mathcal{K}(j)} Z_{jl}(s) \right) \mathrm{d}A_{ik}(s) = 0, \text{ for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}, \tag{2.11}$$

$$\text{Equations associated with the control policy } \pi. \tag{2.12}$$

Equation (2.7) is identical to (2.2). The interpretation of (2.8) is that the busy time for server pool j working on class k at time t accumulates with rate equal to the total number of servers from pool j working on class k customers at time t. Equation (2.10) implies that there can be customers in the queue only when all the servers that can serve that queue are busy. It is called the non-idling condition and indicates that a server can only idle when there is no customer waiting in the queues that he can serve. Equation (2.11) implies that an arriving customer is delayed in the queue only if there is no idle server that can serve that customer at the time of his arrival. Equation (2.12) forces the routing and scheduling decisions to be made according to the selected routing and scheduling policies. Other conditions are self explanatory.

We call $\mathbb{X}_\pi$ the π-parallel server system process (or just π-parallel server system), although $\mathbb{X}_\pi$ is a process defined through the perturbed system. We denote the dimensions of $\mathbb{X}_\pi$ and Z by d and d_z, respectively.

Note that for given a control policy π, and the associated function f_π; see (2.1), it can be

applied to both the parallel server system and the perturbed system. For the parallel system, one can define the corresponding process $\mathbb{X}'_\pi = (A', A'_q, A'_s, Q', Z', T', Y', B', D')$ with each component having the same interpretation as in the perturbed system. Clearly, $A' = A$. But careful readers have noticed that the corresponding equation (2.7) for the departure process does *not* hold. Indeed, $\mathbb{X}'_\pi$ is *sample pathwise* different from the corresponding process $\mathbb{X}_\pi$, although $\mathbb{X}'_\pi$ satisfies all equations (2.3)–(2.12) except (2.7). We have the following result on the equivalence of two processes.

Theorem 2.1. *Under an admissible policy π, $\mathbb{X}_\pi$ is equal to $\mathbb{X}'_\pi$ in distribution when they are given the same initial condition.*

The proof that is placed in Appendix B uses the description of the primitive process presented in the next section and we therefore recommend the reader go over the next section first.

2.2 Primitive processes

The main goal of this thesis is to study state space collapse results in many server diffusion limits. Therefore, we analyze a sequence of systems indexed by r such that the arrival rates grow to infinity as $r \to \infty$. The number of servers also grows to infinity to meet the growing demand. We append "r" to the processes that are associated with the rth system, e.g., $Q^r_k(t)$ is used to denote the number of class k customers in the queue in the rth system at time t. The arrival rate for the ith arrival stream in the rth system is given by λ^r_i and we set $\lambda^r = (\lambda^r_1, \ldots, \lambda^r_I)$. We assume that

$$\lambda^r_i \to \infty, \quad i = 1, \ldots, I, \tag{2.13}$$

as $r \to \infty$. We also assume that all the customer types are asymptotically significant and the growth rate of the arrival rate for each type is proportional, that is,

$$\lambda^r_i / \sum_{i \in \mathcal{I}} \lambda^r_i \to a_i, \tag{2.14}$$

as $r \to \infty$, for some $a_i > 0$ and all $i \in \mathcal{I}$.

Let $\{S_{jk},\, j = 1, \ldots, J, k = 1, 2, \ldots, K\}$ be as given in the previous section and $\{v_{jk}(l) : l = 1, 2, \ldots\}$ be the sequence of interarrival times of the process $\{S_{jk}(t) : t \geq 0\}$. Since S_{jk} is a Poisson process, $\{v_{jk}(l) : l = 1, 2, \ldots\}$ is a set of independent and identically distributed exponential random variables. We define $V_{jk} : \mathbb{N} \to \mathbb{R}$ by

$$V_{jk}(m) = \sum_{l=1}^{m} v_{jk}(l), \quad m \in \mathbb{N},$$

where, by convention, empty sums are set to be zero. The term $V_{jk}(m)$ is the total service requirement of the 1st m class k customers that are served by pool j servers, and V_{jk} is known as the cumulative service time process. By the duality of S_{jk} and V_{jk}, one has

$$S_{jk}(t) = \max\{m : V_{jk}(m) \leq t\}, \quad t \geq 0.$$

It follows from (2.7) that

$$V_{jk}(D^r_{jk}(t)) \leq T^r_{jk}(t) \leq V_{jk}(D^r_{jk}(t) + 1), \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j). \tag{2.15}$$

This condition is identical to the HL condition in a standard multiclass queueing network, where each station has a single server; see, for example, Dai [18].

Next, we give the details of the primitive arrival processes. Let $E_i = \{E_i(t) : t \geq 0\}$ be a delayed renewal process with rate 1 and $E = \{E_i : i \in \mathcal{I}\}$. Let

$$A^r_i(t) = E_i(\lambda^r_i t). \tag{2.16}$$

Let $\{u_i(l) : l = 1, 2, \ldots\}$ be the sequence of interarrival times that are associated with the process E_i. Note that they are independent and $\{u_i(l) : l = 2, 3, \ldots\}$ are identically distributed. We define $U_i : \mathbb{N} \to \mathbb{R}$ by

$$U_i(m) = \sum_{l=1}^{m} u_i(l), \quad m \in \mathbb{N},$$

and so

$$E_i(t) = \max\{m : U_i(m) \leq t\}.$$

We require that the interarrival times of the arrival processes satisfy the following condition that is similar to condition (3.4) in Bramson [13].

$$E\left[u_i(2)^{2+\epsilon}\right] < \infty, \text{ for all } i \in \mathcal{I} \text{ and for some } \epsilon > 0. \tag{2.17}$$

Condition (2.17) is automatically satisfied by the service times since they are assumed be exponentially distributed. For the rest of the paper, we assume that the primitive processes of a parallel server system satisfy (2.17). We also assume that $Q^r(0)$, $Z^r(0)$, S, and E are independent.

We require that the number of servers in the rth system in each pool is selected so that

$$\lim_{r\to\infty} \frac{N_j^r}{|N^r|} = \beta_j, \text{ for all } j \in \mathcal{J} \text{ and for some } \beta_j > 0 \text{ and} \tag{2.18}$$

$$\lim_{r\to\infty} \frac{\lambda_i^r}{|N^r|} = \lambda_i, \text{ for all } i \in \mathcal{I} \text{ and for some } 0 < \lambda_i < \infty. \tag{2.19}$$

We assume that $\{|N^r|\}$ is strictly increasing in r and we set $\lambda = (\lambda_1, \ldots, \lambda_I)$.

The process A_i^r can be taken to be a more general process than the one considered here. One can also consider more general service processes, for example, by taking a series of Poisson processes with the same rate or even with rates converging to a constant in certain manner; i.e., the processes S_{ij}'s may depend on r. We see no potential gain in the applications by doing so; besides, (2.15) and (2.16) allow us to use the established bounds on the primitive processes by Bramson [13] without any modification.

2.3 The static planning problem and asymptotic framework

The static planning problem (SPP) has been used in the literature to determine the optimal nominal allocations of servers' capacities for the service of customer classes; see Harrison [35] and Dai and Lin [20]. Nominal allocations determine the long-run proportion of servers' effort allocated to each class. We take a similar approach to determine the nominal proportion of servers in a server pool that will be allocated to serve each class.

The static planning problem is defined as

$$
\begin{aligned}
&\min \rho \\
&\text{s.t.} \\
&\sum_{k\in\mathcal{K}} \alpha_{ik} = \lambda_i, \text{ for all } i \in \mathcal{I}, \\
&\sum_{j\in\mathcal{J}(k)} \beta_j \mu_{jk} x_{jk} = \sum_{i\in\mathcal{I}} \alpha_{ik}, \text{ for all } k \in \mathcal{K}, \\
&\sum_{k\in\mathcal{K}(j)} x_{jk} \leq \rho, \text{ for all } j \in \mathcal{J}, \\
&x_{jk}, \alpha_{ik} \geq 0, \text{ for all } j \in \mathcal{J}, k \in \mathcal{K}, \text{ and } i \in \mathcal{I}.
\end{aligned}
\tag{2.20}
$$

The quantity α_{ik}/λ_i can be thought of as the long-run proportion of type i customers that are routed to queue k and x_{jk} as the long-run proportion of servers in server pool j that are allocated to serve class k customers. We set $\alpha = \{\alpha_{ik} : i \in \mathcal{I}, j \in \mathcal{K}\}$ and $x = \{x_{jk} : j \in \mathcal{J}, k \in \mathcal{K}\}$.

The objective of the SPP is to minimize the total proportion of required servers in each pool. From this formulation it is clear that referring to x as proportions is a misnomer since $\sum_{k\in\mathcal{K}(j)} x_{jk}$ may be greater than 1. We use the term "proportion" because of Assumption 1 below.

The main difference between our formulation of the SPP and the one in Harrison [35] is that we model routing of customers to queues explicitly as in Stolyar [60]. We pay the price by having one more constraint than his formulation. The main constraint is to be able to serve all the incoming customers. This is formulated in the first and the second constraints. The first constraint assures that all the arriving customers are routed to one of the queues and the second constraint is needed to guarantee that enough service capacity is allocated to all customer classes.

Let (ρ^*, x^*, α^*) be an optimal solution to the SPP. If $\rho^* > 1$, it can be easily shown that the queue length process is not bounded under the fluid limit for r large enough (fluid limits are defined in Appendix A). We will assume for the rest of this paper that $\rho^* \leq 1$.

Now, consider the sequence of parallel server systems described in the previous section,

and the associated SPP with the rth system;

$$\begin{aligned}
&\min \rho^r \\
&\text{s.t.} \\
&\sum_{k\in\mathcal{K}} \alpha^r_{ik} = \lambda^r_i, \text{ for all } i \in \mathcal{I}, \\
&\sum_{j\in\mathcal{J}(k)} N^r_j \mu_{jk} x^r_{jk} = \sum_{i\in\mathcal{I}} \alpha^r_{ik}, \text{ for all } k \in \mathcal{K}, \\
&\sum_{k\in\mathcal{K}(j)} x^r_{jk} \le \rho^r, \text{ for all } j \in \mathcal{J}, \\
&x^r_{jk}, \alpha^r_{ik} \ge 0, \text{ for all } j \in \mathcal{J}, k \in \mathcal{K}, \text{ and } i \in \mathcal{I}.
\end{aligned} \tag{2.21}$$

Let $(\rho^{r,*}, x^{r,*}, \alpha^{r,*})$ be an optimal solution to (2.21). We formulate the many server heavy traffic condition as follows.

Assumption 1. *The SPP (2.20) has a unique optimal solution (ρ^*, x^*, α^*) with λ given by (2.19) and that solution has $\rho^* = 1$ and $\sum_{k\in\mathcal{K}(j)} x^*_{jk} = 1$ for all $j \in \mathcal{J}$. Moreover,*

$$\sqrt{|N^r|}(1 - \rho^{r,*}) \to \theta,$$

for some $\theta \in \mathbb{R}$ as $r \to \infty$.

Even when the SPP (2.20) has an optimal solution with $\rho^* \le 1$, it is not a trivial task to come up with a control policy that will achieve the optimal allocations in the long-run. If ρ^* is close to one, small deviations from the optimal allocations may again cause the queue length to grow without a bound. This phenomenon is closely related to the stability of a control policy in a multiclass queueing network setting. In this paper we only consider control policies that satisfy the following assumption.

Assumption 2. *For a control policy π,*

$$T^r_{jk}(\cdot)/|N^r| \to T^*_{jk}(\cdot) \text{ u.o.c. a.s.}, \tag{2.22}$$

as $r \to \infty$, if $(Q^r(0)/|N^r|, Z^r(0)/|N^r|) \to (0, z)$ a.s., as $r \to \infty$, where $T^*_{jk}(t) = \beta_j x^*_{jk} t$, $z = (z_{jk}, j \in \mathcal{J}, k \in \mathcal{K}(j))$, and $z_{jk} = \beta_j x^*_{jk}$.

We provide a general framework that can be used to check that a control policy satisfies Assumption 2 in Appendix A.

In general, the diffusive scaling is defined in a way that allows one to study the fluctuations of the queue length process around its long-run average. Implicitly given in Assumption 2 and elaborated in Appendix A is that

$$Q^r(\cdot)/|N^r| \to 0 \text{ and } Z^r(\cdot)/|N^r| \to z(\cdot) \text{ u.o.c. a.s.},$$

as $r \to \infty$, where $z(t) = z$, for $t \geq 0$. Hence, we define the diffusive scaling as follows;

$$\hat{Q}^r(t) = \frac{Q^r(t)}{\sqrt{|N^r|}} \text{ and } \hat{Z}^r_{jk}(t) = \frac{Z^r_{jk}(t) - x^*_{jk}N^r_j}{\sqrt{|N^r|}}, \text{ for } t \geq 0. \tag{2.23}$$

CHAPTER III

MAIN RESULTS

In this section, we present a general framework to prove a state space collapse result in the many server diffusion limit of a π-parallel server system process. We first introduce the hydrodynamic model equations. The solutions of these equations play an important role in the general SSC framework. We present our main results in Section 3.2. Naturally, some of the hydrodynamic equations depend on the policy used. We will present the additional hydrodynamic equations for the policies discussed in Chapter 4, here we just note that they can be obtained via standard arguments.

3.1 Hydrodynamic model equations

Consider the process $\tilde{\mathbb{X}}_\pi = (\tilde{A}, \tilde{A}_q, \tilde{A}_s, \tilde{Q}, \tilde{Z}, \tilde{B})$ and the following set of equations:

$$\lambda_i t = \sum_{k\in\mathcal{K}} \tilde{A}_{ik}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \tilde{A}_{ijk}(t), \text{ for all } i \in \mathcal{I}, \tag{3.1}$$

$$\tilde{Q}_k(t) = \tilde{Q}_k(0) + \sum_{i\in\mathcal{I}} \tilde{A}_{ik}(t) - \sum_{j\in\mathcal{J}(k)} \tilde{B}_{jk}(t), \text{ for all } k \in \mathcal{K}, \tag{3.2}$$

$$\tilde{Q}_k(t) \geq 0, \text{ for all } k \in \mathcal{K}, \tag{3.3}$$

$$\tilde{A}_{ik}, \tilde{A}_{ijk}, \tilde{B}_{jk} \text{ are nondecreasing for all } i \in \mathcal{I}, j \in \mathcal{J}, \text{ and } k \in \mathcal{K}, \tag{3.4}$$

$$\tilde{Z}_{jk}(t) = \tilde{Z}_{jk}(0) + \sum_{i\in\mathcal{I}} \tilde{A}_{ijk}(t) + \tilde{B}_{jk}(t) - \mu_{jk} z_{jk} t, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{3.5}$$

$$\sum_{k\in\mathcal{K}(j)} \tilde{Z}_{jk}(t) \leq 0, \text{ for all } j \in \mathcal{J}, \tag{3.6}$$

$$\tilde{Q}_k(t) \left(\sum_{j\in\mathcal{J}(k)} \sum_{l\in\mathcal{K}(j)} \tilde{Z}_{jl}(t) \right) = 0, \text{ for all } k \in \mathcal{K}, \tag{3.7}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)} \left(\sum_{l\in\mathcal{K}(j)} \tilde{Z}_{jl}(s) \right) \mathrm{d}\tilde{A}_{ik}(s) = 0, \text{ for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}, \tag{3.8}$$

$$\text{Additional equations associated with the control policy } \pi, \tag{3.9}$$

where λ_i is defined as in (2.19) and z_{ij} as in Assumption 2. Equations (3.1)-(3.9) are called the hydrodynamic model equations, and they define the *hydrodynamic model* of the π-parallel server system. Any process $\tilde{\mathbb{X}}_\pi$ satisfying (3.1)-(3.9) for all $t \geq 0$ is called a hydrodynamic model solution.

Hydrodynamic model solutions are similar to the fluid model solutions; they are deterministic and absolutely continuous. However, comparison of hydrodynamic model equations (3.1)-(3.9) with the fluid model equations (A.2)-(A.9) reveals major differences. The hydrodynamic counterpart of number of servers in pool j working on class k customers, $\tilde{Z}_{jk}$, can assume negative values unlike the corresponding fluid model process $\bar{Z}_{jk}$, which is always nonnegative. Also, the departure process D (and related processes T and Y) are missing from the hydrodynamic model equations. The reason will become clear when we present the mathematical origins of hydrodynamic model equations but it is also obvious from (3.5) that $\tilde{D}_{jk}(t) = \mu_{jk} z_{jk} t$, $\tilde{T}_{jk}(t) = z_{jk} t$, and $\tilde{Y}_j(t) = 0$ for all $t \geq 0$. This reveals the importance of Assumption 2. Similar to the "efficient" policy concept in conventional heavy traffic, under a policy that does not satisfy this assumption for large enough r fluid scaled queue lengths will be unbounded. Also, the hydrodynamic model solution $\tilde{T}$ cannot be easily characterized for those policies.

The most important property of hydrodynamic model equations is that the hydrodynamic limits satisfy these equations under certain general assumptions. This is similar to the relationship between the fluid model equations and the fluid limits; see Dai [18]. Equation (3.9) is obtained from the policy π. It has to be justified mathematically that the hydrodynamic limits satisfy this equation.

3.2 State space collapse in the diffusion limits

We need a machinery to define a state space collapse in mathematical terms, for this we use a function with the following properties. Let $g : \mathbb{R}^{K+d_z} \to \mathbb{R}^+$ be a nonnegative function that satisfies the following homogeneity condition;

$$g(\alpha x) = \alpha^c g(x), \tag{3.10}$$

for all $x \in \mathbb{R}^{K+d_z}$, $0 \leq \alpha \leq 1$, and for some $c > 0$. We call g an SSC-function. Nonnegativity assumption is made for notational convenience and one can always consider $|g|$ in order to have a nonnegative function if g can take negative values. We make the following assumption about the SSC-function.

Assumption 3. *The function $g : \mathbb{R}^{K+d_z} \to \mathbb{R}^+$ satisfies (3.10) and is continuous on $\mathbb{R}^{K+d_z}$.*

Assumption 3 is needed for a simple reason; we will consider a sequence of stochastic processes that converges to another one and we would like to show that the sequence that consists of the values of g evaluated for each process converges to the value of g evaluated at the limiting process. Assumption 3 makes this possible by virtue of the Continuous Mapping Theorem; see [15]. Condition (3.10) will be needed when we translate the results from hydrodynamic scaled processes to diffusive scaled processes. The class of functions that satisfy Assumption 3 is large enough for most purposes, however, this class can be extended as discussed in Chapter 3.3.

As the machinery to state an SSC result has been set, we are ready to state the conditions on the hydrodynamic model solutions that imply that an SSC result holds in the diffusion limit. The following assumption is analogous to Assumption 3.2 in Bramson [13].

Assumption 4. *Let g be a function that satisfies Assumption 3. There exists a function $H(t)$ with $H(t) \to 0$ as $t \to \infty$ such that*

$$g(\tilde{Q}(t), \tilde{Z}(t)) \leq H(t) \text{ for all } t \geq 0 \tag{3.11}$$

for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$. Furthermore, for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ with $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$, $g(\tilde{Q}(t), \tilde{Z}(t)) = 0$ for $t \geq 0$.

We are ready to state the main result of this thesis.

Theorem 3.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Suppose that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of π-parallel server system satisfies Assumption 4, and*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \text{ in probability} \tag{3.12}$$

as $r \to \infty$. Then, for each $T > 0$,

$$\frac{\left\| g(\hat{Q}^r(t), \hat{Z}^r(t)) \right\|_T}{\left(\left\| \hat{Q}^r(t) \right\|_T \vee \left\| \hat{Z}^r(t) \right\|_T \vee 1 \right)^c} \to 0 \text{ in probability}, \tag{3.13}$$

as $r \to \infty$, where $c > 0$ is given as in (3.10).

Remark 3.2. The result of Theorem 3.1 is still valid if it is only assumed that hydrodynamic limits, not the hydrodynamic model, satisfy Assumption 4. This relaxes the assumption because it will be shown that every hydrodynamic limit over a finite time interval $[0, L]$, for some $L > 0$, is a hydrodynamic model solution on $[0, L]$. The set of hydrodynamic model solutions may contain processes that are not hydrodynamic limits.

Remark 3.3. The SSC result as stated in Theorem 3.1 is called the *multiplicative state space collapse*. If $(\hat{Q}^r, \hat{Z}^r)$ also satisfies

$$\lim_{R\to\infty} \limsup_{r\to\infty} P\left\{\left\|\hat{Q}^r(\cdot)\right\|_T \vee \left\|\hat{Z}^r(\cdot)\right\|_T > R\right\} = 0 \tag{3.14}$$

for all $T > 0$, then one can use this property to remove the denominator from (3.13) and obtain a *strong state space collapse* that is more suitable for applications.

The condition (3.12) can be relaxed as in Theorem 3 in Bramson [13] to only require that $\hat{Q}^r(0)$ and $\hat{Z}^r(0)$ are stochastically bounded. The state space collapse result in this case however does not hold initially at time 0.

Theorem 3.4. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π parallel server system processes. Suppose that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of π-parallel server system satisfies Assumption 4, and $\left|\hat{Q}^r(0)\right| \vee \left|\hat{Z}^r(0)\right|$ is stochastically bounded. Then, for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,*

$$P\left\{\frac{\sup_{L^r/\sqrt{|N^r|}\le t\le T}\left|g(\hat{Q}^r(t), \hat{Z}^r(t))\right|}{\sup_{L^r/\sqrt{|N^r|}\le t\le T}\left(\left|\hat{Q}^r(t)\right| \vee \left|\hat{Z}^r(t)\right| \vee 1\right)^c} > \epsilon\right\} \to 0, \tag{3.15}$$

as $r \to \infty$, where $c > 0$ is given as in (3.10).

Remark 3.5. Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes that satisfy the conditions of Theorem 3.4. If in addition H, given as in Assumption 4, is bounded, then

$$\lim_{R\to\infty} \limsup_{r\to\infty} P\left\{\left\|g(\hat{Q}^r(t), \hat{Z}^r(t))\right\|_{L^r/\sqrt{|N^r|}} > R\right\} = 0. \tag{3.16}$$

The result (3.16) may be used in verifying that

$$\lim_{R\to\infty} \limsup_{r\to\infty} P\left\{\sup_{L^r/\sqrt{|N^r|}\le t\le T}\left(\left|\hat{Q}^r(t)\right| \vee \left|\hat{Z}^r(t)\right|\right) > R\right\} = 0. \tag{3.17}$$

Then, similar to Remark 3.3, one can deduce a strong state space collapse result from Theorem 3.4 using (3.17).

3.3 Extensions

In this section we discuss possible extensions of our SSC result. These extensions involve relaxing the condition (3.10).

3.3.1 A weaker homogeneity condition

Theorem 3.1 can be generalized by relaxing condition (3.10) on the class of SSC functions. We replace condition (3.10) with the following condition: there exist $c_1 > 0$ and $c_2 > 0$ such that

$$\alpha^{c_1} g(x) \leq g(\alpha x) \leq \alpha^{c_2} g(x) \tag{3.18}$$

for all $x \in \mathbb{R}^{K+d_z}$ and $0 \leq \alpha \leq 1$. Under (3.18) Theorem 3.1 holds with c is replaced with c_1.

3.3.2 When the homogeneity condition does not hold

In this section we only assume that the SSC function g satisfies the following condition.

Assumption 5. *Function $g : \mathbb{R}^{K+d_z} \to \mathbb{R}^+$ is a nonnegative function and continuous on $\mathbb{R}^{K+d_z}$.*

When the SSC function g only satisfies Assumption 5 but not Assumption 3 we need the following additional condition on the queue length and busy number of servers processes.

Assumption 6. *For every $T > 0$ (3.14) holds.*

For $T > 0$ we define

$$\mathcal{A}_R^r(T) = \left\{ \left(\left\| \hat{Q}^r(\cdot) \right\|_T \vee \left\| \hat{Z}^r(\cdot) \right\|_T \right) \leq R \right\} \tag{3.19}$$

Remark 3.6. In the current setting we establish the relation between the hydrodynamic limits of the sample paths of $\omega \in \mathcal{A}_R^r(T)$ and a state space collapse result. Therefore, the hydrodynamic limits, hydrodynamic model equations and hydrodynamic model solutions

will depend on R and T. To make this dependence explicit we refer to these hydrodynamic limits, hydrodynamic model equations and hydrodynamic model solutions as the hydrodynamic limits, hydrodynamic model equations and hydrodynamic model solutions on $\mathcal{A}_R^r(T)$. One can show that hydrodynamic model equations (3.1)-(3.8) are still satisfied by all hydrodynamic limits. However, the policy dependent hydrodynamic model equations depend on R and T.

Since the hydrodynamic limits are dependent on R and T we need to modify Assumption 4 as follows.

Assumption 7. *For every $T > 0$, there exists $R_0(T)$ such that for every $R > R_0(T)$, there exists a function $H_{R,T}(t)$ with $H_{R,T}(t) \to 0$ as $t \to \infty$ such that*

$$g\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) \leq H_{R,T}(t) \text{ for all } t \geq 0 \tag{3.20}$$

for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ on $\mathcal{A}_R^r(T)$.

Furthermore, for each hydrodynamic hydrodynamic solution $\tilde{\mathbb{X}}_\pi$ on $\mathcal{A}_R^r(T)$ with $g\left(R\left(\tilde{Q}(0), \tilde{Z}(0)\right)\right) = 0$, $g\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) = 0$, for $t \geq 0$.

We are ready to state the main result of this section.

Theorem 3.7. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Suppose that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 5, Assumption 6 holds, the hydrodynamic model of π-parallel server system satisfies Assumption 7, and*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \text{ in probability} \tag{3.21}$$

as $r \to \infty$. Then, for each $T > 0$,

$$\left\| g(\hat{Q}^r(t), \hat{Z}^r(t)) \right\|_T \to 0 \text{ in probability}, \tag{3.22}$$

as $r \to \infty$.

3.4 State space collapse framework

In this section we prove Theorems 3.1 and 3.4. We begin with introducing the hydrodynamic scaling that will be used to define the hydrodynamic limits. Once we establish the

relationship between the hydrodynamic scaled processes and the hydrodynamic limits we translate condition (3.11) to a condition on the diffusive scaled processes. We finally show that this latter condition implies the desired SSC result in the diffusion limit.

3.4.1 Hydrodynamic scaling and bounds

The hydrodynamic scaling is used by Bramson [13] to establish a relationship between the hydrodynamic and the diffusion limits in conventional heavy traffic asymptotic analysis. We consider a similar time scaling that slows the process down. This allows us to analyze the events that happen instantaneously in the diffusive scale in more detail. This can be achieved by using a scaling similar to the diffusion scaling as given in (2.23) but also scaling the time by $1/\sqrt{|N^r|}$. However, this scaling is not suitable for our purposes. We need the more refined scaling which we call the *hydrodynamic scaling*. For any nonnegative integer m, let

$$x_{r,m} = \left| Q^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right|^2 \vee \left| Z^r\left(\frac{m}{\sqrt{|N^r|}}\right) - \vec{N}^r x^*\right|^2 \vee |N^r|, \tag{3.23}$$

where $\vec{N}^r$ is a diagonal matrix with $\vec{N}^r_{jj'} = N^r_j$ if $j = j'$ and 0 otherwise for $j \in \mathcal{J}$ and $x^* = (x_{jk}, j \in \mathcal{J}, k \in \mathcal{K})$ is given as in Assumption 2. Hence, $Z^r(t) - \vec{N}^r x^*$ is a $J \times K$ matrix with its (j,k)th entry equal to $Z^r_{jk}(t) - x^*_{jk}N^r_j$ if $k \in \mathcal{K}(j)$ and zero otherwise. We define the hydrodynamic scaling by shifting and scaling the processes of $\mathbb{X}^r$ as follows;

$$\begin{aligned}
A^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(A^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - A^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
A^{r,m}_s(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(A^r_s\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - A^r_s\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
A^{r,m}_q(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(A^r_q\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - A^r_q\left(\frac{m}{\sqrt{|N^r|}}\right)\right),
\end{aligned}$$

$$
\begin{aligned}
B^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(B^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-B^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
D^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(D^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-D^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
T^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(T^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-T^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
Y^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(Y^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-Y^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right),\\
Q^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(Q^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)\right), \text{ and}\\
Z^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(Z^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-\vec{N}^r x^*\right).
\end{aligned}
\tag{3.24}
$$

We define

$$
\begin{aligned}
V_{jk}^{r,m}(D_{jk}^{r,m}(t),b) = \frac{1}{\sqrt{x_{r,m}}}\Bigg(&V_{jk}\left(D_{jk}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)+b_1\right)\\
&-V_{jk}\left(D_{jk}^r\left(\frac{m}{\sqrt{|N^r|}}\right)+b_2\right)\Bigg),
\end{aligned}
\tag{3.25}
$$

for $b=(b_1,b_2)\in\mathbb{R}^2$. By (2.15),

$$
V_{jk}^{r,m}(D_{jk}^{r,m}(t),(0,1)) \le T_{jk}^{r,m}(t) \le V_{jk}^{r,m}(D_{jk}^{r,m}(t),(1,0)). \tag{3.26}
$$

Let $\mathbb{X}^{r,m}=(A^{r,m},A_s^{r,m},A_q^{r,m},B^{r,m},T^{r,m},Y^{r,m},Q^{r,m},Z^{r,m})$. We call $\mathbb{X}^{r,m}$ hydrodynamic scaled process. From the definition of $x_{r,m}$ we have that

$$
|\mathbb{X}^{r,m}(0)| \le 1.
$$

It can easily be checked that $\mathbb{X}^{r,m}$ satisfies the following equations for all $t \geq 0$.

$$A_i^{r,m}(t) = \sum_{k\in\mathcal{K}} A_{ik}^{r,m}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} A_{ijk}^{r,m}(t), \forall i \in \mathcal{I}, \tag{3.27}$$

$$Q_k^{r,m}(t) = Q_k^{r,m}(0) + \sum_{i\in\mathcal{I}} A_{ik}^{r,m}(t) - \sum_{j\in\mathcal{J}(k)} B_{jk}^{r,m}(t), \forall k \in \mathcal{K}, \tag{3.28}$$

$$Z_{jk}^{r,m}(t) = Z_{jk}^{r,m}(0) + \sum_{i\in\mathcal{I}} A_{ijk}^{r,m}(t) + B_{jk}^{r,m}(t) - D_{jk}^{r,m}(t), \forall j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{3.29}$$

$$D_{jk}^{r,m}(t) = \frac{S_{jk}(\sqrt{x_{r,m}}T_{jk}^{r,m}(t) + T_{jk}^{r}(m/\sqrt{|N^r|})) - S_{jk}(T_{jk}^{r}(m/\sqrt{|N^r|}))}{\sqrt{x_{r,m}}}, \forall j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{3.30}$$

$$T_{jk}^{r,m}(t) = \frac{N_j^r x_{jk}^*}{|N^r|}t + \frac{\sqrt{x_{r,m}}}{|N^r|}\int_0^t Z_{jk}^{r,m}(s)\mathrm{d}s, \forall j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{3.31}$$

$$Y_j^{r,m}(t) \text{ can only increase when } \sum_{k\in\mathcal{K}(j)} Z_{jk}^{r,m}(t) < 0, \text{ for all } j \in \mathcal{J}, \tag{3.32}$$

$$Q_k^{r,m}(t)\left(\sum_{j\in\mathcal{J}(k)}\sum_{k'\in\mathcal{K}(j)} Z_{jk'}^{r,m}(t)\right) = 0, \forall k \in \mathcal{K}, \tag{3.33}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)}\left(\sum_{k'\in\mathcal{K}(j)} Z_{jk'}^{r,m}(s)\right)\mathrm{d}A_{ik}^{r,m}(s) = 0, \forall i \in \mathcal{I} \text{ and } k \in \mathcal{K}. \tag{3.34}$$

We have the following estimates that are similar to those established in Proposition 5.1 in Bramson [13].

Proposition 3.8. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. Fix $\epsilon > 0$, $L > 0$ and $T > 0$. Then, for large enough r,*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|A^{r,m}(t) - \frac{\lambda^r}{|N^r|}t\right\|_L > \epsilon\right\} \leq \epsilon, \tag{3.35}$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\leq L}|D^{r,m}(t_1) - D^{r,m}(t_2)| > N|t_1 - t_2| + \epsilon\right\} \leq \epsilon, \text{ and} \tag{3.36}$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V_{jk}^{r,m}\left(D_{jk}^{r,m}(t), b\right) - \frac{1}{\mu_{jk}}D_{jk}^{r,m}(t)\right\|_L > \epsilon\right\} \leq \epsilon, \tag{3.37}$$

for all $j \in \mathcal{J}, k \in \mathcal{K}(j)$ and for $b = (0,1)$ or $(1,0)$.

The proof is given in Appendix 3.5.1.1. Using this proposition, one can show that $\mathbb{X}^{r,m}$ is almost Lipschitz as given in the next proposition. In this section and for the remainder of this paper N without a superscript is reused to denote a general constant.

Proposition 3.9. *Let $\{\mathbb{X}^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. Fix $\epsilon > 0$, $L > 0$ and $T > 0$. Then, for large enough r,*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\leq L}|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)| > N|t_1 - t_2| + \epsilon\right\} \leq \epsilon, \tag{3.38}$$

where $N < \infty$ and only depends on (I, J, K, λ).

The proof is given in Appendix 3.5.1.2. For convenience, we assume for the rest of the paper that $N \geq 1$ and $L \geq 1$. Let

$$\mathcal{K}_0^r = \left\{ \max_{m < \sqrt{|N^r|}T} \sup_{t_1, t_2 \leq L} |\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)| \leq N|t_1 - t_2| + \epsilon(r) \right\}, \tag{3.39}$$

where L, N, and T are fixed as before and $\epsilon(r)$ with $\epsilon(r) \to 0$ as $r \to \infty$ is a sequence of real numbers. Similarly, we can replace ϵ in (3.35) and (3.37) by $\epsilon(r)$. We denote these new inequalities obtained from (3.35) and (3.37) by (3.35′) and (3.37′). Let $\mathcal{K}^r$ denote the intersection of $\mathcal{K}_0^r$ with the complements of the events in (3.35′) and (3.37′). As in Bramson [13], when $\epsilon(r) \to 0$ sufficiently slowly as $r \to \infty$, one can show that $P(\mathcal{K}^r) \to 1$ as $r \to \infty$.

We summarize the above discussion in the following corollary for future reference.

Corollary 3.10. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. Fix $L >$ and $T > 0$, and choose N and $\epsilon(r)$ as above. Then, for $\mathcal{K}^r$ defined as above*

$$\lim_{r \to \infty} P(\mathcal{K}^r) = 1. \tag{3.40}$$

3.4.2 Hydrodynamic Limits of π-parallel server systems

In this section, we define the hydrodynamic limits of π-parallel server systems. First, we define a set of functions which contains all the hydrodynamic limits. The following definitions are similar to those in Section 6 of Bramson [13] and the notation is adapted from that paper.

Fix $L > 0$. Let $\tilde{E}$ be the set of right continuous functions with left limits, $x : [0, L] \to \mathbb{R}^d$. Let E' denote those $x \in \tilde{E}$ that satisfies

$$|x(0)| \leq 1$$

and

$$|x(t_2) - x(t_1)| \leq N|t_2 - t_1| \text{ for all } t_1, t_2 \in [0, L],$$

where constant N is chosen as in Proposition 3.9. We set

$$E^r = \{\mathbb{X}^{r,m}, m < \sqrt{|N^r|}T, \omega \in \mathcal{K}^r\}$$

and

$$\mathcal{E} = \{E^r : r \in \mathbb{N}\},$$

where T is fixed, and $\mathcal{K}^r$ is defined as in the previous section. (These quantities are not correlated to the external arrival processes E introduced in Chapter 2.2.)

We define a hydrodynamic limit x of $\mathcal{E}$ to be a point $x \in \tilde{E}$ such that for all $\epsilon > 0$ and $r_0 \in \mathbb{N}$, there exist $r \geq r_0$ and $y \in E^r$, with $\|x(\cdot) - y(\cdot)\|_L < \epsilon$.

Since

$$|\mathbb{X}^{r,m}(0)| \leq 1 \tag{3.41}$$

for all $m < \sqrt{|N^r|}T$ and $r \in \mathbb{N}$, the following result is a corollary of Proposition 4.1 in Bramson [13] and is similar to Proposition 6.1 in that paper. It shows that the hydrodynamic limits are "rich" in the sense that for r large enough every hydrodynamic scaled process is close to a hydrodynamic limit.

Corollary 3.11. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds, π satisfies Assumption 2. Let $\tilde{E}, E^r$, and $\mathcal{E}$ be as specified above. Fix $\epsilon > 0$, $L > 0$ and $T > 0$, and choose r large enough. Then, for $\omega \in \mathcal{K}^r$ and any $m < \sqrt{|N^r|}T$*

$$\left\|\mathbb{X}^{r,m}(\cdot) - \tilde{\mathbb{X}}(\cdot)\right\|_L \leq \epsilon \tag{3.42}$$

for some hydrodynamic limit $\tilde{\mathbb{X}}(\cdot)$ of $\mathcal{E}$ with $\tilde{\mathbb{X}}(\cdot) \in E'$.

The next result is mainly needed to translate the condition on the hydrodynamic model solutions to hydrodynamic limits given in Assumption 4. It also reveals the origin of hydrodynamic model equations.

Proposition 3.12. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. Choose $L > 0$ and let $\tilde{\mathbb{X}}_\pi$ be a hydrodynamic limit of $\mathcal{E}$ over $[0, L]$. $\tilde{\mathbb{X}}_\pi$ satisfies the hydrodynamic model equations (3.1)-(3.9) on $[0, L]$.*

The proof is given in Appendix 3.5.4.

Observe that by (3.10) and definitions of hydrodynamic and diffusion scalings

$$\left|g\left(Q^{r,0}(0), Z^{r,0}(0)\right)\right| \leq \left|g\left(\hat{Q}^{r}(0), \hat{Z}^{r}(0)\right)\right|. \tag{3.43}$$

If condition (3.12) holds, (3.43) implies that $g\left(Q^{r,0}(0), Z^{r,0}(0)\right) \to 0$ in probability as $r \to \infty$. Therefore, we can choose $\epsilon(r)$ with $\epsilon(r) \to 0$ as $r \to \infty$ such that for $\mathcal{L}^r = \mathcal{K}^r \cap \mathcal{G}^r$, where

$$\mathcal{G}^r = \left\{\left|g\left(Q^{r,0}(0), Z^{r,0}(0)\right)\right| \leq \epsilon(r)\right\},$$

we have

$$\lim_{r\to\infty} P\left(\mathcal{L}^r\right) = 1. \tag{3.44}$$

We set

$$E_g^r = \left\{\mathbb{X}^{r,0}(\cdot, \omega), \omega \in \mathcal{L}^r\right\}$$

and

$$\mathcal{E}_g = \left\{E_g^r, r \in \mathbb{N}\right\}.$$

The following proposition is similar to Proposition 6.4 in Bramson [13].

Proposition 3.13. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of the π-parallel server system satisfies Assumption 4. Fix $\epsilon > 0$, $L > 0$ and $T > 0$, and assume that r is large. Then, for $\omega \in \mathcal{K}^r$*

$$g(Q^{r,m}(t), Z^{r,m}(t)) \leq H(t) + \epsilon \tag{3.45}$$

for all $t \in [0, L]$, and $m < \sqrt{|N^r|}T$, with $H(\cdot)$ is given in Assumption 4.

Furthermore, for $\omega \in \mathcal{L}^r$

$$\left\|g(Q^{r,0}(t), Z^{r,0}(t))\right\|_L \leq \epsilon. \tag{3.46}$$

If, in addition, condition (3.12) holds, then (3.44) holds.

The proof is given in Appendix 3.5.2.1.

3.4.3 State space collapse in the diffusion limits

In this section we change the scaling from hydrodynamic to diffusion to prove Theorem 3.1. Once the scaling is changed, a few complications needs to be dealt with regarding the change in the range of the time variable.

We begin with changing the scaling. One can check by employing (2.23) and (3.24) that

$$\begin{aligned} Q_k^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}}\hat{Q}_k^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right) = \frac{1}{y_{r,m}}\hat{Q}_k^r\left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t+m)\right) \text{ and} \\ Z_{jk}^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}}\hat{Z}_{jk}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right) = \frac{1}{y_{r,m}}\hat{Z}_{jk}^r\left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t+m)\right), \end{aligned} \tag{3.47}$$

where

$$y_{r,m} = \sqrt{\frac{x_{r,m}}{|N^r|}} = \left|\hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right| \vee \left|\hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right| \vee 1. \tag{3.48}$$

By changing the scaling in Proposition 3.13 as above, we can rephrase (3.45) and (3.46). However, the domain of the time scales will change and the domain $0 \le t \le L$ for the arguments on the left hand side of (3.47) will correspond to

$$\frac{m}{\sqrt{|N^r|}} \le t \le \frac{1}{\sqrt{|N^r|}}(y_{r,m}L+m) \tag{3.49}$$

for the arguments on the right.

Proposition 3.14. *Let $\{\mathbb{X}^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of the π-parallel server system satisfies Assumption 4, and (3.12) holds. Fix $\epsilon > 0$, $L > 0$ and $T > 0$, and assume that r is large. Then, for $\omega \in \mathcal{K}^r$ and for $H(\cdot)$ given as in Assumption 4*

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \le y_{r,m}^c H\left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t-m)\right) + \epsilon y_{r,m}^c \tag{3.50}$$

for all $t \in [0,T]$ and m satisfying (3.49).

If condition (3.12) holds, then

$$\left\| g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)\right\|_{Ly_{r,0}/\sqrt{|N^r|}} \le \epsilon y_{r,0}^c \tag{3.51}$$

for $\omega \in \mathcal{L}^r$.

Proof. The bounds (3.50) and (3.51) are obtained from (3.45) and (3.46), respectively, by applying (3.47) and using (3.10). □

If we can show that $(\sqrt{|N^r|}t-m)/y_{r,m}$ is large, where $|N^r|$ is the total number of agents in the rth system, we can conclude the proof of Theorem 3.1 by using the convergence property of $H(\cdot)$ as given in Assumption 4. It will be shown that it is enough to have $\sqrt{|N^r|}t - m$ and L large.

Since the value of L is a matter of choice, we can take L sufficiently large and redefine $\mathcal{K}^r$ with the reselected L. Let H be given as in Assumption 4. Since $H(t) \to 0$ as $t \to \infty$, independent of L, for $\epsilon > 0$ fixed, there exists $s^*(\epsilon) > 1$ such that for $t > s^*(\epsilon)$, $H(t) < \epsilon$. We assume for the rest of the paper that

$$L \geq 6Ns^*(\epsilon), \tag{3.52}$$

where N is chosen as in (3.39).

In order to make $\sqrt{|N^r|}t - m$ large, for a fixed $t \in [0, T]$, we take the smallest m that satisfies (3.49), which we denote by $m_r(t)$. We need the following lemmas, whose proofs are given in Appendix 3.5.3, to show that $\sqrt{|N^r|}t - m_r(t)$ is large.

Lemma 3.15. *Let $\{\mathbb{X}^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. For fixed $L > 0$ and $T > 0$, and large enough r*

$$y_{r,m+1} \leq 3Ny_{r,m} \tag{3.53}$$

for $\omega \in \mathcal{K}^r$ and $m < \sqrt{|N^r|}T$, with the constant N chosen as in (3.39).

Let $y_r(m_r(t)) = y_{r,m_r(t)}$.

Lemma 3.16. *Let $\{\mathbb{X}^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. For fixed $L > 0$ and $T > 0$, and large enough r*

$$\sqrt{|N^r|}t - m_r(t) \geq Ly_r(m_r(t))/6N \tag{3.54}$$

for $\omega \in \mathcal{K}^r$ and $t \in \left(Ly_{r,0}/\sqrt{|N^r|}, T\right]$, with the constant N chosen as in (3.39).

Proof of Theorem 3.1. Assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of the π-parallel server system satisfies Assumption 4, and condition (3.12) holds.

Fix $\xi > 0$. By (3.40) and (3.44), there exists $r_0 > 0$ such that

$$P(\mathcal{K}^r) \geq P(\mathcal{L}^r) > 1 - \xi/2 \tag{3.55}$$

for all $r > r_0$. Fix $\epsilon > 0$ and take $L \geq 6Ns^*(\epsilon)$. Then, by (3.50) and Lemma 3.16, for $\omega \in \mathcal{K}^r$, $t \in \left(Ly_{r,0}/\sqrt{|N^r|}, T\right]$, and r large enough

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \leq 2\epsilon \left(y_r(m_r(t))\right)^c. \tag{3.56}$$

But, by (3.48),

$$y_r(m_r(t)) = \left|\hat{Q}^r\left(\frac{m_r(t)}{\sqrt{|N^r|}}\right)\right| \vee \left|\hat{Z}^r\left(\frac{m_r(t)}{\sqrt{|N^r|}}\right)\right| \vee 1 \leq \left\|\hat{Q}^r(\cdot)\right\|_T \vee \left\|\hat{Z}^r(\cdot)\right\|_T \vee 1. \tag{3.57}$$

From (3.51) and (3.57), for $t \in \left[0, Ly_{r,0}/\sqrt{|N^r|}\right]$ and $\omega \in \mathcal{L}^r$

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \leq \epsilon(y_{r,0})^c \leq \epsilon\left(\left\|\hat{Q}^r(\cdot)\right\|_T \vee \left\|\hat{Z}^r(\cdot)\right\|_T \vee 1\right)^c. \tag{3.58}$$

Combining (3.56), (3.57), and (3.58) gives

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \leq 2\epsilon\left(\left\|\hat{Q}^r(\cdot)\right\|_T \vee \left\|\hat{Z}^r(\cdot)\right\|_T \vee 1\right)^c \tag{3.59}$$

for all $t \in [0, T]$ and $\omega \in \mathcal{L}^r$. Finally, by (3.55) and (3.59), for large enough r,

$$P\left\{\frac{\left\|g\left(\hat{Q}^r(\cdot), \hat{Z}^r(\cdot)\right)\right\|_T}{\left(\left\|\hat{Q}^r(\cdot)\right\|_T \vee \left\|\hat{Z}^r(\cdot)\right\|_T \vee 1\right)^c} > 2\epsilon\right\} < \xi.$$

This clearly implies (3.13) since $\epsilon > 0$ and $\xi > 0$ are arbitrary. □

Proof of Theorem 3.4. Suppose that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, the hydrodynamic model of the π-parallel server system satisfies Assumption 4, and $\left|\hat{Q}^r(0)\right| \vee \left|\hat{Z}^r(0)\right|$ is stochastically bounded.

Let

$$u^{r,max} = \max_{i \in \mathcal{I}} \{u_i(m), m = 1, 2, \ldots : U_i(m-1) \leq 2|N^r||\lambda|L\},$$

where $\lambda = (\lambda_1, \ldots, \lambda_I)$ is given by (2.19). In words, $u^{r,max}$ is an upper bound, for r large enough, for the maximum interarrival time for those events that started before time L of the process $\{A_i : i \in \mathcal{I}\}$, since $\lambda_i^r < 2|N^r||\lambda|$ for large enough r. Assume for the moment that

$$u^{r,max}/\sqrt{|N^r|} \to 0 \text{ in probability as } r \to \infty \tag{3.60}$$

and that for some sequence $\{L^r\}$ that satisfies the conditions given in the theorem

$$\left| g\left(\hat{Q}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right), \hat{Z}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right)\right| \to 0 \text{ in probability as } r \to \infty. \tag{3.61}$$

Consider the sequence of processes $\{\mathbb{Y}^r\}$ defined by $\mathbb{Y}^r(\cdot) = \mathbb{X}^r(\frac{L^r}{\sqrt{|N^r|}} + \cdot)$. Then, $\{\mathbb{Y}^r\}$ satisfies (3.12) by (3.61). Also by (3.60), distributions of the first interarrival times of the processes A and S after $L^r/\sqrt{|N^r|}$ satisfy the conditions needed for Proposition 3.8 to be valid. Since the other conditions of Theorem 3.1 are satisfied by $\{\mathbb{Y}^r\}$, the proof above can be repeated to show that (3.13) holds for $\{\mathbb{Y}^r\}$. But, this shows that (3.15) holds for $\{\mathbb{X}^r\}$. Hence, it suffices to show that (3.60) and (3.61) hold.

The limits (3.60) are proven in Lemma 3.17.

Next we prove (3.61). We show that there exists a sequence $\{L^r\}$ with $L^r \to \infty$ as $r \to \infty$ and $L^r = o(\sqrt{|N^r|})$ such that for any $\epsilon > 0$ and $\xi > 0$, there exists r' such that

$$P\left\{\left| g\left(\hat{Q}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right), \hat{Z}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right)\right| > \epsilon\right\} < \xi, \tag{3.62}$$

for all $r > r'$.

Set $\delta_n = 1/n$ and $\tilde{L}^n = (N^n)^{1/4}$ for all $n = 1, 2, \ldots$. Define $\mathcal{K}^r_{\tilde{L}^n}$ as in Section 3.4.1; see (3.39) and the discussion succeeding to (3.39), with L being replaced with $\tilde{L}^n$. Note that, by the definition of $\mathcal{K}^r_{\tilde{L}^n}$ and Proposition 3.14, there exists r_n such that for $r > r_n$

$$P\left\{\mathcal{K}^r_{\tilde{L}^n}\right\} > 1 - 1/n \tag{3.63}$$

and

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \le y^c_{r,m} H\left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t - m)\right) + \delta_n y^c_{r,m} \tag{3.64}$$

holds for all $t \in [0, T]$ and m satisfying (3.49).

Set $L^r = \tilde{L}^1$ and $\tilde{\mathcal{K}}^r = \mathcal{K}^r_{\tilde{L}^1}$ for $r \le r_2$, $L^r = \tilde{L}^n$ and $\tilde{\mathcal{K}}^r = \mathcal{K}^r_{\tilde{L}^n}$ for $r \in (r_n, r_{n+1}]$, and for $n = 2, 3, \ldots$. Note that $L^r = o(\sqrt{|N^r|})$, and $L^r \to \infty$ as $r \to \infty$. Furthermore,

$$\lim_{r\to\infty} P(\tilde{\mathcal{K}}^r) = 1.$$

Fix $\epsilon, \xi > 0$. Let

$$\mathcal{U}^r_R = \left\{ \left| \hat{Q}^r(0) \right| \vee \left| \hat{Z}^r(0) \right| < R \right\}. \tag{3.65}$$

Choose r_0 and $R > 1$ such that for $r \ge r_0$

$$P\left(\mathcal{U}^r_R\right) > 1 - \xi/2.$$

We fix R to this value for the rest of the proof.

Let r_1 be the smallest integer greater than r_0 that satisfies $\delta_{r_1} < \epsilon/(2R^c)$. Choose $r_2 > r_1$ such that for all $r > r_2$, $L^r > 2s^*(\delta_{r_1})R$, where s^* is defined as in (3.52).

For $t \in [R^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$, $m_r(t) = 0$ from (3.49) and

$$\sqrt{|N^r|}t \ge R^{-1}L^r y_{r,0}.$$

Hence, for $r > r_2$, by (3.64),

$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) \le 2\delta_{r_1} y^c_{r,0} < \epsilon \tag{3.66}$$

for all $t \in [R^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$ and $\omega \in \tilde{\mathcal{K}}^r \cap \mathcal{U}^r_R$.

Now observe that for $\omega \in \tilde{\mathcal{K}}^r \cap \mathcal{U}^r_R$, $L^r/\sqrt{|N^r|} \in [R^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$ for all $r \ge 1$. Hence, there exits $r' > r_2$ such that for $r > r'$

$$P\left\{ g\left(\hat{Q}^r(L^r/\sqrt{|N^r|}), \hat{Z}^r(L^r/\sqrt{|N^r|})\right) > \epsilon \right\} < \xi. \tag{3.67}$$

This gives (3.62), thus completes the proof of (3.61). □

Proof of Remark 3.5. Assume that $\{\mathbb{X}^r_\pi\}$ is a sequence of π-parallel server system processes that satisfy the conditions of Theorem 3.4. Also, assume that $g\left(\hat{Q}^r(0), \hat{Z}^r(0)\right)$ is stochastically bounded and H is bounded.

Fix $L > 0$. By assumption there exists a constant $B_0 > 0$ such that $\sup_{t\in[0,\infty)} H(t) < B_0$. By (3.45)

$$g(Q^{r,m}(t), Z^{r,m}(t)) < 2B_0$$

for all $t \in [0, L]$. This implies, similar to (3.50), that

$$\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{Ly_{r,0}/\sqrt{|N^r|}} < 2B_0 y^c_{r,0}. \tag{3.68}$$

for all $\omega \in \mathcal{K}^r$. Let $\mathcal{U}^r_R$ be defined as in (3.65). Since $|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)|$ is stochastically bounded by assumption, for $\epsilon > 0$ fixed, there exists $R > 0$ and $r_1 > 0$ such that $P(\mathcal{U}^r_R) > 1 - \epsilon$, for all $r > r_1$. For each fixed L, on $\mathcal{U}^r \cap \mathcal{K}^r$

$$\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{L/\sqrt{|N^r|}} \leq \|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{Ly_{r,0}/\sqrt{|N^r|}} \leq R.$$

Now choose the sequence $\{L^r\}$ as in the previous proof. Then,

$$\limsup_{r\to\infty} P\left\{\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{L^r/\sqrt{|N^r|}} > R\right\} < \epsilon.$$

Since ϵ and R is arbitrary, this completes the proof. □

3.5 Proofs of the results in Section 3.4

3.5.1 Proofs of the results in Section 3.4.1

3.5.1.1 *Proof of Proposition 3.8*

We observe as in Bramson [13] that it is enough to investigate the processes with index $m = 0$ and then to multiply the ensuing error bounds by the number of processes in each case; $\sqrt{|N^r|}T$. To see this, note that

$$A^{r,m}_i(t) = \frac{1}{\sqrt{x_{r,m}}}\left(E_i\left(\frac{\lambda^r_i}{|N^r|}(\sqrt{x_{r,m}}t + \sqrt{|N^r|}m)\right) - E_i\left(\frac{\lambda^r_i}{|N^r|}\sqrt{|N^r|}m\right)\right).$$

Let $u^{r,m}_i(1)$ be the first residual interarrival time of E_i after time $\frac{\lambda^r_i}{|N^r|}\sqrt{|N^r|}m$.

$$\begin{aligned}
&P\left\{\left\|A^{r,m}_i(t) - \frac{\lambda^r_i}{|N^r|}t\right\|_L > 2\epsilon L\right\} \\
&= P\left\{\left\|\frac{1}{\sqrt{x_{r,m}}}\left(E_i\left(t + \sqrt{|N^r|}m\frac{\lambda^r_i}{|N^r|}\right) - E_i\left(\frac{\lambda^r_i}{|N^r|}\sqrt{|N^r|}m\right)\right) - \frac{t}{\sqrt{x_{r,m}}}\right\|_{L\sqrt{x_{r,m}}\lambda^r_i/|N^r|}\right. \\
&\qquad\qquad \left. > 2\epsilon L\right\} \\
&\leq P\left\{\left\|E_i\left(t + \sqrt{|N^r|}m\frac{\lambda^r_i}{|N^r|}\right) - E_i\left(\frac{\lambda^r_i}{|N^r|}\sqrt{|N^r|}m\right) - (t - u^{r,m}_i(1))\right\|_{L\sqrt{x_{r,m}}\lambda^r_i/|N^r|}\right. \\
&\qquad\qquad \left. > \sqrt{x_{r,m}}\epsilon L\right\} + P\left\{\frac{u^{r,m}_i(1)}{\sqrt{x_{r,m}}} > \epsilon L\right\}.
\end{aligned} \tag{3.69}$$

We show below that the first term on the right hand side is bounded by $\epsilon/(L\sqrt{|N^r|})$. So it remains to be shown that the second term goes to zero as $r \to \infty$. That is, we need to show that $\{u_i^{r,m}(1), i \in \mathcal{I}\}$ satisfies

$$u_i^{r,m}(1)/\sqrt{|N^r|} \to 0 \text{ in probability as } r \to \infty, \text{ for all } i \in \mathcal{I} \tag{3.70}$$

for all $m < \sqrt{|N^r|}T$.

For each departure process, we have

$$D_{jk}^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(S_{jk}\left(T_{jk}^r\left(\frac{\sqrt{x_{r,m}}}{|N^r|}t + \frac{m}{\sqrt{|N^r|}}\right)\right) - S_{jk}\left(T_{jk}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right)\right) \tag{3.71}$$

for $j \in \mathcal{J}, k \in \mathcal{K}(j)$. Hence, by restarting the process at time $m/\sqrt{|N^r|}$, we have that the only condition to be checked is whether the residual time of the first arrival for S_{jk} after time $T_{jk}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \in [0, |N^r|T]$ satisfies a similar condition to (3.70).

The following lemma, taken from Bramson [13], shows that (2.17) holds. Let

$$u_i^{r,T,max} = \max\{|u_i(l)| : U_i(l-1) \le 2|\lambda||N^r|T\}, \text{ for all } i \in \mathcal{I} \text{ and}$$
$$v_{jk}^{r,T,max} = \max\{|v_{jk}(l)| : V_{jk}(l-1) \le |N^r|T\}, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

Lemma 3.17. *Assume that (3.70) holds and that $\lambda^r/|N^r|$ is bounded. Then, for given T,*

$$u_i^{r,T,max}/\sqrt{|N^r|} \to 0 \text{ in probability as } r \to \infty, \text{ for all } i \in \mathcal{I} \text{ and}$$
$$v_{jk}^{r,T,max}/\sqrt{|N^r|} \to 0 \text{ in probability as } r \to \infty, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

Proof. The proofs immediately follow by taking $r = \sqrt{2|\lambda||N^r|}$ and $r = \sqrt{|N^r|}$, respectively, in Lemma 5.1 of Bramson [13]. □

Fix $\epsilon > 0$, $L > 0$, and $T > 0$. We prove each bound separately.

Proof of (3.35). Fix $i \in \mathcal{I}$. Similar to (5.31) in Bramson [13], using Lemma 3.17, for given $\epsilon > 0$ and large enough r,

$$P\left(\|E_i(t) - t\|_{2|\lambda|L\sqrt{x_{r,0}}} \ge 2|\lambda|\epsilon L\sqrt{x_{r,0}}\right) \le \frac{\epsilon}{|2\lambda|L\sqrt{x_{r,0}}}.$$

And for r large enough,

$$\begin{aligned}\frac{1}{\sqrt{x_{r,0}}}\left\|E_i(t)-t\right\|_{2|\lambda|L\sqrt{x_{r,0}}} &\geq \left\|\frac{A_i^r(\frac{\sqrt{x_{r,0}}t}{|N^r|})}{\sqrt{x_{r,0}}}-\frac{\lambda_i^r}{|N^r|}t\right\|_L \\ &= \left\|A_i^{r,0}(t)-\frac{\lambda_i^r}{|N^r|}t\right\|_L.\end{aligned}$$

Hence,

$$P\left\{\left\|A_i^{r,0}(t)-\frac{\lambda_i^r}{|N^r|}t\right\|_L>2\epsilon L\right\}\leq\frac{\epsilon}{2|\lambda|L\sqrt{x_{r,0}}}\leq\frac{2\epsilon}{L\sqrt{|N^r|}}$$

and so

$$P\left\{\left\|A^{r,0}(t)-\frac{\lambda^r}{|N^r|}t\right\|_L>\epsilon L\right\}\leq\frac{2I\epsilon}{2|\lambda|L\sqrt{|N^r|}}.$$

Multiplying the error bounds by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ by a factor of $2I(L\vee T)$ we obtain (3.35). □

Proof of (3.36). Fix $j\in\mathcal{J}$ and $k\in\mathcal{K}(j)$. We first show that for r large enough

$$\begin{aligned}P\left\{\sup_{0\leq t_1\leq t_2\leq L}\left(S_{jk}\left(2\beta_j\sqrt{x_{r,0}}t_2\right)-S_{jk}\left(2\beta_j\sqrt{x_{r,0}}t_1\right)\right)\geq 2\beta_j\sqrt{x_{r,0}}\frac{(t_2-t_1)}{\mu_{jk}}+4\sqrt{x_{r,0}}\beta_jL\epsilon\right\}\\ \leq\frac{4\epsilon}{\beta_jL\sqrt{|N^r|}}.\end{aligned}\tag{3.72}$$

By Proposition 4.3 of Bramson [13] and Lemma 3.17, for large enough r,

$$P\left\{\left\|S_{jk}(t)-\frac{t}{\mu_{jk}}\right\|_{2\beta_jL\sqrt{x_{r,0}}}\geq 2\beta_jL\sqrt{x_{r,0}}\epsilon\right\}\leq 2\frac{\epsilon}{\beta_jL\sqrt{x_{r,0}}}.$$

Then,

$$\begin{aligned}P\Bigg\{\sup_{0\leq t_1\leq t_2\leq L}\Bigg(\left(S_{jk}\left(2\beta_j\sqrt{x_{r,0}}t_2\right)-\frac{2\beta_j\sqrt{x_{r,0}}t_2}{\mu_{jk}}\right)-\left(S_{jk}\left(2\beta_j\sqrt{x_{r,0}}t_1\right)-\frac{2\beta_j\sqrt{x_{r,0}}t_1}{\mu_{jk}}\right)\Bigg)\\ \geq 4\sqrt{x_{r,0}}\beta_jL\epsilon\Bigg\}\leq\frac{4\epsilon}{\beta_jL\sqrt{x_{r,0}}}\leq\frac{4\epsilon}{\beta_jL\sqrt{|N^r|}}.\end{aligned}$$

This gives (3.72). Next, we show that

$$\begin{aligned}P\Bigg\{\sup_{0\leq t_1\leq t_2\leq L}\left(S_{jk}\left(T_{jk}\left(\frac{\sqrt{x_{r,0}}t_2}{|N^r|}\right)\right)-S_{jk}\left(T_{jk}\left(\frac{\sqrt{x_{r,0}}t_1}{|N^r|}\right)\right)\right)\\ \geq\beta_j\sqrt{x_{r,0}}\frac{(t_2-t_1)}{\mu_{jk}}+5\sqrt{x_{r,0}}\beta_jL\epsilon\Bigg\}\leq\frac{5\epsilon}{L\beta_j\sqrt{x_{r,0}}}.\end{aligned}\tag{3.73}$$

We prove (3.73) by showing that the event in (3.73) is included in (3.72). Assume that for $\omega \in \Omega$

$$S_{jk}\left(T_{jk}\left(\frac{\sqrt{x_{r,0}}t_2}{|N^r|}\right)\right) - S_{jk}\left(T_{jk}\left(\frac{\sqrt{x_{r,0}}t_1}{|N^r|}\right)\right) \geq \beta_j\sqrt{x_{r,0}}\frac{(t_2-t_1)}{\mu_{jk}} + 4\sqrt{x_{r,0}}\beta_j L\epsilon \quad (3.74)$$

for some $0 \leq t_1 \leq t_2 \leq L$. Let

$$\tau_l = T_{jk}\left(\frac{\sqrt{x_{r,0}}t_l}{|N^r|}, \omega\right)$$

for $l = 1, 2$. Then, for r large enough

$$0 \leq \tau_1 \leq \tau_2 \leq 2L\sqrt{x_{r,0}}\beta_j \text{ and} \quad (3.75)$$

$$\tau_2 - \tau_1 \leq 2\sqrt{x_{r,0}}\beta_j(t_2 - t_1). \quad (3.76)$$

By (3.74) and (3.76)

$$S_{jk}\left(2\beta_j\sqrt{x_{r,0}}\frac{\tau_2}{2\beta_j\sqrt{x_{r,0}}}\right) - S_{jk}\left(2\beta_j\sqrt{x_{r,0}}\frac{\tau_1}{2\beta_j\sqrt{x_{r,0}}}\right) \geq 2\beta_j\sqrt{x_{r,0}}\frac{\frac{\tau_2}{2\beta_j\sqrt{x_{r,0}}} - \frac{\tau_1}{2\sqrt{\beta_j x_{r,0}}}}{\mu_{jk}} + 4\sqrt{x_{r,0}}\beta_j L\epsilon.$$

By (3.75), $0 \leq \frac{\tau_1}{2\beta_j\sqrt{x_{r,0}}} \leq \frac{\tau_2}{2\beta_j\sqrt{x_{r,0}}} \leq L$. Using this and (3.76), we get that ω also satisfies the inequality in (3.69). Thus we have (3.73). By (3.71), this implies, by reselecting ϵ, that

$$P\left\{\sup_{t_1,t_2\in[0,L]}\left|D^{r,0}(t_2) - D^{r,0}(t_1)\right| \geq N|t_2 - t_1| + \epsilon\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}, \quad (3.77)$$

with $N = \max_{j\in\mathcal{J},k\in\mathcal{K}(j)}\{\beta_j/\mu_{jk}\}$. Multiplying the exceptional probability by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately we obtain (3.36). □

Proof of (3.37). By setting $\epsilon = 1$, $t_2 = L$, and $t_1 = 0$ in (3.77), we have that

$$P\left\{D^r_{jk}\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right) \geq 2NL\sqrt{x_{r,0}}\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \quad (3.78)$$

Off of the exceptional set given in (3.78)

$$D^r_{jk}\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right) + 1 \leq 3NL\sqrt{x_{r,0}}.$$

Let $a = 0$ or 1. It follows from Proposition 4.2 of Bramson [13] that for large enough n

$$P\left\{\left\|V_{jk}(l) - \frac{l}{\mu_{jk}}\right\|_n \geq \epsilon n\right\} \leq \frac{\epsilon}{n}.$$

By setting $n = 3NL\sqrt{x_{r,0}}$, we get

$$P\left\{\left\|V_{jk}(D^r_{jk}(t)+a) - \frac{D^r_{jk}(t)}{\mu_{jk}}\right\|_{\frac{\sqrt{x_{r,0}}}{|N^r|}L} \geq 3NL\sqrt{x_{r,0}}\epsilon\right\} \leq \frac{2\epsilon}{3NL\sqrt{x_{r,0}}} \leq B_2\frac{\epsilon}{\sqrt{|N^r|}}$$

for $B_2 \geq 2/3NL$. By enlarging ϵ appropriately, we get for $\tilde{b} = (1,0)$ or $(0,0)$

$$P\left\{\left\|V^{r,0}_{jk}(D^{r,0}_{jk}(t),\tilde{b}) - \frac{D^{r,0}_{jk}(t)}{\mu_{jk}}\right\|_L \geq \epsilon\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}.$$

Multiplying the exceptional probability by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately we obtain

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V^{r,m}_{jk}(D^{r,m}_{jk}(t),\tilde{b}) - \frac{D^{r,m}_{jk}(t)}{\mu_{jk}}\right\|_L \geq \epsilon\right\} \leq \epsilon. \tag{3.79}$$

For $b = (0,1)$ and $\tilde{b} = (0,0)$, by (3.25)

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V^{r,m}_{jk}(D^{r,m}_{jk}(t),\tilde{b}) - V^{r,m}_{jk}(D^{r,m}_{jk}(t),b)\right\|_L \geq \epsilon\right\} \leq$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left|V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right)\right) - V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right)+1\right)\right| \geq \sqrt{x_{r,m}}\epsilon\right\}. \tag{3.80}$$

Observe that, by (2.15), $V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right)\right) \leq |N^r|T$ and by Lemma 3.17

$$P\left\{v^{r,T,max}_{jk} \geq \sqrt{x_{r,m}}\epsilon\right\} \leq \epsilon \tag{3.81}$$

for large enough r. Thus, we get (3.77) by combining (2.15) with (3.79)-(3.81). □

3.5.1.2 ***Proof of Proposition 3.9***

Proof. We use the bounds established in (3.35)-(3.37). Fix L, T, and $\epsilon > 0$. Choose r large enough so that (3.35)-(3.37) hold with $\epsilon/(3d)$. Let $\mathcal{V}^r$ be the intersection of the complements of the events given in (3.35)-(3.37), so $P\{\mathcal{V}^r\} > 1-\epsilon$. We show that for r large enough and all $\omega \in \mathcal{V}^r$

$$\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\leq L}|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)| \leq \tilde{N}|t_1 - t_2| + \epsilon \tag{3.82}$$

for some $\tilde{N}$ that only depends on (I, J, K, λ). We fix $\omega \in \mathcal{V}^r$ for the rest of the proof and so omit it from the notation. Let $t_1, t_2 \in [0, T]$ and $m \geq 0$. We first show that for any $j \in \mathcal{J}$, and $k \in \mathcal{K}(j)$

$$\left| Z_{jk}^{r,m}(t_2) - Z_{jk}^{r,m}(t_1) \right| \leq N_0 |t_2 - t_1| + \epsilon \tag{3.83}$$

for some $N_0 > 0$. Since $B_{jk}^{r,m}$ is nondecreasing we have by (3.29) that

$$0 \leq B_{jk}^{r,m}(t_2) - B_{jk}^{r,m}(t_1) \leq \sum_{l \in \mathcal{K}(j)} \left(D_{jl}^{r,m}(t_2) - D_{jl}^{r,m}(t_1) \right). \tag{3.84}$$

Combining (3.84) with (3.29) yields

$$\left| Z_{jk}^{r,m}(t_2) - Z_{jk}^{r,m}(t_1) \right| \leq K \left| D^{r,m}(t_2) - D^{r,m}(t_1) \right| + I \left| A^{r,m}(t_2) - A^{r,m}(t_1) \right|.$$

By (3.35), $|A^{r,m}(t_2) - A^{r,m}(t_1)| < 2|\lambda||t_2 - t_1| + \epsilon$ for r large enough. By setting $N_0 = KN + 2I|\lambda|$ and using (3.36), we get (3.83). Equation (3.84) gives that

$$\left| B_{jk}^{r,m}(t_2) - B_{jk}^{r,m}(t_1) \right| \leq N_0 |t_2 - t_1| + \epsilon.$$

Combining this with (3.28) gives

$$\left| Q_{k}^{r,m}(t_2) - Q_{k}^{r,m}(t_1) \right| \leq N_1 |t_2 - t_1| + \epsilon,$$

for $N_1 = (I + J)N_0$. Observe that for any $i \in \mathcal{I}$, $k \in \mathcal{K}$, and $j \in \mathcal{J}(k)$

$$|A_{ik}^{r,m}(t_2) - A_{ik}^{r,m}(t_1)| \leq |A_{i}^{r,m}(t_2) - A_{i}^{r,m}(t_1)|,$$
$$|A_{ijk}^{r,m}(t_2) - A_{ijk}^{r,m}(t_1)| \leq |A_{i}^{r,m}(t_2) - A_{i}^{r,m}(t_1)|.$$

Also, for any $j \in \mathcal{J}$ and $k \in \mathcal{K}(j)$ and for r large enough

$$\|T_{jk}^{r,m}(t_2) - T_{jk}^{r,m}(t_1)\|_L \leq 2\beta_j |t_2 - t_1| \text{ and } \|Y_{j}^{r,m}(t_2) - Y_{j}^{r,m}(t_1)\|_L \leq 2\beta_j |t_2 - t_1|.$$

Note that, by the definition of $\mathcal{V}^r$, the inequalities above hold for all $m < \sqrt{|N^r|}T$. This shows that (3.82) holds, for r large enough, with $\tilde{N} = N_1 \vee 2$. □

3.5.2 Proofs of the results in Section 3.4.2

3.5.2.1 ***Proof of Proposition 3.13***

Proof. Assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, and the hydrodynamic model of the π-parallel server systems satisfies Assumption 4.

Fix $L > 0$ and let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{E}$. Note that since $\tilde{\mathbb{X}}$ is a limit of hydrodynamically scaled processes $|\tilde{\mathbb{X}}(0)| \leq 1$ by (3.41). Also, by Proposition 3.12, $\tilde{\mathbb{X}}$ satisfies the hydrodynamic model equations (3.1)-(3.8) for all $t \in [0, L]$. Thus, using (3.1), (3.2), (3.5), and the fact that $|\tilde{\mathbb{X}}(0)| \leq 1$, one can show that there exists $R_L > 0$ such that

$$\|\tilde{\mathbb{X}}(t)\|_L \leq R_L. \tag{3.85}$$

Fix $\epsilon > 0$. Since g is continuous, there exists $\delta > 0$ such that

$$|g(x) - g(y)| < \epsilon \tag{3.86}$$

if $|x - y| < \delta$ and $x, y \in [-2R_L, 2R_L]^{I+d_z}$.

Fix $0 < \delta < R_L$ as given above and choose r large enough so that (3.42) holds for all $\omega \in \mathcal{K}^r$ and any $m < \sqrt{|N^r|}T$, that is;

$$\left\|\mathbb{X}^{r,m}(\cdot) - \tilde{\mathbb{X}}(\cdot)\right\|_L \leq \delta \tag{3.87}$$

for some hydrodynamic limit $\tilde{\mathbb{X}}$ of $\mathcal{E}$. Hence, by (3.85),

$$\|\mathbb{X}^{r,m}(t)\|_L \leq 2R_L. \tag{3.88}$$

By (3.85)-(3.88) and Assumption 4 we have for all $t \in [0, L]$ that

$$g\left(Q^{r,m}(t), Z^{r,m}(t)\right) \leq H(t) + \epsilon.$$

Result (3.46) is proven similarly. Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{E}_g$. Then, there exits a sequence $\{\mathbb{X}^{r_k,0}\}$, where $\{r_k\}$ is a subsequence of $\{r\}$, such that

$$\|\mathbb{X}^{r_k,0}(\cdot) - \tilde{\mathbb{X}}(\cdot)\| \to 0 \tag{3.89}$$

as $k \to \infty$. But by definition of $\mathcal{E}_g$, $g\left(\tilde{Q}^{r_k,0}(0), \tilde{Z}^{r_k,0}(0)\right) \to 0$. This implies by the continuity of g and (3.89) that $g\left(\tilde{Q}(0), \tilde{Z}(0)\right) = 0$. Thus, by the last statement in Assumption 4,

$$\left\|g\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right\|_L = 0. \tag{3.90}$$

This shows that (3.90) holds for every hydrodynamic limit of $\mathcal{E}_g$

One can show as in Corollary 3.11 that hydrodynamic limits of $\mathcal{E}_g$ are rich in $\mathcal{E}_g$. Hence, for large enough r and $\omega \in \mathcal{L}^r$

$$\left\| \mathbb{X}^{r,0}(\cdot) - \tilde{\mathbb{X}}(\cdot) \right\|_L \leq \delta$$

for some hydrodynamic limit $\tilde{\mathbb{X}} \in \mathcal{E}$ of $\mathcal{E}_g$. Using (3.86) we have

$$g\left(Q^{r,0}(t), Z^{r,0}(t)\right) \leq \epsilon$$

for all $t \in [0, L]$.

The validity of (3.44) when (3.12) holds is already proved before the proposition. □

3.5.3 Proofs of the results in Section 3.4.3

3.5.3.1 ***Proof of Lemma 3.15***

Proof. For $\omega \in \mathcal{K}^r$ and r chosen large enough it follows from (3.39) that

$$|Q^{r,m}(t_2) - Q^{r,m}(t_1)| \leq N|t_2 - t_1| + 1$$

for $t_1, t_2 \in [0, L]$ and $m < \sqrt{|N^r|}T$. Setting $t_1 = 0$ and $t_2 = 1/y_{r,m}$ and applying (3.48) to the above inequality gives

$$\left| Q^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) - Q^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \leq \sqrt{x_{r,m}}\frac{N}{y_{r,m}} + \sqrt{x_{r,m}}$$

and so

$$\left| \hat{Q}^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) \right| - \left| \hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \leq N + y_{r,m} \leq 2Ny_{r,m}.$$

The same argument gives

$$\left| \hat{Z}^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) \right| - \left| \hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \leq 2Ny_{r,m}.$$

Hence

$$\begin{aligned} y_{r,m+1} &\leq \left(\left| \hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \vee \left| \hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \vee 1 \right) + 2Ny_{r,m} \\ &\leq 3Ny_{r,m}, \end{aligned}$$

which yields (3.53). □

3.5.3.2 *Proof of Lemma 3.16*

Proof. Let $t \in \left(Ly_{r,0}/\sqrt{|N^r|}, T\right]$. It follows from the definition of $m_r(t)$ that $m_r(t) \geq 1$. So,

$$\sqrt{|N^r|}t - (m_r(t) - 1) \geq Ly_r(m_r(t) - 1).$$

Setting $m = m_r(t) - 1$ in Lemma 3.15, one has

$$\sqrt{|N^r|}t - m_r(t) \geq Ly_r(m_r(t) - 1) - 1 \geq \frac{L}{3N} y_r(m_r(t)) - 1 \geq \frac{L}{6N} y_r(m_r(t))$$

assuming $L \geq 6N$ as in (3.52). □

3.5.4 Proof of Proposition 3.12

We need the following lemma to prove that the departure process of a hydrodynamic limit satisfies the associated hydrodynamic model equation. Recall that we denote by $\mathcal{A}$ the largest subset of Ω whose elements satisfy (A.14) and $P(\mathcal{A}) = 1$.

Lemma 3.18. *Let $\{\mathbb{X}^r\}$ be a sequence of π parallel server system processes. Assume that Assumption 1 holds and π satisfies Assumption 2. Fix $\varepsilon > 0$, $L > 0$ and $T > 0$. Then, for large enough r and $\omega \in \mathcal{A}$*

$$\max_{m < \sqrt{|N^r|}T} \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^L \left| Z_{jk}^{r,m}(s) \right| ds < \varepsilon, \ \forall\ j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

Proof. Fix $\omega \in \mathcal{A}$ and $\varepsilon > 0$. Let z be given as in Assumption 2. For $m < \sqrt{|N^r|}T$, by (3.23),

$$\frac{\sqrt{x_{r,m}}}{|N^r|} \leq \left\| \frac{Q^r(t)}{|N^r|} \right\|_T \vee \left\| \frac{Z^r(t)}{|N^r|} - \frac{\vec{N}^r}{|N^r|} x^* \right\|_T \vee \frac{1}{\sqrt{|N^r|}}.$$

By Lemma A.3, $\limsup_{r\to\infty} \left\| \frac{Q^r(t)}{|N^r|} \right\|_T \vee \left\| \frac{Z^r(t)}{|N^r|} - \frac{\vec{N}^r}{|N^r|} x^* \right\|_T = 0$. Hence, for r large enough

$$\frac{\sqrt{x_{r,m}}}{|N^r|} \leq \varepsilon. \tag{3.91}$$

Similarly for r large enough

$$\left\| \frac{Q^r(t)}{|N^r|} \right\|_{L\varepsilon+T} \vee \left\| \frac{Z^r(t)}{|N^r|} - \frac{\vec{N}^r}{|N^r|} x^* \right\|_{L\varepsilon+T} < \frac{\varepsilon}{L}. \tag{3.92}$$

Choose r large enough so that (3.91) and (3.92) hold. Then, for such r and for $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$

$$\max_{m<\sqrt{|N^r|}T} \tfrac{1}{|N^r|} \int_0^L \left| Z^r_{jk}\left(\tfrac{\sqrt{x_{r,m}}}{|N^r|}s + \tfrac{m}{\sqrt{|N^r|}}\right) - x^*_{jk}N^r_j \right| \mathrm{d}s \quad \leq L \left\| \frac{Z^r(t)}{|N^r|} - x^*_{jk}\frac{N^r_j}{|N^r|} \right\|_{L\varepsilon+T} < \varepsilon.$$

□

Remark 3.19. Let $\{\epsilon(r)\}$ be a sequence with $\epsilon(r) \to 0$ as $r \to \infty$. Define

$$\Theta^r = \left\{ \max_{m<\sqrt{|N^r|}T} \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^T \left| Z^{r,m}_{jk}(s) \right| \mathrm{d}s < \epsilon(r) \right\}.$$

For $\epsilon(r) \to 0$ slowly enough as $r \to \infty$, $\lim_{r\to\infty} P(\Theta^r) = 1$, by Lemma 3.18. Hence,

$$\lim_{r\to\infty} P(\Theta^r \cap \mathcal{K}^r) = 1, \tag{3.93}$$

by Corollary 3.10, where $\mathcal{K}^r$ defined as in Section 3.4.1. With a slight abuse of notation, we set $\mathcal{K}^r = \Theta^r \cap \mathcal{K}^r$ for simplicity.

Proof of Proposition 3.12. Proof is similar to that of Proposition 6.2 in Bramson [13]. Assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 3, and (3.12) holds.

Fix $\omega \in \mathcal{K}^r$ and let $\mathbb{X}^{r,m}$ be given as in Section 3.4.1. By (3.35′), we have for large enough r that

$$\left\| A^{r,m}(t) - \frac{\lambda^r}{|N^r|}t \right\|_L \leq \epsilon(r). \tag{3.94}$$

Combining (3.26) with (3.37′) gives

$$\left\| T^{r,m}_{jk}(t) - \frac{1}{\mu_{jk}} D^{r,m}_{jk}(t) \right\|_L \leq \epsilon(r). \tag{3.95}$$

Recall that $z_{jk} = \beta_j x^*_{jk}$. Using (3.93), (3.31) and Remark 3.19 gives

$$\left\| T^{r,m}_{jk}(t) - z_{jk}t \right\|_L \leq \epsilon(r). \tag{3.96}$$

Now select any hydrodynamic limit $\tilde{\mathbb{X}}$ of $\mathcal{E}$. For given $\delta > 0$, choose (r, m) so that, $\epsilon(r) \leq \delta$,

$$\left\| \tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t, w) \right\|_L \leq \delta, \tag{3.97}$$

and

$$\left|\frac{\lambda^r}{|N^r|} - \lambda\right| \leq \delta. \tag{3.98}$$

It follows from (3.94) and (3.98) that

$$\left\|\tilde{A}(t) - \lambda t\right\|_L \leq 2\delta \tag{3.99}$$

and from (3.95) and (3.96) that

$$\left\|\tilde{D}_{jk}(t) - \mu_{jk} z_{jk} t\right\|_L \leq 2\delta. \tag{3.100}$$

By combining (3.97), (3.99), (3.27), and (3.28), we get

$$\left\|\lambda_i - \textstyle\sum_{k\in\mathcal{K}} \tilde{A}_{ik}(t) - \sum_{k\in\mathcal{K}, j\in\mathcal{K}(k)} \tilde{A}_{ijk}(t)\right\|_L \leq 2KJ\delta \text{ and} \tag{3.101}$$

$$\left\|\tilde{Q}_k(t) - \tilde{Q}_k(0) - \textstyle\sum_{i\in\mathcal{I}} \tilde{A}_{ik}(t) + \sum_{j\in\mathcal{J}(k)} \tilde{B}_{jk}(t)\right\|_L \leq 4IJ\delta. \tag{3.102}$$

By combining (3.97) with (3.100) and (3.29), we get

$$\left\|\tilde{Z}_{jk}(t) - \tilde{Z}_{jk}(0) - \sum_{i\in\mathcal{I}} \tilde{A}_{ijk}(t) - \tilde{B}_{jk}(t) + \mu_{jk} z_{jk} t\right\|_L \leq 6I\delta. \tag{3.103}$$

Equations (3.100)-(3.103) show that the hydrodynamic limits satisfy (3.1), (3.2), and (3.5). Equations (3.3) and (3.4) are clearly satisfied by the hydrodynamic limits.

The fact that the hydrodynamic limits satisfy (3.7) and (3.8) is proved similarly to the fact that the fluid limits satisfy the fluid analogs of those equations. Hence, we only illustrate the proof of (3.7).

Fix a hydrodynamic limit $\tilde{\mathbb{X}}$. By the definition of a hydrodynamic limit, there exists a sequence (r_l, m_l, ω_l), with $\omega_l \in \mathcal{K}^l$ for all $l \geq 0$, such that

$$\mathbb{X}^{r_l, m_l}(\cdot, \omega_l) \to \tilde{\mathbb{X}}(\cdot) \tag{3.104}$$

u.o.c. as $l \to \infty$. Fix $t > 0$. If $\tilde{Q}_k(t) = 0$, (3.7) holds trivially. Now we assume that $\tilde{Q}_k(t) > a$ for some $a > 0$. By (3.104), there exists an l_0 such that

$$Q_k^{r_l, m_l}(t, \omega_l) > a/2$$

for all $l > l_0$. This implies, by (3.33), that

$$\sum_{j\in\mathcal{J}(k)}\sum_{l\in\mathcal{K}(j)}\tilde{Z}_{jl}^{r_l,m_l}(t,\omega_l)=0.$$

Hence

$$Q_k^{r_l,m_l}(t,\omega_l)\sum_{j\in\mathcal{J}(k)}\sum_{l\in\mathcal{K}(j)}\tilde{Z}_{jl}^{r_l,m_l}(t,\omega_l)=0. \tag{3.105}$$

Convergence in (3.104) implies that

$$Q_k^{r_l,m_l}(t,\omega_l)\sum_{j\in\mathcal{J}(k)}\sum_{l\in\mathcal{K}(j)}\tilde{Z}_{jl}^{r_l,m_l}(t,\omega_l)\to\tilde{Q}_k(t)\sum_{j\in\mathcal{J}(k)}\sum_{l\in\mathcal{K}(j)}\tilde{Z}_{jl}(t)\ \text{ as } l\to\infty.$$

This gives (3.7) by (3.105). □

3.6 Proof of Theorem 3.7

In the rest of this section we assume that Assumption 1 holds, π satisfies Assumption 2, g satisfies Assumption 5, the hydrodynamic limits of π-parallel server system satisfies Assumption 7, Assumption 6 holds and

$$g(\hat{Q}^r(0),\hat{Z}^r(0))\to 0 \text{ in probability}$$

as $r\to\infty$.

3.6.1 Hydrodynamic Limits on $\mathcal{A}_R^r$

Fix $T>0$ and $\epsilon>0$. We will show that for r large enough

$$P\left\{\left\|g(\hat{Q}^r(t),\hat{Z}^r(t))\right\|_T>\epsilon\right\}<\eta,$$

where $\eta>0$ is arbitrary. Note that this implies the conclusion of Theorem 3.7. Hence, we also fix $\eta>0$ for the rest of the proof.

Choose r_0 and $R>R_0$ large enough so that for all $r>r_0$

$$P\left(\mathcal{A}_R^r(T)\right)>1-\eta/2.$$

Fix $R>0$ to this value.

For any nonnegative integer $m < \sqrt{|N^r|}T$, let

$$x_{r,m} = \left| Q^r \left(\frac{m}{\sqrt{|N^r|}} \right) \right|^2 \vee \left| Z^r \left(\frac{m}{\sqrt{|N^r|}} \right) - \vec{N}^r x^* \right|^2 \vee (R^2 |N^r|)$$

The difference between this definition and the definition (3.23) is the last term.

We note that

$$x_{r,m} = R^2 |N^r| \tag{3.106}$$

on $\mathcal{A}_R^r(T)$ for $m < \sqrt{|N^r|}T$. We define the hydrodynamic scaling as in (3.24) and (3.25). Observe that equations (3.27)-(3.34) are still valid. Fix $L > 0$. The results in Proposition 3.8 still hold, hence so does the result in Proposition 3.9.

We redefine $\mathcal{K}^r(T)$ to be the intersection of $\mathcal{K}_0^r$ with $\mathcal{A}_R^r(T)$ and complements of (3.35′) and (3.37′). As in Corollary 3.10

$$\lim_{r \to \infty} P\left(\mathcal{K}^r(T)\right) > 1 - \eta/2.$$

Let

$$E^r = \{\mathbb{X}^{r,m}, m < \sqrt{|N^r|}T, w \in \mathcal{K}^r(T)\}.$$

Corollary 3.11 holds on $\mathcal{K}^r(T)$ with E^r defined as above and

$$\mathcal{E} = \{E^r : r \in \mathbb{N}\}.$$

since (3.41) holds. As described in Remark 3.6, we call the hydrodynamic limits in this case the hydrodynamic limits on $\mathcal{A}_R^r(T)$. Observe that the hydrodynamic limits on $\mathcal{A}_R^r(T)$ also satisfy hydrodynamic model equations (3.1)-(3.8) by Proposition 3.12.

Next we establish a similar result to Proposition 3.13. First note that on $\mathcal{K}^r(T)$

$$\begin{aligned} &g\left(\hat{Q}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right), \hat{Z}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right)\right) = \\ &g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\left(Q^{r,m}(t), Z^{r,m}(t)\right)\right) = g\left(R\left(Q^{r,m}(t), Z^{r,m}(t)\right)\right) \end{aligned} \tag{3.107}$$

for $m < \sqrt{|N^r|}T$ by (3.106) since $\mathcal{K}^r(T) \subset \mathcal{A}_R^r(T)$.

Therefore

$$g\left(\hat{Q}^r\left(0\right),\hat{Z}^r\left(0\right)\right)=g\left(R\left(Q^{r,0}\left(0\right),Z^{r,0}\left(0\right)\right)\right)$$

on $\mathcal{K}^r(T)$.

Let $\mathcal{L}^r(T)=\mathcal{K}^r(T)\bigcap\mathcal{G}^r$ where

$$\mathcal{G}^r=\left\{\left|g\left(R\left(Q^{r,0}(0),Z^{r,0}(0)\right)\right)\right|\leq\epsilon(r)\right\},$$

with $\epsilon(r)\to 0$ slowly enough as $r\to 0$ so that

$$\lim_{r\to\infty}P\left(\mathcal{L}^r\right)>1-\eta/2.$$

As in Proposition 3.13, using (3.11) and the continuity of g we have for $r>r_0$ large enough that

$$g\left(R\left(Q^{r,m}\left(t\right),Z^{r,m}\left(t\right)\right)\right)\leq H_{R,T}(t)+\epsilon,\ t\in[0,L] \tag{3.108}$$

on $\mathcal{K}^r(T)$. Using the second part of Assumption 7, similar to (3.46) we have

$$\left\|g\left(R\left(Q^{r,0}(t),Z^{r,0}(t)\right)\right)\right\|_L\leq\epsilon \tag{3.109}$$

on $\mathcal{L}^r(T)$ for r large enough.

3.6.2 State space collapse in diffusion limits

Let

$$y_{r,m}=\sqrt{\frac{x_{r,m}}{|N^r|}}$$

We begin with changing the scaling using (3.107). As in Proposition 3.14 we have from (3.108) and (3.107) that

$$g\left(Q^r\left(t\right),Z^r\left(t\right)\right)\leq H_{R,T}\left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t-m)\right)+\epsilon$$

for $w\in\mathcal{K}^r(T)$, r large enough and

$$\frac{m}{\sqrt{|N^r|}}\leq t\leq\frac{1}{\sqrt{|N^r|}}(y_{r,m}L+m). \tag{3.110}$$

Also by (3.109) we have that

$$\left\| g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)\right\|_{Ly_{r,0}/\sqrt{|N^r|}} \leq \epsilon$$

on $\mathcal{L}^r(T)$ for r large enough.

Let $m_r(t)$ be the smallest m that satisfies (3.110) with t and $y_r(m_r(t)) = y_{r,m_r(t)}$. Note that on $\mathcal{K}^r(T)$

$$Ly_{r,n} = Ly_{r,m} \text{ for all } n, m < \sqrt{|N^r|}T.$$

Now observe that if $t \in [Ly_{r,0}/\sqrt{|N^r|}, T]$ then $m_r(t) \geq 1$ hence

$$\sqrt{|N^r|}t - (m_r(t) - 1) > Ly_r(m_r(t) - 1) = Ly_r(m_r(t)).$$

Therefore

$$\sqrt{|N^r|}t - m_r(t) > Ly_r(m_r(t)) - 1 > \frac{L}{2}y_r(m_r(t)).$$

for $L > 2$.

Since the value of L is a matter of choice, we can take L sufficiently large and redefine $\mathcal{K}^r(T)$ with the reselected L. Let $H_{R,T}$ be given as in Assumption 7. Since $H_{R,T}(t) \to 0$ as $t \to \infty$ is independent of L, for $\epsilon > 0$ fixed, there exists $s^*(\epsilon) > 1$ such that for $t > s^*(\epsilon)$, $H_{R,T}(t) < \epsilon$. So we set

$$L \geq 6s^*(\epsilon).$$

The proof is completed similar to proof of Theorem 3.1.

CHAPTER IV

APPLICATIONS

In this chapter we present applications of our main results discussed in the previous chapter in three different parallel server systems. First, we study the optimal control of distributed parallel server systems. The discussion is based on the results presented in Tezcan [62]. In Section 4.2 we focus on a class of N-systems. We show that a static priority policy is asymptotically optimal. In Section 4.3 we focus on a V-system with two customer class and prove two SSC results conjectured by Armony and Maglaras [2] and Milner and Olsen [53].

4.1 Optimal control of distributed parallel server systems

In this section, we consider a distributed parallel server (DPS) system where customers arrive at the system according to a Poisson process. A DPS system consists of a single customer class and multiple server pools. Each customer must be routed to a server pool or a queue at his arrival time following a routing policy. Under a routing policy, when there are idle servers in the pool, the customer may be routed for service immediately; otherwise, the customer is routed to the selected queue waiting to be served later. For notational convenience we set $\mu_j = \mu_{jj}$ and assume that

$$\mu_1 < \mu_2 < \ldots < \mu_J. \tag{4.1}$$

Once a customer receives service, he leaves the system. The service times at each server pool are assumed to be i.i.d. and exponentially distributed. A DPS system is illustrated in Figure 1.

The distributed parallel server systems we study is closely related to inverted-V-systems studied in Armony [1]. An inverted-V-system, or $\wedge$-system, also consists of multiple server pools and a single customer class. Unlike a distributed server pool system, in an $\wedge$-system there is only one queue and each server can serve a customer waiting in that queue. Since there is only one queue, there is no routing decision to make in an $\wedge$-system when an arriving

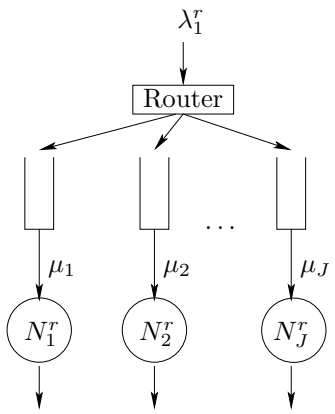

Figure 1: A DPS-parallel server system

customer finds all servers busy. Our main goal is to construct efficient routing policies so that despite the fact that customers are routed upon arrival the system works as efficiently as as a similar $\wedge$-system where customers wait for service in a central queue. This fact is referred as the *complete resource pooling* in the conventional heavy traffic literature, see, for example, Reiman [58].

We focus on two routing policies: the minimum-expected-delay–faster-server-first (MED–FSF) policy and the minimum-expected-delay–load-balancing (MED–LB) policy. The minimum-expected-delay (MED) routing policy is a widely studied and used policy in different applications. Under the MED routing policy, when a customer finds all the servers busy at the time of his arrival, he is routed to the queue with the minimum delay, otherwise he is routed to one of the server pools with idle servers. Kogan et al. [44] demonstrates numerically that the distributed pools system with the MED policy is *not* inferior to the corresponding system with a central queue with respect to the stationary waiting time distribution. The MED policy (and a simpler version, join the shortest queue policy) has been shown to achieve complete resource pooling in distributed parallel server systems under conventional heavy traffic, see Foschini and Salz [26], Reiman [58] and Laws [46]. Winston [70] and Weber [63] showed that when the service times are exponential or have increasing failure rate then it is optimal to assign the job to the shortest queue. Whitt [68], however, showed that for other service-time distributions, join the shortest queue rule need not be optimal. As noted in Section 5.3 of Gans et al. [28], "while there is a fairly extensive literature on load balancing,

little of it appears to be directly applicable to distributed parallel server systems with many servers."

Under both MED-FSF and MED-LB policies, if all servers are busy when a customer arrives at the system, the customer is routed to the queue that has the minimum expected delay. If there is an idle server at his arrival time, then under the MED–FSF policy the customer is routed to the fastest available pool and under the MED–LB to the least utilized available pool.

We analyze DPS systems working under these policies in the Halfin-Whitt many server heavy traffic regime. We consider a sequence of systems indexed by r. The arrival rate to the rth system, λ^r, is equal to r. For simplicity we assume that

$$N_j^r = \lceil \beta_j |N^r| \rceil, \text{ for all } j \in \mathcal{J}, \tag{4.2}$$

where $\beta_j > 0$ is given for each $j \in \mathcal{J}$ with $\sum_{j=1}^J \beta_j = 1$, and for a real number x, $\lceil x \rceil$ is the least integer greater than or equal to x. (It is actually enough to assume that $N_j^r/|N^r| \to \beta_j$ as $r \to \infty$ for all $j \in \mathcal{J}$ for the results in this section to hold.) We define the average service rate $\bar{\mu}$ across all the servers by

$$\bar{\mu} = \sum_{j=1}^{J} \beta_j \mu_j. \tag{4.3}$$

Let the traffic intensity for the rth system be defined by

$$\rho^r = \frac{\lambda^r}{\sum_{j \in \mathcal{J}} N_j^r \mu_j}.$$

We assume that

$$\sqrt{|N^r|}(1 - \rho^r) \to \theta/\bar{\mu} \tag{4.4}$$

for some $\theta > 0$ as $r \to \infty$. Assumption (4.4) implies that the system reaches heavy traffic as $r \to \infty$. It can easily be checked that Assumption 1 holds by (4.4).

Armony [1] shows that the faster-server-first (FSF) policy is asymptotically optimal in the QED regime in the sense that it minimizes the stationary distribution of the waiting time and queue length processes in the limit as the arrival rate goes to infinity. In this

section, we show that the MED–FSF policy is also asymptotically optimal in the QED regime for our distributed systems. Our optimality result is weaker than that in Armony [1] where it is shown that the FSF policy achieves the minimum stationary queue length and waiting time distribution. Here, we show that MED-FSF policy minimizes the stationary queue length distribution and the probability that a customer gets delayed in the queue *before his service starts*.

Although the MED–FSF is asymptotically optimal in minimizing the queue lengths in a call center, all the servers in our distributed system except those with the lowest service rate experience 100% utilization as the offered load gets large. Therefore, this policy punishes those servers who work more efficiently. Employee burnout increases with overwork and employee turnout may increase if there is a sense of unfairness in a work environment. In addition to these problems, a company using MED–FSF is likely to loose its most efficient agents since they are the ones being overworked. Therefore, MED–FSF policy may increase the operating costs of a call center in the long-run. One solution to this problem is to overstaff those faster pools and give them more breaks to lower the utilization of faster servers to an acceptable level. However, this adds another level of decision making procedure to call center management.

Considering these disadvantages of the MED–FSF policy, we propose the MED–LB policy which routes the calls among call centers "fairly". We show that the MED–LB policy asymptotically balances the load of the servers; i.e., the utilizations of all the servers in the system become equal as the offered load gets large. We further show that, operating under the MED–LB policy, the distributed system achieves the *resource pooling* effect in that the stationary distribution of the total queue length and the waiting time processes are approximated by those in an $M/M/n$ system. The arrival rate and the total number of servers in the latter system are the same as those in the distributed system. However, the service rate for each server is equal to the average rate among all servers in the distributed system. The performance of distributed systems is usually approximated by this single server pool system in practice. It is shown here for the first time that this approximation is asymptotically correct. The same result also holds for the $\wedge$-systems under the LB

policy. One can come up with other policies that would yield balanced utilizations across server pools in our distributed system but the MED–LB policy is easy to implement and its performance can be accurately estimated using the Erlang C formula. We also discuss how the MED–LB policy can be modified to distribute the total available idle time in desired proportions among all the server pools.

Using our asymptotic results, we derive approximations for the performances of the systems under the MED–FSF and MED–LB policies. Since all our results are asymptotic results, we conduct simulation experiments to illustrate the accuracy of our results. We compare the performance of the distributed systems under the MED–FSF and MED–LB policies with that of the corresponding $\wedge$-systems and test the accuracy of the approximations obtained from the asymptotic analysis in several distributed systems. We conduct additional simulation experiments to test if the MED–LB and LB policies balance the utilizations of the servers in relatively small systems. The simulation results show that our asymptotic results are also observed in systems with sizes comparable to existing call centers and the asymptotic approximations provide accurate estimates for the stationary delay probability and expected waiting time.

The main results of this section can be summarized as follows.

1. Under the MED–FSF policy, the stationary delay probability and stationary queue lengths are asymptotically minimized among all adapted policies. Furthermore, under the MED–FSF policy, the distributed system performs as good as a corresponding $\wedge$-system.

2. The MED–LB policy in a distributed system asymptotically balances the utilizations of the server pools. Also, under the MED–LB policy, a distributed system performs as good as the corresponding $\wedge$-system. Both systems perform as good as a corresponding $M/M/n$ system.

3. As above results are derived through many server diffusion limits, we obtain formulas for approximate performance analysis of a distributed system under both the MED–FSF and MED–LB policies.

While establishing the diffusion limits, we use Theorem 3.1 to prove a state space collapse (SSC) result in the many server heavy traffic analysis of parallel server systems.

4.1.1 The queueing model and the asymptotic framework

As described in the previous section we consider a distributed parallel server system that consists of $J \geq 2$ server pools and a single customer class. Server pool j consists of N_j homogeneous servers and has its own dedicated queue. Customers arrive at the system according to a Poisson process with rate λ. Each customer must be routed to a server pool or a queue at his arrival time following a certain routing policy. Under a routing policy, an arriving customer may be routed to an idle server if there are idle servers at the time of his arrival or to a queue waiting to be served later. Once a customer receives service he leaves the system. The service time of each server in pool j is assumed to be exponentially distributed with rate μ_j. A server residing in server pool j can only handle customers who are routed to the jth queue. Once a customer joins a queue he cannot swap to other queues nor can he renege. Customers in the same queue are served on a first-in-first-out (FIFO) basis. We refer to such a system as a *distributed parallel server system* or a distributed system. The corresponding $\wedge$-system of a distributed system has the same parameters with the distributed system except that it has only one queue and a customer who finds all the servers busy at the time of his arrival is routed to this queue waiting to be served by one of the servers in the system later.

The customers that are routed to the jth queue or server pool are called the class j customers. For notational convenience, we define $\mathcal{J} = \{1, \ldots, J\}$ the set of server pools. Since each customer class is associated with a unique server pool this set will also give the set of indices for the customer classes. Since there is only one customer type we omit the subscript "1" from the notation in this section, e.g., $Q_i(t)$ denotes the total number of type 1 customers waiting in the ith queue.

We focus on two non-idling routing policies: the MED–LB and the MED–FSF policies. Under the MED policy when a customer arrives to the system to find all the servers busy he is routed to the queue with the minimum expected delay (or waiting time). The expected

waiting time in queue j at time t is taken to be given by $Q_j(t)/(N_j\mu_j)$, which actually turns out to be asymptotically correct. Under both the MED–LB and the MED–FSF policies, if all servers are busy when a customer arrives at the system, the customer is routed to one of the queues according to the MED policy. If there is an idle server at his arrival time, then under the MED–FSF policy the customer is routed to the fastest available pool and under the MED–LB to the least utilized available pool, where the utilization of the server pool j at time t is given by $Z_j(t)/N_j$. We assume that ties are broken arbitrarily. In an $\wedge$-system there is only one queue hence there is no routing decision to make when an arriving customer finds all servers busy. Thus, the MED–FSF and MED–LB routing policies in our distributed system reduces to the FSF and LB policies in the corresponding $\wedge$-system.

We use $X^r(t)$ to denote the total number of customers in the system at time t and we set $X^r = \{X^r(t), t \geq 0\}$. The diffusion-scaled total number of customers in the system, $\hat{X}^r$, is defined by

$$\hat{X}^r(t) = \frac{X^r(t) - |N^r|}{\sqrt{|N^r|}}. \tag{4.5}$$

In order to gain insight on DPS systems' performance we analyze the weak limits of $\hat{Q}^r$, $\hat{Z}^r$ and $\hat{X}^r$ as $r \to \infty$.

Let $W^r(t)$ denote the amount of time a customer will wait before his service starts if he arrives at time t and $W^r = \{W^r(t) : t \geq 0\}$. The process W^r is known as the virtual waiting time process. We define the diffusion-scaled virtual waiting time process $\hat{W}^r$ by

$$\hat{W}^r(t) = \sqrt{|N^r|}W^r(t). \tag{4.6}$$

We are also interested in the asymptotic behavior of the stationary distribution of $(\hat{Q}^r, \hat{Z}^r, \hat{X}^r, \hat{W}^r)$ as $r \to \infty$. For a routing policy π, we denote the stationary probability distribution of $(\hat{Q}^r, \hat{Z}^r, \hat{X}^r, \hat{W}^r)$ by $\mathbb{P}_{\pi^r}$ when it exists. For notational convenience, we denote by

$$(\hat{Q}^r(\infty, \pi), \hat{Z}^r(\infty, \pi), \hat{X}^r(\infty, \pi), \hat{W}^r(\infty, \pi))$$

a random variate that has the distribution $\mathbb{P}_{\pi^r}$. We call $\mathbb{P}(W^r(\infty) > 0)$ the stationary delay probability in the rth system. If a stationary distribution of a process Y does not exist, we

set

$$\mathbb{P}\{Y(\infty) > x\} = 1, \tag{4.7}$$

for all $x \in \mathbb{R}$.

4.1.2 Main Results

Our main results are based on the asymptotic analysis of the stochastic process $(\hat{Q}^r, \hat{Z}^r, \hat{X}^r, \hat{W}^r)$ and its stationary behavior as $r \to \infty$. The proofs of the results in this section are presented in Section 4.1.4.4.

We first focus on the MED–FSF policy and show that it minimizes the stationary distribution of the queue lengths and the stationary delay probability among all adapted policies as described in the following theorem.

Theorem 4.1.1. *Consider a sequence of MED–FSF distributed server systems. Assume that (4.2) and (4.4) hold. Then, for any adapted routing policy* π

$$\lim_{r\to\infty} \mathbb{P}\{\hat{X}^r(\infty, \textit{MED–FSF}) > x\} \leq \liminf_{r\to\infty} \mathbb{P}\{\hat{X}^r(\infty, \pi) > x\}, \tag{4.8}$$

for all $x \in \mathbb{R}$ *and*

$$\lim_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \textit{MED–FSF}) > 0\} \leq \liminf_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \pi) > 0\}, \tag{4.9}$$

In Theorem 4.1.1 we only require that π is adapted and do not assume that it is non-idling or serves the customers on a FIFO basis. Theorem 4.1.1 is proved by comparing the limit of the sequence of the stationary distributions of the distributed systems with that of the corresponding $\wedge$-system. We show that the MED–FSF policy achieves the same asymptotic performance as it does in an identical $\wedge$-system. Using this result and the asymptotic optimality of the FSF policy in an $\wedge$-system we prove that the MED–FSF policy is asymptotically optimal as described in Theorem 4.1.1.

Let $X^r_\wedge(t)$ be the number of customers in the corresponding $\wedge$-system at time t and

$$\hat{X}^r_\wedge(t) = \frac{X^r_\wedge(t) - |N^r|}{\sqrt{|N^r|}} \tag{4.10}$$

We denote the the virtual waiting time process in these systems by $W^r_\wedge$ and define the diffusion-scaled waiting process by $\hat{W}^r_\wedge(t) = \sqrt{|N^r|}W^r_\wedge(t)$. The proof of Theorem 4.1.1 also yields the following result.

Theorem 4.1.2. *Consider a sequence of MED–FSF distributed server systems and the sequence of corresponding FSF $\wedge$-systems. Assume that (4.2) and (4.4) hold. Then*

$$\lim_{r\to\infty} \mathbb{P}\{\hat{X}^r(\infty, \textit{MED–FSF}) > x\} = \lim_{r\to\infty} \mathbb{P}\{\hat{X}^r_\wedge(\infty, \textit{FSF}) > x\} = F(x), \tag{4.11}$$

for all $x \in \mathbb{R}$ and

$$\lim_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \textit{MED–FSF}) > w\} = \lim_{r\to\infty} \mathbb{P}\{\hat{W}^r_\wedge(\infty, \textit{FSF}) > w\} = F(\bar{\mu}w), \tag{4.12}$$

for all $w \geq 0$, where $F(x) = \int_{-\infty}^{x} f(u)du$ and f is the density function defined by

$$f(x) = \begin{cases} \frac{\theta}{\sqrt{\bar{\mu}}} \exp\{-\theta x/\sqrt{\bar{\mu}}\}\alpha, & \textit{if } x \geq 0 \\ \frac{\sqrt{\frac{\mu_1}{\bar{\mu}}}\phi\left(\sqrt{\frac{\mu_1}{\bar{\mu}}}x + \frac{\theta}{\sqrt{\mu_1}}\right)}{\Phi\left(\frac{\theta}{\sqrt{\mu_1}}\right)}(1-\alpha), & \textit{if } x < 0 \end{cases} \tag{4.13}$$

where

$$\alpha = \left[1 + \frac{\theta/\sqrt{\mu_1}\Phi\left(\theta/\sqrt{\mu_1}\right)}{\phi\left(\theta/\sqrt{\mu_1}\right)}\right]^{-1} = \mathbb{P}\left\{X(\infty) > 0\right\}. \tag{4.14}$$

Remark 4.1. [1] shows that the FSF routing policy is asymptotically optimal for the $\wedge$-systems in the sense that

$$\lim_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \text{FSF}) > w\} \leq \liminf_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \pi) > w\}, \tag{4.15}$$

for *all* $w \geq 0$ and any adapted HL policy π. Note that this is stronger than (4.9) since our result only holds for $w = 0$. The main reason is that for $\wedge$-systems working under an HL policy (4.9) implies (4.15) since customers are served on a FIFO basis. In a distributed system a server can idle even though there are customers in other queues and a customer arriving at that instant *overtakes* the customers that are already in queue and starts his service before them. Therefore, customers in a distributed system under, for example, a non-idling routing policy are not served on a FIFO basis. Hence, (4.9) does not imply (4.15) for distributed systems.

Next we obtain approximations for the performance of the rth system under the FSF and MED–FSF policies using (4.11) and (4.12).

Corollary 1. *Consider a sequence of MED–FSF distributed server systems. Under the assumptions of Theorem 4.1.2*

$$\mathbb{P}(W^r(\infty) > 0) \to \alpha \quad \text{and} \quad \left| \mathbb{E}[\hat{W}^r(\infty)] - \alpha \frac{\sqrt{|N^r|\bar{\mu}}}{\sqrt{r}\theta} \right| \to 0 \tag{4.16}$$

as $r \to \infty$, where α is given by (4.14) and $\bar{\mu}$ is given by (4.3).

Remark 4.2. The results in (4.16) also hold for the sequence of corresponding FSF $\wedge$-systems, see Lemma 4.1 in Armony [1].

The last result we obtain for the MED–FSF DSP systems involves the differences between the utilizations of server pools.

Theorem 4.1.3. *Consider a sequence of MED–FSF distributed server systems. Under the assumptions of Theorem 4.1.2, for $i = 2, 3, \ldots, J$*

$$\lim_{r\to\infty} \sqrt{|N^r|}\mathbb{E}\left[\frac{Z_i^r(\infty)}{N_i^r}\right] = 1 \text{ and} \tag{4.17}$$

$$\lim_{r\to\infty} \sqrt{|N^r|}\mathbb{E}\left[\frac{Z_1^r(\infty)}{N_1^r}\right] = 1 - \frac{(1-\alpha)}{\beta_1}\left(\frac{\phi\big(\theta/\sqrt{\mu_1}\big)}{\Phi\big(\theta/\sqrt{\mu_1}\big)} + \frac{\theta\sqrt{\bar{\mu}}}{\mu_1}\right) \tag{4.18}$$

where $x^- = \min\{x, 0\}$ and $X(\infty)$ is a random variable with distribution F defined in Theorem 4.1.2.

Now we focus on the distributed systems operating under the MED–LB policy and the corresponding $\wedge$-systems. We have the following result on the utilizations of the server pools under the MED–LB policy.

Theorem 4.1.4. *Consider a sequence of MED–LB distributed server systems. Assume that (4.2) and (4.4) hold. Then*

$$\lim_{r\to\infty} \sqrt{|N^r|}\left|\mathbb{E}\left[\frac{Z_i^r(\infty)}{N_i^r} - \frac{Z_j^r(\infty)}{N_j^r}\right]\right| = 0. \tag{4.19}$$

for $i, j \in \mathcal{J}$. If in addition

$$\left(\hat{Q}^r(0), \hat{Z}^r(0)\right) \Rightarrow \left(\hat{Q}(0), \hat{Z}(0)\right), \tag{4.20}$$

as $r \to \infty$ *for a random vector* $(\hat{Q}(0), \hat{Z}(0))$ *and*

$$\left| \frac{\hat{Z}_j^r(0)}{\beta_j} - \frac{\hat{Z}_{j'}^r(0)}{\beta_{j'}} \right| \to 0 \tag{4.21}$$

for all $j, j' \in \mathcal{J}$ *in probability as* $r \to \infty$*, then for any* $T > 0$

$$\sqrt{|N^r|} \left\| \frac{Z_i^r(t)}{N_i^r} - \frac{Z_j^r(t)}{N_j^r} \right\|_T \to 0 \text{ in probability as } r \to \infty \tag{4.22}$$

in probability as $r \to \infty$ *for* $i, j \in \mathcal{J}$*.*

Remark 4.3. The results of Theorem (4.1.4) also hold for the sequence of corresponding LB $\wedge$-systems.

We next prove that, under the MED–LB policy, a distributed system performs as well as the corresponding $\wedge$-system and both systems performs as good as a corresponding $M/M/n$ system.

Consider a sequence of $M/M/n$ systems with the arrival rate and the number of servers in the rth system is equal to those of the rth distributed system. Assume that the service rate of each server in this system is equal to the average service rate $\bar{\mu}$ in the distributed system given by (4.3). Let $\underline{X}^r(t)$ denote the the total number of customers in the rth $M/M/n$ system at time t. We define the diffusion-scaled total number of customers process in these systems by

$$\hat{\underline{X}}^r(t) = \frac{\underline{X}^r(t) - |N^r|}{\sqrt{|N^r|}}. \tag{4.23}$$

We use $\hat{\underline{X}}^r(\infty)$ denote the weak limit of $\hat{\underline{X}}^r(t)$ as $t \to \infty$, which exists for each r by (4.4) and standard results on the existence of a stationary distribution of an $M/M/n$ system. Let $\underline{W}^r(t)$ denote the virtual waiting time for the rth single server system, $\hat{\underline{W}}^r(t) = \sqrt{|N^r|}\underline{W}^r(t)$ and $\underline{W}^r(\infty)$ denote the weak limit of $\underline{W}^r(t)$ as $t \to \infty$.

Theorem 4.1.5. *Consider a sequence of MED–LB distributed systems and the sequence of corresponding LB* $\wedge$*-systems. Assume that (4.2) and (4.4) hold. Let* $\underline{X}^r$ *and* $\underline{W}^r$ *be defined as above. Then,*

$$\lim_{r\to\infty} \mathbb{P}\{\hat{X}^r(\infty) > x\} = \lim_{r\to\infty} \mathbb{P}\{\hat{X}_\wedge^r(\infty) > x\} = \lim_{r\to\infty} \mathbb{P}\{\hat{\underline{X}}^r(\infty) > x\} = F(x) \tag{4.24}$$

for all $x \in \mathbb{R}$ *and*

$$\lim_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty) > w\} = \lim_{r\to\infty} \mathbb{P}\{\hat{W}^r_{\wedge}(\infty) > w\} = \lim_{r\to\infty} \mathbb{P}\{\underline{\hat{W}}^r(\infty) > w\} = F(\bar{\mu}w), \quad (4.25)$$

for all $w \geq 0$, *where* $F(x) = \int_{-\infty}^{x} f(u)du$ *and* f *is the density function defined by*

$$f(x) = \begin{cases} \frac{\theta}{\sqrt{\bar{\mu}}} \exp\{-\theta x/\sqrt{\bar{\mu}}\}\alpha, & \text{if } x \geq 0 \\ \frac{\phi\left(x+\frac{\theta}{\sqrt{\bar{\mu}}}\right)}{\Phi\left(\frac{\theta}{\sqrt{\bar{\mu}}}\right)}(1-\alpha), & \text{if } x < 0 \end{cases} \quad (4.26)$$

where

$$\alpha = \left[1 + \frac{\theta/\sqrt{\bar{\mu}}\Phi\left(\theta/\sqrt{\bar{\mu}}\right)}{\phi\left(\theta/\sqrt{\bar{\mu}}\right)}\right]^{-1} = \mathbb{P}\left\{X(\infty) > 0\right\}. \quad (4.27)$$

.

Remark 4.4. The results of Corollary 1 also hold for the distributed systems operating under the MED–LB policy and the corresponding $\wedge$-systems with α given by (4.27).

4.1.3 Simulation Experiments

Since the results presented in Section 4.1.2 are asymptotical results, in this section we conduct simulation experiments to evaluate the quality of those results. We consider five cases. In each case, we simulate a distributed system and the corresponding $\wedge$-system. The parameters of all five cases are displayed in Table 1 (the time unit is taken to be one minute.)

The first four cases correspond to systems that have three server pools and the last case corresponds to a system that has eight server pools. The parameters of the first three cases are selected to investigate the effect of the offered load, defined by $\lambda/\bar{\mu}$, on the quality of our results. We set the arrival rate in the second and the third case to be 10 and 40 times the arrival rate in the first case, respectively, to observe this effect. Balanced server assignment among all pools may affect the quality of our asymptotic results. In order to observe the effect of unbalanced server staffing, in the fourth case, one of server pools is set to have significantly fewer servers than the other pools. Finally, in the last case, we consider a system with eight pools to observe the effect of the number of pools on the quality of our results.

Case	J	λ	N	μ
1	3	50	(13,7,9)	(1.48, 1.77, 2.4)
2	3	500	(125,63,89)	(1.48, 1.77, 2.4)
3	3	2000	(497,255,347)	(1.48, 1.77, 2.4)
4	3	500	(195 95 22)	(1.48, 1.77, 2.4)
5	8	500	(45,45,68,70,76,81,112,126)	(0.72,0.95,0.85,0.8,0.86,0.9,0.88,0.67)

Table 1: The simulation data to evaluate asymptotic results.

In all the simulation experiments, each performance measure estimate is presented with its 95% confidence interval. The length of each simulation run is selected to allow 12 million customers to arrive to the system. Also a warm-up period of 1.2 million customer arrivals is used. We divide the total simulation length to ten time intervals of equal length to apply batch means technique, see [45]. The confidence intervals that are reported along with the estimates are obtained using the batch means. Also, when two or more policies are compared the simulations are run using common random number generators so that the interarrival times and the service requirement of the nth customers in all the simulations are the same for $n = 1, 2, \ldots$.

4.1.3.1 Simulation results of the MED–FSF policy

In this section, we focus on the FSF routing policy. Theorem 4.1.2 says that when the offered load is high, a distributed system operating under the MED-FSF routing policy has a similar performance to the corresponding $\wedge$-system operating under the FSF routing policy. In this section we test this result in five systems with different parameters.

Table 2 displays simulation results as well as analytical approximations for all five cases. The results under the MED-FSF policy are displayed in the left half of Table 2. (The results under the MED-LB policy, displayed in the right half of Table 2, will be discussed in the next section.) For each case, the simulation estimates of the delay probability ($\mathbb{P}(W > 0)$), in percentage, and the average waiting time ($\mathbb{E}[W]$) in seconds are presented for the distributed system, in Row DS, and for the corresponding $\wedge$-system, in Row $\wedge$. The half-widths of confidence intervals are presented in parentheses next to the simulation estimates. The differences between the estimates from the distributed system and the ones from the

		MED-FSF		MED-LB	
Case	$\mathbb{P}(W>0)(\%)$	$\mathbb{E}[W]$	$\mathbb{P}(W>0)(\%)$	$\mathbb{E}[W]$	
1	DS	61.60 (0.43)	12.14 (0.31)	66.53 (0.27)	13.38 (0.15)
	$\wedge$	63.88 (0.41)	11.87 (0.30)	65.80 (0.29)	12.09 (0.14)
	DS $-\wedge$	-3.60	0.27	1.10	1.29
	Approx.	61.50	11.60	63.89	12.24
	DS $-$ Approx.	0.16	0.54	4.0	1.14
2	DS	60.64 (0.88)	3.64 (0.16)	64.83 (1.27)	3.76 (0.31)
	$\wedge$	61.60 (0.89)	3.63 (0.16)	65.68 (1.28)	3.72 (0.30)
	DS $-\wedge$	-1.60	0.01	-1.30	0.04
	Approx.	60.84	3.65	64.24	3.85
	DS $-$ Approx.	-0.33	0.01	0.91	0.09
3	DS	61.52 (1.71)	1.83(0.20)	64.48 (1.77)	1.92 (0.19)
	$\wedge$	61.02 (1.77)	1.83 (0.20)	65.01 (1.76)	1.92 (0.19)
	DS $-\wedge$	0.80	0	-0.82	0
	Approx.	61.65	1.89	64.97	1.99
	DS $-$ Approx.	-0.21	-0.06	-0.75	-0.07
4	DS	61.88 (0.93)	3.92 (0.18)	63.12 (0.92)	3.99 (0.18)
	$\wedge$	62.88 (0.90)	3.91 (0.18)	64.10 (0.92)	3.98 (0.18)
	DS $-\wedge$	-1.60	0.01	-1.50	0.01
	Approx.	62.64	3.97	64.15	4.07
	DS $-$ Approx.	-1.20	-0.05	-1.60	-0.08
5	DS	43.95 (1.06)	2.80 (0.16)	46.97 (1.04)	2.99 (0.16)
	$\wedge$	47.00(1.08)	2.78(0.16)	50.11 (1.00)	2.95 (0.16)
	DS $-\wedge$	-6.50	0.02	-6.20	0.04
	Approx.	46.32	2.75	50.19	2.98
	DS $-$ Approx.	-5.10	0.05	-6.40	0.01

Table 2: The results of simulation experiments.

corresponding $\wedge$-system are presented in Row DS $-\wedge$.

We first note that as the offered load gets high the performance of the distributed system under the MED–FSF becomes very close to that of the corresponding $\wedge$-system. In the first case the percentage difference between the estimated delay probability in these two models is around 3.6%, in the second case it decreases to 1.5% and decreases to less than 1% in the third case. The differences of the expected waiting times are even smaller. By comparing the results of the fourth case with those of the second case we observe that having a significantly smaller server pool does not affect the percentage differences. However, the number of server pools in the system has a big impact. The percentage difference between the delay probabilities are twice as much as the difference in the first case, which is the

second largest difference among all the cases.

Formulas in (4.16) give analytical approximations for delay probability and average waiting time. For each parameter case, these approximate values are presented in Row Approx. The differences between the analytical approximations and simulate estimates from the distributed systems are presented in Row DS − Approx. In first three cases, the percentage difference of delay probability is at most 0.33%. Even in Case 4 when the server pools are not balanced, the approximation for delay probability is quite close to the simulation estimate (the difference is 1.2%). In Case 5 when there are eight server pools, and the arrival rate is equal to 500 as in the second case, the approximation performs significantly worse, 5.1% in percentage difference. Overtaking as described in Remark 4.1 also explains why the quality of the approximation for the delay probability in the distributed systems diminishes when the number of server pools is increased because as the number of server pools increases so does the probability that an arriving customer will find an idle server given that there are customers waiting in the system. Note that the difference in the average waiting times is only 0.05 seconds.

From the simulation results it is clear that the distributed systems operating under the MED–FSF policy *do perform* as good as the corresponding $\wedge$-systems in terms of the delay probability and average waiting time in the queue.

4.1.3.2 Simulation results of the MED–LB policy

In this section, we focus on the MED-LB routing policy. Recall that the simulation results for the distributed systems operating under the MED-LB routing policy and for the corresponding $\wedge$-systems operating under the LB routing policy are presented in the right half of Table 2.

Theorem 4.1.5 asserts that when the offered load is high the performance of a distributed system operating under the MED-LB policy is similar to that of the corresponding $\wedge$-system under the LB policy. This is clearly observed in the first four cases; the largest percentage difference between delay probabilities is 1.5% and the largest difference between average waiting times is 1.14 seconds. In cases 2-4 the difference in waiting times is less than 0.1

seconds. The number of server pools has a big impact on the percentage difference of the delay probability for the systems under MED–LB as well; the percentage difference is 6.2% in Case 5, four times more than the second largest difference. However, the difference in waiting time is again very small, only 0.01 seconds.

Next we assess the quality of the approximations provided by Remark 4.4 and (4.16) using the simulation results. First we note that in cases 2-5 the differences between the average waiting times estimated for the distributed system and the ones provided by the approximation are less than 0.1 seconds. This difference is 1.14 seconds in the first experiment. Therefore, in terms of the average waiting time approximations provide very accurate estimates. The percentage difference of the delay probabilities is less than 1.7% in first four cases, but in case 5 it is significantly larger (6.4%). Again, this is due to overtaking that is explained in Remark 4.1.

We conclude that the distributed systems operating under the MED-LB policy performs as good as the corresponding $\wedge$-systems operating under the LB policy.

Theorem 4.1.5 asserts that the difference between the utilizations of the servers is $o(1/\sqrt{|N^r|})$ when the offered load is high for the distributed systems operating under the MED–LB policy. To evaluate the quality of this asymptotic result we conduct additional simulation experiments. Also, to illustrate that the difference between the utilizations of server pools is high under MED–FSF policy we simulate the same systems under this policy.

We consider three cases with two server pools. The parameters of these experiments are selected to investigate the effects of different service rates and unbalanced staffing levels on the difference of the utilizations. The values of the parameters for these cases is displayed in Table 3. The first case is a homogeneous system in the sense that the number of servers in each pool is the same and the service rates of all the servers are equal. We do not simulate this system under the FSF policy since the service times for both pools are equal. In the second case we set the service rate of the first server pool to be 2/3 of the service rate of the second server pool. In the third case, we set the number of servers in the first pool to be significantly less than that in the second pool, and set the service rates as in the second case.

Case	J	λ	N	μ
6	2	97	(50, 50)	(1,1)
7	2	97	(50, 50)	(0.8, 1.2)
8	2	97	(29, 64)	(0.8, 1.2)

Table 3: The simulation data to test Theorem 4.1.4

	MED-FSF		MED-LB	
Case	$\wedge$	DS	$\wedge$	DS
6	NA	NA	0.30 (0.01)	0.33 (0.01)
7	5.20 (0.11)	5.15 (0.11)	0.55 (0.01)	0.58 (0.01)
8	6.93 (0.15)	6.81 (0.15)	0.67 (0.02)	0.72 (0.03)

Table 4: The percentage differences between utilizations.

Table 4 presents the results of the simulation experiments. We also consider the corresponding $\wedge$-model in each case. We display the percentage difference between the average utilization of the first and second server pools in $\wedge$-systems under the $\wedge$ column and for the distributed systems under the DS column. For all estimates, we show the half-width of the 95% confidence intervals.

Observe that under the LB and MED–LB policies the percentage differences between the utilizations of the servers in all the systems are very small; less than 0.8%. Hence, even for these relatively small systems the LB policy seems to balance the load of the server pools. We observe that the differences in the utilizations are more in the systems with different service rates and unbalanced staffing levels than that in the homogenous system. Also the differences are slightly lower in $\wedge$-systems than those in the distributed systems. When we compare the percentage differences under the FSF and MED–FSF policies with the LB and MED–LB policies, we see that they increase about 10 times. This verifies once again how the FSF policy may be unfair in routing calls to server pools.

4.1.4 Proofs of the main results

In this section we prove the results stated in Section 4.1.2. In Section 2.1 we discuss the dynamics of distributed systems to mathematically characterize the behavior of these systems. and introduce the notation used in the rest of this section.

In Section 4.1.4.2 we provide asymptotic upper bounds for $Q^r(t)$ and $N_j^r - Z_j^r(t)$ in a

sequence of distributed systems operating under a non-idling routing policy. These bounds are later used to show that $\hat{Q}^r(t)$ and $\hat{Z}^r(t)$, as defined in (2.23), are stochastically bounded. Also, geometric Lyapunov functions are built using these results; see Section 4.1.4.4. In Section 4.1.4.3, we analyze the limits of the fluid scaled processes, $Q^r(\cdot)/|N^r|$ and $Z^r(\cdot)/|N^r|$, as $r \to \infty$. We establish the invariant states of the fluid limits. This result is needed to justify that the diffusion scaling is properly defined, to use the results of [22] and to establish the diffusion limits. In Section 4.1.4.4 we study the weak limits of the diffusion scaled processes $\hat{Q}^r(\cdot)$ and $\hat{Z}^r(\cdot)$ of a distributed system operating under the MED–FSF or the MED–LB policies as $r \to \infty$. We also show that the stationary distribution of $\hat{X}^r$ converges weakly to the stationary distribution of the corresponding diffusion limit as $r \to \infty$. Finally, we provide the proofs of Theorems 4.1.1, 4.1.2, 4.1.4 and 4.1.5 and Corollary 1 in Section 4.1.4.4 using the results in Sections 4.1.4.2-4.1.4.4.

4.1.4.1 The dynamics of distributed parallel server systems

We next present the additional equations that must be satisfied by the distributed systems operating under the MED–FSF and MED–LB policies. First we focus on the non-idling routing policies.

For a non-idling routing policy π, in addition to equations (2.2)-(2.12) in [22], $\mathbb{X}_\pi$ must also satisfy the following condition:

$$\int_0^\infty \mathbf{1}\left\{\sum_{j=1}^{J} Z_j^r(s) < N^r\right\} \mathrm{d}A^{q,r}(s) = 0. \tag{4.28}$$

This condition implies that an arriving customer who finds idle servers will be routed to one of the idle servers. In addition to (4.28) we assume that for $\pi \in \Pi$, there exists $a_\pi^r > 0$, for each $r > 0$, such that

$$A_j^{q,r}(t) \text{ can only increase when } Q_j^r(t) \le a_\pi^r Q_{j'}^r(t) \text{ for all } j' \in \mathcal{J}. \tag{4.29}$$

This implies that as long as the number of customers in one of the queues is more than a_π^r times the queue length of another queue, arriving customers are not routed to the former queue. Note that the MED–FSF and MED–LB policies satisfy (4.29).

Recall that we assume that the service rates are increasing with the index of the server pool as stated in assumption (4.1). Under the MED–FSF policy the following must hold.

$$A_j^{q,r}(t) \text{ can only increase when } \frac{Q_j^r(t)+Z_j^r(t)-N_j^r}{N_j^r\mu_j} \leq \min_{j'\in\mathcal{J}}\left\{\frac{Q_{j'}^r(t)+Z_{j'}^r(t)-N_{j'}^r}{N_{j'}^r\mu_{j'}}\right\} \quad (4.30)$$

and

$$A_j^{s,r}(t) \text{ can only increase when } \sum_{j'=j+1}^{J}\left(Z_{j'}^r(t)-N_j^r\right)=0 \text{ for } j\in\mathcal{J}. \quad (4.31)$$

By the non–idling condition (4.28), $A_j^{q,r}(t)$ can only increase when all the servers are busy. Hence, (4.30) is invoked when there are no idle servers in the system. And if $Z_j^r(t)=N_j^r$, then $Q_j^r(t)/(N_j^r\mu_j)$ gives the expected delay time of a customer joining the jth queue before his service starts. The condition (4.31) implies that customers can be routed to server pool j only when all the faster servers in the system, servers in pools $j+1$ through J, are busy.

Under the MED–LB policy the following must hold in addition to (4.30).

$$A_j^{s,r}(t) \text{ can only increase when } \frac{Z_j^r(t)}{N_j^r} \leq \min_{j'\in\mathcal{J}}\left\{\frac{Z_{j'}^r(t)}{N_{j'}^r}\right\}. \quad (4.32)$$

In this case, $Z_j^r(t)/N_j^r$ gives the proportion of busy servers. Hence, (4.32) implies that the server pool with the lowest proportion of busy servers receives the arrival. The ties in (4.30) and (4.32) are broken arbitrarily.

We set $\mu_{\min}=\min_{j\in\mathcal{J}}\{\mu_j\}$, $N_{\min}^r=\min_{j\in\mathcal{J}}\{N_j^r\}$,

$$\breve{S}_j(t)=S_j(t)-t \text{ and } \breve{A}_j^r(t)=A_j^r(t)-\lambda^r t. \quad (4.33)$$

For $T>0$, we define

$$\mathcal{M}_T^{\cap}=\cap_{r=1}^{\infty}\left(\{\|A^r(t)-A^r(t-)\|_T\leq 1\}\cap_{j=1}^{J}\left\{\|S_j(t)-S_j(t-)\|_{|N^r|T}\leq 1\right\}\right). \quad (4.34)$$

The set $\mathcal{M}_T^{\cap}$ is the set of sample paths for which only one arrival or departure at any given instant in $[0,T]$ from the system can happen. By Lemma 9,

$$\mathbb{P}\left(\mathcal{M}_T^{\cap}\right)=1 \text{ for any } T>0. \quad (4.35)$$

Let

$$\lambda = \lim_{r\to\infty} \lambda^r/|N^r|.$$

By (4.4),

$$\lambda = \sum_{j=1}^{J} \beta_j \mu_j = \bar{\mu}.$$

Notation: We denote the set of non-idling routing policies that satisfy (4.29) by Π and, with a slight abuse of terminology, we also refer to these routing policies as non-idling routing policies. We use the convention

$$\inf\{\emptyset\} = \infty \quad \text{and} \quad \sup\{\emptyset\} = -\infty.$$

We also use the big and little-o notation: For two sequences $\{x_n\}$ and $\{y_n\}$, we say x_n is $O(y_n)$ and write $x_n = O(y_n)$, if there exists n_0 and M such that $|x_n| \leq M|y_n|$ for $n > n_0$. We say x_n is $o(y_n)$ and write $x_n = o(y_n)$, if $\lim_{n\to\infty} |x_n|/|y_n| = 0$. For two random two random variables Υ_1 and Υ_2, $\Upsilon_1 \sim \Upsilon_2$ means they have the same distribution.

4.1.4.2 Asymptotic bounds on Q^r and Z^r under a non-idling routing policy

In this section we derive asymptotic bounds on Q^r and Z^r. These bounds are used to show that $\hat{Q}^r$ and $\hat{Z}^r$ are stochastically bounded in each finite interval $[0, T]$ which is required in Section 4.1.4.4 to prove our SSC results. They are also used to define Lyapunov functions that are used to prove the convergence of stationary distributions in Section 4.1.4.4.

Let $x = (x_1, \ldots, x_J) \in \mathbb{R}^J$. We define $\varphi_i^r : \mathbb{R}^J \to \mathbb{R}$, for $i = 1, 2$, by

$$\varphi_1^r(x) = \sum_{j=1}^{J} \left(N_j^r - x_j\right) \quad \text{and} \quad \varphi_2^r(x) = \sum_{j=1}^{J} x_j. \tag{4.36}$$

Clearly $\varphi_1^r(Z_j^r(t)) \geq 0$ and $\varphi_2^r(Q_j^r(t)) \geq 0$ for all $t \geq 0$. We present bounds for $\varphi_1^r(Z_j^r(t))$ and $\varphi_2^r(Q_j^r(t))$ in terms of their initial states and certain functions of primitive processes. The proofs of the results in this section are placed in Appendix C.1. First we present bounds for $\varphi_1^r(Z_j^r(t))$. Recall that $\mathcal{M}^{\cap}$ is defined by (4.34).

Theorem 4.1.6. *Let $\mathbb{X}^r$ be a distributed parallel server system operating under a non-idling routing policy. Assume that (4.2) and (4.4) hold. Then, there exists r_0 such that for every $t_0 > 0$, $\omega \in \mathcal{M}^{\cap}$ and $r > r_0$ if*

$$\varphi_1^r(Z^r(0)) > \frac{4\theta\sqrt{\lambda^r}(t_0 \vee 1)}{\mu_{\min} \wedge 1}, \tag{4.37}$$

then

$$\begin{aligned}\varphi_1^r(Z^r(t_0)) \leq & \varphi_1^r(Z^r(0)) - \theta\sqrt{\lambda^r}t_0 + |o(\sqrt{|N^r|})| + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} \\ & + 2\|A^r(t) - \lambda^r t\|_{t_0},\end{aligned} \tag{4.38}$$

otherwise

$$\begin{aligned}\varphi_1^r(Z^r(t_0)) \leq & 2J + \varphi_1^r(Z^r(0)) + \theta\sqrt{\lambda^r}t_0 + |o(\sqrt{|N^r|})| + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} \\ & + 2\|A^r(t) - \lambda^r t\|_{t_0}.\end{aligned} \tag{4.39}$$

Next we present bounds for $\varphi_2^r(Q_j^r(t))$. One of the terms in these bounds is ζ^r that is defined by

$$\begin{aligned}\zeta^r(t_0) = & \sup_{\substack{s_1 \leq s_2 \in [0,t_0] \\ \upsilon_1,\ldots,\upsilon_J \in [0,t_0] \\ \upsilon_j + (s_2 - s_1) \leq t_0}} \left\{ \left(\theta\sqrt{\lambda^r} - N_{\min}^r \mu_{\min}\right)(s_2 - s_1) + \left|\breve{A}^r(s_2) - \breve{A}^r(s_1)\right| \right. \\ & \left. + \sum_{j=1}^{J} \left|\breve{S}_j\left(|N^r|\left(\upsilon_j + (s_2 - s_1)\right)\right) - \breve{S}_j\left(|N^r|\upsilon_j\right)\right| \right\},\end{aligned} \tag{4.40}$$

where $\breve{S}_j$ and $\breve{A}$ are given by (4.33).

Theorem 4.1.7. *Let $\mathbb{X}^r$ be a distributed parallel server system operating under a non-idling routing policy. Assume that (4.2) and (4.4) hold. Then, there exists r_0 such that for every $t_0 > 0$, $\omega \in \mathcal{M}^{\cap}$ and $r > r_0$, if*

$$\varphi_2^r(Q^r(0)) > \theta\sqrt{\lambda^r}(t_0 \vee 1), \tag{4.41}$$

then

$$\begin{aligned}\varphi_2^r(Q^r(t_0)) \leq & \varphi_2^r(Q^r(0)) - \theta\sqrt{\lambda^r}t_0 + |o(\sqrt{|N^r|})| + 2t_0(\mu_{\max} \vee 1)(J + \zeta^r(t_0)) \\ & + 4\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} + 2\|A^r(t) - \lambda^r t\|_{t_0},\end{aligned} \tag{4.42}$$

otherwise

$$\begin{aligned}\varphi_2^r(Q^r(t_0)) &\leq \varphi_2^r(Q^r(0)) + 2J + 2(t_0 \vee 1)(\mu_{\max} \vee 1)(J + \zeta^r(t_0)) + |o(\sqrt{|N^r|})| \\ &\quad + 4\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} + 2\|A^r(t) - \lambda^r t\|_{t_0} \end{aligned} \tag{4.43}$$

Remark 4.5. In an $\wedge$-system there is only one queue, hence, the processes A, A_q, Q are one dimensional and so have only one subsrcipt but otherwise all the components of $(A^r, A_q^r, A_s^r, Q^r, Z^r, T^r, Y^r, B$ have the same interpretations they have for the distributed systems.

The results of Theorems 4.1.6 and 4.1.7 also hold for an $\wedge$-system $\mathbb{X}_\wedge^r = (A^r, A_q^r, A_s^r, Q^r, Z^r, T^r, Y^r, B^r, D^r)$ operating under a non-idling and HL routing policy with $\varphi_2^r : \mathbb{R} \to \mathbb{R}$ is defined by $\varphi_2^r(x) = x$.

4.1.4.3 Fluid limits

In this section we study the fluid limits of the distributed systems. The results in this section are needed to show that the diffusion scaling introduced in Section 4.1.1 is properly defined. By using the result on the invariant states of the fluid limits in this section we verify that when (4.20) holds the fluid limits are time-invariant. The results in this section are used in the proofs of our SSC results to verify that Assumption 1 of [22] holds and in proving the weak convergence of $\hat{X}^r$.

The fluid scaling $\bar{\mathbb{X}}^r(\cdot)$ is defined by $\bar{\mathbb{X}}^r(\cdot) = \mathbb{X}^r(\cdot)/|N^r|$. The following notation and definitions are introduced in [22] but we repeat them here for completeness. The process $\bar{\mathbb{X}}(\cdot)$ is called a fluid limit of $\{\mathbb{X}^r\}$ if there exists a sequence $\{r_l\}$, with $r_l \to \infty$ as $l \to \infty$, and $\omega \in \mathcal{A}$ such that $\bar{\mathbb{X}}^{r_l}(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}(\cdot, \omega)$, where $\mathcal{A} \subset \Omega$ is taken as in Theorem A.1 in [22] and $\mathbb{P}\{\mathcal{A}\} = 1$. The existence of the fluid limits are established and the fluid model equations that are satisfied by every fluid limit are presented in Theroem A.1 in Dai and Tezcan [22]. We call the vector (q, z) an invariant state of the fluid limits if for any fluid limit $\bar{\mathbb{X}}$, $\bar{Q}(0) = q$ and $\bar{Z}(0) = z$ implies $\bar{Q}(t) = q$ and $\bar{Z}(t) = z$ for all $t > 0$.

The following result characterizes the invariant states of the fluid limits of the MED–FSF and MED–LB distributed server pool systems.

Lemma 1. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF or MED–LB distributed server pool systems. Assume that (4.2) and (4.4) hold and that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Let $q_1(a) = a$, for $a \geq 0$, $q_j(a) = a\mu_j\beta_j/(\mu_1\beta_1)$, $q(a) = (q_1(a), q_2(a), \ldots, q_J(a))$, $z_j = \beta_j$ and $z = (z_1, \ldots, z_J)$. Then $\mathcal{M} = \{(q(a), z) : a \geq 0\}$ is the set of all the invariant states of the fluid limits of $\{\mathbb{X}^r\}$.*

A proof is presented in Appendix C.2.

Remark 4.6. It can similarly be proved that if $\{\mathbb{X}^r\}$ is a sequence of LB or FSF $\wedge$-systems and $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$, then $\mathcal{M} = \{(a, z) : a \geq 0\}$ is the set of all the invariant states of the fluid limits of $\{\mathbb{X}^r\}$.

4.1.4.4 Diffusion limits

In this section we establish the weak limits of $\hat{Q}^r$, $\hat{Z}^r$ and $\hat{W}^r$ as $r \to \infty$. By (4.20), $\bar{Q}^r(0) \to 0$ and $\bar{Z}^r(0) \to z$ as $r \to \infty$, where z is given as in Lemma 1. Hence, the diffusion scalings defined in (2.23) give the fluctuations around the fluid limits.

In the following section we establish two SSC results for the distributed systems operating under the MED–FSF and MED–LB policies. Then in Section 4.1.4.4 we establish the diffusion limits using these SSC results. We focus on the stationary distributions of these processes in Sections 4.1.4.4 and 4.1.4.4.

State space collapse We first give an intuitive explanation of our SSC results and illustrate the results in a distributed system with two server pools. These results are proven to hold for systems with arbitrary number of server pools. The proofs of the propositions in this section are presented in Appendix C.3.1.

The MED policy routes the customers to the queue with the minimum expected delay when all the servers are busy, where the expected delay of a queue, say j, at time t is defined by $Q_j^r(t)/(\mu_j N_j^r)$. Assume that all the servers are busy at time t and

$$Q_1^r(t)/(\mu_1 N_1^r) \ll Q_2^r(t)/(\mu_2 N_2^r). \tag{4.44}$$

Since the arrival rate is greater than the total service rate of servers in the first pool, one would expect to see that the number of customers in queue 1 will increase and the number of

customers in queue 2 will decrease starting from time t, barring, of course, some stochastic fluctuations. Hence, the value of $Q_2^r(t)/(\mu_2 N_2^r) - Q_1^r(t)/(\mu_1 N_1^r)$ is expected to decrease and as long as (4.44) holds. Under the MED–FSF policy, if a server in pool 2 becomes available in a two pool distributed system at time t then he receives the next arriving customer after time t (recall that we assume $\mu_1 \leq \mu_2$). Since $\lambda^r \gg \mu_2 N_2^r$ for r large enough the idle time of servers in higher priority pools becomes "very" small for r large enough. In general we have the following result.

Proposition 4.7. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that (4.2), (4.4) and (4.20) hold. Then, for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \leq t \leq T} \left| \frac{\hat{Q}_j^r(t)}{\beta_j \mu_j} - \frac{\hat{Q}_{j'}^r(t)}{\beta_{j'} \mu_{j'}} \right| \vee \left| \sum_{j=2}^{J} \hat{Z}_j^r(t) \right| > \epsilon \right\} \to 0, \tag{4.45}$$

as $r \to \infty$.

If in addition

$$\left| \sum_{j=2}^{J} \hat{Z}_j^r(0) \right| \to 0 \tag{4.46}$$

and

$$\left| \frac{\hat{Q}_j^r(0)}{\beta_j \mu_j} - \frac{\hat{Q}_{j'}^r(0)}{\beta_{j'} \mu_{j'}} \right| \to 0 \tag{4.47}$$

in probability as $r \to \infty$ for $j, j' \in \mathcal{J}$, then for every $T > 0$

$$\left\| \frac{\hat{Q}_j^r(t)}{\beta_j \mu_j} - \frac{\hat{Q}_{j'}^r(t)}{\beta_{j'} \mu_{j'}} \right\|_T \vee \left\| \sum_{j=2}^{J} \hat{Z}_j^r(t) \right\|_T \to 0 \tag{4.48}$$

for all $j, j' \in \mathcal{J}$ in probability as $r \to \infty$.

Remark 4.8. It can be similarly shown that if $\{\mathbb{X}^r\}$ is a sequence of FSF $\wedge$-systems and (4.2),(4.4) and (4.20) hold, then for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \leq t \leq T} \left| \sum_{j=2}^{J} \hat{Z}_j^r(t) \right| > \epsilon \right\} \to 0,$$

as $r \to \infty$. If in addition (4.46) holds, then for every $T > 0$

$$\left\| \sum_{j=2}^{J} \hat{Z}_j^r(t) \right\|_T \to 0$$

in probability as $r \to \infty$.

Next we consider the SSC under the MED–LB policy. In a distributed parallel server system with two server pools under the MED–LB policy if the percentage of busy servers in the first pool is less than that in the second pool, the differences will decrease since the server pool with higher percentage of busy servers will not receive any arrivals. We have the following result for the distributed systems under the MED-LB policy.

Proposition 4.9. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. Assume that (4.2),(4.4) and (4.20) hold. Then, for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \frac{\hat{Q}_j^r(t)}{\beta_j \mu_j} - \frac{\hat{Q}_{j'}^r(t)}{\beta_{j'} \mu_{j'}} \right| \vee \left| \frac{\hat{Z}_j^r(t)}{\beta_j} - \frac{\hat{Z}_{j'}^r(t)}{\beta_{j'}} \right| > \epsilon \right\} \to 0, \tag{4.49}$$

as $r \to \infty$.

If in addition (4.21) and (4.47) hold, then for every $T > 0$

$$\left\| \frac{\hat{Q}_j^r(t)}{\beta_j \mu_j} - \frac{\hat{Q}_{j'}^r(t)}{\beta_{j'} \mu_{j'}} \right\|_T \vee \left\| \frac{\hat{Z}_j^r(t)}{\beta_j} - \frac{\hat{Z}_{j'}^r(t)}{\beta_{j'}} \right\|_T \to 0 \tag{4.50}$$

for all $j, j' \in \mathcal{J}$ in probability as $r \to \infty$.

Remark 4.10. It can be similarly shown that if $\{\mathbb{X}^r\}$ is a sequence of LB $\wedge$-systems and (4.2),(4.4) and (4.20) hold, then for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \frac{\hat{Z}_j^r(t)}{\beta_j} - \frac{\hat{Z}_{j'}^r(t)}{\beta_{j'}} \right| > \epsilon \right\} \to 0,$$

as $r \to \infty$. If in addition (4.21) holds, then for every $T > 0$

$$\left\| \frac{\hat{Z}_j^r(t)}{\beta_j} - \frac{\hat{Z}_{j'}^r(t)}{\beta_{j'}} \right\|_T \to 0$$

for all $j, j' \in \mathcal{J}$ in probability as $r \to \infty$.

Diffusion limits of the total queue length and the virtual waiting time processes: The SSC results established in (4.48) and (4.50) reveal that under the MED–FSF and MED–LB policies the individual queue lengths and number of customers in service in each pool can be estimated from the total number of customers in the system with an error that goes to zero in probability as $r \to \infty$. Hence, it is enough to focus on the total number of customers in the system instead of analyzing each queue and number of customers in service in a server pool separately. To this end let $\hat{X}^r(t)$ be defined as in (2.23) and $\hat{X}^r = \{X^r(t) : t \geq 0\}$. We have the following weak limits for $\hat{X}^r$ under the MED–FSF and MED–LB policies. The proofs of these theorems are presented in Appendix C.3.2.

Proposition 4.11. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that (4.2),(4.4), (4.20) and (4.48) hold. Then*

$$\hat{X}^r \Rightarrow \hat{X}, \text{ as } r \to \infty,$$

where $\hat{X}$ is the unique solution to the following stochastic differential equation (SDE)

$$d\hat{X}(t) = h(\hat{X})dt + \sqrt{2\mu}db(t), \tag{4.51}$$

where b is a standard Brownian Motion and

$$h(x) = \begin{cases} -\theta\sqrt{\mu}, & \text{if } x \geq 0, \\ -\theta\sqrt{\mu} - \mu_1 x, & \text{if } x < 0. \end{cases}$$

Remark 4.12. By Theorem 11.4.5 in [65] and by Theorem 4.7, under the conditions of Proposition 4.11,

$$(\hat{Q}^r, \hat{Z}^r) \Rightarrow (\hat{Q}, \hat{Z}),$$

where $\hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_J)$ and $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_J)$ with

$$\hat{Q}_j(t) = \frac{\mu_j \beta_j}{\sum_{\ell=1}^{J} \mu_\ell \beta_\ell} (\hat{X}(t))^+ \text{ for } j \in \mathcal{J}. \tag{4.52}$$

$$\hat{Z}_1(t) = (\hat{X}(t))^- \text{ and } \hat{Z}_j(t) = 0 \text{ for } 2 \leq j \leq J. \tag{4.53}$$

Proposition 4.13. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. Assume that (4.2),(4.4), (4.20) and (4.50) hold. Then*

$$\hat{X}^r \Rightarrow \hat{X}, \text{ as } r \to \infty,$$

where $\hat{X}$ is the unique solution to the following SDE

$$d\hat{X}(t) = h(\hat{X})dt + \sqrt{2\mu}db(t), \tag{4.54}$$

where b is a standard Brownian Motion and

$$h(x) = \begin{cases} -\theta\sqrt{\mu}, & \text{if } x \geq 0, \\ -\theta\sqrt{\mu} - \mu x, & \text{if } x < 0. \end{cases} \tag{4.55}$$

Remark 4.14. By Theorem 11.4.5 in [65] and Theorem 4.9, under the conditions of Proposition 4.13,

$$(\hat{Q}^r, \hat{Z}^r) \Rightarrow (\hat{Q}, \hat{Z}),$$

where $\hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_J)$ and $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_J)$ with

$$\hat{Q}_j(t) = \frac{\mu_j\beta_j}{\sum_{\ell=1}^J \mu_\ell\beta_\ell}(\hat{X}(t))^+ \text{ for } j \in \mathcal{J} \text{ and} \tag{4.56}$$

$$\hat{Z}_j(t) = \frac{\beta_j}{\sum_{\ell=1}^J \mu_\ell\beta_\ell}(\hat{X}(t))^- \text{ for } j \in \mathcal{J}. \tag{4.57}$$

Remark 4.15. It can be similarly shown that

1. if $\{\mathbb{X}^r\}$ is a sequence of FSF $\wedge$-systems satisfying (4.2),(4.4), (4.20) and (4.49), then

   $$\hat{X}^r_\wedge \Rightarrow \hat{X}, \text{ as } r \to \infty,$$

   and (4.53) holds, where $\hat{X}^r_\wedge$ is given by (4.10) and $\hat{X}$ is the unique solution to the SDE (4.51).

2. if $\{\mathbb{X}^r\}$ is a sequence of LB $\wedge$-systems satisfying (4.2),(4.4), (4.20) and (4.51) then

   $$\hat{X}^r_\wedge \Rightarrow \hat{X}, \text{ as } r \to \infty,$$

   and (4.57) holds, where $\hat{X}^r_\wedge$ is given by (4.10) and $\hat{X}$ is the unique solution to the SDE (4.54).

Next we focus on the virtual waiting time process. Let $W_j^r(t)$ be the virtual waiting time for queue j at time t in the rth system, i.e., the time a customer would wait before its service

is started if he joins queue j at time t under a HL routing policy, and $W_j^r = \{W_j^r(t) : t \geq 0\}$ be the virtual waiting time process for the jth queue. Then

$$W_j^r(t) = \inf\{s \geq 0 : D_j^r(s+t) \geq Q_j^r(0) + Z_j^r(0) + A_j^{q,r}(t) + A_j^r(t) - (N_j^r - 1)\}.$$

Let $\kappa^r(t)$ denote the index of the server pool or the queue that a customer arriving at time t would be routed to. Obviously, $\kappa^r(t)$ depends on the routing policy. For example, under the MED–LB policy

$$\kappa^r(t) = \begin{cases} \{j : Q_j^r(t)/(\mu_j N_j^r) < Q_l^r(t)/(\mu_l N_l^r) \text{ for all } l \in \mathcal{J} \setminus j\}, & if \sum_{j\in\mathcal{J}} Z_j^r(t) = |N^r|, \\ \{j : Z_j^r(t)/N_j^r < Z_l^r(t)/N_l^r \text{ for all } l \in \mathcal{J} \setminus j\}, & if \sum_{j\in\mathcal{J}} Z_j^r(t) < |N^r|. \end{cases}$$

From the definition of $\kappa^r(t)$ it follows that

$$W^r(t) = W^r_{\kappa^r(t)}(t).$$

We show that weak limit of W^r can be expressed as a simple function of X.

Theorem 4.1.8. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF (MED–LB) distributed server pool systems. Under the conditions of Proposition 4.11 (resp. Proposition 4.13)*

$$\hat{W}^r \Rightarrow \frac{\left[\hat{X}\right]^+}{\mu} \tag{4.58}$$

as $r \to \infty$, where $\hat{X}$ is the unique solution to the SDE (4.51) (resp. the SDE (4.54)).

Remark 4.16. It can be similarly shown that if $\{\mathbb{X}^r\}$ is a sequence of FSF (LB) $\wedge$-systems that satisfies the conditions of the first part (resp. second) of Remark 4.15 then (4.58) holds with $\hat{W}^r$ is replaced by $\hat{W}^r_\wedge$.

Stationary distributions of the diffusion limits: Our asymptotic optimality and equivalence results are stated in terms of the stationary distributions. The main reason is that the staffing decisions in call centers are usually made using stationary values of the performance measures. Hence, it is of practical value to study the convergence of the stationary probabilities of the queue length and waiting time processes. In this section we present the steady state probabilities for $\hat{X}$, the limiting diffusion process in Propositions 4.11 and 4.13.

Theorem 4.1.9. *Let $\hat{X}(\cdot)$ be the diffusion process that is the unique solution to the SDE (4.51). Then the steady-state distribution of $\hat{X}(\cdot)$ has a density f given by (4.13).*

Theorem 4.1.10. *Let $\hat{X}(\cdot)$ be the diffusion process that is the unique solution to the SDE (4.54). Then the steady-state distribution of $\hat{X}(\cdot)$ has a density f given by (4.26).*

Theorems 4.1.9 and 4.1.10 can be proven similarly to Proposition 3.5 in [1].

Convergence of stationary distributions: In this section we show that $\hat{X}^r(\infty)$ converges weakly to the stationary distribution of its weak limit under the MED–FSF and MED–LB policies as $r \to \infty$. In order to prove the convergence we first show that the stationary distribution exists and then show that the sequence of stationary distributions are tight.

In order to show the existence of the stationary distribution of $\hat{X}^r$ for each r, we consider the stability of a distributed server pool system under a non-idling routing policy. We show that (Q^r, Z^r) has a stationary distribution whenever a natural traffic condition is satisfied.

Theorem 4.1.11. *Let $\pi \in \Pi$ and $\mathbb{X}_\pi^r$ be a π distributed server pool system. If*

$$\lambda^r < \sum_{j \in \mathcal{J}} \mu_j N_j^r \tag{4.59}$$

then the process (Q^r, Z^r) has a unique stationary distribution.

We present a proof in Appendix C.3.3.1. The proof is based on the relationship established in [18] between the stability of the corresponding conventional fluid limit and the positive recurrence of the underlying Markov chain.

Next we show that the sequence of the stationary distributions of the process $\{\hat{Q}^r, \hat{Z}^r\}$ is tight under any non-idling routing policy. Recall that a sequence of random variables, $\{L^r\}$, taking values in a metric space $(\mathcal{S}, \varrho)$ is said to be tight if for every $\epsilon > 0$ there exists a compact set $\mathcal{K} \subset \mathcal{S}$ such that $\inf_r \mathbb{P}\{L^r \in \mathcal{K}\} > 1 - \epsilon$ [24].

Theorem 4.1.12. *Let $\pi \in \Pi$ and $\{\mathbb{X}_\pi^r\}$ be a sequence of π distributed server pool systems. If (4.2) and (4.4) hold, then the sequence $\{\hat{Q}^r(\infty, \pi), \hat{Z}^r(\infty, \pi)\}$ is tight.*

A proof is presented in Appendix C.3.3.2. The proof is based on the results of [27]. In particular, we define two functions and show that they are geometric Lyapunov functions for these systems. Then, we use Theorem 5 in their paper to conclude the proof.

Recall that $\hat{X}^r(\infty)$ and $\hat{W}^r(\infty)$ denote the stationary distribution of the processes $\hat{X}^r$ and $\hat{W}^r$, respectively.

Theorem 4.1.13. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. If (4.2) and (4.4) hold, then*

$$\hat{X}^r(\infty) \Rightarrow \hat{X}(\infty) \text{ and} \tag{4.60}$$

$$\hat{W}^r(\infty) \Rightarrow \frac{[X(\infty)]^+}{\bar{\mu}}, \tag{4.61}$$

where $\hat{X}(\infty)$ has the density given by (4.13).

Theorem 4.1.14. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. If (4.2) and (4.4) hold, then*

$$\hat{X}^r(\infty) \Rightarrow \hat{X}(\infty) \text{ and} \tag{4.62}$$

$$\hat{W}^r(\infty) \Rightarrow \frac{[X(\infty)]^+}{\bar{\mu}}, \tag{4.63}$$

where $\hat{X}(\infty)$ has the density given by (4.26).

The proofs of Theorems 4.1.13 and 4.1.14 are presented in Appendix C.3.3.3.

Remark 4.17. It can be similarly shown that

1. if $\{\mathbb{X}^r\}$ is a sequence of FSF $\wedge$-systems that satisfy (4.2) and (4.4), then (4.60) and (4.61) hold with $\hat{X}^r$ and $\hat{W}^r$ are replaced by $\hat{X}^r_\wedge$ and $\hat{W}^r_\wedge$, respectively,

2. if $\{\mathbb{X}^r\}$ is a sequence of LB $\wedge$-systems that satisfy (4.2) and (4.4), then (4.62) and (4.63) hold with $\hat{X}^r$ and $\hat{W}^r$ are replaced by $\hat{X}^r_\wedge$ and $\hat{W}^r_\wedge$, respectively.

Proofs of the results in Section 4.1.2: Next we prove Theorems 4.1.1, 4.1.2, 4.1.4 and 4.1.5 and Corollary 1.

Proof of Theorem 4.1.1. Consider a sequence of MED–FSF distributed server systems described in Section 4.1.1. Assume that (4.2) and (4.4) hold.

Fix an adapted routing policy π and consider a sequence of π distributed server systems and a sequence of $\wedge$-systems with the rth systems in both sequences having the same arrival and service rates and number of servers in each pool. Let $Q^r_\wedge(t)$ denote the number of customers in the queue and $Z^r_{j,\wedge}(t)$ denote the number of customers in service in the jth pool at time t in the rth $\wedge$-system. Recall that $Q^r = (Q^r_1, \ldots, Q^r_J)$ and $Z^r = (Z^r_1, \ldots, Z^r_J)$ are the number of customers in the queue and in the service processes in a distributed system.

We claim that there exists an adapted routing policy π' for the $\wedge$-systems such that for a distributed system operating under the policy π and the corresponding $\wedge$-system operating under the policy π'

$$Q^r_\wedge(\infty, \pi') \sim \sum_{j \in \mathcal{J}} Q^r_j(\infty, \pi) \text{ and } Z^r_{j,\wedge}(\infty, \pi') \sim Z^r_j(\infty, \pi) \text{ for all } j \in \mathcal{J}. \tag{4.64}$$

The policy π' is constructed from the policy π as follows. Consider the distributed system and the corresponding $\wedge$-system and assume that the interarrival times of the customers to each system is equal and the service requirement of the kth customer arriving to each system is the same. The routing policy π dictates the order customers are served in the distributed system. Assume that the system is *initially empty*. The customers in the $\wedge$-system can be served in the same order and in the same server pool as follows; start the service of the kth arriving customer in the $\wedge$-system when the kth arriving customer's service in the distributed system starts and in both systems route the customer to the same server pool. Denote the routing policy in the $\wedge$-system by π'. Then,

$$Q^r_\wedge(\cdot, \pi') = \sum_{j \in \mathcal{J}} Q^r_j(\cdot, \pi) \text{ and } Z^r_{j,\wedge}(\cdot, \pi') = Z^r_j(\cdot, \pi) \text{ for all } j \in \mathcal{J} \text{ a.s.} \tag{4.65}$$

Hence, (4.64) holds. Also, π' is adapted to $(Q^r_\wedge, Z^r_\wedge)$ since π is adapted to (Q^r, Z^r).

Let

$$\hat{X}^r_\wedge(t) = (Q^r_\wedge(t) + \sum_{j \in \mathcal{J}} (Z^r_{j,\wedge}(t) - N^r_j))/\sqrt{|N^r|}$$

and $\hat{X}_\wedge(\infty, \pi')$ be the weak limit of $\hat{X}_\wedge(t, \pi')$ as $t \to \infty$ if it exists, and taken as in (4.7) otherwise. Then, by (4.65), $\hat{X}^r_\wedge(\infty, \pi') \sim \hat{X}^r(\infty, \pi)$.

Now, consider the preemptive FSF (P–FSF) policy in the same sequence of $\wedge$-systems. A preemptive policy allows a customer to be handed-off to another server, who will resume the service from the point it has been discontinued. Under the P-FSF policy in an $\wedge$-system if an arriving customer finds more than one available server he is served by the faster one. Also slower servers hand off a customer whenever a faster server becomes available.

By Proposition 3.1 of [1] in an $\wedge$-system

$$\mathbb{P}\{X^r_\wedge(\infty, \text{P–FSF}) > x\} \leq \mathbb{P}\{X^r_\wedge(\infty, \pi') > x\}$$

for any adapted policy π' and for all $x \in \mathbb{R}$. (In [1], π' is assumed to be HL but it is not required in the proof of their Proposition 3.1.) By Proposition 4.5 of [1], the argument above, and Theorem 4.1.13, we get (4.8).

Observe from Theorem 4.1.13 that

$$\lim_{r\to\infty} \mathbb{P}\{\hat{W}^r(\infty, \text{MED–FSF}) > 0\} = \lim_{r\to\infty} \mathbb{P}\{X^r(\infty, \text{MED–FSF}) > 0\},$$

since $\hat{X}(\infty)$ is a continuous random variable. We get (4.9) by combining the PASTA property of an adapted policy [71] with (4.8).

□

Proof of Theorem 4.1.2. The result follows from Proposition 4.11, Remark 4.15, Theorem 4.1.8, Remark 4.16, Theorem 4.1.9, Theorem 4.1.13 and Remark 4.17. □

Proof of Corollary 1. Note that by (4.12) we have that

$$\mathbb{P}(W^r(\infty) > 0) \to \alpha$$

as $r \to \infty$, where α is given by (4.14). Now consider $\mathbb{E}[Q^r(\infty)]$, the expected queue length in the steady state. By (4.1.13) and (C.55)

$$\mathbb{E}[Q^r(\infty)/\sqrt{|N^r|}] \to \mathbb{E}[(\hat{X}(\infty))^+] \tag{4.66}$$

as $r \to \infty$, where

$$\mathbb{E}[(\hat{X}(\infty))^+] = \alpha\frac{\sqrt{\overline{\mu}}}{\theta} \tag{4.67}$$

by Theorem 4.1.9. By combining (4.66), (4.67), (4.58), (4.2) and (4.4), we obtain

$$\left|\mathbb{E}[\hat{W}^r(\infty)] - \frac{\sqrt{|N^r|}\mathbb{E}[(\hat{X}(\infty))^+]}{\sqrt{r}}\right| = \left|\mathbb{E}[\hat{W}^r(\infty)] - \alpha\frac{\sqrt{|N^r|\bar{\mu}}}{\sqrt{r}\theta}\right| \to 0,$$

where α is given by (4.14). □

Proof of Theorem 4.1.3. For $i \geq 2$, by Proposition 4.7 and Theorem 4.1.12

$$\left|\frac{\hat{Z}_i^r(\infty)}{\beta_i}\right| \to 0 \text{ in probability as } r \to \infty.$$

and

$$\left|\hat{Z}_1^r(\infty)\right| \to \left(\hat{X}(\infty)\right)^- \text{ in probability as } r \to \infty.$$

By (C.55) the sequence

$$\left\{\left|\frac{\hat{Z}_i(\infty)}{\beta_i}\right|\right\}$$

is uniformly integrable. This yields (4.17) by Theroem 4.5.4 of [16]. □

Proof of Theorem 4.1.4. The convergence in (4.22) follows from Proposition 4.9. Next we prove (4.19). For $i, j \in \mathcal{J}$ define

$$\Upsilon_{ij}^r = \left|\frac{\hat{Z}_i^r(\infty)}{\beta_i} - \frac{\hat{Z}_j^r(\infty)}{\beta_j}\right|.$$

Observe that by Proposition 4.9 and Theorem 4.1.12

$$\Upsilon_{ij}^r \to 0 \text{ in probability as } r \to \infty. \tag{4.68}$$

By (C.55) the sequence $\{\Upsilon_{ij}^r\}$ is uniformly integrable. This yields (4.19) by (4.68) and Theroem 4.5.4 of [16]. □

Proof of Theorem 4.1.5. The last equality in (4.24) follows from Theorem 1 of [32] and the last equality in (4.25) can be proved using arguments similar to those in the proof of Theorem 3 in [29]. The first and the second equalities in (4.24) and (4.25) follows from Proposition 4.13, Remark 4.15, Theorem 4.1.8, Remark 4.16, Theorem 4.1.10, Theorem 4.1.14 and Remark 4.17. □

4.1.5 Concluding Remarks

Under the FSF policy all the servers except those with the lowest service rate are utilized 100% and under the LB policy the utilizations of all the servers are equal. The LB policy can be modified to distribute the available percentage of idle time, which is equal to $(1-\rho^r)$ in the rth system, in desired proportions among all the server pools. To illustrate this let $d = (d_1, \ldots, d_J)$. Under the modified LB with parameter d (MLB_d) policy (also under the MED–MLB_d policy) when there are idle servers in the system a customer arriving arriving to the system at time t is routed to the server pool with minimum

$$\frac{Z_j^r(t) - N_j^r}{d_j N_j^r}.$$

Note that if $d_1 = d_2 = \ldots = d_J$ this policy reduces to the original LB policy. If $d_j < d_k$ for $k, j \in \mathcal{J}$, the utilization of the server pool j will be more than the utilization of the server pool k . Therefore, the utilizations of all servers can be controlled by assigning appropriate values to d. One can show similar to (4.22) that

$$\sqrt{|N^r|}\left\|\frac{Z_i^r(t) - N_i^r}{d_i N_i^r} - \frac{Z_j^r(t) - N_j^r}{d_j N_j^r}\right\|_T \to 0 \text{ in probability as } r \to \infty.$$

Similar results that are established for the systems operating under the LB and MED–LB routing policies in Section 4.1.2 can also be shown to hold under the MLB_d and MED–MLB_d routing policies. In particular, under the MLB_d and MED–MLB_d routing policies (4.16) hold with

$$\alpha = \left[1 + \frac{\theta/\sqrt{\tilde{\mu}}\Phi\left(\theta/\sqrt{\tilde{\mu}}\right)}{\phi\left(\theta/\sqrt{\tilde{\mu}}\right)}\right]^{-1},$$

where

$$\tilde{\mu} = \frac{\sum_{j=1}^J d_j \beta_j \mu_j}{\sum_{j=1}^J \beta_j \mu_j}.$$

Let the limiting stationary delay probability under a policy π be denoted by α_π. If $1 = d_1 \geq d_2 \geq \ldots \geq d_J$, then it can easily be shown under the assumption (4.1) that

$$\alpha_{\text{FSF}} \leq \alpha_{\text{MLB}_\text{d}} \leq \alpha_{\text{LB}}. \tag{4.69}$$

If $\mu_1 < \mu_2 \leq \mu_3 \leq \ldots \leq \mu_J$ and $1 = d_1 > d_2 \geq \ldots \geq d_J$, then the inequalities in (4.69) are strict and as $d_2 \downarrow 0$, $\alpha_{\mathrm{MLB_d}} \downarrow \alpha_{\mathrm{FSF}}$, and as $d_J \uparrow 1$, $\alpha_{\mathrm{MLB_d}} \uparrow \alpha_{\mathrm{LB}}$.

In some applications one or more pools have significantly fewer servers than the other pools. In this case our approximations may perform poorly because our results are based on the asymptotic analysis when all the server pools have significantly many servers. Extensions of the MED–LB routing policy shall be considered for these systems.

The minimum expected delay-longest idle agent (MED–LI) routing policy is a commonly used policy to balance utilizations of the servers in call center industry. Under the MED–LI policy an arriving customer who finds idle servers upon arrival is routed to the agent who has been idle the longest at the decision instant. If all the servers are busy at the time of an arrival, the customer is routed to the queue with the minimum expected delay. The MED–LI policy is available by default in most commercial automatic call distributers; see, for example, Cisco Intelligent Contact Management [17]. In future research, the performance of the MED–LB policy will be compared with that of the MED–LI policy.

4.2 Asymptotically optimal control policies for N-systems

In this section we consider parallel servers systems with an N-design, or N-systems for short, with many servers. An N-system consists of two customer classes and two server pools. The servers in the first pool can only serve class 1 customers whereas the second pool can serve either class. An N-system is illustrated in Figure 2. We assume that all the service times are exponentially distributed and $\mu_{11} < \mu_{21} = \mu_{22}$. For the rest of this section we set $\mu_{11} = \mu_1$ and $\mu_{21} = \mu_{22} = \mu_2$.

Our objective is to minimize the total holding cost during a finite time interval. Let h_i denote the holding cost for each unit time a customer is held in queue i. We assume that $h_1 > h_2$; i.e.; it costs more to have a class 1 customer wait in the queue than a class 2 customer. The objective function we consider is

$$\min_{\pi \in \Pi} \int_0^T (h_1 Q_1(t) + h_2 Q_2(t)) \mathrm{d}t.$$

Above, Π denotes the class of admissible policies described in Section 2.1.

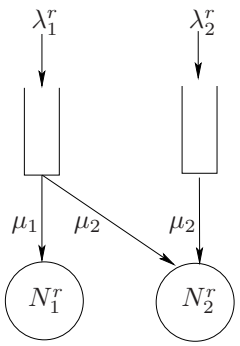

Figure 2: An N-system

We focus on N-systems where the first server pool cannot handle all the load from class 1 customers and needs help from the second server pool. For the admissible scheduling policies described in Section 2.1, a server from the second pool can be assigned to serve a class 1 customer in two situations: first, when a class 1 customer arrives to the system to find idle agents in that pool and second, when a server in pool 2 finishes service and finds a customer in the first queue.

Consider the following static priority policy, which we denote by π^* in the rest of this section;

a. If $Q_1^r(t) \geq 1$, servers in pool 2 give priority to class 1 customers, i.e.; whenever a server finishes a service and finds a customer waiting in class 1 queue he starts serving the class 1 customer who has been waiting the longest. If there are no customers in class 1 at that instant then he checks class 2 queue.

b. For class 1 customers, server pool 2 has priority over server pool 1, i.e.; upon arrival if a class 1 customer finds idle servers in both pools he starts his service in the second server pool.

c. Servers do not idle when there is a customer they can handle is waiting in the queue.

The main result of this section is that π^* is asymptotically optimal for the N-systems under the Halfin-Whitt many server regime.

Although, the class of N-systems we consider in this section is admittedly seems to be

arbitrary and constrained to satisfy certain conditions, our results can be extended in several directions that we do not explore here. For example, if the service rate of the first server pool is greater than that of the second server pool and $h_2 > h_1$ then another static priority policy can be shown to be asymptotically optimal. However, our analysis technique cannot be used when $\mu_{21} \neq \mu_{22}$. The main objective of this section is to show that how our SSC framework can be used to identify asymptotically optimal policies for parallel server systems with many servers. In conventional heavy traffic analysis, N-systems provided an important stepping stone for the analysis of more general systems, see Harrison and Lopez [41], Bell and Williams [10], and Ata and Kumar [4]. Future research will address extending our analysis in this section to more general parallel server systems.

The rest of this section is organized as follows. In the following section we review the related literature and motivate our results. We give a detailed description of N-systems we study in Section 4.2.2. In Section 4.2.3, we present our main results. The remaining sections are devoted to the proofs of our main results.

4.2.1 Previous Work and motivation

N-systems have been analyzed under the conventional heavy traffic in the literature extensively. In Harrison [39], it is shown through simulation experiments that the policy π^* explained above can be unstable even when the system have enough capacity to serve all the customers. Surprisingly, we show that this policy is asymptotically optimal under the Halfin-Whitt many server regime.

Next, we present the results of a simulation experiment that is similar to Harrison's to illustrate this phenomenon. Let $N_1 = N_2 = 1$, $\lambda_1 = 1.4\rho$, $\lambda_2 = 0.7\rho$, $\mu_1 = 0.7$, and $\mu_2 = 1$. We take the unit time to be one minute. Theoretically, if $\rho < 1$ then the system has enough capacity to be stable. However, under the static priority policy described above this will not be the case. We set $\rho = 95\%$ and simulate this N-system for 1000 hours. Figure 4.2.1 shows the queue lengths from this simulation as a function of time. It is clearly seen from this figure that the static priority policy is not even stable in this case.

On the other hand, the static priority policy performs a lot better in the presence of

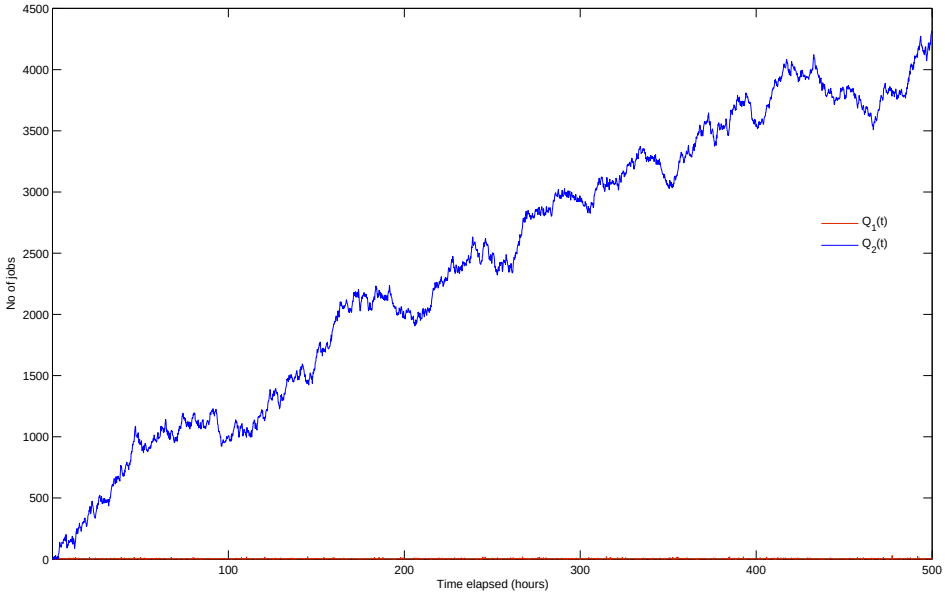

Figure 3: Behavior of the N-system with one server in each pool under π^*.

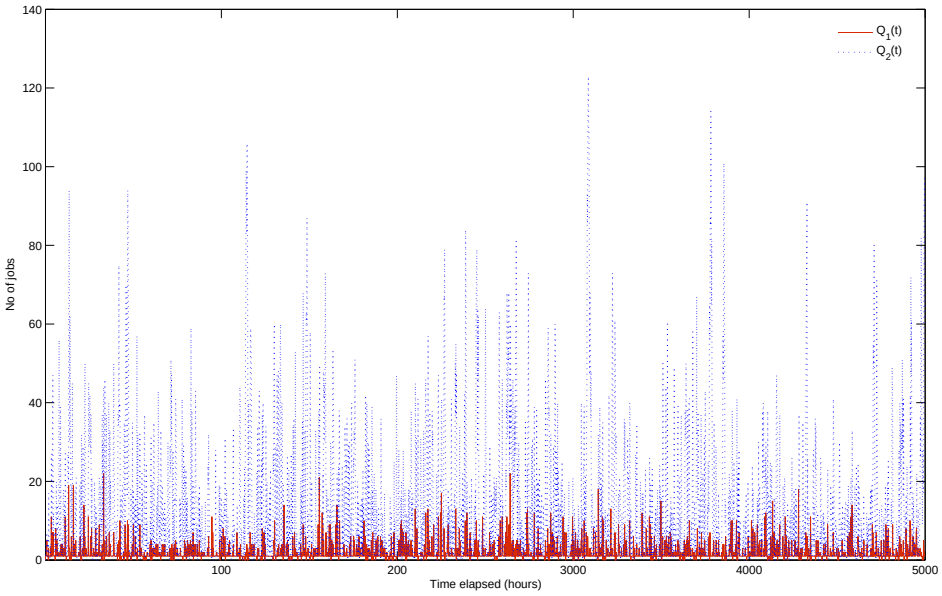

Figure 4: Behavior of the N-system with 20 servers in each pool under π^*.

many servers than it does in a single server setting. To illustrate, we simulate a similar N-system with many servers. For this simulation experiment, we multiply the arrival rates in the system simulated above by 20 and set the number of servers in both pools equal to 20. Note that the total load on the system divided by the total capacity of the system is the same with that of the model simulated previously. Figure 4.2.1 shows the queue lengths from this simulation as a function of time. It is clear from this figure that this realization of the system is stable. Also, the number of customers in the first queue is very small. As illustrated by this example, even when the number of servers and the arrival rates in the system are increased by only 20 times the N-system under the static priority policy becomes stable.

Remark 4.18. We do not claim that the static priority policy is stable for an N-system with many servers whenever $\rho < 1$. In fact, for each N-system with multiple servers one can find a value $\bar{\rho}$ such that for $\bar{\rho} < \rho < 1$ that system is unstable. However, we conjecture that $\sqrt{|N^r|}(1 - \bar{\rho}) \to 0$ as the number of servers go to infinity.

For N-systems under conventional heavy traffic, Harrison [39] proposed a discrete-review policy to minimize the holding costs and proved that it is asymptotically optimal in the conventional heavy traffic limit. Harrison and Lopez [41] extended his results to parallel server systems.

To mitigate the problems with the static priority policy, Bell and Williams [9] proposed a threshold type static priority policy and showed that it is also asymptotically optimal in conventional heavy traffic limit. They extend their results to more general systems in Bell and Williams [10]. Their policy requires a threshold value and only when the number of customers in the first queue exceeds this value servers in the second pool give priority to the customer in the first queue. They show that the threshold value can be taken to be in the order of $\log(1/(1 - \rho))$, where ρ is the optimal solution to the static planning problem that is similar to (2.21). To illustrate the performance of their policy, we simulate the N-system with one server in each pool we considered above with threshold value equal to 5. Figure 4.2.1 shows the queue lengths from this simulation as a function of time. As opposed to the N-system under the static priority policy, the number of customers in the

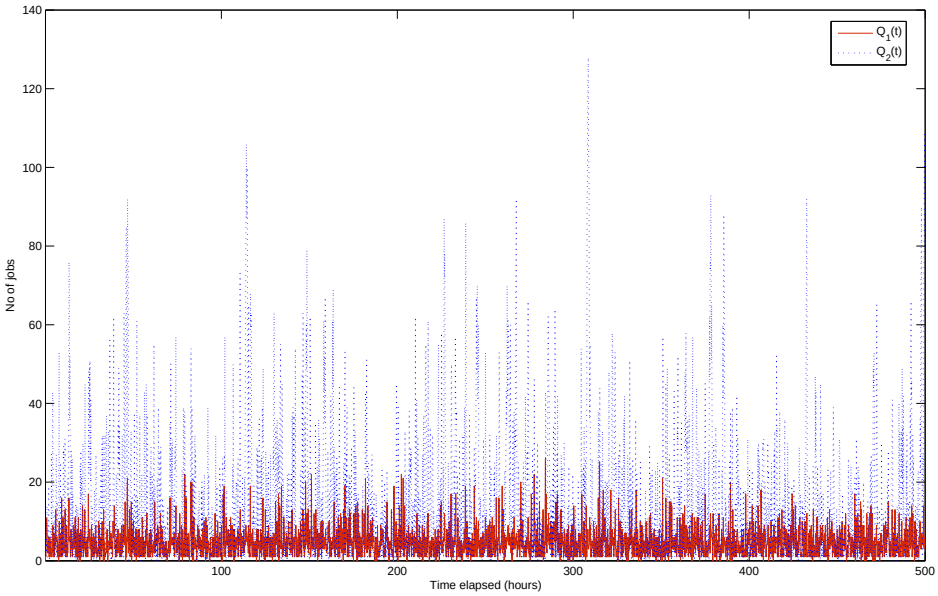

Figure 5: Behavior of the N-system with one server in each pool under Bell-Williams policy.

second queue in this case does not present a growing trend. Our result reveals that one does not need a threshold when the number of servers get large. However, in the light of Remark 4.18 one may want to use a threshold value in the order of $\log(1/(1-\rho))$. It can be shown that the static priority policy with a threshold value that is $o(\sqrt{1/(1-\rho)})$ is still asymptotically optimal

As mentioned above, we show that π^* is asymptotically optimal for the N-systems under the Halfin-Whitt regime. Other policies, which are more complicated than π^*, have been shown to be asymptotically optimal for similar systems. Harrison and Zeevi [38] and Atar et al. [7] studied V-systems and Atar [6, 5] studied tree-like parallel server systems. These papers considered dynamic control in the Halfin-Whitt regime. They focused on a formal derivation of a diffusion control problem (DCP) and obtain optimal control policies. However, specification of the optimal policies uses the solution of a set of partial differential equations (PDE's) and this set of PDE's can only be solved numerically. Also, the parameters of these PDE's depend on the system parameters therefore the control policies they obtain is sensitive to the changes in these parameters. In Atar et al. [7] and Atar [6, 5],

they derive control policies for the actual system from the optimal control of the DCP and prove that these policies are asymptotically optimal. Since the policies can only be specified numerically, approximations to the system performance cannot be obtained under these policies. Our purpose is to come up with a simple optimal policy that can also be used to approximate the optimal system performance.

4.2.2 Model description

As noted above, we analyze a sequence of N-systems indexed by r under the Halfin-Whitt regime. Recall that $\lambda^r = (\lambda_1^r, \lambda_2^r)$ denotes the arrival rates to each class, $N^r = (N_1^r, N_2^r)$ denotes the number of servers in each server pool and $|N^r|$ denote the total number of servers in the rth system in this sequence.

We assume that the system reaches heavy traffic as r gets large. We also assume that the total capacity of the servers in pool 1 is not enough to handle all class 1 customers. Therefore, second server pool must serve some of the class 1 customers for stability of the system. Specifically, we assume that there exist $\lambda_i > 0$, $k = 1, 2$, $\beta_j > 0$, $j = 1, 2$, such that

$$\frac{\lambda_k^r}{|N^r|} \to \lambda_k, \text{ as } r \to \infty \text{ for } k = 1, 2 \text{ and } N_j^r = |N^r|\beta_j, \tag{4.70}$$

and $x_{21}, x_{22} > 0$ that satisfy $x_{21} + x_{22} = 1$ and

$$\lambda_1 = \beta_1\mu_1 + \beta_2 x_{21}\mu_2, \ \lambda_2 = \beta_2 x_{22}\mu_2. \tag{4.71}$$

In addition we assume that

$$\lambda_1^r = \mu_1 N_1^r + \mu_2 x_{21} N_2^r - \sqrt{|N^r|}\theta_1 \text{ and} \tag{4.72}$$

$$\lambda_2^r = \mu_2 x_{22} N_2^r - \sqrt{|N^r|}\theta_2. \tag{4.73}$$

for $\theta_1, \theta_2 \in \mathbb{R}$ and we set $\theta = \theta_1 + \theta_2$. Under assumptions (4.70)-(4.73), Assumption 1 in Section 2.3 holds.

Since there is no routing we only use one subscript with each queue. The notation used in this section differs from that introduced in Chapter 2 only in that we omit the subscript "i" in the notation used in this section. For example, instead of $A_{121}(t)$ we use $A_{21}(t)$ to denote the number of class 1 customers whose service started in the second server pool

immediately at the time of their arrival who arrived to the system by time t. For notational convenience we denote the total number of class k customers in the system at time t by $X_k^r(t)$ hence

$$X_k^r(t) = Q_k^r(t) + \sum_{j \in J(k)} Z_{jk}^r(t).$$

By (4.70)-(4.73), the diffusion scaling defined at the end of Section 2.3 becomes

$$\hat{Z}_{11}^r(t) = \frac{Z_{11}^r(t) - N_1^r}{\sqrt{|N^r|}},\ \hat{Z}_{21}^r(t) = \frac{Z_{21}^r(t) - x_{21}N_2^r}{\sqrt{|N^r|}},\ \hat{Z}_{22}^r(t) = \frac{Z_{22}^r(t) - x_{22}N_2^r}{\sqrt{|N^r|}}$$

For the rest this chapter we assume that

$$(\hat{Q}^r(0), \hat{Z}^r(0)) \Rightarrow (\hat{Q}(0), \hat{Z}(0)) \text{ and } (\bar{Q}^r(0), \bar{Z}^r(0)) \to (\bar{Q}(0), \bar{Z}(0)) \text{ a.s.} \tag{4.74}$$

as $r \to \infty$, where $\bar{Q}(0) = (0, 0)$ and $\bar{Z}(0) = (\beta_1, x_{21}\beta_2, x_{22}\beta_2)$.

4.2.3 Main results

As mentioned above we focus on minimizing the linear holding cost in the queues. The instantaneous diffusion scaled holding cost at time t under policy π is defined by

$$H^{r,\pi}(t) = \sum_{i \in \mathcal{I}} h_i \hat{Q}_i^r(t)$$

We define the cumulative cost process under an admissible policy π by

$$\zeta^{r,\pi}(t) = \int_0^t H^{r,\pi}(s)ds \text{ for } t \geq 0.$$

Let the process Y^* be the solution of the following SDE

$$Y^*(t) = \hat{X}_1(0) + \hat{X}_2(0) - \theta t + r_1 W_1(t) + r_2 W_2(t) + \mu_1 \int_0^t (Y^*(s))^- ds,\ t \geq 0,$$

where

$$r_i = (\lambda_i C_{U,i}^2 + \lambda_i)^{1/2} \tag{4.75}$$

and $C_{U,i}^2$ is the coefficient of variation of the interarrival times of the process E_i defined in Chapter 2.2 and for real x $x^- = (-x) \wedge 0$ and $x^+ = x \wedge 0$. We define

$$\zeta^*(t) = \int_0^t h_2 (Y^*(s))^+ ds. \tag{4.76}$$

The following result shows that ζ^* provides a lower bound for the cumulative holding costs under any admissible policy as $r \to \infty$.

Theorem 4.19. *Let π be an admissible policy and $\{\mathbb{X}^{r,\pi}\}$ be a sequence of N-systems working under π. Assume that (4.70)-(4.74) hold. Then for each fixed $x > 0$*

$$\liminf_{r\to\infty} \mathbb{P}\{\zeta^{r,\pi}(T) > x\} \geq \mathbb{P}\{\zeta^*(T) > x\}.$$

With Theorem 4.19 the following result shows that π^* is asymptotically optimal.

Theorem 4.20. *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-systems working under the static priority policy π^*. Assume that (4.70)-(4.74) hold. For each fixed $t > 0$ and $x > 0$*

$$\lim_{r\to\infty} \mathbb{P}\{\zeta^{r,\pi^*}(t) > x\} = \mathbb{P}\{\zeta^*(t) > x\}. \tag{4.77}$$

In the proof of Theorem 4.20 we use the SSC framework established in Chapter 3 to prove that in the limit as $r \to \infty$

i. the queue length process for the first queue is always zero,

ii. the servers in the second pool never idle,

iii. the servers in the first pool may only idle if there are no customers in either queue.

Using the last two SSC results we will show that the total number of customers in the system is minimized. The first SSC result ensures that customers are held only in the cheap queue.

4.2.4 Analysis of an admissible policy

In this section we prove Theorem 4.19. The proof consists of two steps. First, we define a mapping which we use to obtain a lower bound for the total number of customers in the system. In the second step, we prove that this mapping indeed provides a lower bound for the total cost under any admissible policy.

4.2.4.1 A minimizing mapping

In this section, we define a mapping on $D[0,\infty)$ that we use to obtain a lower bound on the total number of customers in the system. For $x \in D[0,\infty)$, the mapping $\psi : D[0,\infty) \to D[0,\infty)$ is defined by

$$\psi(x) = y$$

where

$$y(t) = x(t) + \mu_1 \int_0^t (y(s))^- ds. \tag{4.78}$$

Next we establish the basic properties of ψ

Lemma 4.21. *For each $x \in D[0,\infty)$ there exists a unique y that satisfies (4.78). Furthermore, ψ is Lipschitz continuous.*

Proof. Let $x \in D[0,\infty)$ and $\mu_1 > 0$. For $n \geq 0$, let

$$y^{n+1}(t) = x(t) + \mu_1 \int_0^t (y^n(s))^- ds \tag{4.79}$$

and

$$Y^{(n)}(t) = \|y^{n+1}(s) - y^n(s)\|_t.$$

This gives us

$$Y^{(n+1)}(t) = \|y^{n+1}(s) - y^n(s)\|_t \leq \mu_1 \int_0^t \left|y^{n+1}(s) - y^n(s)\right| ds = \mu_1 \int_0^t Y^{(n)}(s) ds.$$

Hence, by Lemma 11.3 in Mandelbaum et. al [51],

$$Y^{n+1}(t) \leq \mu_1 \frac{T^n}{n!} \sup_{0 \leq s \leq t} Y^{(0)}(s)$$

Therefore, similar to (11.22) in Mandelbaum et. al [51], $\{y^n(\cdot)\}$ is a Cauchy sequence hence converges to a limit y uniformly on compact sets. This proves existence.

To prove continuity, assume that $x_i(t) \in D[0,\infty)$ for $i = 1, 2$. Then for any $T > 0$

$$|\psi(x_1)(t) - \psi(x_2)(t)| \leq |x_1(t) - x_2(t)| + \mu_1 \int_0^t |\psi(x_1)(s) - \psi(x_2)(s)| ds$$

By Corollory 11.2 in in Mandelbaum et. al [51],

$$\|\psi(x_1)(t) - \psi(x_2)(t)\|_T \leq \mu_1 \|x_1(t) - x_2(t)\|_T e^T.$$

□

Next we show that, for each $x \in D[0,\infty)$, y is the minimum function that satisfies ceratin conditions. In the following section we deal with the total time allocated to each class. Let $T_i \in C[0,\infty)$, $i = 1, 2$ and

$$T_i(t;s) = T_i(t) - T_i(s).$$

In the next section we use $\hat{T}_{jk}(t)$ to denote the diffusion scaled deviation of the total time allocated from the jth server pool to the kth customer class from its nominal value $x_{jk}\beta_j t$.

Given $x \in D[0,\infty)$, and $\mu_1 < \mu_2$ assume that $(\tilde{y}, T_1, T_2) \in D^3[0,\infty)$ satisfy

$$\tilde{y}(t) = x(t) - \mu_1 T_1(t) - \mu_2 T_2(t) \tag{4.80}$$

$$T_1(t;s) \leq 0, T_2(t;s) \leq 0 \tag{4.81}$$

$$T_1(t;s) + T_2(t;s) \leq -\int_s^t (\tilde{y}(s))^- ds \tag{4.82}$$

for all $t \geq 0$.

Theorem 4.22. *Let $x \in D[0,\infty)$ and assume that $(\tilde{y}, T_1, T_2) \in D[0,\infty) \times C[0,\infty) \times C[0,\infty)$ satisfy (4.80)-(4.82), then for each fixed $T > 0$*

$$\sup_{0 \leq t \leq T} \{\psi(x)(t) - \tilde{y}(t)\} \leq 0. \tag{4.83}$$

In particular,

$$\int_0^T \psi(x)(t)\,dt \leq \int_0^T \tilde{y}(t)\,dt.$$

Proof. Let $x \in D[0,\infty)$ and assume that $\tilde{y} \in D[0,\infty)$ and $(T_1, T_2) \in C^2[0,\infty)$ satisfy (4.80)-(4.82).

First we show that

$$\psi(x)(t) - \psi(x)(t-) = \tilde{y}(t) - \tilde{y}(t-) \text{ for all } t \in [0,T]. \tag{4.84}$$

We prove (4.83) using this result.

Fix $T > 0$. Since $x \in D[0,\infty)$, by Lemma 1 in Billingsley [11], there exists $M > 0$ such that

$$\|x(t)\|_T < M.$$

By (4.78)

$$\psi(x)(t) \geq x(t) > -M.$$

Therefore,

$$\|\psi(x)(t)\|_T \leq \|x(t)\|_T + \mu_2 \int_0^T |(\psi(x)(s))^-| ds \leq (M + \mu_2 M)T$$

This gives that

$$\mu_1 \int_0^t (\psi(x)(s))^- ds$$

is continuous on $[0, T]$. Since T_1 and T_2 are continuous (4.84) holds.

Now assume that there exists $t_0 \in [0, T]$ such that

$$\tilde{y}(t_0) < \psi(x)(t_0). \tag{4.85}$$

Let

$$s_0 = \sup\{0 \leq s \leq t : \tilde{y}(s) - \psi(x)(s) = 0\},$$

so that

$$\tilde{y}(t) < \psi(x)(t) \text{ for all } t \in (s_0, t_0], \text{ and } \tilde{y}(s_0) = \psi(x)(s_0) \tag{4.86}$$

since $\tilde{y}(s) - \psi(x)(s)$ is continuous by (4.84) and $x \in D[0, \infty)$.

Then,

$$\begin{aligned}
\tilde{y}(t_0) - \psi(x)(t_0) &= \mu_1 \int_{s_0}^{t_0} (\psi(x)(s))^- ds - \mu_1 T_1(t_0, s_0) - \mu_2 T_2(t_0, s_0) \\
&\geq \mu_1 \int_{s_0}^{t_0} (\psi(x)(s))^- ds - \mu_1 \left(T_1(t_0, s_0) + T_2(t_0, s_0)\right) \\
&\geq \mu_1 \int_{s_0}^{t_0} (\psi(x)(s))^- - (\tilde{y}(s))^- ds \\
&\geq 0.
\end{aligned}$$

by (4.86). This contradicts with (4.85). □

4.2.4.2 Proof of Theorem 4.19

Fix an admissible policy π. Note that by our definition of an admissible policy $Z^r_{jk} \in D[0,\infty)$ a.s. for each r. Let

$$\hat{T}^r_{jk}(t) = \sqrt{|N^r|}\left(\frac{T^r_{jk}(t)}{|N^r|} - x_{jk}\beta_j t\right) = \int_0^t \hat{Z}^r_{jk}(s)ds.$$

Observe that $\hat{T}^r_{jk}$ satisfies the following conditions; for all $0 \le s \le t$

$$-N^r_1(t-s)/\sqrt{|N^r|} \le \hat{T}^{r,\pi}_{11}(t;s) \le 0 \tag{4.87}$$

$$-N^r_2(t-s)/\sqrt{|N^r|} \le \hat{T}^{r,\pi}_{21}(t;s) + \hat{T}^{r,\pi}_{22}(t;s) \le 0 \tag{4.88}$$

$$\hat{T}^{r,\pi}_{11}(t;s) + \hat{T}^r_{21}(t;s) \le \int_s^t \hat{X}^{r,\pi}_1(u)\mathrm{d}u \tag{4.89}$$

$$\hat{T}^{r,\pi}_{22}(t;s) \le \int_s^t \hat{X}^{r,\pi}_2(u)\mathrm{d}u. \tag{4.90}$$

Proof of Theorem 4.19. The idea of the proof is similar to that of Proposition 2 in Ata and Kumar [4]. Fix an admissible policy $\pi > 0$, $T > 0$ and $x > 0$. Choose a subsequence r_j such that

$$\lim_{r_j\to\infty} \mathbb{P}\{\zeta^{r_j,\pi}(T) > x\} = \liminf_{r\to\infty} \mathbb{P}\{\zeta^{r,\pi}(T) > x\}$$

Let $\{T^{r,\pi}\}$ be the sequence of allocation processes under policy π. Since

$$\bar{T}^{r,\pi}_{jk}(t) - \bar{T}^{r,\pi}_{jk}(s) \le \frac{N^r_j}{|N^r|}(t-s),$$

the sequence

$$\left\{\left(\hat{A}^{r_j}(\cdot), \hat{S}^{r_j}(\cdot), \bar{T}^{r,\pi}(\cdot)\right)\right\}$$

is tight and any weak limit of this sequence has continuous paths almost surely. In particular, the limit is of the following form

$$\left(\hat{A}^*(\cdot), \hat{S}^*(\cdot), \bar{T}^{\pi}(\cdot)\right), \tag{4.91}$$

where $\hat{A}^*$ and $\hat{S}^*(\cdot)$ are driftless Brownian motions of appropriate dimension, $\bar{T}^\pi$ is a non-decrasing process with

$$\bar{T}^\pi(t) - \bar{T}^\pi(s) \le (t-s)e, \quad \text{for } 0 \le s \le t \text{ a.s.}$$

for all $t \geq 0$.

Let $\{\bar{T}^{r',\pi}\}$ be a further subsequence of $\{\bar{T}^{r_j,\pi}\}$ which converges weakly to a limit as in (4.91). By appealing to the Skorohod representation theorem, we may choose an equivalent distributional representation (which we will denote by putting a " $\sim$ " above the symbols) such that the sequence of random processes

$$\left(\tilde{\hat{A}}^{r',\pi}(\cdot), \tilde{\hat{S}}^{r',\pi}(\cdot), \tilde{\bar{T}}^{r',\pi}(\cdot)\right),$$

as well as the limit

$$\left(\tilde{\hat{A}}^{*}(\cdot), \tilde{\hat{S}}^{*}(\cdot), \tilde{\bar{T}}^{\pi}(\cdot)\right)$$

are defined on a new probability space, say $\left(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}\right)$, so that $\tilde{P}$ a.s.

$$\left(\tilde{\hat{A}}^{r',\pi}(\cdot), \tilde{\hat{S}}^{r',\pi}(\cdot), \tilde{\bar{T}}^{r',\pi}(\cdot)\right) \rightarrow \left(\tilde{\hat{A}}^{*}(\cdot), \tilde{\hat{S}}^{*}(\cdot), \tilde{\bar{T}}^{\pi}(\cdot)\right), \tag{4.92}$$

u.o.c. as $r' \rightarrow \infty$. We can also assume without loss of generality that there exists a sequence of random vectors in $(\tilde{Q}^r(0), \tilde{Z}^r(0))$ in this new space that is independent from all stochastic processes in 4.92 such that $(\tilde{Q}^r(0), \tilde{Z}^r(0))$ has the same distribution with $(Q^r(0), Z^r(0))$.

We define the following processes on this new probability space:

$$\begin{aligned}
\tilde{\bar{A}}_k^{r',\pi}(t) &= \frac{1}{\sqrt{|N^r|}} \tilde{\hat{A}}_k^{r',\pi}(t) + \frac{\lambda_k^r t}{|N^r|} && (4.93)\\
\tilde{\bar{S}}_{jk}^{r',\pi}(t) &= \frac{1}{\sqrt{|N^r|}} \tilde{\hat{S}}_{jk}^{r',\pi}(t) + \mu_{jk} t. && (4.94)\\
\tilde{\hat{T}}_{jk}^{r',\pi}(t) &= \sqrt{|N^r|}\left(\tilde{\bar{T}}_{jk}^{r',\pi}(t) - \beta_i x_{jk} t\right) && (4.95)\\
\tilde{\hat{X}}_1^{r',\pi}(t) &= \tilde{\hat{X}}_1^r(0) + \tilde{\hat{A}}_1^{r',\pi}(t) - \tilde{\hat{S}}_{11}^{r',\pi}(\tilde{\bar{T}}_{11}^{r',\pi}(t)) - \tilde{\hat{S}}_{21}^{r',\pi}(\tilde{\bar{T}}_{21}^{r',\pi}(t))\\
&\quad - \mu_1 \tilde{\hat{T}}_{11}^{r',\pi}(t) - \mu_2 \tilde{\hat{T}}_{21}^{r',\pi}(t) && (4.96)\\
\tilde{\hat{X}}_2^r(t) &= \tilde{\hat{X}}_2^r(0) + \tilde{\hat{A}}_2^{r',\pi}(t) - \tilde{\hat{S}}_{22}^{r',\pi}(\tilde{\bar{T}}_{22}^{r',\pi}(t)) - \mu_2 \tilde{\hat{T}}_{22}^{r',\pi}(t), && (4.97)
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\hat{X}}_1^r(0) &= \tilde{\hat{Q}}_1^r(0) + \tilde{\hat{Z}}_{11}^r(0) + \tilde{\hat{Z}}_{21}^r(0),\\
\tilde{\hat{X}}_2^r(0) &= \tilde{\hat{Q}}_2^r(0) + \tilde{\hat{Z}}_{22}^r(0)
\end{aligned}$$

and

$$\tilde{\hat{Q}}_k^r(0) = \tilde{Q}_k^r(0)/\sqrt{|N^r|} \text{ and } \tilde{\hat{Z}}_{jk}^r(0) = \frac{\tilde{Z}_{jk}^r(0) - x_{jk}\beta_j}{\sqrt{|N^r|}}$$

We note that processes defined by (4.93)-(4.97) have the same joint distribution as the corresponding scaled processes in the original probability space for each r. Also, since in the original space $\hat{T}^{r,\pi}$ satisfies (4.87)-(4.90) and $\left(\hat{X}_1^{r,\pi}(\cdot), \hat{X}_2^{r,\pi}(\cdot), \hat{T}^{r,\pi}(\cdot)\right)$ and $\left(\tilde{\hat{X}}_1^{r,\pi}(\cdot), \bar{\tilde{X}}_2^{r,\pi}(\cdot), \tilde{\bar{T}}^{r,\pi}(\cdot)\right)$ are equal in distribution, we have

$$\tilde{\bar{T}}^{r,\pi}(\cdot) \text{ is nondecreasing a.s.}$$

$$\tilde{\bar{T}}^{r,\pi}(t;s) \leq (t-s)e \text{ for } s \leq t \text{ a.s.}$$

and $\tilde{\bar{T}}^{r,\pi}$ satisfies (4.87)-(4.90) with $\hat{X}_1^{r,\pi}$ and $\hat{X}_2^{r,\pi}$ are replaced by $\tilde{\hat{X}}_1^{r,\pi}$ and $\tilde{\hat{X}}_2^{r,\pi}$, respectively:

$$-N_1^r(t-s)/\sqrt{|N^r|} \leq \tilde{\hat{T}}_{11}^{r,\pi}(t;s) \leq 0 \tag{4.98}$$

$$-N_2^r(t-s)/\sqrt{|N^r|} \leq \tilde{\hat{T}}_{21}^{r,\pi}(t;s) + \tilde{\hat{T}}_{22}^{r,\pi}(t;s) \leq 0 \tag{4.99}$$

$$\tilde{\hat{T}}_{11}^{r,\pi}(t;s) + \tilde{\hat{T}}_{21}^{r}(t;s) \leq \int_s^t \tilde{\hat{X}}_1^{r,\pi}(u)\mathrm{d}u \tag{4.100}$$

$$\tilde{\hat{T}}_{22}^{r,\pi}(t;s) \leq \int_s^t \tilde{\hat{X}}_2^{r,\pi}(u)\mathrm{d}u. \tag{4.101}$$

We have that

$$\left(\tilde{\bar{S}}^{r',\pi}(\cdot), \tilde{\bar{A}}^{r',\pi}(\cdot)\right) \to (\mu(\cdot), \lambda(\cdot)) \tag{4.102}$$

where

$$\begin{aligned} \mu(t) &= (\mu_1 t, \mu_2 t, \mu_2 t) \text{ and} \\ \lambda(t) &= (\lambda_1 t, \lambda_2 t) \end{aligned}$$

for all $t \geq 0$. We also define

$$\begin{aligned}
\tilde{\bar{X}}_1^{r,\pi}(t) &= \frac{1}{\sqrt{|N^r|}}\left(\tilde{\hat{X}}_1^{r,\pi}(t)\right) + \beta_1 + x_{21}\beta_2 \\
&= \tilde{\hat{X}}_1^{r}(0) + \tilde{\hat{A}}_1^{r',\pi}(t) - \tilde{\hat{S}}_{11}^{r',\pi}\left(\tilde{\hat{T}}_{11}^{r',\pi}(t)\right) - \tilde{\hat{S}}_{21}^{r',\pi}\left(\tilde{\hat{T}}_{21}^{r',\pi}(t)\right) \\
&\quad - \frac{\lambda_1^r t}{|N^r|} + \beta_1\mu_1 + x_{21}\beta_2\mu_2 \\
\tilde{\bar{X}}_2^{r,\pi}(t) &= \frac{1}{\sqrt{|N^r|}}\left(\tilde{\hat{X}}_2^{r,\pi}(t)\right) + x_{22}\beta_2 \\
&= \tilde{\hat{X}}_2^{r}(0) + \tilde{\hat{A}}_2^{r',\pi}(t) - \tilde{\hat{S}}_{22}^{r',\pi}\left(\tilde{\hat{T}}_{22}^{r',\pi}(t)\right) - \frac{\lambda_2^r t}{|N^r|} + x_{22}\beta_2\mu_2
\end{aligned}$$

Observe that this scaling corresponds to the fluid scaling in the original probability space. By (4.102) and (4.92)

$$\tilde{\bar{X}}_i^{r,\pi}(\cdot) \to \tilde{\bar{X}}_i^{\pi}(\cdot)$$

a.s. u.o.c., where

$$\begin{aligned}
\tilde{\bar{X}}_1^{\pi}(t) &= \tilde{\bar{X}}_1(0) + \lambda_1 t - \mu_1\tilde{\bar{T}}_{11}^{\pi}(t) - \mu_2\tilde{\bar{T}}_{21}^{\pi}(t) \\
\tilde{\bar{X}}_2^{\pi}(t) &= \tilde{\bar{X}}_2(0) + \lambda_2 t - \mu_2\tilde{\bar{T}}_{22}^{\pi}(t).
\end{aligned}$$

We note that

$$\tilde{\bar{T}}_{11}^{\pi}(t) \leq \beta_1\ \tilde{\bar{T}}_{21}^{\pi}(t) + \tilde{\bar{T}}_{22}^{\pi}(t) \leq \beta_2 \tag{4.103}$$

by (4.98) and (4.99). Also, by (4.74)

$$\tilde{\bar{X}}_1(0) = \beta_1 + \beta_2 x_{21} \text{ and } \tilde{\bar{X}}_2(0) = \beta_2 x_{22} \tag{4.104}$$

Let

$$\begin{aligned}
\tilde{x}^{r,\pi}(t) &= \tilde{\hat{X}}_1^{r}(0) + \tilde{\hat{A}}_1^{r',\pi}(t) - \tilde{\hat{S}}_{11}^{r',\pi}(\tilde{\hat{T}}_{11}^{r',\pi}(t)) - \tilde{\hat{S}}_{21}^{r',\pi}(\tilde{\hat{T}}_{21}^{r',\pi}(t)) + \tilde{\hat{X}}_2^{r}(0) \qquad (4.105) \\
&+ \tilde{\hat{A}}_2^{r',\pi}(t) - \tilde{\hat{S}}_{22}^{r',\pi}(\tilde{\hat{T}}_{22}^{r',\pi}(t)), \\
\tilde{Y}^{r,\pi}(t) &= \tilde{\hat{X}}_1^{r,\pi}(t) + \tilde{\hat{X}}_2^{r,\pi}(t) = \tilde{x}^{r,\pi}(t) - \mu_1\tilde{\hat{T}}_{11}^{r',\pi}(t) - \mu_2\left(\tilde{\hat{T}}_{21}^{r',\pi}(t) + \tilde{\hat{T}}_{22}^{r',\pi}(t)\right) \qquad (4.106)
\end{aligned}$$

and

$$\tilde{\xi}^{r,\pi}(T) = \int_0^T \left(\tilde{Y}^{r,\pi}(s)\right)^+ ds.$$

Also, we define

$$\begin{aligned}
\tilde{\bar{T}}^*_{11}(t) &= x_{11}\beta_1 t,\ \tilde{\bar{T}}^*_{22}(t) = x_{21}\beta_2 t,\ \tilde{\bar{T}}^*_{22}(t) = x_{22}\beta_2 t, \\
\tilde{x}^{r,*}(t) &= \tilde{\hat{X}}^r_1(0) + \tilde{\hat{A}}^{r',\pi}_1(t) - \tilde{\hat{S}}^{r',\pi}_{11}(\tilde{\bar{T}}^*_{11}(t)) - \tilde{\hat{S}}^{r',*}_{21}(\tilde{\bar{T}}^*_{21}(t)) + \tilde{\hat{X}}^r_2(0) \\
&+ \tilde{\hat{A}}^{r',\pi}_2(t) - \tilde{\hat{S}}^{r',\pi}_{22}(\tilde{\bar{T}}^*_{22}(t)),
\end{aligned}$$

$$\tilde{Y}^{r,*}(t) = \psi(\tilde{x}^{r,*})(t),$$

and

$$\tilde{\xi}^{r,*}(T) = \int_0^T \left(\tilde{Y}^{r,*}(s)\right)^+ ds. \tag{4.107}$$

We have by (4.92) that

$$\begin{aligned}
&\tilde{\hat{A}}^{r',\pi}_1(\cdot) - \tilde{\hat{S}}^{r',\pi}_{11}(\tilde{\bar{T}}^*_{11}(\cdot)) - \tilde{\hat{S}}^*_{21}(\tilde{\bar{T}}^*_{21}(\cdot)) \to \tilde{\hat{A}}^*_1(\cdot) - \tilde{\hat{S}}^*_{11}(\tilde{\bar{T}}^*_{11}(\cdot)) - \tilde{\hat{S}}^*_{21}(\tilde{\bar{T}}^*_{21}(\cdot)), \\
&\tilde{\hat{A}}^{r',\pi}_2(\cdot) - \tilde{\hat{S}}^{r',\pi}_{22}(\tilde{\bar{T}}^{r',*}_{22}(\cdot)) \to \tilde{\hat{A}}^*_2(\cdot) - \tilde{\hat{S}}^*_{22}(\tilde{\bar{T}}^*_{22}(\cdot))
\end{aligned}$$

$\tilde{P}$-a.s. u.o.c.

Let process $\tilde{Y}^*$ be the solution of the following SDE

$$\begin{aligned}
\tilde{Y}^*(t) &= \hat{X}_1(0) + \hat{X}_2(0) + \tilde{\hat{A}}^*_1(t) - \tilde{\hat{S}}^*_{11}(\tilde{\bar{T}}^*_{11}(t)) - \tilde{\hat{S}}^*_{21}(\tilde{\bar{T}}^*_{21}(t)) \\
&+ \tilde{\hat{A}}^*_2(t) - \tilde{\hat{S}}^*_{22}(\tilde{\bar{T}}^*_{22}(t)) + \mu_1 \int_0^t (Y^*(s))^- ds,\ t \geq 0,
\end{aligned}$$

We define

$$\tilde{H}^*(t) = h_2(\tilde{Y}^*(t))^+,\ t \geq 0,$$

and

$$\tilde{\zeta}^*(t) = \int_0^t \tilde{H}^*(s) ds,\ t \geq 0.$$

Clearly, $\tilde{\zeta}^*$ has the same distribution with ζ^* defined by (4.76). Also

$$h_2 \tilde{\xi}^{r,*}(T) \to \tilde{\zeta}^*(T)$$

$\tilde{P}$-a.s. as $r \to \infty$ by Lemma 4.21 and the continuous mapping theorem.

Next, we divide the sample paths into two sets based on their fluid limits. By (4.103) and (4.104)

$$\tilde{\bar{X}}_1^{\pi}(T) \geq \beta_1 + x_{21}\beta_2 \text{ and } \tilde{\bar{X}}_2^{\pi}(T) \geq x_{22}\beta_2.$$

Define

$$\mathcal{V}_T = \left\{ w \in \tilde{\Omega} : \tilde{\bar{X}}_1^{\pi}(T) = \beta_1 + x_{21}\beta_2, \tilde{\bar{X}}_2^{\pi}(T) = x_{22}\beta_2 \right\}.$$

Observe that for $w \in \tilde{\Omega}$ and $0 \leq t \leq T$

$$\tilde{\bar{T}}_{11}^{\pi}(t) = x_{11}\beta_1 t,\ \tilde{\bar{T}}_{21}^{\pi}(t) = x_{21}\beta_2 t,\ \tilde{\bar{T}}_{22}^{\pi}(t) = x_{22}\beta_2 t.$$

Next we apply Theorem 4.22 to $\tilde{Y}^{r,\pi}$, $\tilde{\hat{T}}_{11}^{r',\pi}$, and $\tilde{\hat{T}}_{2}^{r',\pi}(\cdot) = \tilde{\hat{T}}_{21}^{r',\pi}(\cdot) + \tilde{\hat{T}}_{22}^{r',\pi}(\cdot)$. By (4.98) and (4.99), (4.81) holds. By (4.106), (4.80) holds. By (4.100) and (4.101),

$$\tilde{\hat{T}}_{11}^{r,\pi}(t;s) + \tilde{\hat{T}}_{2}^{r}(t;s) \leq \int_s^t \left(\tilde{\hat{X}}_1^{r,\pi}(u) + \tilde{\hat{X}}_2^{r,\pi}(u) \right) \mathrm{d}u.$$

Combining this with (4.98) and (4.99) yields

$$\tilde{\hat{T}}_{11}^{r,\pi}(t;s) + \tilde{\hat{T}}_{2}^{r}(t;s) \leq + \int_s^t \left(\tilde{\hat{X}}_1^{r,\pi}(u) + \tilde{\hat{X}}_2^{r,\pi}(u) \right)^{-} \mathrm{d}u.$$

Therefore, (4.82) holds as well.

Since ψ is continuous by Lemma 4.21, by using Theorem 4.22, we have that for all $\omega \in \mathcal{V}_T$

$$\lim_{r' \to \infty} \tilde{\xi}^{r',\pi}(T) \geq \lim_{r' \to \infty} \tilde{\xi}^{r',*}(T). \tag{4.108}$$

Next we define

$$\mathcal{U}_T = \tilde{\Omega} \setminus \mathcal{V}_T.$$

Fix $\omega \in \mathcal{U}_T$. By continuity of $\tilde{\bar{X}}_1^{\pi}$ and $\tilde{\bar{X}}_2^{\pi}$ we can choose T_0 such that

$$\tilde{\bar{X}}_1^{\pi}(t) + \tilde{\bar{X}}_2^{\pi}(t) - (\beta_1 + x_{21}\beta_2 + x_{22}\beta_2) > 0 \tag{4.109}$$

for all $t \in [T_0, T]$. Then,

$$\tilde{\xi}^{r,\pi}(T) = \sqrt{|N^{r'}|} \int_0^T \left(\tilde{\bar{X}}_1^{r',\pi}(s) + \tilde{\bar{X}}_2^{r',\pi}(s) - (\beta_1 + x_{21}\beta_2 + x_{22}\beta_2) \right)^{+} \mathrm{d}s$$

Combining this with (4.109), we have that

$$\liminf_{r'\to\infty} \tilde{\xi}^{r,\pi}(T) = \infty$$

for all $\omega \in \mathcal{U}_T$. Finally, this with (4.108) gives that

$$\liminf_{r'\to\infty} \tilde{\xi}^{r',\pi}(T) \geq \lim_{r'\to\infty} \tilde{\xi}^{r,*}(T) \tag{4.110}$$

$\tilde{P}$-a.s.

Define

$$\xi^{r,\pi}(T) = \int_0^T (Y^{r,\pi}(s))^+ \, ds$$

and note that $\xi^{r,\pi}$ and $\tilde{\xi}^{r,\pi}$ have the same distribution. Also,

$$\zeta^{r,\pi}(T) = \int_0^T \left(h_1 \hat{Q}_1^{r,\pi}(s) + h_2 \hat{Q}_2^{r,\pi}(s)\right) ds \quad \geq \quad h_2 \xi^{r,\pi}(T).$$

Hence, for every $x > 0$

$$\begin{aligned}
\lim_{r'\to\infty} P\left\{\zeta^{r',\pi}(T) > x\right\} &\geq \lim_{r'\to\infty} \tilde{P}\left\{h_2 \xi^{r,\pi}(T) > x\right\} = \lim_{r'\to\infty} \tilde{P}\left\{h_2 \tilde{\xi}^{r,\pi}(T) > x\right\} \\
&= \lim_{r'\to\infty} \tilde{E}\left[1\{h_2 \tilde{\xi}^{r,\pi}(T) > x\}\right] \\
&\geq \tilde{E}\left[\liminf_{r'\to\infty} 1\{h_2 \tilde{\xi}^{r,\pi}(T) > x\}\right] \\
&\geq \tilde{E}\left[\liminf_{r'\to\infty} 1\{h_2 \tilde{\xi}^{r,*}(T) > x\}\right] \\
&= \tilde{P}\left\{\tilde{\zeta}^{*}(T) > x\right\}
\end{aligned}$$

□

4.2.5 Analysis of the static priority policy

In this section we prove Theorem 4.20. We begin with analyzing the fluid limits of the static priority policy. Then, we establish a SSC result using the framework in Chapter 3. Finally, we prove the asymptotic optimality of π^* using these two results.

First note that, by (4.74)

$$\bar{Q}^{r,\pi^*}(0) \to 0, \ \bar{Z}_{11}^{r,\pi^*}(0) \to \beta_1, \ \bar{Z}_{21}^{r,\pi^*}(0) \to x_{21}\beta_2, \ \bar{Z}_{22}^{r,\pi^*}(0) \to x_{22}\beta_2$$

a.s. as $r \to \infty$. the following result establishes the fluid limits N-systems working under π^*.

Proposition 4.23 (Fluid limits)**.** *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-systems working under the static priority policy π^*. Assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Let $q = (0,0)$ and $z = (z_{11}, z_{21}, z_{21}) = (\beta_1, x_{21}\beta_2, x_{22}\beta_2)$. Then, (q,z) is an invariant state of the fluid limits of $\{\mathbb{X}^{r,\pi^*}\}$.*

The proof is placed in Section 4.2.5.1. Next we prove three SSC results that will be used for the optimality proof.

Proposition 4.24 (State Space Collapse)**.** *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-systems working under the static priority policy π^*. Assume that (4.70)-(4.73) hold. Then for each $T > 0$, there exists $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, such that*

i.

$$\left\|\hat{Q}^{r,\pi^*}(\cdot)\right\|_{L^r/\sqrt{|N^r|}} \vee \left\|\hat{Z}^{r,\pi^*}(\cdot)\right\|_{L^r/\sqrt{|N^r|}} \tag{4.111}$$

satisfies the compact containment condition,

ii.

$$\left|\hat{X}^{r,\pi^*}(L^r/\sqrt{|N^r|})\right| \Rightarrow \left|\hat{X}^{r,\pi^*}(0)\right|, \tag{4.112}$$

as $r \to \infty$,

iii.

$$\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left\{ \left|\hat{Q}_1^{r,\pi^*}(t)\right| \vee \left|\hat{Z}_{11}^{r,\pi^*}(t) + (\hat{X}_1^{r,\pi^*}(t) + \hat{X}_2^{r,\pi^*}(t))^-\right| \right.$$
$$\left. \vee \left|\hat{Z}_{21}^{r,\pi^*}(t) + \hat{Z}_{22}^{r,\pi^*}(t)\right| \right\} \to 0, \tag{4.113}$$

in probability as $r \to \infty$.

Now we are ready to prove Theorem 4.20.

Proof of Theorem 4.20. Assume that the assumptions of Proposition 4.28 hold. We omit pi^* from the notation below. Let $x^r(t)$ be defined by

$$\begin{aligned} x^r(t) &= \hat{X}_1^r(0) + \hat{X}_2^r(0) + \hat{A}_1^r(t) + \hat{A}_2^r(t) - \hat{S}_{11}^r\left(\bar{T}_{11}^r(t)\right) - \hat{S}_{21}^r\left(\bar{T}_{21}^r(t)\right) \\ &- \hat{S}_{22}^r\left(\bar{T}_{22}^r(t)\right) - \theta t. \end{aligned} \tag{4.114}$$

and

$$\hat{Y}^r(t) = \hat{X}_1^r(t) + \hat{X}_2^r(t).$$

For a process y^r associated with the rth system, we set

$$\underline{y}^r(t) = y^r\left(t + \frac{L^r}{\sqrt{|N^r|}}\right)$$

Note that

$$\zeta^{r,\pi^*}(T) = \zeta^{r,\pi^*}\left(\frac{L^r}{\sqrt{|N^r|}}\right) + \underline{\zeta}^{r,\pi^*}\left(T - \frac{L^r}{\sqrt{|N^r|}}\right). \tag{4.115}$$

We first handle the first term on the right hand side above.

$$\begin{aligned}\zeta^{r,\pi^*}\left(\frac{L^r}{\sqrt{|N^r|}}\right) &= \int_o^{\frac{L^r}{\sqrt{|N^r|}}} (h_1\hat{Q}_1^{r,\pi^*}(t) + h_2\hat{Q}_2^{r,\pi^*}(t))dt \\ &\leq \left(h_1\|Q_1(t)\|_{\frac{L^r}{\sqrt{|N^r|}}} + h_2\|Q_2(t)\|_{\frac{L^r}{\sqrt{|N^r|}}}\right)\frac{L^r}{\sqrt{|N^r|}}.\end{aligned}$$

This shows by (4.111) that

$$\zeta^{r,\pi^*}\left(\frac{L^r}{\sqrt{|N^r|}}\right) \to 0 \tag{4.116}$$

as $r \to \infty$.

Next, we handle the second term on the R.H.S. of (4.115). By (4.112)

$$\underline{\hat{X}}^r(0) \Rightarrow \hat{X}^r(0).$$

as $r \to \infty$. Hence,

$$\underline{x}^r \Rightarrow W$$

as $r \to \infty$, where W is a Brownian motion with drift θ and variance $r_1 + r_2$, see (4.75).

By (4.113)

$$\underline{\hat{Z}_{21}^r}(\cdot) + \underline{\hat{Z}_{22}^r}(\cdot) \Rightarrow 0 \text{ and} \tag{4.117}$$

$$(\underline{\hat{Z}_{11}^r}(\cdot) + (\underline{\hat{X}_1^r}(\cdot) + \underline{\hat{X}_2^r}(\cdot))^-) \Rightarrow 0 \tag{4.118}$$

as $r \to \infty$. Therefore, by continuity of ψ we have that

$$\underline{\hat{Y}}^{r,\pi^*}(\cdot) \Rightarrow Y^*(\cdot)$$

as $r \to \infty$. Also, since $\underline{\hat{Q}_1^{r,\pi^*}}(\cdot) \Rightarrow 0$ as $r \to \infty$ and by (4.117) and (4.118)

$$h_1 \underline{\hat{Q}_1^{r,\pi^*}}(\cdot) + h_2 \underline{\hat{Q}_2^{r,\pi^*}}(\cdot) \Rightarrow h_2 Y^*(\cdot)$$

as $r \to \infty$. By the continuity of the integration operator, see Theorem 11.5.1 Whitt [65], we have

$$\underline{\zeta^{r,\pi^*}}(T) \Rightarrow \zeta^*(T) \tag{4.119}$$

as $r \to \infty$. By Theorem 4.4 in Billingsley [11], (4.116), and (4.119)

$$\zeta^{r,\pi^*}\left(T + \frac{L^r}{\sqrt{|N^r|}}\right) \Rightarrow \zeta^*(T)$$

as $r \to \infty$. This yields the desired result by Theorem 4.19, since

$$\zeta^{r,\pi^*}(T) \leq \zeta^{r,\pi^*}\left(T + \frac{L^r}{\sqrt{|N^r|}}\right)$$

a.s. □

4.2.5.1 Fluid Limits of N-systems under π^*

Before we establish the fluid limits of N-systems working under π^* we first provide a few properties of the policy π^*. We first note that policy π^* is non-idling so $\mathbb{X}^r$ also satisfies equation (2.10). In addition, from the description of π^*, under π^* $\mathbb{X}^r$ satisfies the following system equations.

Recall that under our static priority policy class 1 customers have priority over class 2 customers in the second server pool. Hence, a class 2 customer in the queue can start receiving service if there are no class 1 customers waiting in the queue for service. Recall that $B_{jk}^r(t)$ denote the number of class k customers who started their service in server pool j before time t after waiting in the queue. Therefore

$$B_{22}^r(t) \text{ can only increase when } Q_1^r(t) = 0. \tag{4.120}$$

Since the second server pool has priority over the first server pool, an arriving customer will start his service in the first server pool only when all the servers in the second pool

are busy. Recall that $A^r_{jk}(t)$ denote the number of class k customers whose service started immediately at the time of their arrival in server pool j by time t. Hence

$$A^r_{11}(t) \text{ and } B^r_{11}(t) \text{ can only increase when } Z^r_{21}(t) + Z^r_{22}(t) = N^r_2, \tag{4.121}$$

If there are class 2 customers waiting in the queue to receive service then all the servers in the second pool should be busy at that instant since the static priority policy is non-idling. We showed in Lemma 3 (see Appendix B) that only one event can happen at any instant w.p. 1. Therefore, when a class 1 customer arrives to the system at an instant when there are class 2 customers waiting in the queue, his service cannot start in the second pool immediately. Therefore, for $s < t$

$$\text{If } Q^r_2(\tau) > 0 \text{ for all } \tau \in [s,t], \text{ then } A^r_{21}(t) - A^r_{21}(s) = 0. \tag{4.122}$$

w.p. 1.

We first characterize the fluid model equations of π^* and then establish the invariant states of the fluid limits.

Proposition 4.25. *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Then, every fluid limit $\bar{\mathbb{X}}^{\pi^*}$ of $\{\mathbb{X}^{r,\pi^*}\}$ satisfies the following equations in addition to the fluid model equations (A.2)-(A.11). For every regular point $t > 0$ of $\bar{\mathbb{X}}^{\pi^*}$*

$$\dot{\bar{B}}_{21}(t) = \beta_2\mu_2 \text{ if } \bar{Q}_1(t) > 0 \tag{4.123}$$

$$\dot{\bar{B}}_{11}(t) = \beta_1\mu_1 \text{ if } \bar{Q}_1(t) > 0 \tag{4.124}$$

$$\dot{\bar{A}}_{21}(t) + \dot{\bar{A}}_{22}(t) = \lambda_1 + \lambda_2 \text{ if } \bar{Z}_{21}(t) + \bar{Z}_{22}(t) < \beta_2 \tag{4.125}$$

$$\dot{\bar{B}}_{22}(t) = \beta_2\mu_2 \text{ if } \bar{Z}_{11}(t) < \beta_1 \text{ and } \bar{Q}_2(t) > 0. \tag{4.126}$$

Remark 4.26. Intuitive explanations of these fluid model equations can be given as follows. Equation (4.123) ((4.124)) implies that all the servers in the second pool (resp., first pool) serve a class 1 customer as soon as they finish serving the current customer. It can be shown that (4.123) follows from (4.120) and (4.124) follows from the non-idling condition.

Equation (4.125) implies that when there are idle servers in the second pool all the arriving customers will start their service in the second pool. It can be shown that (4.125) follows from (4.121) and the non-idling condition.

Equation (4.126) implies that when there are class 2 customers in the queue and there are idle servers in the first pool, servers in the second pool serve class 2 customers as soon as they finish service. We show below that (4.126) follows from (4.122).

Proof of Proposition 4.25. We restrict our attention to the set of sample paths that do not have a service completion from server pool 2 and an arrival to class 1 together at any time instant. Again by Lemma 3 in Appendix B the set that satisfy this condition has probability one.

We only give the proof of (4.126), other fluid model equations are proved similarly.

Let $\bar{\mathbb{X}}^{\pi^*}$ be a fluid limit of $\{\mathbb{X}^{r,\pi^*}\}$ and assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Also assume that

$$\bar{Z}_{11}(t) < \beta_1 - \epsilon \text{ and } \bar{Q}_2(t) > \epsilon$$

for a regular point $t > 0$ of $\bar{\mathbb{X}}^{\pi^*}$ and for some $\epsilon > 0$.

Since $\bar{\mathbb{X}}^{\pi^*}$ is a fluid limit of $\{\mathbb{X}^{r,\pi^*}\}$, there exists a subsequence, which we denote again by r for notational convenience, such that $\omega \in \Omega$ such that

$$\bar{\mathbb{X}}^{r,\pi^*}(\cdot,\omega) \to \bar{\mathbb{X}}^{\pi^*}(\cdot) \text{ u.o.c.} \tag{4.127}$$

as $r \to \infty$ and $\bar{\mathbb{X}}^{\pi^*}$ satisfies fluid model equations (A.2)-(A.11). Then, there exists $\delta > 0$ and r_0 such that

$$Z^r_{11}(s) < N^r_1(\beta_1 - \epsilon/2) \text{ and } \bar{Q}^r_2(s) > \epsilon/2 \tag{4.128}$$

for all $s \in [t-\delta, t+\delta]$ and for $r > r_0$. Therefore, by (2.10) in Section 2.1

$$Q_1^{r,\pi^*}(s) = 0 \tag{4.129}$$

for all $s \in [t-\delta, t+\delta]$ and for $r > r_0$. Hence, by (4.122),

$$A^r_{21}(t+\delta) - A^r_{21}(t-\delta) = 0$$

for all $r > r_0$, since $\bar{Q}_2^r(s) > \epsilon/2$. Therefore, by (4.127)

$$\dot{\bar{A}}_{21}(t) = 0. \tag{4.130}$$

Also

$$B_{21}^r(t+\delta) - B_{21}^r(t-\delta) = 0$$

for $r > r_0$, by (4.129) and non-negativity of $Q_1^{r,\pi^*}(\cdot)$. Hence,

$$\dot{\bar{B}}_{21}(t) = 0. \tag{4.131}$$

Since π^* is non-idling and $\bar{Q}_2(t) > \epsilon$, by (A.8)-(A.11)

$$\dot{\bar{Z}}_{21}(t) + \dot{\bar{Z}}_{22}(t) = 0.$$

By (A.4)

$$\dot{\bar{Z}}_{21}(t) + \dot{\bar{Z}}_{22}(t) = \dot{\bar{B}}_{21}(t) + \dot{\bar{A}}_{21}(t) + \dot{\bar{B}}_{22}(t) + \dot{\bar{A}}_{22}(t) - \dot{\bar{D}}_{21}(t) - \dot{\bar{D}}_{22}(t).$$

Hence, (4.131) and (4.130) imply that

$$\dot{\bar{B}}_{22}(t) = \dot{\bar{D}}_{21}(t) + \dot{\bar{D}}_{22}(t).$$

And so

$$\dot{\bar{B}}_{22}(t) = \beta_2\mu_2,$$

by (A.8) and the fact that $\bar{Q}_2(t) > 0$. □

Proof of Proposition 4.23. By Definition A.2, we need to show that if $(\bar{Q}(0), \bar{Z}(0)) = (q, z)$, then $(\bar{Q}(t), \bar{Z}(t)) = (q, z)$ for all $t \geq 0$.

Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Also assume that $(\bar{Q}(0), \bar{Z}(0)) = (q, z)$. We proceed in several steps:

(1) We first show that $\bar{Q}_1(t) = 0$ for all $t > 0$. Note that whenever $\bar{Q}_1(t) > 0$

$$\dot{\bar{Q}}_1(t) = \lambda_1 - \mu_1\beta_1 - \mu_2\beta_2 = -\mu_2 x_{22}\beta_2$$

by (4.123). Hence

$$\bar{Q}_1(t) = 0 \text{ for all } t \geq 0. \tag{4.132}$$

(2) Next we show that $\bar{Z}_{21}(t) + \bar{Z}_{22}(t) = \beta_2$ for all $t \geq 0$. Similar to part (1), if $\bar{Z}_{21}(t) + \bar{Z}_{22}(t) < \beta_2$, then

$$\dot{\bar{Z}}_{21}(t) + \dot{\bar{Z}}_{22}(t) = \lambda_1 + \lambda_2 - \mu_2\beta_2 = \mu_1\beta_1$$

by (4.125). Since $\bar{Z}_{21}(t) + \bar{Z}_{22}(t) \leq \beta_2$ this proves that

$$\bar{Z}_{21}(t) + \bar{Z}_{22}(t) = \beta_2 \text{ for all } t \geq 0. \tag{4.133}$$

(3) Now we show that $(\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t) = 0$ for all $t > 0$. If $(\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t) < 0$ then $\bar{Z}_{11}(t) < \beta_1$ and $\bar{Q}_2(t) > 0$. First note that $\bar{Q}_1(t) = 0$ by (A.6). Also, by (4.126)

$$\dot{\bar{Q}}_2(t) = \lambda_2 - \mu_2\beta_2 = -\mu_2 x_{21}\beta_2$$

and $\dot{\bar{A}}_{21}(t) = 0$. This gives that $\dot{\bar{A}}_{11}(t) = \lambda_1$ by (A.2). Hence, by (A.4)

$$\dot{\bar{Z}}_{11}(t) = \lambda_1 - \dot{D}_{11}(t) > \lambda_1 - \mu_1\beta_1 = \mu_2 x_{21}\beta_2.$$

These imply that

$$\begin{aligned} \frac{d}{dt}\left((\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t)\right) &= \dot{\bar{Z}}_{11}(t)\bar{Q}_2(t) + (\bar{Z}_{11}(t) - \beta_1)\dot{\bar{Q}}_2(t) \\ &> \mu_2 x_{21}\beta_2(\bar{Q}_2(t) + (\beta_1 - \bar{Z}_{11}(t)). \end{aligned}$$

If $(\beta_1 - \bar{Z}_{11}(t))\bar{Q}_2(t) > 1$ then $(\bar{Q}_2(t) + (\beta_1 - \bar{Z}_{11}(t)) \geq 1$. Otherwise, if $(\beta_1 - \bar{Z}_{11}(t))\bar{Q}_2(t) \leq 1$, then $(\bar{Q}_2(t) + (\beta_1 - \bar{Z}_{11}(t)) \geq (\beta_1 - \bar{Z}_{11}(t))\bar{Q}_2(t)$. Thus, if $(\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t) < 0$, then

$$\frac{d}{dt}\left((\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t)\right) > 0.$$

Therefore

$$(\bar{Z}_{11}(t) - \beta_1)\bar{Q}_2(t) = 0 \text{ for all } t \geq 0. \tag{4.134}$$

(4) Now assume that $\bar{Z}_{22}(t) < x_{22}\beta_2 - \epsilon$ for some $\epsilon > 0$. Since

$$\dot{Q}_2(t) = \lambda_2 - \mu_2\bar{Z}_{22}(t)$$

this implies by the continuity of $\bar{\mathbb{X}}$, (A.3) and (A.4) that

$$\bar{Q}_2(t) > 0.$$

By (4.133) and (4.134) this implies

$$\begin{aligned} \bar{Z}_{21}(t) &= x_{21}\beta_2 + \epsilon, \\ \bar{Z}_{11}(t) &= \beta_1. \end{aligned}$$

Also, by continuity of $\bar{Q}_2(t)$ and (4.135), there exists $\delta > 0$ such that $\bar{Q}_2(s) > 0$ for all $s \in [t-\delta, t+\delta]$. Therefore, again by (4.134),

$$\bar{Z}_{11}(s) = \beta_1,$$

for all $s \in [t-\delta, t+\delta]$. Hence,

$$\dot{\bar{Z}}_{11}(t) = 0.$$

However, by (4.132), (A.3) and (A.4)

$$\dot{\bar{Z}}_{21}(t) + \dot{\bar{Z}}_{11}(t) = \lambda_1 - \mu_1 \bar{Z}_{11}(t) - \mu_2 \bar{Z}_{21}(t) < -\epsilon\mu_2.$$

And so, by using (4.133) we have that

$$\dot{\bar{Z}}_{22}(t) > \epsilon\mu_2.$$

This implies that whenever $\bar{Z}_{22}(t) < x_{22}\beta_2$, $\dot{\bar{Z}}_{22}(t) > 0$. Hence $\bar{Z}_{22}(t) \geq x_{22}\beta_2$.

It can be shown similarly that $\bar{Z}_{21}(t) \geq x_{21}\beta_2$. Thus, $\bar{Z}_{21}(t) = x_{21}\beta_2$ and $\bar{Z}_{22}(t) = x_{22}\beta_2$ for all $t \geq 0$.

(5) Since $\bar{Z}_{22}(t) = x_{22}\beta_2$ for all $t \geq 0$, $\bar{Q}_2(t) = 0$ for all $t \geq 0$ by (A.3) and (A.4).

(6) Finally, since $\bar{Z}_{21}(t) \geq x_{21}\beta_2$ and $\bar{Q}_1(t) = 0$ for all $t \geq 0$, $\bar{Z}_{11}(t) \geq \beta_1$.

□

4.2.5.2 Hydrodynamic Limits

Let $\mathbb{X}^{r,m}$ be the hydrodynamically scaled version of $\mathbb{X}^{r,\pi^*}$, see Chapter 3.4.1, $\tilde{\mathbb{X}}^{\pi^*}$ denote a hydrodynamic limit of $\mathbb{X}^{r,\pi^*}$. In this section we establish the additional hydrodynamic equations that must be satisfied by the hydrodynamic limits of $\mathbb{X}^{r,\pi^*}$.

From (4.120)-(4.122), it is readily obtained that the components of $\mathbb{X}^{r,m}$ satisfy the following equations in addition to equations (3.24)-(3.26),

$$B_{22}^{r,m}(t) \text{ can only increase when } Q_1^{r,m}(t) = 0, \tag{4.135}$$

$$A_{11}^{r,m}(t) \text{ and } B_{11}^{r,m}(t) \text{ can only increase when } Z_{21}^{r,m}(t) + Z_{22}^{r,m}(t) = 0, \tag{4.136}$$

$$\text{If } Q_2^{r,m}(\tau) > 0 \text{ for all } \tau \in [s,t], \text{ then } A_{22}^{r,m}(t) - A_{22}^{r,m}(s) = 0 \tag{4.137}$$

Proposition 4.27. *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that (4.74) holds. Then, every hydrodynamic limit $\tilde{\mathbb{X}}^{\pi^*}$ of $\{\mathbb{X}^{r,\pi^*}\}$ satisfies the following equations in addition to the hydrodynamic model equations (3.1)-(3.8). For every regular point $t > 0$ of $\tilde{\mathbb{X}}^{\pi^*}$*

$$\dot{\tilde{B}}_{21}(t) = \mu_2\beta_2 \text{ if } \bar{Q}_1(t) > 0, \tag{4.138}$$

$$\dot{\tilde{B}}_{11}(t) = \mu_1\beta_1 \text{ if } \bar{Q}_1(t) > 0, \tag{4.139}$$

$$\dot{\tilde{A}}_{21}(t) + \dot{\tilde{A}}_{22}(t) = \lambda_1 + \lambda_2 \text{ if } \tilde{Z}_{21}(t) + \tilde{Z}_{22}(t) < 0, \tag{4.140}$$

$$\dot{\tilde{B}}_{22}(t) = \mu_2\beta_2 \text{ if } \tilde{Z}_{11}(t) < 0 \text{ and } \tilde{Q}_2(t) > 0. \tag{4.141}$$

Proof. We only give a proof of (4.141). The other hydrodynamic equations are proved similarly.

Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that (4.74) holds. Let $\tilde{\mathbb{X}}^{\pi^*}$ be a hydrodynamic limit of $\{\mathbb{X}^{r,\pi^*}\}$.

Fix $t > 0$. Assume that $\tilde{Z}_{11}(t) < 0$ $\tilde{Q}_2(t) > 0$. Then, by the continuity of $\tilde{\mathbb{X}}^{\pi^*}$, there exists an $\varepsilon > 0$ and a $\tau > 0$ such that $\tilde{Q}_2(s) > \varepsilon$ and $\tilde{Z}_{11}(t) < -\varepsilon$ for all $s \in [t-\tau, t+\tau]$. Fix $0 < \delta < \epsilon/2$ and choose r large enough, together with an integer m and $\omega \in \mathcal{K}^r$, so that (3.97) holds for δ.

It follows from (3.97) that

$$Z_{11}^{r,m}(t) < -\varepsilon/4 \text{ and } Q_2^{r,m}(s) > \varepsilon/4 \tag{4.142}$$

for all $s \in [t-\tau, t+\tau]$. By (3.7), this implies

$$Q_1^{r,m}(s) = 0 \tag{4.143}$$

for all $s \in [t-\tau, t+\tau]$. Hence, by (4.137),

$$A_{21}^{r,m}(t+\tau) - A_{21}^{r,m}(t-\tau) = 0.$$

This implies that

$$\dot{\tilde{A}}_{21}(t) = 0.$$

Also, by (3.28), (4.142) and (4.143) imply that

$$\dot{\tilde{B}}_{21}(t) = 0.$$

Combining this with (3.1)-(3.8) we get the the desired result. □

4.2.5.3 State Space Collapse Result

Using the results established in the previous section about the hydrodynamic limits and Theorem 3.1, we first prove a multiplicative state space collapse result. Before we prove Proposition 4.24, we establish an intermediate result as a corollary to the multiplicative state space collapse results. Proposition 4.24 readily follows from this corollary. We define

$$\tilde{X}_1(t) = \tilde{Q}_1(t) + \tilde{Z}_{11}(t) + \tilde{Z}_{21}(t) \text{ and} \tag{4.144}$$

$$\tilde{X}_2(t) = \tilde{Q}_2(t) + \tilde{Z}_{22}(t) \tag{4.145}$$

Proposition 4.28 (Multiplicative SSC result)**.** *Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that (4.74) holds. Then for each $T > 0$, there exists $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, such that for*

$$B_T(L^r) = \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(\left| \hat{Q}^r(t) \right| \vee \left| \hat{Z}^r(t) \right| \vee 1 \right)$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \hat{Q}_1^r(t) \right|}{B_T(L^r)} \to 0$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(\left| \hat{Z}_{11}^r(t) + \left(\hat{X}_1^r(t) + \hat{X}_2^r(t) \right)^- \right| \right)}{B_T(L^r)} \to 0$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(|\hat{Z}_{21}^r(t) + \hat{Z}_{22}^r(t)| \right)}{B_T(L^r)} \to 0$$

in probability as $r \to \infty$.

Proof. As mentioned before, we use Theorem 3.4 to prove this result. We focus on the term in the middle and comment at the end of the proof how the other two terms can be handled.

Let $\{\mathbb{X}^{r,\pi^*}\}$ be a sequence of N-models working under the static priority policy π^*. Assume that (4.74) holds. Let $g : \mathbb{R}^5 \to \mathbb{R}^+$ be defined by

$$g(q,z) = |z_1 + (q_1 + q_2 + z_1 + z_2 + z_3)^-|$$

where $q = (q_1, q_2) \in \mathbb{R}^2$ and $z = (z_1, z_2, z_3)$. Recall that $\hat{Q}^r(t) = (\hat{Q}^r_1(t), \hat{Q}^r_2(t))$ and $\hat{Z}^r(t) = (\hat{Z}^r_{11}(t), \hat{Z}^r_{21}(t), \hat{Z}^r_{22}(t))$. Therefore,

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) = \left| \hat{Z}^r_{11}(t) + \left(\hat{X}^r_1(t) + \hat{X}^r_2(t) \right)^- \right|$$

We next show that assumptions of Theorem 3.4 hold. Assumption 1 holds by (4.70)-(4.73). Assumption 2 holds by Proposition 4.23 above and Lemma A.3. Also for $\alpha \in (0,1)$

$$g(\alpha q, \alpha z) = |\alpha z_1 + \alpha(q_1 + q_2 + z_1 + z_2 + z_3)^-| = \alpha g(q,z).$$

Since g is clearly continuous, g satisfies Assumption 3. It remains to be shown that hydrodynamic limits of $\{\mathbb{X}^{r,\pi^*}\}$ and g satisfies Assumption 4.

Let $\{\tilde{\mathbb{X}}^{\pi^*}\}$ be a hydrodynamic limit of $\{\mathbb{X}^{r,\pi^*}\}$. We note that by (3.3) and (3.6)

$$\tilde{Q}(t) \geq 0, \ \tilde{Z}_{11}(t) \leq 0 \text{ and } \tilde{Z}_{21}(t) + \tilde{Z}_{22}(t) \leq 0. \tag{4.146}$$

By (4.144)

$$\tilde{X}_1(t) + \tilde{X}_2(t) = \tilde{Q}_1(t) + \tilde{Q}_2(t) + \tilde{Z}_{11}(t) + \tilde{Z}_{21}(t) + \tilde{Z}_{21}(t). \tag{4.147}$$

By (3.1)-(3.6), for every regular point $t > 0$ of $\{\tilde{\mathbb{X}}^{\pi^*}\}$,

$$\dot{\tilde{X}}_1(t) + \dot{\tilde{X}}_2(t) = 0. \tag{4.148}$$

Assume that for a regular point $t > 0$ of $\{\tilde{\mathbb{X}}^{\pi^*}\}$

$$g(\tilde{Q}(t), \tilde{Z}(t)) > 0. \tag{4.149}$$

We next show that if (4.149) holds

$$\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) < -\epsilon. \tag{4.150}$$

We fix a regular point $t > 0$ and handle two possible cases separately.

(1) First assume that

$$\tilde{Z}_{11}(t) + \left(\tilde{X}_1(t) + \tilde{X}_2(t)\right)^- > 0. \tag{4.151}$$

By (4.147), (4.151) and (4.146) imply that

$$\tilde{Z}_{21}(t) + \tilde{Z}_{22}(t) < 0. \tag{4.152}$$

Therefore, by (3.7)

$$\tilde{Q}_1(t) = 0 \text{ and } \tilde{Q}_2(t) = 0. \tag{4.153}$$

Combining (4.140) with (4.152), (4.153) and (3.6) yields that

$$\dot{\tilde{Z}}_{11}(t) = -\mu_1\beta_1.$$

This with (4.148) gives that

$$\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) = \dot{\tilde{Z}}_{11}(t) + \frac{d}{dt}\left(\tilde{X}_1(t) + \tilde{X}_2(t)\right)^- = -\mu_1\beta_1 \tag{4.154}$$

(2) Now assume that

$$\tilde{Z}_{11}(t) + \left(\tilde{X}_1(t) + \tilde{X}_2(t)\right)^- < 0. \tag{4.155}$$

By (4.147) and (3.7), (4.155) and (4.146) imply that

$$\tilde{Z}_{11}(t) < 0 \text{ and } \tilde{Q}_2(t) > 0.$$

Therefore, by (3.7),

$$\tilde{Q}_1(t) = 0 \text{ and } \tilde{Z}_{21}(t) + \tilde{Z}_{21}(t) = 0.$$

Using (4.141) we have that

$$\dot{\tilde{Z}}_{11}(t) = \lambda_1 - \mu_1 x_1 = \mu_2 x_2 \beta_2.$$

This with (4.148) and (4.155) gives that

$$\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) = -\dot{\tilde{Z}}_{11}(t) + \frac{d}{dt}\left(\tilde{X}_1(t) + \tilde{X}_2(t)\right)^- = -\mu_2\beta_2.$$

This with (4.154) gives (4.150). Hence, g and the hydrodynamic limits of $\{\mathbb{X}^{r,\pi^*}\}$ satisfy Assumption 4. This proves the result for the middle term.

To handle the first term we define $g : \mathbb{R}^5 \to \mathbb{R}^+$ by

$$g(q, z) = |q_1|,$$

where $q = (q_1, q_2) \in \mathbb{R}^2$ and $z = (z_1, z_2, z_3)$ and use (4.138) and (4.139) to verify that Assumption 4 holds for this g and the hydrodynamic limits of $\{\mathbb{X}^{r,\pi^*}\}$.

To handle the last term we define $g : \mathbb{R}^5 \to \mathbb{R}^+$ by

$$g(q, z) = |z_2 + z_3|,$$

where $q = (q_1, q_2) \in \mathbb{R}^2$ and $z = (z_1, z_2, z_3)$ and use (4.140) to verify that Assumption 4 holds for this g and the hydrodynamic limits of $\{\mathbb{X}^{r,\pi^*}\}$.

□

The following result is obtained by algebraic manipulations and the proof is placed at the end of this section

Corollary 4.29. *Under the assumptions of Proposition 4.28, for $T > 0$ and L^r is given as in Proposition 4.28 and*

$$\tilde{B}_T(L^r) = \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(\left| \hat{X}^r(t) \right| \vee 1 \right),$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \hat{Q}^r_1(t) \right|}{\tilde{B}_T(L^r)} \to 0,$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(\left| \hat{Z}^r_{11}(t) + \left(\hat{X}^r_1(t) + \hat{X}^r_2(t) \right)^- \right| \right)}{\tilde{B}_T(L^r)} \to 0, \text{ and}$$

$$\frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(|\hat{Z}^r_{21}(t) + \hat{Z}^r_{22}(t)| \right)}{\tilde{B}_T(L^r)} \to 0$$

in probability as $r \to \infty$.

Proof of Corollary 4.29. Fix $T > 0$ assume that assumptions of Proposition 4.28 hold. Then, there exists a sequence $\epsilon_r \to 0$ as $r \to \infty$ such that

$$P\Bigg\{ \frac{\sup_{L^r/\sqrt{|N^r|}\le t\le T} \left|\hat{Q}_1^r(t)\right|}{B_T(L^r)} \vee \frac{\sup_{L^r/\sqrt{|N^r|}\le t\le T} \left(\left| \hat{Z}_{11}^r(t) + \left(\hat{X}_1^r(t) + \hat{X}_2^r(t)\right)^- \right| \right)}{B_T(L^r)} \\ \vee \frac{\sup_{L^r/\sqrt{|N^r|}\le t\le T} \left(|\hat{Z}_{21}^r(t) + \hat{Z}_{22}^r(t)| \right)}{B_T(L^r)} > \epsilon_r \Bigg\} < \epsilon_r. \tag{4.156}$$

Let $\mathcal{D}^r$ denote the complement of the event in the above expression. We assume that

$$\sup_{L^r/\sqrt{|N^r|}\le t\le T} \left(\left|\hat{Q}^r(t)\right| \vee \left|\hat{Z}^r(t)\right| \right) > 1,$$

as otherwise the result follows immediately. For notational simplicity for all the processes above, with a slight abuse of notation, we set $\underline{x}^r(t) = x^r\left(t + \frac{L^r}{\sqrt{|N^r|}}\right)$ and $T^r = T - \frac{L^r}{\sqrt{|N^r|}}$.

Choose r large enough so that $\epsilon_r < 1/100$. Then on $\mathcal{D}^r$

$$\epsilon_r > \frac{\left\|\underline{\hat{Q}_1^r}(t)\right\|_{T^r}}{\left\|\underline{\hat{Q}^r}(t)\right\|_{T^r} \vee \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r}} > \frac{\left\|\underline{\hat{Q}_1^r}(t)\right\|_{T^r}}{\left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} \vee \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r}}.$$

Therefore, on $\mathcal{D}^r$

$$\left\|\underline{\hat{Q}_1^r}(t)\right\|_{T^r} < \epsilon_r \left(\left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r} \right). \tag{4.157}$$

Next we establish a bound for $\underline{\hat{Z}_{11}^r}$. By (4.156) and (4.157) on $\mathcal{D}^r$

$$\left\| \underline{\hat{Z}_{11}^r}(t) + \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t)\right)^- \right\|_{T^r} < \epsilon_r \left(\left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} \vee \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r} \right)$$

If $\left\|\underline{\hat{Z}_{11}^r}(t)\right\|_{T^r} \ge \left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} \vee \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r}$, then

$$\left\|\underline{\hat{Z}_{11}^r}(t)\right\|_{T^r} < 2 \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} \right)$$

for r large enough. Hence, in this case

$$\left\| \underline{\hat{Z}_{11}^r}(t) + \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t)\right)^- \right\|_{T^r} < 2\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\hat{X}_2^r(t)\right\|_{T^r} \right).$$

Therefore,

$$\left\| \underline{\hat{Z}_{11}^r}(t) + \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t)\right)^- \right\|_{T^r} < \\ 2\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r} \right). \tag{4.158}$$

Combining this with (4.157) we get

$$\left\|\underline{\hat{Q}_1^r}(t)\right\|_{T^r} < 3\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right) \quad (4.159)$$

Next we establish a bound for $\|\hat{Q}_2^r(t)\|_{T^r}$. By (2.23),

$$\underline{\hat{Q}_2^r}(t) = \underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t) - \underline{\hat{Q}_1^r}(t) - \underline{\hat{Z}_{11}^r}(t) - \underline{\hat{Z}_{21}^r}(t) - \underline{\hat{Z}_{22}^r}(t).$$

Since π^* is non-idling, $\underline{\hat{Z}_{21}^r}(t) + \underline{\hat{Z}_{22}^r}(t) = 0$ if $\hat{Q}_2^r(t) > 0$. Therefore,

$$0 \le \underline{\hat{Q}_2^r}(t) \le \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t) - \underline{\hat{Q}_1^r}(t) - \underline{\hat{Z}_{11}^r}(t)\right) \vee 0. \quad (4.160)$$

We have by (4.158) and (4.159) that

$$\begin{aligned}\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t) - \underline{\hat{Q}_1^r}(t) - \underline{\hat{Z}_{11}^r}(t) < \underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t) + \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t)\right)^- \\ 5\epsilon_r \left(\left\|\sqrt{\underline{\hat{X}_1^r}}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Q}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right).\end{aligned}$$

This gives by (4.160) for r large enough that

$$\begin{aligned}\|\underline{\hat{Q}_2^r}(t)\|_{T^r} \le 2\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + 2\left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} \\ + 10\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right).\end{aligned}$$

Combining this with (4.158) and (4.159) yields

$$\begin{aligned}&\left\|\underline{\hat{Z}_{11}^r}(t) + \left(\underline{\hat{X}_1^r}(t) + \underline{\hat{X}_2^r}(t)\right)^-\right\|_{T^r} \\ &\quad < 4\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right) \text{ and } \quad (4.161) \\ &\left\|\underline{\hat{Q}_1^r}(t)\right\|_{T^r} < 4\epsilon_r \left(\left\|\underline{\hat{X}_1^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right) \quad (4.162)\end{aligned}$$

for r large enough.

Next we establish a bound for $\|\underline{\hat{Z}_{21}^r}(t)\|_{T^r}$. Observe that

$$\underline{\hat{Z}_{21}^r}(t) = \underline{\hat{X}_1^r}(t) - \underline{\hat{Z}_{11}^r}(t) - \underline{\hat{Q}_1^r}(t).$$

This gives by (4.161) and (4.162) that

$$\begin{aligned}\|\underline{\hat{Z}_{21}^r}(t)\|_{T^r} \le 2\|\underline{\hat{X}_1^r}(t)\|_{T^r} + 2\|\underline{\hat{X}_1^r}(t)\|_{T^r} \\ + 4\epsilon_r \left(\left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{X}_2^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{21}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}_{22}^r}(t)\right\|_{T^r}\right).\end{aligned}$$

Hence, for r large enough

$$\begin{aligned}\|\underline{\hat{Z}^r_{21}}(t)\|_{T^r} &\leq 4\|\underline{\hat{X}^r_1}(t)\|_{T^r} + 4\|\underline{\hat{X}^r_1}(t)\|_{T^r} \\ &+ 8\epsilon_r \left(\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + \left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r_{22}}(t)\right\|_{T^r}\right). \qquad (4.163)\end{aligned}$$

Combining this with (4.161) and (4.162) yields

$$\begin{aligned}&\left\|\underline{\hat{Z}^r_{11}}(t) + \left(\underline{\hat{X}^r_1}(t) + \underline{\hat{X}^r_2}(t)\right)^-\right\|_{T^r} \\ &\qquad < 20\epsilon_r \left(\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + \left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r_{22}}(t)\right\|_{T^r}\right) \qquad (4.164) \\ &\left\|\underline{\hat{Q}^r_1}(t)\right\|_{T^r} < 20\epsilon_r \left(\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + \left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r_{22}}(t)\right\|_{T^r}\right) \qquad (4.165) \\ &\|\underline{\hat{Q}^r_2}(t)\|_{T^r} \leq 2\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + 2\left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} \\ &\qquad + 50\epsilon_r \left(\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + \left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r_{22}}(t)\right\|_{T^r}\right). \qquad (4.166)\end{aligned}$$

Finally, by (4.156), on $\mathcal{D}^r$

$$\|\underline{\hat{Z}^r_{21}}(t) + \underline{\hat{Z}^r_{22}}(t)\|_{T^r} < \epsilon_r \left(\left\|\underline{\hat{Q}^r}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r}(t)\right\|_{T^r}\right).$$

Thus, by (4.164)-(4.166)

$$\|\underline{\hat{Z}^r_{22}}(t)\|_{T^r} < 10\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + 10\left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + 100\epsilon_r \left(\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + \left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r} + \left\|\underline{\hat{Z}^r_{22}}(t)\right\|_{T^r}\right).$$

Therefore,

$$\|\underline{\hat{Z}^r_{22}}(t)\|_{T^r} < 20\left\|\underline{\hat{X}^r_1}(t)\right\|_{T^r} + 20\left\|\underline{\hat{X}^r_2}(t)\right\|_{T^r}.$$

This combined with (4.163)-(4.166) yields the desired result. □

Using this result we can now prove that the compact containment condition is satisfied by $\left\|\hat{X}^r(t)\right\|_T$ for each $T > 0$.

Proposition 4.30. *Under the assumptions of Proposition 4.28,* $\left\{\left\|\hat{X}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\}$ *satisfies the compact containment condition, where* L^r *is taken as in Proposition 4.28.*

Proof. Assume that assumptions of Proposition 4.28 hold. We omit π^* from the notation below for simplicity. Fix $T > 0$.

By (3.5),

$$\{\|\hat{Q}_1^r(t)\|_{L^r/\sqrt{|N^r|}}\},\ \{\|\hat{Z}_{11}^r(t)+\left(\hat{X}_1^r(t)+\hat{X}_2^r(t)\right)^-\|_{L^r/\sqrt{|N^r|}}\},\ \text{and}$$
$$\{\|\hat{Z}_{21}^r(t)+\hat{Z}_{22}^r(t)\|_{L^r/\sqrt{|N^r|}}\}, \tag{4.167}$$

satisfy the compact containment condition, since for each SSC function used to prove the SSC result H is a decreasing function. Therefore, we can find sequence of random variables $\{\Gamma^r\}$ that satisfies the compact containment condition such that

$$|\hat{Z}_{11}^r(t)| \le \Gamma^r + |\hat{X}_1^r(t)| + |\hat{X}_2^r(t)| \tag{4.168}$$

for all $t \in [0, L^r/\sqrt{|N^r|}]$.

Observe that by definition of fluid and diffusion scalings and (4.70)-(4.73)

$$\hat{X}_1^r(t)+\hat{X}_2^r(t) \;=\; x^r(t)-\mu_1\int_0^t \hat{Z}_{11}^r(s)ds-\mu_2\int_0^t\left(\hat{Z}_{21}^r(s)+\hat{Z}_{22}^r(s)\right)ds, \tag{4.169}$$

where $x^r(t)$ is defined by (4.114).

Using (4.169) and (4.168) we get for all $t \in [0, L^r/\sqrt{|N^r|}$

$$\begin{aligned}
|\hat{X}_1^r(t)|+|\hat{X}_2^r(t)| &\le |x^r(t)|+\mu_1\int_0^t\left(\left|\hat{X}_1^r(s)\right|+\left|\hat{X}_2^r(s)\right|+\Gamma^r(s)\right)ds \\
&\qquad +\mu_2\int_0^t\left(|\hat{Z}_{21}^r(s)+\hat{Z}_{22}^r(s)|\right)ds \\
&\le |x^r(t)|+\mu_1\int_0^t\left(\left|\hat{X}_1^r(s)\right|+\left|\hat{X}_2^r(s)\right|\right)ds+\frac{L^r}{\sqrt{|N^r|}}\mu_2\|\hat{Z}_{21}^r(s)+\hat{Z}_{22}^r(s)\|_{L^r/\sqrt{|N^r|}} \\
&\qquad +\frac{L^r}{\sqrt{|N^r|}}\mu_1\Gamma^r.
\end{aligned}$$

By Gronwall's inequality, see [51],

$$\begin{aligned}
&\sup_{0\le t\le \frac{L^r}{\sqrt{|N^r|}}}\left\{|\hat{X}_1^r(t)|+|\hat{X}_2^r(t)|\right\} \le \\
&\left(\|x^r(t)\|_{\frac{L^r}{\sqrt{|N^r|}}}+\frac{L^r}{\sqrt{|N^r|}}\left(\mu_2\|\hat{Z}_{21}^r(s)+\hat{Z}_{22}^r(s)\|_{L^r/\sqrt{|N^r|}}+\mu_1\Gamma^r\right)\right)\exp\left\{\mu_1\frac{L^r}{\sqrt{|N^r|}}\right\}.
\end{aligned}$$

Thus, $\left\|\hat{X}^r(t)\right\|_{L^r/\sqrt{|N^r|}}$ satisfies the compact containment condition by (4.167) and (4.168).

□

Proposition 4.31. *Under the assumptions of Proposition 4.28,*

$$\left\{\left\|\hat{Q}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\} \text{ and } \left\{\left\|\hat{Z}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\}$$

satisfy the compact containment condition, where L^r is taken as in Proposition 4.28.

Proof. Assume that assumptions of Proposition 4.28 hold. We omit π^* from the notation below for simplicity. Fix $T > 0$.

$$\left\{\left\|\hat{Q}_1^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\} \tag{4.170}$$

satisfies the compact containment condition by (4.167) and

$$\left\{\left\|\hat{Z}_{11}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\} \tag{4.171}$$

satisfies the compact containment condition by Proposition 4.30 and (4.167). Since

$$\hat{Z}_{21}^r(t) = \hat{X}_1^r(t) - \hat{Q}_1^r(t) - \hat{Z}_{11}^r(t)$$

$$\left\{\left\|\hat{Z}_{21}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\}$$

satisfies the compact containment condition by Proposition 4.30, (4.170) and (4.171). Then,

$$\left\{\left\|\hat{Z}_{22}^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\}$$

satisfies the compact containment condition by (4.167). Finally,

$$\left\{\left\|\hat{Q}_2^r(t)\right\|_{L^r/\sqrt{|N^r|}}\right\}$$

satisfies the compact containment condition by Proposition 4.30. □

Proposition 4.32. *Under the assumptions of Proposition 4.28,*

$$\hat{X}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) \Rightarrow \hat{X}(0)$$

as $r \to \infty$, where L^r is taken as in Proposition 4.28

Proof. Assume that assumptions of Proposition 4.28 hold. We omit π^* from the notation below for simplicity. Fix $T > 0$.

$$\begin{aligned}\hat{X}_1^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) - \hat{X}_1^r(0) &= \hat{A}_1^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) - \hat{S}_{11}^r\left(\bar{T}_{11}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right) - \hat{S}_{21}^r\left(\bar{T}_{21}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right) \\ &- \mu_1\int_0^{\frac{L^r}{\sqrt{|N^r|}}} \hat{Z}_{11}^r(s)ds - \mu_2\int_0^{\frac{L^r}{\sqrt{|N^r|}}} \hat{Z}_{21}^r(s)ds \\ &+ L^r\left(\frac{\lambda_1^r}{|N^r|} - \mu_1\beta_1 - \mu_2 x_{21}\beta_2\right).\end{aligned}$$

The first three terms on the RHS above go to zero in probability since $\hat{A}_1^r(\cdot)$, $\hat{S}_{11}^r\left(\bar{T}_{11}^r(\cdot)\right)$, and $\hat{S}_{21}^r\left(\bar{T}_{21}^r(\cdot)\right)$ converge weakly to Brownian motions jointly. Also, since $\left\{\|\hat{Z}_{11}^r(s)\|_{\frac{L^r}{\sqrt{|N^r|}}}\right\}$ and $\left\{\|\hat{Z}_{21}^r(s)\|_{\frac{L^r}{\sqrt{|N^r|}}}\right\}$ satisfy the compact containment condition, and $\frac{L^r}{\sqrt{|N^r|}} \to 0$, terms in the second line goes to zero. The last term goes to zero by (4.72). Hence

$$\left|\hat{X}_1^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) - \hat{X}_1^r(0)\right| \to 0$$

in probability as $r \to \infty$. We get the desired result by (4.74) and Theorem 4.1 in Billingsley [11]. □

Proposition 4.33. *Under the assumptions of Proposition 4.28,* $\left\{\left\|\hat{X}^r(t)\right\|_T\right\}$ *satisfies the compact containment condition.*

Proof. Assume that assumptions of Proposition 4.28 hold. We omit π^* from the notation below for simplicity. Fix $T > 0$.

Recall that $x(t;s) = x(t) - x(s)$ for a process x. By definition of fluid and diffusion scalings and (4.70)-(4.73)

$$\begin{aligned}\hat{X}_1^r(t) + \hat{X}_2^r(t) &= \hat{X}_1^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) + \hat{X}_2^r\left(\frac{L^r}{\sqrt{|N^r|}}\right) + x^r\left(t;\frac{L^r}{\sqrt{|N^r|}}\right) \\ &- \mu_1\int_{\frac{L^r}{\sqrt{|N^r|}}}^t \hat{Z}_{11}^r(s)ds - \mu_2\int_{\frac{L^r}{\sqrt{|N^r|}}}^t \left(\hat{Z}_{21}^r(s) + \hat{Z}_{22}^r(s)\right)ds. \quad (4.172)\end{aligned}$$

where $x^r(t)$ is defined by (4.114). Recall that

$$\underline{y^r}(t) = y^r\left(t + \frac{L^r}{\sqrt{|N^r|}}\right)$$

for a sequence of processes $\{y^r(\cdot)\}$. For notational convenience we set

$$T^r = T - \frac{L^r}{\sqrt{|N^r|}}.$$

Choose $\epsilon_i^r(t)$, for $i = 1, 2$ such that

$$\hat{Z}_{21}^r(t) + \hat{Z}_{22}^r(t) = \epsilon_1^r(t) \left(\left\| \underline{\hat{X}_1^r}(t) \right\|_{T^r} + \left\| \underline{\hat{X}_2^r}(t) \right\|_{T^r} + 1 \right) \tag{4.173}$$

$$\hat{Z}_{11}^r(t) = -\left(\hat{X}_1^r(t) + \hat{X}_2^r(t) \right)^- + \epsilon_2^r(t) \left(\left\| \underline{\hat{X}_1^r}(t) \right\|_{T^r} + \left\| \underline{\hat{X}_2^r}(t) \right\|_{T^r} + 1 \right) \tag{4.174}$$

for all $t \in \left[\frac{L^r}{\sqrt{|N^r|}}, T \right]$. Note that, by Corollary 4.29, for $i = 1, 2$

$$\|\underline{\epsilon_i^r}(t)\|_{T^r} \to 0 \tag{4.175}$$

in probability as $r \to \infty$. Let

$$\breve{x}^r(t) = \hat{X}_1^r \left(\frac{L^r}{\sqrt{|N^r|}} \right) + \hat{X}_2^r \left(\frac{L^r}{\sqrt{|N^r|}} \right) + x^r \left(t; \frac{L^r}{\sqrt{|N^r|}} \right)$$

Using (4.172), (4.173) and (4.174) we get

$$\begin{aligned}
|\hat{X}_1^r(t)| + |\hat{X}_2^r(t)| &\leq \breve{x}^r(t) + \mu_1 \int_{\frac{L^r}{\sqrt{|N^r|}}}^t \left| \left(\hat{X}_1^r(s) + \hat{X}_2^r(s) \right)^- \right| ds \\
&+ \left(\mu_1 \|\underline{\epsilon_1^r}(t)\|_{T^r} + \mu_2 \|\underline{\epsilon_2^r}(t)\|_{T^r} \right) \left(\|\underline{\hat{X}_1^r}(t)\|_{T^r} + \|\underline{\hat{X}_2^r}(t)\|_{T^r} + 1 \right) T \\
&\leq |\breve{x}^r(t)| + \mu_1 \int_{\frac{L^r}{\sqrt{|N^r|}}}^t \left(\left| \hat{X}_1^r(s) \right| + \left| \hat{X}_2^r(s) \right| \right) ds \\
&+ \left(\mu_1 \|\underline{\epsilon_1^r}(t)\|_T + \mu_2 \|\underline{\epsilon_2^r}(t)\|_{T^r} \right) \left(\|\underline{\hat{X}_1^r}(t)\|_{T^r} + \|\underline{\hat{X}_2^r}(t)\|_{T^r} + 1 \right) T
\end{aligned}$$

By Gronwall's inequality, see [51],

$$\begin{aligned}
&\sup_{\frac{L^r}{\sqrt{|N^r|}} \leq t \leq T} \left\{ |\hat{X}_1^r(t)| + |\hat{X}_2^r(t)| \right\} \leq \\
&\left(\|\breve{x}^r(t)\|_T + \left(\mu_1 \|\underline{\epsilon_1^r}(t)\|_{T^r} + \mu_2 \|\underline{\epsilon_2^r}(t)\|_T \right) \left(\|\underline{\hat{X}_1^r}(t)\|_{T^r} + \|\underline{\hat{X}_2^r}(t)\|_{T^r} + 1 \right) T \right) \exp\{\mu_1 T\}
\end{aligned}$$

Since

$$\sup_{\frac{L^r}{\sqrt{|N^r|}} \leq t \leq T} \left\{ |\hat{X}_1^r(t)| + |\hat{X}_2^r(t)| \right\} \geq 0.5 \left(\|\underline{\hat{X}_1^r}(t)\|_{T^r} + \|\underline{\hat{X}_2^r}(t)\|_{T^r} \right),$$

we have

$$\begin{aligned}
&\left(0.5 - \left(\mu_1 \|\underline{\epsilon_1^r}(t)\|_T + \mu_2 \|\underline{\epsilon_2^r}(t)\|_T \right) T \exp\{\mu_1 T\} \right) \left(\|\underline{\hat{X}_1^r}(t)\|_{T^r} + \|\underline{\hat{X}_2^r}(t)\|_{T^r} \right) \leq \\
&\qquad \left(\|\breve{x}^r(t)\|_T + \left(\mu_1 \|\underline{\epsilon_1^r}(t)\|_T + \mu_2 \|\underline{\epsilon_2^r}(t)\|_T \right) T \right) \exp\{\mu_1 T\}
\end{aligned}$$

By (4.175), for arbitrary $\epsilon > 0$ we can choose R large enough so that, for $r > r_0(\epsilon, R)$

$$P\left\{\mu_1\|\underline{\epsilon_1^r}(t)\|_T + \mu_2\|\underline{\epsilon_2^r}(t)\|_T > \frac{1/4}{T\exp\{\mu_1 T\}}\right\} < \epsilon/4 \text{ and}$$

Similary, by (4.114), Proposition 4.23, Proposition 4.30, we have for $r > r_1(\epsilon, R)$ that

$$P\left\{\|\breve{x}^r(t)\|_T \exp\{\mu_1 T\} > (R-1)/4\right\} < \epsilon/4.$$

Hence, for $r > r_0(\epsilon, R) \vee r_1(\epsilon, R)$

$$P\left\{\|\underline{\hat{X}_1^r}(t)\|_T + \|\underline{\hat{X}_2^r}(t)\|_{T^r} > R\right\} < \epsilon.$$

Combining this with Proposition 4.30 gives the desired result. □

Proof of Proposition 4.24. We proved part (i) in Proposition 4.31, part (ii) in Proposition 4.32. Part (iii) follows from Propositions 4.28 and 4.33 and Remark 3.3. □

4.3 Open problems from the literature

In this section we illustrate two applications of Theorem 3.4 in a V-model parallel server system that is studied in Armony and Maglaras [3] and Milner and Olsen [53]. A V-model consists of a single server pool and multiple customer classes. We will focus on the case with two customer classes and assume that service rates of the customer classes are equal. A V-model parallel server system is illustrated in Figure 6.

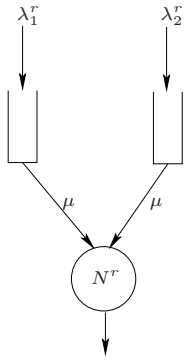

Figure 6: A V-model parallel server system

4.3.1 Armony-Maglaras threshold policy

Armony and Maglaras [3] uses a V-model system to study a contact center with two channels, one for real-time telephone service, and another for a postponed call-back service offered with a guarantee on the maximum delay until a reply is received. We assume that the second customer class consists of those customers who call for the call-back option.

Armony and Maglaras [3] proposed the following policy.

Threshold Rule. If $Q_2(t) > \sqrt{|N^r|}\theta$, give priority to class 2, otherwise give priority to class 1.

Let

$$\lambda^r = |N^r|\mu(1 - \frac{\beta}{\sqrt{|N^r|}}). \tag{4.176}$$

We assume that the arrival rates for each customer class is given according to

$$\lambda_1^r = \eta\lambda^r \text{ and } \lambda_2^r = (1-\eta)\lambda^r. \tag{4.177}$$

for some $\eta \in (0,1)$. Let

$$\hat{X}^r(t) = \hat{Q}_1^r(t) + \hat{Q}_2^r(t) + \hat{Z}_{11}^r(t) + \hat{Z}_{21}^r(t). \tag{4.178}$$

We assume that

$$(\hat{Q}^r(0), \hat{Z}^r(0)) \Rightarrow \left(\hat{Q}(0), \hat{Q}(0)\right) \tag{4.179}$$

By Theorem 2 in Halfin and Whitt [32], $\hat{X}^r$ converges weakly to a diffusion process X as $r \to \infty$.

We show that the following SSC collapse result, Proposition 3.1 in Armony and Maglaras [3], holds.

Proposition 4.34. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (4.179) holds and*

$$(\hat{Q}_1^r(0), \hat{Q}_2^r(0)) \Rightarrow \left((X(0)-\theta)^+, (X(0)^+ \wedge \theta)\right) \tag{4.180}$$

as $r \to \infty$. Then

$$(\hat{Q}_1^r(\cdot), \hat{Q}_2^r(\cdot)) \Rightarrow \left((X(\cdot)-\theta)^+, (X(\cdot)^+ \wedge \theta)\right)$$

as $r \to \infty$.

The proof presented in Armony and Maglaras [3] contains a step that cannot be rigorously proved, see inequality (29) in that paper. In this study, we will present an alternative proof using Theroem 3.4. Using Proposition 4.34 one can prove the asymptotic optimality of the threshold policy, see Proposition 3.4 in Armony and Maglaras [3].

4.3.1.1 Analysis of Armony-Maglaras threshold policy

Let $\{\mathbb{X}^r\}$ be a sequence of V-systems working under the Armony-Maglaras threshold policy. We start our analysis by presenting the additional equations that must be satisfied by $\mathbb{X}^r$.

Since class 2 jobs get priority when the number of class 2 jobs in the queue exceeds $\sqrt{|N^r|}\theta$,

$$B_{11}^r(t) + A_{11}^r(t) \text{ can only increase when } Q_2(t)^r < \sqrt{|N^r|}\theta. \tag{4.181}$$

Also

$$B_{21}^r(t) \text{ can only increase when } Q_2(t)^r \geq \sqrt{|N^r|}\theta. \tag{4.182}$$

The following proposition characterizes the fluid limits of the V-parallel server systems working under the Armony-Maglaras threshold policy.

Proposition 4.35. *Let* $\{\mathbb{X}^r\}$ *be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that the conditions of Theorem A.1 are satisfied.*

i. In addition to the fluid limit equations (A.2)-(A.9), each fluid limit $\bar{\mathbb{X}}$ *of* $\mathbb{X}^r$ *satisfies*

$$\dot{\bar{A}}_{11}(t) + \dot{\bar{B}}_{11}(t) = 0 \text{ when } \bar{Q}_2(t) > 0,$$

ii. Let $\vec{q}_r = (q_1, q_2)$, *where* $q_1 = r \geq 0$ *and* $q_2 = 0$ *and* $z = \{z_{11}, z_{12}\}$, *where* $z_{1i} = \frac{\lambda_i}{\mu_i}$ *for* $i = 1, 2$. *Then,* $\mathcal{M} = \{(\vec{q}_r, z) : r \geq 0\}$ *is the set of all the invariant states of the fluid limits of* $\mathbb{X}^r$.

Proof is placed in Section 4.3.1.2.

By (4.179),

$$(\bar{Q}^r(0), \bar{Z}^r(0)) \Rightarrow (0, z)), \tag{4.183}$$

where $z = (\lambda_1/\mu_1, \lambda_2/\mu_2)$, hence $\mathbb{X}^r$ satisfies Assumption 2. Note that, $\mathbb{X}^r$ satisfies Assumption 1 by (4.176) and (4.177).

We prove the SSC result using Theorem 3.7. Therefore, we next show that Assumption 6 holds. The following result is established by Halfin and Whitt [32].

Theorem 4.36. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy and $\hat{X}^r$ be defined as in (4.178). If (4.179) holds then*

$$\hat{X}^r(\cdot) \Rightarrow \hat{X}(\cdot)$$

where

$$\hat{X}(t) = \hat{X}(t) + W(t) - \beta t - \mu \int_0^t (\hat{X}(s))^- ds$$

and W is a Borwnian motion with sero drift and variance 2μ.

It can be easily showed using Theorem 4.36 that

$$\lim_{R\to\infty} \lim_{r\to\infty} P\left\{ \left\| \hat{X}^r(t) \right\|_T > R \right\} = 0. \tag{4.184}$$

Next, we show that Assumption 6 holds using this result.

Proposition 4.37. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy.*

$$\lim_{R\to\infty} \lim_{r\to\infty} P\left\{ \left\| \hat{Q}^r(t) \right\|_T \vee \left\| \hat{Z}^r(t) \right\|_T > R \right\} = 0, \tag{4.185}$$

i.e., $\{\mathbb{X}^r\}$ satisfies Assumption 6.

A proof is presented in Section 4.3.1.2.

Now we define the SSC function for this setting. Let $q = (q_1, q_2) \in \mathbb{R}^2$, $z = (z_1, z_2) \in \mathbb{R}^2$, $x = q_1 + q_2 + z_1 + z_2$ and $g : \mathbb{R}^4 \to \mathbb{R}$ be defined by

$$g(q, z) = q_1 - (x - \theta)^+$$

Clearly $|g|$ is continuous but it does not satisfy (3.10). Therefore, we use Theorem 3.7.

Proof of Proposition 4.34. Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (4.176), (4.177), (4.179) and (4.180) holds.

From the results above and definition of g in order to invoke Theorem 3.7 it is enough to show that Assumption 7 holds.

We first establish the additional hydrodynamic equations that must be satisfied by $\{\mathbb{X}^r\}$. First note that, by (4.181) and (4.182)

$$B_{11}^r(t) + A_{11}^r(t) \text{ can only increase when}$$
$$g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) = \hat{Q}_1^r(t) - (\hat{X}^r(t) - \theta)^+ > 0 \tag{4.186}$$

and

$$B_{21}^r(t) \text{ can only increase when } g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right) = \hat{Q}_1^r(t) - (\hat{X}^r(t) - \theta)^+ \leq 0, \tag{4.187}$$

since, if $\hat{Q}_1^r(t) > (\hat{X}^r(t) - \theta)^+$, then $\hat{X}^r(t) = \hat{Q}_1^r(t) + \hat{Q}_2^r(t)$, because the policy is non-idling. Therefore, $\hat{Q}_2^r(t) \leq \theta$ in this case. Similary, if $\hat{Q}_1^r(t) \leq (\hat{X}^r(t) - \theta)^+$ and $\hat{Q}_2^r(t) > 0$, then $\hat{Q}_2^r(t) \geq \theta$.

Equations (4.186) and (4.187) imply that

$$B_{11}^{r,m}(t) + A_{11}^{r,m}(t) \text{ can only increase when}$$
$$g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\left(Q^{r,m}(t), Z^{r,m}(t)\right)\right) > 0 \text{ and} \tag{4.188}$$
$$B_{21}^{r,m}(t) \text{ can only increase when } g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\left(Q^{r,m}(t), Z^{r,m}(t)\right)\right) \leq 0, \tag{4.189}$$

where the hydrodynamic scaled process $\mathbb{X}^{r,m}$ is defined by (3.24).

Fix $R > 0$ and $T > 0$ and let $\mathcal{A}^r(T)$ be defined as in (3.19). Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit on $\mathcal{A}_R^r(T)$. Recall that we have showed that $\tilde{\mathbb{X}}$ satisfies (3.1)-(3.8). We next characterize the additional equations associated with the policy. We claim that

$$\dot{\tilde{B}}_{11}(t) = \mu \text{ when } g\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) > 0 \text{ and } \tilde{Q}_1(t) > 0 \tag{4.190}$$
$$\dot{\tilde{B}}_{12}(t) = \mu \text{ when } g\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) < 0 \text{ and } \tilde{Q}_2(t) > 0. \tag{4.191}$$

To show this assume that

$$g\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) > 2\epsilon \text{ and } \tilde{Q}_1(t) > 2\epsilon \tag{4.192}$$

for some $\epsilon > 0$. By continuity of g and $\tilde{\mathbb{X}}$ there exists $\delta > 0$ such that

$$g\left(R\left(\tilde{Q}(s), \tilde{Z}(s)\right)\right) > \epsilon \text{ and } \tilde{Q}_1(s) > \epsilon$$

for all $s \in [t-\delta, t+\delta]$.

Pick r large enough together with an integer m and $w \in \mathcal{A}^r(T)$ so that

$$\|\tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t)\| < \epsilon/2.$$

This gives that

$$g\left(R\left(Q^{r,m}(s), Z^{r,m}(s)\right)\right) > \epsilon/2 \text{ and } Q_1^{r,m}(s) > \epsilon/2,$$

since $\sqrt{\frac{x_{r,m}}{|N^r|}} = R$ on $\mathcal{A}^r(T)$. By (4.188)

$$B_{12}^{r,m}(t+\delta) - B_{12}^{r,m}(t-\delta) = 0,$$

and so

$$\dot{\tilde{B}}_{12}(t) = 0,$$

Now, by (3.5)

$$\dot{\tilde{Z}}_{12}(t) = -(1-\eta)\mu \text{ and}$$
$$\dot{\tilde{Z}}_{11}(t) = \dot{\tilde{B}}_{11}(t) - \mu\eta\mu.$$

Equations (3.5), (3.8) and (4.192) give that

$$\dot{\tilde{Z}}_{11}(t) + \dot{\tilde{Z}}_{11}(t) = 0.$$

Hence

$$\dot{\tilde{B}}_{11}(t) = \mu.$$

Condition (4.191) is proved similarly.

Next we prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}|g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)| < 0 \tag{4.193}$$

for every regular point t of $|g|$ whenever $|g\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)| > 0$.

Let $\tilde{X}_i(t) = \tilde{Q}_i(t) + \tilde{Z}_{1i}(t)$ and $\tilde{X}(t) = \tilde{X}_1(t) + \tilde{X}_2(t)$. Then, by (3.5) and (3.5), $\tilde{X}_i(t) = \tilde{X}_i(0)$ for all $t \geq 0$, hence

$$\dot{\tilde{X}}(t) = 0 \text{ for all } t \geq 0. \tag{4.194}$$

First assume that $g_1\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) > 0$. Then, by (4.190)

$$\dot{\tilde{Q}}_1(t) = \lambda_1 - \mu = -(1-\eta)t.$$

Hence,

$$\dot{g}_1\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) = R\dot{\tilde{Q}}_1(t) - \frac{\mathrm{d}}{\mathrm{d}t}(R\tilde{X}(t) - \theta)^+ = -(1-\eta)t.$$

by (4.194).

Similarly, if $g_1\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) < 0$ then

$$\dot{g}_1\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) = \dot{\tilde{Q}}_1(t) = \eta t.$$

This proves (4.193). By (4.193) Assumption 7 holds and this completes the proof. □

4.3.1.2 Proofs of Propositions 4.35 and 4.37

Proof of Proposition 4.35. We prove the proposition in two parts.

i. Let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that $\{\bar{\mathbb{X}}^r(\cdot, \omega)\}$, for some $\omega \in \mathcal{A}$, where $\mathcal{A}$ is defined as in proof of Theorem A.1 that satisfies (A.14), converges u.o.c. to $\bar{\mathbb{X}}$. Assume that $\bar{Q}_2(t) > 0$.

By the continuity of $\bar{Q}$ there exists $\varepsilon > 0$ and $\delta > 0$ such that $\bar{Q}_2(s) > \varepsilon$ for all $s \in [t-\delta, t+\delta]$. Since $\{\bar{\mathbb{X}}^r(\cdot, \omega)\}$ converges u.o.c. to $\bar{\mathbb{X}}$, $\bar{Q}_2^r(s) > \varepsilon/4$ for all $s \in [t-\delta, t+\delta]$ and r large enough. Hence, $A_{11}^r(\cdot, \omega)$ and $B_{11}^r(\cdot, \omega)$ are flat on $[t-\delta, t+\delta]$ by (4.181). Hence

$$\dot{\bar{A}}_{11}(t) + \dot{\bar{B}}_{11}(t) = 0. \tag{4.195}$$

ii. Fix $(\vec{q}_r, z) \in \mathcal{M}$. We show that if $\bar{Q}(0) = \vec{q}_r$ and $\bar{Z}(0) = z$ then $\bar{Q}(t) = \vec{q}_r$ and $\bar{Z}(t) = z$ for all $t > 0$. So assume that $\bar{Q}(0) = \vec{q}_r$ and $\bar{Z}(0) = z$ for a fluid model solution.

We start by showing that $\bar{Z}_{12}(t) \geq z_{12}$. Let $f_1(t) = (\bar{Z}_{12}(t) - z_{12})^-$. It is enough to show, by virtue of Lemma 5.2 of Dai [18], that $\dot{f}_1(t) \leq 0$ whenever $f_1(t) > 0$ for a regular point $t > 0$. Assume that f_1 is differentiable at time $t > 0$ and that $f_1(t) > 0$, i.e., $\bar{Z}_{12}(t) < z_{12}$. Note that by (A.4),

$$\dot{\bar{Z}}_{12}(t) = \dot{\bar{B}}_{12}(t) + \dot{\bar{A}}_{212}(t) - \mu_{12}\bar{Z}_{12}(t).$$

If $\bar{Q}_2(t) > 0$, then by (4.195) , (A.4) and (A.8), $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \dot{\bar{D}}_{11}(t) + \dot{\bar{D}}_{12}(t) = \mu_1\bar{Z}_{11}(t) + \mu_2\bar{Z}_{21}(t)$. Also $\bar{Q}_2(t) > 0$ implies $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$, by (A.6). Hence, $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) > \mu_{12}\bar{Z}_{12}(t)$, which implies $\dot{\bar{Z}}_{12}(t) > 0$ and $\dot{f}_1(t) < 0$. If $\bar{Q}_2(t) = 0$, then we claim that $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_2$. If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) < 1$, this trivially follows from (A.6). If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$ then we use the fact that $\dot{\bar{Q}}_2(t) = 0$, since it achieves its minimum at t. This implies by (A.3) that $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_2$. Hence, if $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) < z_{12}$, then $\dot{f}_1(t) \leq 0$. Hence, if $\bar{Z}_{12}(0) \geq z_{12}$ then $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$.

Next, we show that if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$. Assume that $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$ and that $\bar{Q}_2(t) > 0$. By the previous argument we have that $\bar{Z}_{12}(t) \geq z_{12}$. By (A.3) and (A.4), $\dot{\bar{Q}}_2(t) + \dot{\bar{Z}}_{12}(t) \leq \lambda_1 - \mu_{12}\bar{Z}_{12}(t) \leq 0$. By (4.195), $\dot{\bar{Z}}_{12}(t) \geq 0$ when $\bar{Q}_2(t) > 0$. Hence, if $\bar{Q}(t) > 0$ and it is differentiable at $t > 0$ then $\dot{\bar{Q}}(t) \leq 0$. Hence, if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$.

Now we are ready to show that if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) = z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$ for all $t \geq 0$. Assume that $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, and $\bar{Z}_{12}(t) > z_{12}$ for a regular point $t > 0$. By the previous paragraph $\bar{Q}_2(t) = 0$, hence $\dot{\bar{Q}}_2(t) = 0$ by a similar argument above. If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$, $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_1$ by the fact that $\dot{\bar{Q}}_2(t) = 0$, and by equations (A.2) and (A.3). If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) < 1$, then $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_1$ by (A.2),(A.3), (A.5), and (A.9). Since $\dot{\bar{Z}}_{12}(t) = \dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) - \mu_{12}\bar{Z}_{12}(t)$, by (A.4), $\dot{\bar{Z}}_{12}(t) < 0$ if $\bar{Z}_{12}(t) > z_{12}$. Hence, if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) = z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$ for all $t \geq 0$.

Next we show that if $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$ then $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$ for all $t \geq 0$. Let $t > 0$ be a regular point. By the

arguments above, we have that $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$. Hence, $\bar{Z}_{11}(t) \leq z_{11}$ by the definition of the fluid scaling. So assume that $\bar{Z}_{11}(t) < z_{11}$. This implies $\bar{Q}_{11}(t) = 0$. Hence, $\dot{\bar{A}}_{111}(t)+\dot{\bar{B}}_{11}(t) = \lambda_1$. This gives that $\dot{\bar{Z}}_{11}(t) > 0$, since $\dot{\bar{Z}}_{11}(t) = \dot{\bar{A}}_{111}(t)+\dot{\bar{B}}_{11}(t)-\mu_{11}\bar{Z}_{11}(t)$.

Finally, we show that if $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, $\bar{Q}_1(0) = r$, and $\bar{Z}_{11}(0) = z_{11}$ then $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, $\bar{Q}_1(t) = r$, and $\bar{Z}_{11}(t) = z_{11}$ for all $t \geq 0$. Let $t > 0$ be a regular point. By the arguments above, we have that $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$. Assume that $\bar{Q}_1(t) > r$. Then $\dot{\bar{Q}}_1(t) = \dot{\bar{A}}_{11}(t) - \dot{\bar{B}}_{11}(t) = 0$, since by (A.2) $\dot{\bar{A}}_{11}(t) + \dot{\bar{A}}_{111}(t) = \lambda_1$ and by (A.4) $\dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) = \mu_{11}z_{11} = \lambda_1$, when $\bar{Z}_{11}(t) = z_{11}$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Q}_{12}(t) = 0$. □

Proof of Proposition 4.37. Proof is similar to that of Lemma 3.2 of Puhalskii and Reiman [56]. Let

$$\hat{X}_k^r(t) = \frac{Q_k^r(t) + Z_{1k}^r(t) - |N^r|\lambda_k/\mu}{\sqrt{|N^r|}} \tag{4.196}$$

for $k = 1, 2$. We claim that

$$|\hat{Z}_{1k}^r(t)| \leq |\hat{X}_k^r(t)| + \left(\hat{X}^r(t)\right)^+. \tag{4.197}$$

To prove this assume that $\hat{Z}_{1k}^r(t) < 0$, otherwise the result is obvious. If $\hat{Z}_{11}^r(t)+\hat{Z}_{12}^r(t) < 0$, then $\hat{Q}_k^r(t) = 0$ so the result follows. Assume that $\hat{Z}_{11}^r(t) + \hat{Z}_{12}^r(t) = 0$. Without loss of generality we can assume that $k = 1$. Since $\hat{Z}_{11}^r(t) < 0$, $\hat{Z}_{12}^r(t) = -\hat{Z}_{11}^r(t)$ and $\hat{Q}_2^r(t) \geq 0$, so (4.197) follows.

By (4.196), for $k = 1, 2$

$$\begin{aligned}\hat{X}_k^r(t) &= \hat{X}_k^r(0) + \left(\frac{A_k^r(t) - \lambda_k^r t}{\sqrt{|N^r|}}\right) - \sqrt{|N^r|}\left(\frac{S_{1k}\left(|N^r|\int_0^t \bar{Z}_{1k}^r(s)\mathrm{d}s\right)}{|N^r|} - \mu_{1k}\int_0^t \bar{Z}_{1k}^r(s)\mathrm{d}s\right)\\ &\quad + \sqrt{|N^r|}t\left(\frac{\lambda_k^r}{|N^r|} - \lambda_k\right) - \mu_{1k}\int_0^t \hat{Z}_{1k}^r(s)\mathrm{d}s.\end{aligned}$$

Let

$$\hat{A}_k^r(t) = \frac{A_k^r(t) - \lambda_k^r t}{\sqrt{|N^r|}} \text{ and } \hat{S}_k^r(t) = \sqrt{|N^r|}\left(\frac{S_{1k}\left(|N^r|\int_0^t \bar{Z}_{1k}^r(s)\mathrm{d}s\right)}{|N^r|} - \mu\int_0^t \bar{Z}_{1k}^r(s)\mathrm{d}s\right).$$

Note that

$$\hat{A}_k^r(\cdot) \Rightarrow W_k^a(\cdot) \text{ and } \hat{S}_k^r(t) \Rightarrow W_k^s(\cdot) \tag{4.198}$$

as $r \to \infty$ by Proposition 4.35, (4.183), and Donsker's Theorem, see Billingsley [11], where W_k^a and W_k^b are Brownians motion with zero drift and variance λ_k.

Now observe that

$$\begin{aligned}
|X_1^r(t)| + |X_2^r(t)| &\leq |\hat{X}_1^r(0)| + |\hat{X}_2^r(0)| + \left|\hat{A}_1^r(t) + \hat{S}_1^r(t)\right| + \left|\hat{A}_2^r(t) + \hat{S}_2^r(t)\right| \\
&+ \mu \int_0^t \left(|\hat{Z}_{11}^r(s)| + |\hat{Z}_{21}^r(s)|\right) \mathrm{d}s \\
&\leq |\hat{X}_1^r(0)| + |\hat{X}_2^r(0)| + \left|\hat{A}_1^r(t)\right| + \left|\hat{A}_2^r(t)\right| + \left|\hat{S}_1^r(t)\right| + \left|\hat{S}_2^r(t)\right| \\
&+ 3\mu \int_0^t \left(|X_1^r(s)| + |X_2^r(s)|\right) \mathrm{d}s,
\end{aligned}$$

where the last inequality follows from (4.197). This with Gronwall's inequality and (4.198) gives

$$\lim_{R\to\infty} \lim_{r\to\infty} P\left\{\left\|\hat{X}_1^r(t)\right\|_T \vee \left\|\hat{X}_2^r(t)\right\|_T > R\right\} = 0.$$

This gives (4.185) since $\hat{Q}_i^r(t) \geq 0$ for all $t \geq 0$, $r \geq 0$, and $k = 1, 2$. □

4.3.2 Milner-Olsen threshold policy

In Milner and Olsen [53], a call center with contract and non-contract customers is studied. The first customer class consist of those customers under contract who are given a guarantee for the percentile of delay. The second customer class is not given a guarantee on the service level. In the model studied in Armony and Maglaras [3], the service level guarantee for the second customer class is based on the waiting time not the percentile of the delay. We again assume that (4.176),(4.177) and (4.179) hold.

The following policy is proposed by Milner and Olsen [53] and they conjecture that it satisfies the service level guarantee for the first customer class and minimize the expected waiting time of the customers in the second class.

Milner-Olsen Policy: Let $0 < Q_L < Q_H$. When a server becomes available if $Q_L < \hat{Q}_1^r(t)$ and $\hat{Q}_1^r(t) + \hat{Q}_2^r(t) < Q_H$ then class 1 customers are served if available, otherwise class 2 customers (if any) are served. However, if when a server becomes available either $\hat{Q}_1^r(t) \leq Q_L$, $\hat{Q}_1^r(t) + \hat{Q}_2^r(t) \leq Q_L$, or $\hat{Q}_1^r(t) + \hat{Q}_2^r(t) \geq Q_H$ then class 2 customers are served, if possible, otherwise class 1 customers (if any) are routed to the free server.

The Milner-Olsen policy can also be defined as follows. Let $q = (q_1, q_2) \in \mathbb{R}^2$, $z = (z_1, z_2) \in \mathbb{R}^2$, $Q = q_1 + q_2$ and $g_1 : \mathbb{R}^4 \to \mathbb{R}$ be defined by

$$g_1(Q) = \begin{cases} Q & \text{if } Q \leq Q_L \\ Q_L & \text{if } Q_L < Q < Q_H \\ Q & \text{if } Q_L \leq Q \end{cases}$$

Then, if $\hat{Q}_1^r(t) > g_1(\hat{Q}_1^r(t) + \hat{Q}_2^r(t))$ give priority to class 1 customers otherwise class 2 customers get priority.

As in the previous section, by Theorem 2 in Halfin and Whitt [32], $\hat{X}^r$ converges weakly to a diffusion process $\hat{X}$ as $r \to \infty$ if $\hat{X}^r(0)$ converges weakly, see Theorem 4.36. Let

$$\hat{Q}_1(t) = g_1(\hat{X}(t)).$$

We claim $\hat{Q}_1$ is not in $D[0, \infty)$. To see this assume that $\hat{X}(0) = Q_H$. Then, on any finite interval $[0, T]$, $\hat{X}(0) - Q_H$ changes sign infinitely many times with positive probability, see Section 3.5.C and Problem 2.7.8 in Karatzas and Shreve [43]. Hence, $\hat{Q}_1$ has no right limit at 0. Therefore, in order to study the convergence of $\hat{Q}_1^r$ to $\hat{Q}_1$ a broader space than $D[0, \infty)$ must be considered with an appropriate metric. Instead, we study another policy to approximate the performance of the policy proposed by Milner and Olsen which ensures that $\hat{Q}_1 \in D[0, \infty)$.

We define a new function g_1^ϵ for $\epsilon > 0$ as follows.

$$g_1^\epsilon(Q) = \begin{cases} Q & \text{if } Q(t) \leq Q_L \\ Q_L & \text{if } Q_L < Q \leq Q_H - \epsilon \\ Q_L + \frac{Q_H - Q_L}{\epsilon}(Q - Q_H + \epsilon) & \text{if } Q_H - \epsilon < Q < Q \\ Q & \text{if } Q_L \leq Q \end{cases}$$

We define a new policy using g_1^ϵ as follows. Let

$$g^\epsilon(q, z) = q_1 - g_1^\epsilon(Q).$$

If $\hat{Q}_1^r(t) > g_1^\epsilon(\hat{Q}^r(t), \hat{Z}^r(t)) > 0$ give priority to class 1 customers otherwise give priority to class 2 customers. We call this policy the Milner-Olsen$^\epsilon$ threshold policy.

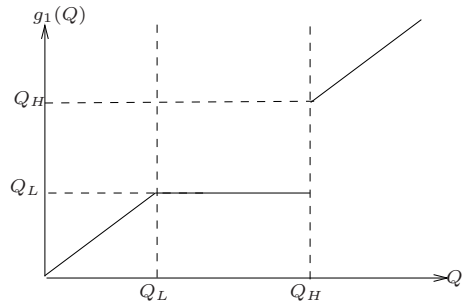

Figure 7: Graph of g_1

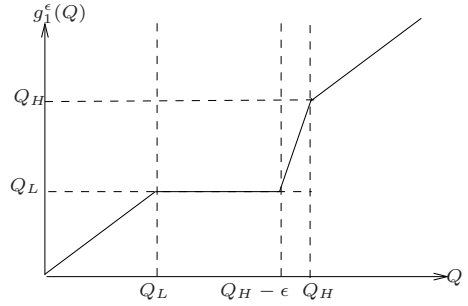

Figure 8: Graph of g_1^ϵ

Figure 4.3.2 shows Q versus $g_1(Q)$ and Figure 4.3.2 shows Q versus $g_1^\epsilon(Q)$. Clearly, g_1 is not continuous but $g_1^\epsilon(Q)$ is. We show that by continuity of g_1^ϵ $\hat{Q}_1^\epsilon$ is in $D[0,\infty)$.

We will prove the following state space collapse result using Theroem 3.4.

Proposition 4.38. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Milner-Olsen$^\epsilon$ threshold policy. Assume that*

$$(\hat{Q}_1^r(0), \hat{Q}_2^r(0)) \Rightarrow \left(g_1^\epsilon\left(\hat{Q}_1(0)+\hat{Q}_2(0)\right), \hat{Q}_1(0)+\hat{Q}_2(0) - g_1^\epsilon\left(\hat{Q}_1(0)+\hat{Q}_2(0)\right)\right)$$

as $r \to \infty$. Then

$$\left(\hat{Q}_1^r(\cdot), \hat{Q}_2^r(\cdot)\right) \Rightarrow \left(g_1^\epsilon\left(\hat{X}(\cdot)^+\right), \hat{X}(\cdot)^+ - g_1^\epsilon\left(\hat{X}(\cdot)^+\right)\right)$$

as $r \to \infty$.

Remark 4.39. It can be shown that; see Corollary 2 in Halfin and Whitt [32], for every $\eta > 0$, there exists $\epsilon > 0$ such that

$$P\left\{g_1(\hat{Q}(t)) - g_1^\epsilon(\hat{Q}(t)) > 0\right\} \leq \eta.$$

4.3.2.1 Analysis of Milner-Olsen$^\epsilon$ threshold policy

Let $\{\mathbb{X}^r\}$ be a sequence of V-systems working under the Milner-Olsen$^\epsilon$ threshold policy. We start our analysis by presenting the additional equations that must be satisfied by $\mathbb{X}^r$.

According to the Milner-Olsen$^\epsilon$ policy for all $s \leq t$

$$\begin{aligned} B_{11}^r(t) - B_{11}^r(s) &= D_{11}^r(t) + D_{12}^r(t) - D_{11}^r(s) - D_{12}^r(s) \\ &\text{if } g^\epsilon\left(\hat{Q}^r(u), \hat{Z}^r(u)\right) > 0 \text{ for all } u \in [s,t] \end{aligned} \tag{4.199}$$

and

$$\begin{aligned} B_{21}^r(t) - B_{21}^r(s) &= D_{11}^r(t) + D_{21}^r(t) - D_{11}^r(s) + D_{21}^r(s) \\ &\text{if } g^\epsilon\left(\hat{Q}^r(u), \hat{Z}^r(u)\right) < 0 \text{ for all } u \in [s,t]. \end{aligned} \tag{4.200}$$

The following proposition characterizes the fluid limits of V-parallel server systems working under the Milner-Olsen$^\epsilon$ threshold policy.

Proposition 4.40. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Milner-Olsen$^\epsilon$ threshold policy. Assume that the conditions of Theorem A.1 are satisfied.*

i. In addition to the fluid limit equations (A.2)-(A.9), each fluid limit $\bar{\mathbb{X}}$ of $\mathbb{X}^r$ satisfies

$$\dot{\bar{B}}_{11}(t) = 0 \text{ if } \bar{Q}_1(t) + \bar{Q}_2(t) > 0,$$

ii. Let $\vec{q}_r = (q_1, q_2)$, where $q_1 = r \geq 0$ and $q_2 = 0$ and $z = \{z_{11}, z_{12}\}$, where $z_{1i} = \frac{\lambda_i}{\mu_i}$ for $i = 1, 2$. Then, $\mathcal{M} = \{(\vec{q}_r, z) : r \geq 0\}$ is the set of all the invariant states of the fluid limits of $\mathbb{X}^r$.

The proof is similar to that of Proposition 4.35 hence it is omitted.

By (4.179),

$$(\bar{Q}^r(0), \bar{Z}^r(0)) \Rightarrow (0, z)),$$

where $z = (\lambda_1/\mu_1, \lambda_2/\mu_2)$, hence $\mathbb{X}^r$ satisfies Assumption 2. Note that, $\mathbb{X}^r$ satisfies Assumption 1 by (4.176) and (4.177).

We again prove the SSC result using Theorem 3.7. Therefore, we next show that Assumption 6 holds. We first note that the Milner-Olsen$^\epsilon$ threshold policy is non-idling. Therefore, the results of Theorem 4.36 and (4.184) hold for $\hat{X}^r$ in the current model too. Using these results we can also show that Proposition 4.37 also holds under the Milner-Olsen$^\epsilon$ threshold policy. Now we are ready to prove Proposition 4.38.

Proof of Proposition 4.38. Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Milner-Olsen$^\epsilon$ threshold policy.

We define the SSC function for this model by $|g^\epsilon|(\cdot) = g^\epsilon(\cdot)$. Note that $|g^\epsilon|$ is continuous. Therefore, it satisfies Assumption 5. Also, $\mathbb{X}^r$ satisfies Assumption 1 by (4.176) and (4.177). Also, by Proposition 4.40 Assumption 2 holds. In addition, $\{\mathbb{X}^r\}$ satisfies Assumption 6 by Proposition 4.37. Therefore, it remains to check that $\{\mathbb{X}^r\}$ satisfies Assumption 7.

First we establish the hydrodynamic limits. Equations (4.199) and (4.200) imply that,

for all $s \leq t$

$$B_{11}^{r,m}(t) - B_{11}^{r,m}(s) = D_{11}^{r,m}(t) + D_{12}^{r,m}(t) - D_{11}^{r,m}(s) + D_{12}^{r,m}(s)$$
$$\text{if } g^{\epsilon}\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\left(\hat{Q}^{r,m}(u), \hat{Z}^{r,m}(u)\right)\right) > 0 \text{ for all } u \in [s,t] \quad (4.201)$$
$$B_{12}^{r}(t) - B_{12}^{r}(s) = D_{11}^{r}(t) + D_{12}^{r}(t) - D_{11}^{r}(s) + D_{12}^{r}(s)$$
$$\text{if } g^{\epsilon}\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\left(\hat{Q}^{r,m}(u), \hat{Z}^{r,m}(u)\right)\right) < 0 \text{ for all } u \in [s,t] \quad (4.202)$$

where the hydrodynamic scaled process $\mathbb{X}^{r,m}$ is defined by (3.24).

Fix $R > 0$ and $T > 0$ and let $\mathcal{A}^r(T)$ be defined as in (3.19). Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit on $\mathcal{A}_R^r(T)$. Recall that we have showed that $\tilde{\mathbb{X}}$ satisfies (3.1)-(3.8). We next characterize the additional equations associated with the policy. We claim that

$$\dot{\tilde{B}}_{11}(t) = \mu \text{ when } g^{\epsilon}\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) > 0 \quad (4.203)$$
$$\dot{\tilde{B}}_{12}(t) = \mu \text{ when } g^{\epsilon}\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) < 0 \text{ and .} \quad (4.204)$$

To see this assume that

$$g^{\epsilon}\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) < -2\eta$$

for some $\eta > 0$. By continuity of g^{ϵ} and $\tilde{\mathbb{X}}$ there exists $\delta > 0$ such that

$$g^{\epsilon}\left(R\left(\tilde{Q}(s), \tilde{Z}(s)\right)\right) < -\eta$$

for all $s \in [t-\delta, t+\delta]$.

Pick r large enough together with an integer m and $w \in \mathcal{A}^r(T)$ so that

$$\|\tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t)\| < \epsilon/2.$$

This gives that

$$g^{\epsilon}\left(R\left(Q^{r,m}(s), Z^{r,m}(s)\right)\right) < -\eta/2,$$

since $\sqrt{\frac{x_{r,m}}{|N^r|}} = R$ on $\mathcal{A}^r(T)$. By (4.201)

$$B_{12}^{r,m}(t+\delta) - B_{12}^{r,m}(t-\delta) = 0,$$

and so

$$\dot{\tilde{B}}_{12}(t) = 0,$$

Now, by (3.5)

$$\dot{\tilde{Z}}_{12}(t) = -(1-\eta)\mu \text{ and}$$
$$\dot{\tilde{Z}}_{11}(t) = \dot{\tilde{B}}_{11}(t) - \mu\eta\mu.$$

Equations (3.5), (3.8) and (4.192) give that

$$\dot{\tilde{Z}}_{11}(t) + \dot{\tilde{Z}}_{11}(t) = 0.$$

Hence

$$\dot{\tilde{B}}_{11}(t) = \mu.$$

Condition (4.204) is proved similarly.

Next we prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}|g^\epsilon|\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)| < 0 \tag{4.205}$$

whenever $|g^\epsilon|\left(\hat{Q}^r(t), \hat{Z}^r(t)\right)| > 0$. We first note that (4.194) also holds for this case.

First assume that $g^\epsilon\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) < 0$. Then, by (4.203)

$$\dot{\tilde{Q}}_1(t) = \lambda_1 - \mu = -(1-\eta)t.$$

Hence,

$$\dot{g}^\epsilon\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) = R\dot{\tilde{Q}}_1(t) - \frac{\mathrm{d}}{\mathrm{d}t}(R\tilde{X}(t) - \theta)^+ = (1-\eta).$$

by (4.194).

Similarly, if $g_1\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) > 0$ then

$$\dot{g}^\epsilon\left(R\left(\tilde{Q}(t), \tilde{Z}(t)\right)\right) = \dot{\tilde{Q}}_1(t) = -\eta.$$

This proves (4.205). By (4.205) Assumption 7 holds and this completes the proof. □

CHAPTER V

CONCLUSIONS

In this thesis, we focused on the asymptotic analysis of parallel server systems under the Halfin-Whitt regime. Parallel server systems we studied are quite general and can be used to analyze systems with many servers that are especially seen in the service sector.

The main contribution of this study is two-fold; (i) we take the first step to establish a general theory to validate many-server approximations and establish a framework for showing SSC results that should facilitate the asymptotic analysis of multiclass parallel server systems and (ii) we prove new SSC results for three parallel server systems.

Before studying many-server diffusion limits of parallel server systems, we first formulated the many-server heavy traffic regime using a static planning problem (SPP) similar to that in Harrison [35]. Using this formulation we found the optimal allocation of servers' capacities to customer classes. We also provided a basic result based on fluid limits to check if allocations under a policy is in accordance with the solution of that SPP. We focused our attention to those policies whose allocations are optimal in this sense. Using these allocations, we also were able to define a general diffusion scaling under which meaningful limits can be obtained.

Once we set the background for more detailed diffusion analysis, we provided a necessary condition for a multiplicative SSC result to hold in the diffusion limits. This necessary condition is related to the hydrodynamic limits. Hydrodynamic limits have similar properties to the fluid limits but they are different. Most importantly, they are deterministic and almost everywhere differentiable. These properties enable one to check whether the necessary condition for a SSC holds in a straightforward manner. We also provided two extensions to our main result. These extensions involve relaxing the homogeneity condition on the space collapse function.

SSC results play an important role in establishing diffusion limits both under the conventional heavy traffic and the Halfin-Whitt many-server diffusion limits. We illustrate the application of our results in three different systems and prove several SSC results. To the best of our knowledge, these results are new and our framework made the asymptotic analysis of these systems possible.

We first focus on a distributed parallel server system operating under two different policies; MED–FSF and MED–LB policies. We proves that the MED-FSF policy is asymptotically optimal in the sense that it stochastically minimizes the stationary distribution of the number of customers in the system and the stationary probability that a customer gets delayed in the queue. However, all the servers in our distributed system except those with the lowest service rate experience 100% utilization. Under the MED-LB policy, on the other hand, the utilizations of all the server pools are equal. We also showed that under both policies the system performs as well as a corresponding single queue system.

The second system we consider is known as the N-system. These systems have been extensively studied in the literature under the conventional heavy traffic and served as a stepping stone for the analysis of more general parallel server systems. We showed that when the service times only depend on the server pool providing service a static priority rule is asymptotically optimal in an N-system. The optimality is in terms of stochastically minimizing linear holding costs during a finite time interval.

Finally, we study two results conjectured in the literature for V-systems. First, we prove a state space collapse result conjectured in Armony and Maglaras [3]. Then, we propose a policy whose asymptotic performance is arbitrarily close to the conjectured performance of the policy proposed by Milner and Olsen [53] and prove a state space collapse result under this policy.

The results presented in this thesis can be extended in several directions. The conditions for a SSC result we provided here are only sufficient. It is important to establish necessary and sufficient conditions for a SSC result. Another direction that will be explored in the future is to establish a deterministic relationship between waiting times and queue lengths. Distribution of waiting times are commonly used to measure the performance of service

systems. However, in many-server diffusion analysis, queue length processes are easier to analyze than waiting time processes. Also, we provided a necessary condition only for a multiplicative SSC result. Generally, a strong SSC result is needed in diffusion analysis. Sufficient conditions for a policy to satisfy the compact containment condition will also help facilitate the many-server diffusion analysis.

In our analysis of DPS systems, we assumed that all the server pools are large. A current trend in call center industry is to employ agents working from home. In one extreme case, a DPS system becomes totally distributed in the sense that as the number of servers go to infinity so does the number of queues that customers are routed to. In this case we believe that more sophisticated load balancing policies, which involve periodic load balancing, would be needed. The systems that fall in between these two extreme cases are also of practical importance.

We showed that a static priority policy is asymptotically optimal for N-systems that satisfy certain conditions. However, this analysis can be extended to other N-systems as long as service rates only depend on the server pool providing the service. Also, when service rates of two customer classes from the second server pool only differ by a "reasonable" amount, the "perturbation approach" in Maglaras and Zeevi [48] can be used to provide close approximations for the system performance.

Extending our optimality result to more general parallel server system is another important extension and is currently being pursued. Another interesting direction that could be explored is to establish necessary condition for the stability of a static priority policy when the number of servers is fixed. This would be important in applications as to see for what utilization levels a threshold value should be used.

APPENDIX A

FLUID LIMITS OF PARALLEL SERVER SYSTEMS

In this chapter we study the fluid limits and present the fluid model equations of parallel server systems. We establish a general framework that can be used to check if Assumption 2 is satisfied by a control policy.

Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes and

$$\bar{\mathbb{X}}_\pi^r(t) = \mathbb{X}_\pi^r(t)/N^r. \tag{A.1}$$

We call this scaling the fluid scaling and $\bar{\mathbb{X}}_\pi^r$ the fluid scaled process. $\bar{\mathbb{X}}_\pi$ is called a fluid limit of $\{X_\pi^r\}$ if there exists an $\omega \in \Omega$ and a sequence $\{r_l\}$ with $r_l \to \infty$ as $l \to \infty$, such that $\bar{\mathbb{X}}_\pi^{r_l}(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}_\pi$ as $l \to \infty$. The following theorem is analogous to Theorem 4.1 in Dai [18].

Theorem A.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes. Assume that (2.18) and (2.19) hold and $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Then, $\{\bar{\mathbb{X}}_\pi^r\}$ is a.s. pre-compact in the Skorohod space $\mathbb{D}^d[0,\infty)$ endowed with the u.o.c. topology. Thus, the fluid*

limits exist and each fluid limit, $\bar{\mathbb{X}}_\pi$, of $\left\{\bar{\mathbb{X}}_\pi^r\right\}$ satisfies the following equations for all $t \geq 0$.

$$\lambda_i t = \sum_{k\in\mathcal{K}} \bar{A}_{ik}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \bar{A}_{ijk}(t), \text{ for all } i \in \mathcal{I}, \tag{A.2}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \sum_{i\in\mathcal{I}} \bar{A}_{ik}(t) - \sum_{j\in\mathcal{J}(k)} \bar{B}_{jk}(t), \text{ for all } k \in \mathcal{K}, \tag{A.3}$$

$$\bar{Z}_{jk}(t) = \bar{Z}_{jk}(0) + \sum_{i\in\mathcal{I}} \bar{A}_{ijk}(t) + \bar{B}_{jk}(t) - \mu_{jk}\bar{T}_{jk}(t), \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j) \tag{A.4}$$

$$\bar{T}_{jk}(t) = \int_0^t \bar{Z}_{jk}(s)ds, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{A.5}$$

$$\bar{Q}_k(t)\left(\sum_{j\in\mathcal{J}(k)}\left(\beta_j - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}(t)\right)\right) = 0, \text{ for all } k \in \mathcal{K}, \tag{A.6}$$

$$\bar{Y}_j(t) = \beta_j t - \sum_{k\in\mathcal{K}(j)} \bar{T}_{jk}(t), \text{ for all } j \in \mathcal{J}, \tag{A.7}$$

$$\int_0^t \sum_{k\in\mathcal{K}(j)} \bar{Q}_j(s)d\bar{Y}_j(s) = 0, \text{ for all } j \in \mathcal{J}, \tag{A.8}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)} \left(\beta_j - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}(s)\right) d\bar{A}_{ik}(s) = 0, \text{ for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K} \tag{A.9}$$

$$\bar{A}, \bar{A}_q, \bar{A}_s, \bar{T}, \bar{Y}, \text{ and } \bar{B} \text{ are nondecreasing}, \tag{A.10}$$

$$\bar{Q}(t) \geq 0, \bar{Z}_{jk}(t) \geq 0, \text{ and } \sum_{k\in\mathcal{K}(j)} \bar{Z}_{jk}(t) \leq \beta_j, \text{ for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j). \tag{A.11}$$

Definition A.2. We call the vector (q, z) a steady state of the fluid limits if for any fluid limit $\bar{\mathbb{X}}_\pi$, $\bar{Q}(0) = q$ and $\bar{Z}(0) = z$ implies $\bar{Q}(t) = q$ and $\bar{Z}(t) = z$ for all $t > 0$.

We denote the set of all the steady states of the fluid limits of $\{\mathbb{X}_\pi^r\}$ by $\mathcal{M}_\pi$. The following result presents a condition that is equivalent to Assumption 2.

Lemma A.3. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π parallel server system processes that satisfies the conditions of Theorem A.1 and Assumption 1. A control policy π satisfies Assumption 2 if and only if $(0, z) \in \mathcal{M}_\pi$, where $z_{jk} = \beta_j x^*_{jk}$ and x^* is given as in Assumption 1.*

Proof. Assume that π satisfies Assumption 2 and that $(Q^r(0)/N^r, Z^r(0)/N^r) \to (0, z)$ a.s. as $r \to \infty$. By (2.22), (A.4), and (A.5)

$$\sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \sum_{i\in\mathcal{I}} \bar{A}_{ijk}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \bar{B}_{jk}(t) = \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \mu_{jk}\beta_j x^*_{jk} t. \tag{A.12}$$

By (A.2)

$$\sum_{i\in\mathcal{I}} \lambda_i t = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} \bar{A}_{ik}(t) + \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \bar{A}_{ijk}(t). \tag{A.13}$$

By Assumption 1, $\sum_{i\in\mathcal{I}}\lambda_i = \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}(k)}\mu_{jk}\beta_j x^*_{jk}t$. This combined with (A.12) and (A.13) implies that

$$\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}(k)}\bar{B}_{jk}(t) = \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}}\bar{A}_{ik}(t).$$

Hence, by (A.3)

$$\sum_{k\in\mathcal{K}}\bar{Q}_k(t) = \sum_{k\in\mathcal{K}}\bar{Q}_k(0) + \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{I}}\bar{A}_{ik}(t) - \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}(k)}\bar{B}_{jk}(t) = \sum_{k\in\mathcal{K}}\bar{Q}_k(0) = 0.$$

Thus, $(0,z)\in\mathcal{M}_\pi$.

Now assume that $(0,z)\in\mathcal{M}_\pi$ and that $(Q^r(0)/N^r, Z^r(0)/N^r)\to(0,z)$ a.s. as $r\to\infty$. Then, by (A.5), $T^*_{jk}(t) = \beta_j x^*_{jk}t$ for all $t\geq 0$.

□

Proof of Theorem A.1. Assume that (2.18) and (2.19) hold. Let $\mathcal{A}\subset\Omega$ be such that $\{\bar{Q}^r(0)\}$ is bounded and the following FSLLN holds,

$$\frac{E(N^r\cdot)}{N^r}\to\nu(\cdot) \text{ and } \frac{S_{jk}(N^r\cdot)}{N^r}\to\alpha_{jk}(\cdot) \text{ u.o.c.} \tag{A.14}$$

as $r\to\infty$, where $\alpha_{jk}(t) = \mu_{jk}t$, for all $j\in\mathcal{J}, k\in\mathcal{K}(j)$, and $\nu(t) = te$, where e is the I-dimensional row vector of ones. Note that we can take $P(\mathcal{A}) = 1$ from (2.13).

Consider a sequence of numbers which we again denote, with a slight abuse of notation, by $\{r\}$. We show that $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$ has a convergent subsequence, for all $\omega\in\mathcal{A}$. We fix $\omega\in\mathcal{A}$ for the rest of the proof.

Observe that

$$\left|\frac{T^r(t_2,\omega)}{N^r} - \frac{T^r(t_1,\omega)}{N^r}\right| \leq |t_2 - t_1|,$$

for all $0\leq t_1 < t_2$. Hence, $\left\{\bar{T}^r(\cdot,w)\right\}$ is tight; there exists a subsequence $\{r_l\}$ such that $\bar{T}^{r_l}(\cdot,\omega)$ converges u.o.c. to some continuous function $\bar{T}$. Since, $\bar{D}^{r_l}_{jk}(t) = \frac{1}{N^r}S_{jk}\left(N^r\bar{T}^{r_l}_{jk}(t)\right)$,

$$\bar{D}^{r_l}_{jk}(\cdot) \text{ converges u.o.c. to } \bar{D}_{jk}(\cdot), \tag{A.15}$$

where $\bar{D}_{jk}(t) = \mu_{jk}\bar{T}(t)$, by random time change theorem; see Theorem 5.3 in [15]. By (A.14), $\bar{A}^{r_l}(\cdot,\omega)$ converges u.o.c. to $\bar{A}(t) = \lambda t$, hence, it is pre-compact in $D^I[0,\infty)$. But,

for all $0 \leq t_1 < t_2$ and $j \in \mathcal{J}$,

$$\bar{A}^{r_l}_{ik}(t_2,\omega) - \bar{A}^{r_l}_{ik}(t_1,\omega) \leq \bar{A}^{r_l}(t_2,\omega) - \bar{A}^{r_l}(t_1,\omega).$$

By Theorem 12.3 of Billingsley [11], this implies that $\{\bar{A}^{r_l}_{ik}(\cdot,\omega)\}$ is also pre-compact. By the same argument, so is $\{\bar{A}^{r_l}_{ijk}(\cdot,\omega)\}$, for all $i \in \mathcal{I}$, $k \in \mathcal{K}$, and $j \in \mathcal{J}(k)$.

Without loss of generality we can assume that

$$(\bar{A}^{r_l}(\cdot,\omega), \bar{A}^{r_l}_s(\cdot,\omega), \bar{A}^{r_l}_q(\cdot,\omega), \bar{D}^{r_l}(\cdot,\omega), \bar{T}^{r_l}(\cdot,\omega), \bar{Y}^{r_l}(\cdot,\omega)) \tag{A.16}$$

converges u.o.c. as $l \to \infty$. We next show that $\bar{B}^{r_l}$ is pre-compact. Fix $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$, and $0 \leq t_1 \leq t_2$. We omit ω from the notation below. By (2.5), we have that

$$\bar{B}^{r_l}_{jk}(t_2) - \bar{B}^{r_l}_{jk}(t_1) \leq \sum_{k' \in \mathcal{K}(j)} \left(\bar{D}^{r_l}_{jk'}(t_2) - \bar{D}^{r_l}_{jk'}(t_1) \right). \tag{A.17}$$

Since $\bar{B}^{r_l}_{jk}$ is nondecreasing, again by Theorem 12.3 of Billingsley [11], (A.17) implies that $\left\{ \bar{B}^{r_l}_{jk}(\cdot,\omega) \right\}$ is pre-compact. This implies by (2.5) and (A.16) that $\left\{ \bar{Z}^{r_l}_{jk}(\cdot,\omega) \right\}$ is pre-compact. Finally, combining these results with (2.4), we have that $\left\{ \bar{Q}^{r_l}_k(\cdot,\omega) \right\}$ is pre-compact. Thus, $\left\{ \bar{\mathbb{X}}^r \right\}$ has a convergent subsequence.

Next, we show that every fluid limit satisfies (A.2)-(A.9). Let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$, for some $\omega \in \mathcal{A}$, converges u.o.c. to $\bar{\mathbb{X}}$. $\bar{\mathbb{X}}$ satisfies (A.2) by (2.3), the convergence of $\bar{A}^r_i(\cdot,\omega)$ to $\bar{A}_i(t) = \lambda t$ and the fact that $\bar{A}^r_{ik}(\cdot,\omega)$ and $\bar{A}^r_{ijk}(\cdot,\omega)$ are both convergent. Equation (A.3) is satisfied by $\bar{\mathbb{X}}$ by (2.4) and the convergence of $\bar{Q}^r(0,\omega)$, $\bar{A}^r_{kj}(\cdot,\omega)$, and $\bar{B}^r_{jk}(\cdot,\omega)$. Equation (A.4) follows from (A.15), the convergence of $\bar{Z}^r_{jk}(0,\omega)$, $\bar{A}^r_{ijk}(\cdot,\omega)$, and $\bar{B}^r_{jk}(\cdot,\omega)$. Equation (A.5) follows from (2.8) and the convergence of $\bar{Z}^r_{jk}(\cdot,\omega)$ to $\bar{Z}_{jk}(\cdot)$ u.o.c.

We next show that $\bar{\mathbb{X}}$ satisfies (A.6). Fix $t > 0$. If $\bar{Q}_k(t) = 0$, then (A.6) is satisfied trivially, so assume that $\bar{Q}_k(t) > 0$. By the continuity of $\bar{Q}_k$, there exist $t > \delta > 0$ and $\varepsilon > 0$ such that $\bar{Q}_k(s) > \varepsilon$ for all $s \in [t-\delta, t+\delta]$. Since $\bar{Q}^r_k$ converges u.o.c. to $\bar{Q}_k$, for large enough r

$$\bar{Q}^r_k(s,\omega) > \varepsilon/4 \quad \text{for all } s \in [t-\delta, t+\delta].$$

By (2.10), this gives

$$\sum_{j\in\mathcal{J}(k)}\left(\frac{N_j^r}{N^r}-\sum_{l\in\mathcal{K}(j)}\bar{Z}_{jl}^r(s,\omega)\right)=0 \ \text{ for all } s\in[t-\delta,t+\delta].$$

Using the fact that $\bar{\mathbb{X}}^r$ converges u.o.c. to $\bar{\mathbb{X}}$ again, we have that

$$\sum_{j\in\mathcal{J}(k)}\left(\beta_j-\sum_{l\in\mathcal{K}(j)}\bar{Z}_{jl}(s)\right)=0 \ \text{ for all } s\in[t-\delta,t+\delta],$$

thus proving (A.6). It can be shown similarly that $\bar{\mathbb{X}}$ satisfies (A.8) and (A.9). □

Remark A.4. It follows from (A.2)-(A.5) and the proof of Theorem A.1 that each component of $\bar{\mathbb{X}}$ is Lipschitz continuous and so they are absolutely continuous and differentiable almost everywhere with respect to the Lebesgue measure on $[0,\infty)$.

APPENDIX B

EQUIVALENCE OF THE ORIGINAL AND PERTURBED SYSTEMS

The key to the equivalence between the original and the perturbed systems is the assumption of exponential service times. Without the memoryless property, the residual service times would upset the equivalence we show in this appendix.

For notational simplicity we focus on systems with no routing. We assume that a type i customer will be automatically routed to queue k at the time of his arrival. Therefore $I = K$ and we omit the subscript i from the notation throughout this section. For the rest of this section we fix a policy π. Recall that each policy is associated with a transaction function f_π. For simplicity, we assume that f_π is a deterministic function, but it can be taken as a random variable that is independent of $\{\mathbb{X}(t) : t \geq 0\}$.

The following result will be needed in the following sections.

Lemma B.1. *Let $(\Omega, P, \mathcal{F})$ be a probability space and $A, B, \mathcal{N} \in \mathcal{F}$ with $P(\mathcal{N}) = 0$. Then*

$$P\left(A|B \cup \mathcal{N}\right) = P\left(A|B\right),$$

where

$$P\left(A|B\right) P\left(B\right) = P\left(A \cap B\right).$$

Proof is elementary, hence omitted.

B.1 Piecewise-deterministic Markov processes

To prove the equivalence of the original system with the perturbed system we model both systems as piecewise-deterministic Markov processes (PDP's) that are introduced by Davis [23]. Below, we give a brief overview of PDP's. A thorough treatment of the subject can be found in Davis [23].

For our purposes, it is enough to define a PDP on a state space $E \subset R^p$ partitioned into a boundary E_δ and interior E_o. We let $\mathcal{E}$ denote the Borel subsets of E, and we will let $P(E)$ be the space of probability measures on the measurable space $(E, \mathcal{E})$, endowed with the topology of weak convergence. Under suitable regularity conditions a PDP can be uniquely determined by a vector field $h : E \to \mathbb{R}^p$, an intensity function $\lambda : E \to \mathbb{R}_+$, and a transition measure $\varpi : E \to P(E)$.

Between jumps a PDP $x(t)$ obeys $dx(t)/dt = h(x(t))$, and jumps occur at rate $\lambda(x)$ when the process is at state x, independently of the process history. If a jump occurs at $x \in E_o$ or the process reaches the boundary at $x \in E_\delta$, the process is transferred immediately to a new state given randomly by probability measure $\varpi(dx|x)$. We assume that $\varpi(E_o|x) = 1$. We use σ_n to denote the the nth jump time of the PDP process $x(t)$.

Let $N(t) = \sum_{i=1}^{\infty} I_{\sigma_i \leq t}$. Under the assumption

$$E[N(t)] < \infty \text{ for all } t$$

it can be shown that $\{x(t)\}$ is a strong Markov process, see Davis [23]

B.2 Construction of parallel server systems

In this section we construct the processes associated with a parallel server system. Recall that arrivals to class k are given by a renewal process E_k and we assume that the interarrival times of E_k are given by the i.i.d. sequence $\{u_k(m) : m = 1, 2, \ldots\}$ for each $k \in \mathcal{K}$. We also assume that $u_k(1) > 0$ w.p. 1.

We also assume that a sequence of i.i.d. exponential random variables $\{v'_k(m) : m = 1, 2, \ldots\}$ gives the service times, where $v'_k(m)$ is the service request for the mth class k customer. We assume that customers already present in the system at time 0 are ordered in some manner, so if there are $Z'_k(0)$ class k customers in the system at time 0, $(v'_k(1), \ldots, v'_k(Z'_k(0)))$ gives the service times of these customers. The actual service time of a customer naturally depends on the server pool handling his request. The service time of the mth customer is given by $v'_k(m)/\mu_{jk}$ if he is handled by a server in pool j. We assume that the service and interarrival times are independent.

Let $X'(t) = (Q'(t), Z'(t), b'(t))$ denote the state of the system at time t, where Q' and Z' have the same interpretations as before and $b'(t) = (b'_1(t), \ldots, b'_K(t))$ with $b'_k(t)$ is the remaining time before the next class k customer will arrive from outside at time t. We assume for simplicity that $b_k(0) > 0$ for $k \neq k'$. Recall that, we use $\{\sigma_n\}$ to denote the increasing sequence of event times, either an arrival to or departure from the system.

Let $\tilde{v}_{jk\ell}(t)$ denote the remaining service time of the the class k customer being served by the ℓth server among the servers who serve class k customers in server pool j at time t for $\ell \leq Z'_{jk}(t)$. For notational simplicity we set

$$\begin{aligned} \mathcal{V}(z) &= \{(j,k,\ell) : k \in \mathcal{K}, j \in \mathcal{J}(k), \ell \in \{1, \ldots, z_{jk}\}\}, \\ \mathcal{V}_{k'}(z) &= \{(j,k',\ell) : j \in \mathcal{J}(k), \ell \in \{1, \ldots, z_{jk}\}\}, \end{aligned}$$

where $z \in \mathbb{N}^{J \times K}$ so that if $(j,k,l) \in \mathcal{V}(z)$ then $\tilde{v}_{jk\ell}(t)$ is a well defined random variable. Recall that $s_n = \sigma_n - \sigma_{n-1}$, where σ_n is defined in Chapter 2 and observer that

$$s_n = \min\{b'(\sigma_n), \tilde{v}_{jk\ell}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(Z'(\sigma_n))\}.$$

We note that how the servers are indexed in a pool does not matter. However, once a customer begins service with a particular server, he stays in service with the same server until his service is completed. In particular, that server cannot switch to serving any new arrival.

The following lemma uses our assumption that service times are independent and exponentially distributed to show a parallel server system satisfies a Markov property.

Lemma 2. *Let $x = (q, z, b)$, with $q \in \mathbb{N}^K$, $z \in \mathbb{N}^{J \times K}$, and $b \in \mathbb{R}_+^K$. The following statements are true for each n.*

(i) Only one of the following events may happen at time σ_n; only one service completion or arrival to one or more customer classes.

(ii) Let $t_{jk\ell} \geq 0$ and $\tilde{\tilde{v}}_{jk\ell}$ for $(j,k,\ell) \in \mathcal{V}(z)$ be an exponential r.v. with rate μ_{jk} that are

independent from $\{X(t) : t \geq 0\}$. *Then*

$$P\left\{\tilde{v}_{jk\ell}(\sigma_n) > t_{jk\ell}; \forall (j,k,\ell) \in V(z) | X'(\sigma_n) = x\right\} = \prod_{(j,k,\ell)\in V(z)} P\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\}, \quad \text{(B.1)}$$

for any nonempty $V(z) \subset \mathcal{V}(z)$.

Proof. We prove the result by induction. For $n = 0$, (i) holds since we assume that $b_k(0) > 0$ for all $k \in \mathcal{K}$. The second result (ii) holds since at time zero all the service times are assumed to have exponential distribution. Assume that (i) and (ii) hold for n.

Next we prove that (i) holds for $n+1$. By our induction argument the event(s) at time σ_n is either a departure or one or more arrival events. Also, at time σ_n all the remaining service times have exponential distribution by our induction argument. Therefore, the remaining service times have a continuous distribution, and so the probability that either more than one departure or a simultaneous departure and arrival occur at time σ_{n+1} is 0. Hence, (i) holds for $n + 1$.

Now, let $x(m) = (q(m), z(m), b(m))$, with $q(m) \in \mathbb{N}^K$, $z(m) \in \mathbb{N}^{J\times K}$, and $b(m) \in \mathbb{R}_+^K$, for $m = 1, 2, \ldots$. For simplicity we take $V(z) = \mathcal{V}(z)$. Then

$$\begin{aligned}
&P\left\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_{n+1}) = x(n+1)\right\} \\
&= \int_{\mathbb{N}^K\times\mathbb{N}^{J\times K}\times\mathbb{R}_+^K} P\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_m) = x(m); m = n, n+1\} \\
&\qquad\qquad dP\left\{X'(\sigma_n) = x(n) | X'(\sigma_{n+1}) = x(n+1)\right\}.
\end{aligned}$$

Recall that we assume f_π is a deterministic function. Assume that the event $\{X'(\sigma_n) = x(\sigma_n), X'(\sigma_{n+1}) = x(\sigma_{n+1})\}$ is given so we know $Z_{jk}(t) + Q_k(t)$ at times σ_n and σ_{n+1}. Therefore, we can determine w.p. 1 if an arrival to class k has happened or a class k customer's service is completed at time σ_{n+1}.

Since at time σ_{n+1}, there is either a service completion or arrival events; either

$$q_k(n+1) + \sum_{j\in\mathcal{K}} z_{jk}(n+1) = q_k(n) + \sum_{j\in\mathcal{K}} z_{jk}(n) - 1 \quad \text{(B.2)}$$

for some $k \in \mathcal{K}$ or

$$q_k(n+1) + \sum_{j\in\mathcal{K}} z_{jk}(n+1) = q_k(n) + \sum_{j\in\mathcal{K}} z_{jk}(n) + 1 \quad \text{(B.3)}$$

for $k \in \tilde{\mathcal{K}} \subset \mathcal{K}$.

Assume that (B.3) holds for some $\tilde{\mathcal{K}} \subset \mathcal{K}$. Then,

$$\begin{aligned}
&\{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\} \\
&\quad \equiv \{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\} \cup \mathcal{N} \\
&\quad \equiv \{X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\} \cup \mathcal{N}, \qquad \text{(B.4)}
\end{aligned}$$

for some P-null set $\mathcal{N}$. Note that from the discussion above there cannot be a service completion event at the same instant. Then, by (B.4) and Lemma B.1, when $x(n)$ and $x(n+1)$ satisfy (B.3)

$$\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\big\} = \\
&\quad P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \\
&\qquad\qquad |X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\big\} = \\
&\quad P\big\{\tilde{v}_{jk\ell}(\sigma_n) > t_{jk\ell} + s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n)) \cap \mathcal{V}(z(n+1)) \\
&\qquad\qquad |X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\big\} \\
&\quad P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \setminus \mathcal{V}(z(n)) \\
&\qquad\qquad |X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\big\}
\end{aligned}$$

Above, $\mathcal{V}(z(n)) \cap \mathcal{V}(z(n+1))$ gives the set of customers whose service have started on or before σ_n and $\mathcal{V}(z(n+1)) \setminus \mathcal{V}(z(n))$ gives the set of customers whose service have started on σ_{n+1}, since we assume that an admissible policy is non-preemptive. The second equality follows from that fact that if a class k customer's service starts at time σ_{n+1} at service pool j it has exponential distribution with rate μ_{jk}, independent from other service times. Now,

$$\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_n) > t_{jk\ell} + s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n)) \cap \mathcal{V}(z_{n+1}) \\
&\qquad\qquad |X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\big\} \\
&= \prod_{(j,k,\ell) \in \mathcal{V}(z(n)) \cap \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\}
\end{aligned}$$

by induction hypothesis. And

$$P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \setminus \mathcal{V}(z(n)) \\ |X'(\sigma_n) = x(n), \tilde{v}_{jk\ell}(\sigma_n) > s_n; \forall (j,k,\ell) \in \mathcal{V}(z(n))\big\}$$

$$= \prod_{(j,k,\ell)\in\mathcal{V}(z(n+1))\setminus\mathcal{V}(z(n))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\} \tag{B.5}$$

because of the independence of the service times. Hence, we get the desired result in this case.

Now assume that (B.2) holds for k'. When (B.2) holds there are two possible cases that must be analyzed separately:

(a) $z_{jk'}(n+1) = z_{jk'}(n)$ for all $j \in \mathcal{J}(k')$ and $q_{k'}(n+1) = q_{k'}(n) - 1$. In this case one of the server pools have completed serving a class k' customer, but we cannot extract from (B.2) which one it is.

(b) $z_{j'k'}(n+1) = z_{j'k'}(n)$ for some $j' \in \mathcal{J}(k')$ and $q_{k'}(n+1) = q_{k'}(n) - 1$. In this case a server in pool j' has completed serving a class k' customer.

Below, we prove the result for (a), the proof of in case (b) is similar. We define

$$\tau_{k'}(\sigma_n) = \min\{\tilde{v}_{jk'\ell}(\sigma_n); \{jk'\ell\} \in \mathcal{V}_{k'}(z(n))\}. \tag{B.6}$$

Then, if (a) holds

$$\begin{aligned}
&\{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\} \\
&\quad \equiv \{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1), \tau_{k'}(\sigma_n) = s_n\} \cup \mathcal{N} \\
&\quad \equiv \{X'(\sigma_n) = x(n), \tau_{k'}(\sigma_n) = s_n\} \cup \mathcal{N}
\end{aligned} \tag{B.7}$$

for some P-null set $\mathcal{N}$. Therefore, in this case

$$\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\big\} = \\
&P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_n) = x(n), \tau_{k'}(\sigma_n) = s_n\big\} \\
&= \sum_{(j'k'\ell')\in\mathcal{V}_{k'}(z(n))} P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \\
&\qquad |\tilde{v}_{jk\ell}(\sigma_n) > \tilde{v}_{j'k'\ell'}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(z(n)) \setminus \{j',k',l'\}, X'(\sigma_n) = x(n)\big\} \\
&\qquad P\big\{\tilde{v}_{j'k'\ell'}(\sigma_n) = s_n\big\}
\end{aligned} \tag{B.8}$$

In the above display, the first equality follows from (B.7) and the second equality follows from the induction argument and from the independence of service times. Now, for any $\{j', k', \ell'\} \in \mathcal{V}_{k'}(z(n))$

$$
\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \\
&\qquad\qquad |\tilde{v}_{jk\ell}(\sigma_n) > \tilde{v}_{j'k'\ell'}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(z(n)) \setminus \{j', k', l'\}, X'(\sigma_n) = x(n)\big\} \\
&= P\big\{\tilde{v}_{jk\ell}(\sigma_n) > t_{jk\ell} + s_n; \forall (j,k,\ell) \in \big(\mathcal{V}(z(n)) \setminus \{j', k', \ell'\}\big) \cap \mathcal{V}(z(n+1)) \\
&\qquad\qquad |\tilde{v}_{jk\ell}(\sigma_n) > \tilde{v}_{j'k'\ell'}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(z(n)) \setminus \{j', k', l'\}, X'(\sigma_n) = x(n)\big\} \\
&\quad \times P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \setminus \big(\mathcal{V}(z(n)) \setminus \{j', k', \ell'\}\big)\big\}. \qquad \text{(B.9)}
\end{aligned}
$$

In the above display $(\mathcal{V}(z(n)) \setminus \{j', k', \ell'\}) \cap \mathcal{V}(z(n+1))$ gives the set of customers whose service has started on or before σ_n not including $\{j', k', \ell'\}$, since an admissible policy is non-preemptive. The set $\mathcal{V}(z(n+1)) \setminus \mathcal{V}(z(n)) \setminus \{j', k', \ell'\}$ is the set of customers whose services have started on σ_{n+1}. The equality follows from the independence of service times.

For all $(j,k,\ell) \in \mathcal{V}(z(n+1)) \setminus (\mathcal{V}(z(n)) \setminus \{j', k', \ell'\})$ the remaining service times at σ_{n+1} have exponential distribution since their service starts at time σ_{n+1}. For those who started service before or on σ_n, we have by the induction argument and the memoryless property of exponential distribution

$$
\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_n) > t_{jk\ell} + s_n; \forall (j,k,\ell) \in \big(\mathcal{V}(z(n)) \setminus \{j', k', \ell'\}\big) \cap \mathcal{V}(z(n+1)) \\
&\qquad |\tilde{v}_{jk\ell}(\sigma_n) > \tilde{v}_{j'k'\ell'}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(z(n)) \setminus \{j', k', l'\}, X'(\sigma_n) = x(n)\big\} \\
&= \prod_{(j,k,\ell) \in (\mathcal{V}(z(n)) \setminus \{j', k', \ell'\}) \cap \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\}. \qquad \text{(B.10)}
\end{aligned}
$$

Hence, when the event at σ_{n+1} is a class k' service completion, by (B.8)-(B.10)

$$
\begin{aligned}
&P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) | X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\big\} = \\
&\quad \sum_{(j'k'\ell') \in \mathcal{V}_{k'}(z(n))} P\big\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(z(n+1)) \\
&\qquad\qquad |\tilde{v}_{jk\ell}(\sigma_n) > \tilde{v}_{j'k'\ell'}(\sigma_n); \forall (j,k,\ell) \in \mathcal{V}(z(n)) \cap (j'k'\ell'), X'(\sigma_n) = x(n)\big\} \\
&\qquad\qquad\qquad P\big\{\tilde{v}_{j'k'\ell'}(\sigma_n) = s_n\big\} \\
&= \sum_{(j'k'\ell') \in \mathcal{V}_{k'}(z(n))} \left(\prod_{(j,k,\ell) \in \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\} \right) P\big\{\tilde{v}_{j'k'\ell'}(\sigma_n) = s_n\big\} \\
&= \prod_{(j,k,\ell) \in \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\} \qquad \text{(B.11)}
\end{aligned}
$$

Finally, by (B.5) and (B.11), we have that

$$
\begin{aligned}
&P\left\{\tilde{v}_{jk\ell}(\sigma_{n+1}) > t_{jk\ell}; \forall (j,k,\ell) \in \mathcal{V}(\sigma_{n+1})) | X'(\sigma_{n+1}) = x(n)\right\} \\
&= \prod_{(j,k,\ell) \in \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk\ell} > t_{jk\ell}\big\} \\
&\quad \int_{\mathbb{N}^K \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K} P\left\{X'(\sigma_n) = x(n) | X'(\sigma_{n+1}) = x(n+1)\right\} dF_{X'(\sigma_n)}(x(n)),
\end{aligned}
$$

where each $\tilde{\tilde{v}}_{jk\ell}$ is an exponential random variable with rate μ_{jk} and is independent of $\mathcal{V}(z(n+1))$ which gives the desired result since

$$
\int_{\mathbb{N}^K \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K} dP\left\{X'(\sigma_n) = x(n) | X'(\sigma_{n+1}) = x(n+1)\right\} = 1.
$$

□

Using Lemma 2 we can show that $X'(t) = (Q'(t), Z'(t), b'(t))$ is a PDP. First note that

$$
\frac{db'_k(t)}{dt} = -1, \text{ for } t \in (\sigma_i, \sigma_{i+1}), k = 1, 2, \ldots, K, \qquad \text{(B.12)}
$$

$$
\frac{dq'_k(t)}{dt} = \frac{dz'_{jk}(t)}{dt} = 0, \text{ for } k = 1, 2, \ldots, K, \text{ and } j \in \mathcal{J}(k). \qquad \text{(B.13)}
$$

$$
\qquad \text{(B.14)}
$$

Assume that $X(\sigma_n) = x(n)$. Let $t_n^* = \min\{b'_k(n) : k = 1, 2, \ldots, K\}$ and

$$
G_{x(n)}(t) = \begin{cases} \exp\{-\sum_{k\in\mathcal{K}, j\in\mathcal{J}(k)} z_{jk}\mu_{jk}t\}, & \text{if } t < t_n^* \\ 0, & t \geq t_n^*. \end{cases} \qquad \text{(B.15)}
$$

By Lemma 2, since for $t > 0$ $P\left\{\tilde{\tilde{v}}_{jkl} > t\right\} = e^{-\mu_{jk}t}$ for z_{jk} denoting the number of class k customers being served by a server in pool j at time σ_n, $P(s_n > t|X(\sigma_n) = x(n)) = G_{x(n)}(t)$. Hence, the jump rate $\lambda(\cdot)$ is given by

$$\lambda(x(n)) = \sum_{k\in\mathcal{K}, j\in\mathcal{J}(k)} z_{jk}\mu_{jk}.$$

Let $q(m) \in \mathbb{N}^K$, $z(m) \in \mathbb{N}^{J\times K}$, and $b(m) \in \mathbb{R}_+^K$ for $m = 1, 2$. Define a probability measure ϖ on $(\mathbb{N}^K \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K) \times (\mathbb{N}^K \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K)$ as follows:

$$\varpi((q(2), z(2), b(2)); (q(1), z(1), b(1))) = F_k(dB) \tag{B.16}$$

if $b_k(1) = 0, b_k(2) = B$ for $B \in \mathbb{R}$ and $(q(2), z(2)) = f_\pi(q(1), z(1), e_k, E_0)$, where F_k is the law of the interarrival times of class k customers and e_k and E_0 are defined as in Chapter 2. The distribution given in (B.16) specifies the behavior of the system when there is an arrival to class k. Let

$$\varpi((q(2), z(2), b(2)); (q(1), z(1), b(1))) = \frac{z_1(jk)\mu_{jk}}{\sum_{k\in\mathcal{K}, j\in\mathcal{J}(k)} z_1(jk)\mu_{jk}} \tag{B.17}$$

if $b_1(k) > 0, b_2(k) = b_1(k)$ for all $k = 1, 2, \ldots, K$ and $(q(2), z(2)) = f_\pi(q(1), z(1), e_0, E_{jk})$. The right hand side of (B.17) gives the probability that the service of a class k customer is completed in server pool j given that there is a service completion in the system. From Lemma 2, this characterization is true since each remaining service time at any instant has exponential distribution and is independent from the other service times and from the remaining time for an arrival.

It is clear that $X'(t) = (Q'(t), Z'(t), b'(t))$ is a PDP with intensity function λ, transition measure ϖ, and evolution equation (B.12). Also, since $\mathbb{E}[A_k(t)] < \infty$ for all $t \geq 0$ and remaining service times at event times have exponential distribution, $\mathbb{E}[N(t)] < \infty$ for all $t \geq 0$, therefore, a parallel service system is a regular strong PDP.

B.3 Construction of perturbed systems

Fix an admissible policy π. Next we build another process that has the same intensity function, evolution equation and transition measure as the parallel server system described

in the previous section. Then, we show that this new process has the same departure process pathwise as our perturbed system.

Let $\{v_{jk}(l) : l = 1, 2, \ldots\}$ be a sequence of independent and identically distributed exponential random variables for $k \in \mathcal{K}$ and $j \in \mathcal{J}(k)$. Also assume that these sequences are mutually independent and each one is independent from the sequence of interarrival times for each customer class. By the description of our perturbed system in Section 2.1, the service completion time of the mth class k job in server pool j is given by

$$d_m = \inf \left\{ t : \int_0^t Z_{jk}(s)ds = V_{jk}(m) \right\} \tag{B.18}$$

where

$$V_{jk}(m) = \sum_{l=1}^{m} v_{jk}(l)$$

We next show that when departures are modeled as above, $X(t) = (Q(t), Z(t), b(t))$ is a piecewise-deterministic Markov process with the same parameters as $X'(t) = (Q'(t), Z'(t), b'(t))$. Let

$$\tilde{v}_{jk}(\sigma_n) = V_{jk}(D_{jk}(\sigma_n) + 1) - \sigma_n,$$

where $D_{jk}(t)$ is the number of class k service completions in server pool j by time t; see Chapter 2. Therefore, $\tilde{v}_{jk}(\sigma_n)$ is the remaining interarrival time for the $(D_{jk}(\sigma_n) + 1)$st arrival of the process S_{jk} at time σ_n. For notational simplicity, we set

$$\mathcal{Y} = \{(j, k) : k \in \mathcal{K}, j \in \mathcal{J}(k)\} \text{ and}$$

$$\mathcal{Y}_k = \{(j, k) : j \in \mathcal{J}(k)\}$$

Lemma 3. *Let $x = (q, z, b)$, with $q \in \mathbb{N}^K$, $z \in \mathbb{N}^{J \times K}$, and $b \in \mathbb{R}_+^K$. The following statements are true for each n.*

(i) Only one of the following events may happen at time σ_n; only one service completion or arrival to one or more customer classes.

(ii) For any non-empty $\tilde{\mathcal{Y}} \subset \mathcal{Y}$

$$P\left\{ \tilde{v}_{jk}(\sigma_n) > t_{jk}; \forall (j, k) \in \tilde{\mathcal{Y}} | X(\sigma_n) = x \right\} = \prod_{(j,k) \in \tilde{\mathcal{Y}}} P\{\tilde{\tilde{v}}_{jk} > t_{jk}\}, \tag{B.19}$$

for $t_{jk} > 0$, where $\tilde{\tilde{v}}_{jk}$ is an exponential r.v. with rate μ_{jk}. In addition,

$$P(s_n > t | X(\sigma_n) = x(n)) = G_{x(n)}(t), \tag{B.20}$$

where $G_{x(n)}(t)$ is given by (B.15).

Proof. The proof is similar to that of Lemma 2 by using an induction argument. We set $\tilde{\mathcal{Y}} = \mathcal{Y}$ for simplicity.

For $n = 0$, the result holds trivially from our assumptions. Assume that the results hold for n. The proof of (i) for $n+1$ is similar to that in the proof of Lemma 2, hence omitted here.

Assume that statements in (i) and (ii) are true for n and let $x(m) = (q(m), z(m), b(m))$, with $q(m) \in \mathbb{N}^K$, $z(m) \in \mathbb{N}^{J \times K}$, and $b(m) \in \mathbb{R}_+^K$, for $m = 1, 2, \ldots$. Then

$$\begin{aligned}
&P\left\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y} | X(\sigma_{n+1}) = x(n+1)\right\} \\
&= \int_{\mathbb{N}^K \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K} P\left\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y} | X(\sigma_m) = x(m); m = n, n+1\right\} \\
&\qquad\qquad dP\left\{X(\sigma_n) = x(n) | X(\sigma_{n+1}) = x(n+1)\right\}.
\end{aligned}$$

Since f_π is a deterministic function and only one event can happen at time σ_n, the event $\{X(\sigma_n) = x(\sigma_n), X(\sigma_{n+1}) = x(\sigma_{n+1})\}$ yields which event happened at time σ_{n+1}. As in the proof of Lemma 2, one can analyze an arrival event and a departure event similarly. The analysis for the arrival event is similar to that in the proof of Lemma 2. Hence, we focus on a departure event.

Assume that (B.2) holds for k'. As in the proof of Lemma 2, one should analyze two possible cases separately. Again we only focus on case (a). Define

$$\tau_{k'}(\sigma_n) = \min\{\tilde{v}_{jk'}(\sigma_n)/z_{jk'}(n); (j,k') \in \mathcal{Y}\} \tag{B.21}$$

to be the remaining time for the first service completion after time σ_n, where we take $x/0 = \infty$ for a real number x. Then,

$$\begin{aligned}
&\{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1)\} \\
&\qquad \equiv \{X'(\sigma_n) = x(n), X'(\sigma_{n+1}) = x(n+1), \tau_{k'}(\sigma_n) = s_n\} \cup \mathcal{N} \\
&\qquad \equiv \{X'(\sigma_n) = x(n), \tau_{k'}(\sigma_n) = s_n\} \cup \mathcal{N}
\end{aligned} \tag{B.22}$$

for some P-null set $\mathcal{N}$. Therefore, by Lemma B.1, when $x(n)$ and $x(n+1)$ satisfy (B.2)

$$P\big\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y} | X(\sigma_n) = x(n), X(\sigma_{n+1}) = x(n+1)\big\} =$$

$$P\big\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y} | X(\sigma_n) = x(n), \tau_{k'}(\sigma_n) = s_n\big\} =$$

$$\sum_{(j',k') \in \mathcal{Y}_{k'}, z_{j'k'}(n) > 0} P\big\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y}$$

$$|X(\sigma_n) = x(n), \frac{\tilde{v}_{j'k'}(\sigma_n)}{z_{j'k'}(n)} = s_n; (j,k) \in \mathcal{Y}\big\}$$

$$P\left\{\frac{\tilde{v}_{j'k'}(\sigma_n)}{z_{j'k'}(n)} = s_n\right\}.$$

Now, for $j' \in \mathcal{Y}_{k'}$ with $z_{j'k'}(n) > 0$

$$P\big\{\tilde{v}_{jk}(\sigma_{n+1}) > t_{jk}; \forall (j,k) \in \mathcal{Y}$$

$$|X(\sigma_n) = x(n), \frac{\tilde{v}_{j'k'}(\sigma_n)}{z_{j'k'}(n)} = s_n; (j,k) \in \mathcal{Y}\big\}$$

$$= P\big\{\tilde{v}_{jk}(\sigma_n) > t_{jk} + z_{jk}(n)s_n; \forall (j,k) \in \mathcal{Y} \setminus (j',k')$$

$$|X(\sigma_n) = x(n), \tilde{v}_{jk}(\sigma_n) > z_{jk}(n)s_n; (j,k) \in \mathcal{Y} \setminus (j',k')\big\} P\big\{\tilde{v}_{j'k'}(\sigma_{n+1}) > t_{j'k'}\big\}$$

$$= \prod_{(j,k) \in \mathcal{V}(z(n+1))} P\big\{\tilde{\tilde{v}}_{jk} > t_{jk}\big\},$$

where the last inequality follows from the induction argument, independence of the interarrival times of $S_{j'k'}$ and memoryless property of exponential distribution. This gives (B.19). Equation (B.20) follows from (B.19) since, given $X(\sigma_{n+1}) = x(n+1)$

$$s_{n+1} = \min\{b_k, \tilde{v}_{jk}(\sigma_{n+1})/z_{jk}\}.$$

Therefore, for $t < b_k$ for all $k \in \mathcal{K}$

$$P\left\{s_{n+1} > t | X(\sigma_{n+1}) = x(n+1)\right\}$$

$$= P\left\{\tilde{v}_{jk}(\sigma_{n+1})/z_{jk} > t, (j,k) \in \mathcal{Y} | X(\sigma_{n+1}) = x(n+1)\right\}.$$

□

Proof of Theorem 2.1. Using Lemma 3 one can show that the system defined above has the same PDP characterization as the original system, that is; they have the same transition measure, intensity function and evolution equation. Also, by a similar argument to that

at the end of the last section, $\mathbb{E}[N(t)] < \infty$ for all $t \geq 0$ for the system constructed above too. Hence, by Theorem 5.5 in Davis [23], both Markov processes have the same generator. Therefore, they have the same finite-dimensional distribution by Proposition 1.6 in Chapter 4 of Ethier and Kurtz [24].

Now let S_{jk} be Poisson process with interarrival times given by the sequence $\{v_{jk}(l) : l = 1, 2, \ldots\}$. Note that the mth service completion time from $S_{jk}\left(\int_0^t Z_{jk}(s)ds\right)$ is also given by (B.18). Hence, the departure processes for the system generated above are the same pathwise as the Poisson process characterization in the perturbed system.

□

APPENDIX C

PROOFS OF THE RESULTS IN SECTION 4.1

C.1 Proofs of the results in Section 4.1.4.2

Proof of Theorem 4.1.6. For notational simplicity we set $\varphi_1^r(t) = \varphi_1^r(Z^r(t))$. Fix $t_0 > 0$, $r > 0$ and $\omega \in \mathcal{M}^{\cap}$.

1. Assume that (4.37) holds. Let $t_1(\omega) = \inf\{t : \varphi_1^r(t,\omega) < \varphi_1^r(0,\omega)/2\}$. We investigate two possible cases; $t_1(\omega) \le t_0$ and $t_1(\omega) > t_0$, separately. We omit ω from the notation in the rest of the proof.

   Case 1. First assume that $t_1 \le t_0$. Set $s_0 = t_1$ and define

   $$s_{2i+1} = \inf\{t > s_{2i} : \varphi_1^r(t) = 0\} \text{ and } s_{2i+2} = \inf\{t > s_{2i+1} : \varphi_1^r(t) > 0\} \text{ for } i = 0, 1, \ldots \quad \text{(C.1)}$$

   For any $t \in [s_{2i+1}, s_{2i+2})$, $\varphi_1^r(t) = 0$, so assume that $t_0 \in [s_{2i}, s_{2i+1})$ for some i. Note that

   $$\varphi_1^r(t_0) \le \varphi_1^r(s_{2i}) - (A^r(t_0) - A^r(s_{2i})) + \sum_{j=1}^{J} \left(S_j\left(B_j^r(t_0)\right) - S_j\left(B_j^r(s_{2i})\right)\right),$$

   since all arrivals are routed to one of the queues that have idle servers. Hence,

   $$\begin{aligned}
   \varphi_1^r(t_0) &\le \varphi_1^r(s_{2i}) - \left(\breve{A}^r(t_0) - \breve{A}^r(s_{2i})\right) + \sum_{j=1}^{J} \left(\breve{S}_j\left(B_j^r(t_0)\right) - \breve{S}_j\left(B_j^r(s_{2i})\right)\right) \\
   &\quad + \sum_{j=1}^{J} \mu_j \int_{s_{2i}}^{t_0} Z_j^r(s)\mathrm{d}s - \lambda^r(t_0 - s_{2i}) \\
   &\le \varphi_1^r(s_{2i}) + \theta\sqrt{\lambda^r}(t_0 - t_1) + |o(\sqrt{|N^r|})| + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} \\
   &\quad + 2\|A^r(t) - \lambda^r t\|_{t_0}. \qquad \text{(C.2)}
   \end{aligned}$$

   Since, by Lemma 9, $\varphi_1^r(s_{2i}) < (\varphi_1^r(0)/2) \vee J$ and $\varphi_1^r(0)/2 > J$ for r large enough,

   $$\begin{aligned}
   \varphi_1^r(t_0) &\le \varphi_1^r(0)/2 + \theta\sqrt{\lambda^r}(t_0 - t_1) + |o(\sqrt{|N^r|})| + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} \\
   &\quad + 2\|A^r(t) - \lambda^r t\|_{t_0}
   \end{aligned}$$

Now add and subtract $\varphi_1^r(0)/2$ to the right hand side above to get (4.38).

CASE 2. Now assume that $t_1 > t_0$, then

$$\begin{aligned}
\varphi_1^r(t_0) &\leq \varphi_1^r(0) - A^r(t_0) + \sum_{j=1}^{J} S_j\left(B_j^r(t_0)\right) \\
&= \varphi_1^r(0) - \breve{A}^r(t_0) + \sum_{j=1}^{J} \breve{S}_j\left(B_j^r(t_0)\right) + \sum_{j=1}^{J} \mu_j \int_0^{t_0} Z_j^r(s)\mathrm{d}s - \lambda^r t_0 \\
&\leq \varphi_1^r(0) - \left(\mu_{\min}\varphi_1^r(0)/2 - \sqrt{\lambda^r}\theta\right) t_0 + |o(\sqrt{|N^r|})| + \sum_{j=1}^{J} \left\|S_j^r(t) - \mu_j t\right\|_{|N^r|t_0} \\
&\quad + \left\|A^r(t) - \lambda^r t\right\|_{t_0}.
\end{aligned}$$

By (4.37), the last inequality gives (4.38).

2. Now assume that

$$\varphi_1^r(Z^r(0)) \leq \frac{4\theta\sqrt{\lambda^r}(t_0 \vee 1)}{\mu_{\min} \wedge 1}$$

First assume that $\varphi_1^r(0,\omega) > 0$. Set $s_0 = 0$ and define s_{2i+1} and s_{2i+2} as in (C.1). For any $t \in [s_{2i+1}, s_{2i+2})$, $\varphi_1^r(t) = 0$, so assume that $t_0 \in [s_{2i}, s_{2i+1})$ for some i. Then, we have that (C.2) holds with $t_1 = 0$. Since $\varphi_1^r(s_{2i}) < \varphi_1^r(0) \vee J$, (C.2) yields (4.39). If $\varphi_1^r(0,\omega) = 0$, then define $t_1 = \inf\{t > 0 : \varphi_1^r(t,\omega) > 0\}$. Since $\varphi_1^r(t_1,\omega) < 2J$, we get the result from a similar discussion used in the case with $\varphi_1^r(0,\omega) > 0$ above.

□

Proof of Theorem 4.1.7. For notational simplicity we set $\varphi_2^r(t) = \varphi_2^r(Q^r(t))$. Fix $t_0 > 0$, $r > 0$ and $\omega \in \mathcal{M}^{\cap}$.

1. Assume that $\varphi_2^r(0) > \theta\sqrt{\lambda^r}(t_0 \vee 1)$.

   Let

$$t_1 = \inf\{t \geq 0 : \varphi_1^r(t) = 0\} \text{ and } t_2 = \inf\{t \geq 0 : \varphi_2^r(t) = 0\}$$

   We will study the following three possible cases separately; (1) $t_1 \leq t_0 \leq t_2$, (2) $t_1 > t_0$ and $t_2 > t_0$, (3) $t_2 \leq t_0$.

CASE 1. Assume that $t_1 \leq t_0 \leq t_2$. Let $s_0 = t_1$ and define

$$s_{2i+1} = \inf\{t > s_{2i} : \varphi_1^r(t) > 0\} \text{ and } s_{2i+2} = \inf\{t > s_{2i+1} : \varphi_1^r(t) = 0\} \text{ for } i = 0, 1, \ldots.$$

One of the following must hold; $s_1 > t_0$ or $s_{2k+2} < t_0 < s_{2k+3}$ or $s_{2k+1} \leq t \leq s_{2k+2}$ for some $k \geq 0$.

If $s_1 > t_0$ then

$$\begin{aligned}
\varphi_2^r(t_0) &\leq \varphi_2^r(t_1) + (A^r(t_0) - A^r(t_1)) + \sum_{j=1}^{J} (S_j(t_0) - S_j(t_1)) \\
&\leq \varphi_2^r(t_1) - \sqrt{\lambda^r}\theta(t_0 - t_1) + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} \\
&\quad + 2\|A^r(t) - \lambda^r t\|_{t_0}. \qquad \text{(C.3)}
\end{aligned}$$

By definition of t_1,

$$\varphi_2^r(t_1) \leq \varphi_2^r(0) - N_{\min}^r \mu_{\min} t_1 + 2\sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0}. \qquad \text{(C.4)}$$

For r large enough $N_{\min}^r/\sqrt{\lambda^r} > \theta/\mu_{\min}$, hence for such r we get (4.42) by combining (C.3) and (C.4).

Now assume that either $s_{2k+2} < t_0 < s_{2k+3}$ or $s_{2k+1} \leq t \leq s_{2k+2}$ for some $k \geq 0$.

We define $\Delta_i = [s_{2i+1}, s_{2i+2}]$, for $i = 0, 1, \ldots$. For any $i \geq 0$ and $t \in [s_{2i+1}, s_{2i+2}]$, there exists at least one $j_t \in \mathcal{J}$ such that $Q_{j_t}^r(t) < N_{j_t}^r$. Define

$$a_j^i = \inf\{t \geq s_{2i+1} : Q_j^r(t) < N_j^r\} \wedge s_{2i+2}.$$

We assume for simplicity that $s_{2i+1} = a_1^i \leq a_2^i \leq \ldots \leq a_J^i \leq s_{2i+2}$. If $Q_j^r(t) < N_j^r$ for some $t \in [s_{2i+1}, s_{2i+2})$, then $Q_j^r(s_{2i+2}) < N_j^r + 2$, by Lemma 9 and since no arrivals join queue j during $[s_{2i+1}, s_{2i+2})$ when $Q_j^r(t) \geq N_j^r$. Hence, if $a_J^i < s_{2i+2}$ then $\varphi_2^r(s_{2i+2}) < 2J$. Now assume that $a_J^i = s_{2i+2}$. Then for any $t \in [s_{2i+1}, s_{2i+2})$

$$\begin{aligned}
\varphi_1^r(t) &\leq \varphi_1^r(s_{2i+1}) - (A^r(t) - A^r(s_{2i+1})) + \sum_{j=1}^{J-1} \left(S_j(B_j^r(t)) - S_j(B_j^r(s_{2i+1}))\right) \\
&\leq 2J - \left(\breve{A}^r(t) - \breve{A}^r(s_{2i+1})\right) + \sum_{j=1}^{J-1} \left(\breve{S}_j(B_j^r(t)) - \breve{S}_j(B_j^r(s_{2i+1}))\right) \\
&\quad + \left(\theta\sqrt{\lambda^r} - N_{\min}^r \mu_{\min}\right)(t - s_{2i+1}) \\
&\leq 2J + \zeta^r(t_0). \qquad \text{(C.5)}
\end{aligned}$$

Choose $l \leq k$ so that Δ_l is the last interval such that $a_j^l < s_{2l+2} \wedge t_0$, for all $j \in \mathcal{J}$, if there exists such l or set $l = k+1$. Hence, Δ_l is the last interval during which all the queues become empty at least once.

a. Assume that $l \leq k$. If $s_{2l+1} \leq t_0$, then

$$\varphi_2^r(t_0) < 2J. \tag{C.6}$$

If $s_{2l+2} < t_0$, then (C.5) holds for all $t \in [s_{2l+2}, t_0]$ since if $t \in [s_{2i+2}, s_{2i+3}]$, then $\varphi_1^r(t) = 0$ and otherwise we have from the discussion above that (C.5) holds. Hence, for $Q_\Sigma^r(t) = \sum_{j=1}^J (Q_j^r(t) + Z_j^r(t))$

$$\begin{aligned} Q_\Sigma^r(t_0) - Q_\Sigma^r(s_{2l+2}) &\leq A^r(t_0) - A^r(s_{2l+2}) - \sum_{j\in\mathcal{J}} \left(S(B_j^r(t_0)) - S(B_j^r(s_{2l+2}))\right) \\ &\leq \left(\mu_{\max}(2J + \zeta^r(t_0)) - \sqrt{\lambda^r}\theta\right)(t_0 - s_{2l+2}) \\ &\quad +2\sum_{j=1}^J \left\|S_j(t) - \mu_j t\right\|_{|N^r|t_0} + 2\left\|A^r(t) - \lambda^r t\right\|_{t_0}. \end{aligned}$$

Since $\varphi_1^r(t_0) \leq 2J + \zeta^r(t_0)$, we have from the last inequality that

$$\begin{aligned} \varphi_2^r(t_0) - \varphi_2^r(s_{2l+2}) &\leq \left(\mu_{\max}(2J + \zeta^r(t_0)) - \sqrt{\lambda^r}\theta\right)(t_0 - s_{2l+2}) \\ &\quad +2\sum_{j=1}^J \left\|S_j(t) - \mu_j t\right\|_{|N^r|t_0} + 2\left\|A^r(t) - \lambda^r t\right\|_{t_0} \\ &\quad +2J + \zeta^r(t_0). \end{aligned} \tag{C.7}$$

Since $\varphi_2^r(s_{2l+2}) < 2J$ by definition of l, we get (4.42) from (C.7).

b. If $l = k+1$, then (C.5) holds for all $t \in [t_1, t_0]$. Hence,

$$\begin{aligned} \varphi_2^r(t_0) - \varphi_2^r(t_1) &\leq \left(\mu_{\max}(2J + \zeta^r(t_0)) - \sqrt{\lambda^r}\theta\right)(t_0 - t_1) \\ &\quad +2\sum_{j=1}^J \left\|S_j(t) - \mu_j t\right\|_{|N^r|t_0} + 2\left\|A^r(t) - \lambda^r t\right\|_{t_0} \\ &\quad +2J + \zeta^r(t_0). \end{aligned} \tag{C.8}$$

For r large enough $N_{\min}^r/\sqrt{\lambda^r} > \theta/\mu_{\min}$, hence for such r we get (4.42) by combining (C.8) and (C.4).

CASE 2. Now assume that $t_1 > t_0$ and $t_2 > t_0$. Then, none of the arrivals join a queue that has customers waiting during $[0, t_0]$. Let $a_j = \inf\{t > 0 : Q_j^r(t) < N_j^r\}$. As in Case 1, if $a_j \leq t_0$ for all $j \in \mathcal{J}$, then $\varphi_2^r(t_0) < 2J$. If $a_j > t_0$ for some $j \in \mathcal{J}$ then

$$\varphi_2^r(t_0) - \varphi_2^r(0) \quad \leq \quad -N_{\min}^r \mu_{\min} t_0 + 2 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{|N^r|t_0} . \tag{C.9}$$

Hence for r large enough, we get (4.42).

CASE 3. If $t_2 < t_0$, define

$$t_2' = \sup\{t > t_2 : \varphi_2^r(t) > 0\}.$$

If $t_2' \geq t_0$, then $\varphi_2^r(t_0) = 0$, so assume that $t_2' < t_0$. We have that $0 < \varphi_2^r(t_2') < 2J$. We get (4.42) from the discussion below.

2. Now assume that

$$\varphi_2^r(Q^r(0)) \leq \theta \sqrt{\lambda^r}(t_0 \vee 1),$$

and define

$$\tilde{t} = \sup\{t_0 \geq t \geq 0 : \varphi_2^r(t) = 0\}.$$

If $\tilde{t} = -\infty$, set $\tilde{t} = 0$. Observe that $\theta\sqrt{\lambda^r}(t_0 \vee 1) \geq \varphi_2^r(\tilde{t}) > 0$ and $\varphi_2^r(t) > 0$ for all $[\tilde{t}, t_0]$. By considering the path from $\tilde{t}$ to t_0, one can use the same arguments above, but this time only cases 1 and 2 will apply. It is obvious that (C.3)-(C.9) hold for this case as well, since they do not depend on the initial value of φ_2^r. Hence, we get (4.43).

□

C.2 Proofs of the results in Section 4.1.4.3

We first establish the additional fluid model equations that must be satisfied by the fluid limits of the MED–FSF and MED–LB distributed server pool systems. Then, using these equations, we determine the set of invariant states of the fluid limits for both systems.

Lemma 4. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Every fluid limit $\bar{\mathbb{X}}$ of $\{\mathbb{X}^r\}$ satisfies the following*

equations in addition to the fluid model equations (A.2)-(A.11) in [22]. For every $j \in \mathcal{J}$ and a regular point t of $\bar{\mathbb{X}}$

$$\dot{\bar{A}}_j^q(t) = 0 \text{ when } \frac{\bar{Q}_j(t)}{\beta_j \mu_j} > \frac{\bar{Q}_{j'}(t)}{\beta_{j'} \mu_{j'}} \text{ for some } j' \in \mathcal{J} \text{ and} \tag{C.10}$$

$$\sum_{j'=j}^{J} \dot{\bar{A}}_{j'}^s(t) = \lambda \quad \text{if} \quad \sum_{j'=j}^{J} (\bar{Z}_{j'}(t) - \beta_{j'}) < 0 \tag{C.11}$$

Proof. Let $\bar{\mathbb{X}}$ be a fluid limit. Fix $t > 0$ and assume that $\frac{\bar{Q}_j(t)}{\beta_j \mu_j} > \frac{\bar{Q}_{j'}(t)}{\beta_{j'} \mu_{j'}}$ for some $j' \in \mathcal{J}$. By continuity of $\bar{\mathbb{X}}$, there exists $\delta > 0$ such that

$$\frac{\bar{Q}_j(s)}{\beta_j \mu_j} > \frac{\bar{Q}_{j'}(s)}{\beta_{j'} \mu_{j'}}$$

for all $s \in [t - \delta, t + \delta]$. Let $\omega \in \mathcal{A}$, for $\mathcal{A}$ given as in Section 4.1.4.3. Assume for notational simplicity that $\bar{\mathbb{X}}^r(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}(\cdot, \omega)$. Note that by (4.2)

$$N_\ell^r = \beta_\ell |N^r| + o(|N^r|) \quad \text{for all } \ell \in \mathcal{J}.$$

Hence, for r large enough

$$\frac{\bar{Q}_j(s)}{\beta_j \mu_j + o(|N^r|)/|N^r|} > \frac{\bar{Q}_{j'}(s)}{\beta_{j'} \mu_{j'} - o(|N^r|)/|N^r|},$$

so

$$\frac{Q_j^r(s)/|N^r|}{\beta_j \mu_j + o(|N^r|)/|N^r|} > \frac{Q_{j'}^r(s)/|N^r|}{\beta_{j'} \mu_{j'} - o(|N^r|)/|N^r|}$$

Thence

$$\frac{Q_j^r(s)}{N_j^r \mu_j} \geq \frac{Q_j^r(s)}{|N^r| \beta_j \mu_j + o(|N^r|)} > \frac{Q_{j'}^r(s)}{|N^r| \beta_{j'} \mu_{j'} - o(|N^r|)} \geq \frac{Q_{j'}^r(s)}{N_{j'}^r \mu_{j'}} \tag{C.12}$$

for large enough r for all $s \in [t - \delta, t + \delta]$. This implies by (4.30) for r large enough that $A_j^{q,r}(s)$ is flat on $[t - \delta, t + \delta]$. Hence $\dot{\bar{A}}^q(t) = 0$. Fluid limit equation (C.11) is proved similarly. □

Lemma 5. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. Assume that $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Every fluid limit $\bar{\mathbb{X}}$ of $\{\mathbb{X}^r\}$ satisfies (C.10) and the following equation in addition to the fluid model equations (A.2)-(A.8) in [22]. For every $j \in \mathcal{J}$ and a regular point t of $\bar{\mathbb{X}}$*

$$\dot{\bar{A}}_j^s(t) = 0 \text{ when } \frac{\bar{Z}_j(t)}{\beta_j} > \frac{\bar{Z}_{j'}(t)}{\beta_{j'}} \text{ for some } j' \in \mathcal{J}. \tag{C.13}$$

The proof is similar to that of Lemma 4.

Proof of Lemma 1. Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems and $h_1(t) = \sum_{j=1}^{J} |\bar{Z}_j(t) - \beta_j|$. Note that if $h_1(t) > 0$ and t is a regular point of $\bar{\mathbb{X}}$ then $\dot{h}_1(t) \leq 0$ from the fluid model equation (C.11), and equations (A.2) and (A.9) in [22]. Hence, if $h_1(0) = 0$, then $h_1(t) = 0$ for all $t \geq 0$ by virtue of Lemma 2.4.5 of [19].

Now let $h_2(t) = \max_{j\in\mathcal{J}}\{\bar{Q}_j(t)/\beta_j\} - \min_{j\in\mathcal{J}}\{\bar{Q}_j(t)/\beta_j\}$ and assume that $\bar{Z}_j(0) = \beta_j$ for all $j \in \mathcal{J}$. If t is a regular point of $\bar{\mathbb{X}}$ and $h_2(t) > 0$ then $\dot{h}_2(t) \leq 0$, by (C.10), (C.11), (A.2)-(A.11) in [22] and Lemma 2.8.6 of [19]. Hence, if $h_1(0) = h_2(0) = 0$, then $(\bar{Q}(0), \bar{Z}(0)) = (\bar{Q}(t), \bar{Z}(t))$. Note that $h_1(0) = h_2(0) = 0$ if and only if $(q, z) \in \mathcal{M}$. Therefore, (q, z) is an invariant state if and only if $(q, z) \in \mathcal{M}$. For the MED–LB policy the result is proved similarly using (C.13). □

C.3 Proofs of the results in Section 4.1.4.4

C.3.1 Proofs of Propositions 4.7 and 4.9

We use the framework of [22]. We need to check that Assumptions 1 through 4 in that paper are satisfied by the sequuence of MED–FSF and MED–LB distributed systems. It can easily be checked using (4.2) and (4.4) that the static planning problem (2.20) in [22] has a unique optimal solution with $x_j^* = 1$ for all $j \in \mathcal{J}$, and so, by (4.4), Assumption 1 in that paper is satisfied. Assumption 2 in [22] is satisfied by both policies by Lemma 1 and by Theorem A.1 in the same paper. So we focus on Assumptions 2 and 4. In this section we first define the appropriate SSC functions, see Section 4.1 of [22] for more details on SSC functions, then we show that Assumption 3 in that paper is satisfied by the MED–LB and MED–FSF distributed systems.

Hydrodynamic scaling and hydrodynamic limits are introduced in [22]. They showed that hydrodynamic limits satisfy a set of equations that are called hydrodynamic model equations. To check Assumption 4 in [22], one needs to show that the hydrodynamic model solutions, which are solutions of the hydrodynamic model equations, satisfy certain conditions. We start with characterizing the additional hydrodynamic model equations for the MED–FSF and MED–LB policies; see Lemmas 6 and 7. Then we show that hydrodynamic

model equations satisfy Assumption 4 in [22] in Propositions C.1 and C.2. This gives us the multiplicative state space collapse results. To prove the strong state space collapse results stated in Theorems 4.7 and 4.9, we show in Theorem C.3.1 that condition (4.15) in [22] is satisfied by $\hat{Q}^r$ and $\hat{Z}^r$ under any completely non-idling policy.

Recall that $\lambda = \lim_{r\to\infty} \lambda^r/|N^r| = \bar{\mu}$.

Lemma 6. *Let $\{\mathbb{X}^r\}$ be a sequence of MED-FSF distributed server pool systems. Every hydrodynamic limit $\tilde{\mathbb{X}}$ of $\{\mathbb{X}^r\}$ satisfies the following equations in addition to equations (4.9)-(4.14) in [22]. For every $j \in \mathcal{J}$*

$$\dot{\tilde{A}}_j^q(t) = 0 \text{ when } \frac{\tilde{Q}_j(t)}{\beta_j\mu_j} > \frac{\tilde{Q}_{j'}(t)}{\beta_{j'}\mu_{j'}} \text{ for some } j' \in \mathcal{J} \text{ and} \tag{C.14}$$

$$\dot{\tilde{A}}_j^{s,+}(t) = \lambda \text{ when } \sum_{l=j}^{J} \tilde{Z}_l(t) < 0, \tag{C.15}$$

where $\tilde{A}_j^{s,+}(t) = \sum_{\ell=j}^{J} \tilde{A}_\ell^s(t)$.

Proof. Let $\mathbb{X}^{r,m}$ be the Hydrodynamically scaled version of $\mathbb{X}^r$ as defined in Section 5.1 in [22]. Then, by (4.30) and (4.31) $X^{r,m}$ satisfies the following equations.

$$A_j^{q,r,m}(t) \quad \text{can only increase when } \frac{Q_j^{r,m}(t)}{N_j^r\mu_j} \le \frac{Q_{j'}^{r,m}(t)}{N_{j'}^r\mu_{j'}} \text{ for all } j' \in \mathcal{J} \text{ and} \tag{C.16}$$

$$A_j^{s,r,m}(t) \quad \text{can only increase when } \sum_{l=j+1}^{J} Z_l^{r,m}(t) = 0. \tag{C.17}$$

Let $\tilde{\mathbb{X}}$ be a cluster point of $\{\mathbb{X}^{r,m}\}$, for some $L > 0$.

Fix $T > 0$. By definition of a hydrodynamic limit, given $\epsilon > 0$ one can choose (r, m, ω) such that

$$\left\| \tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t,w) \right\|_L \le \epsilon. \tag{C.18}$$

The rest of the proof is similar to the proof of the first part of Lemma 1. Fix $L > t > 0$ and $j \in \mathcal{J}$. Assume that

$$\frac{\tilde{Q}_j(t)}{\beta_j\mu_j} > \frac{\tilde{Q}_{j'}(t)}{\beta_{j'}\mu_{j'}} \text{ for some } j' \in \mathcal{J}.$$

By continuity of $\tilde{\mathbb{X}}$, there exists a $\delta > 0$ such that

$$\frac{\tilde{Q}_j(s)}{\beta_j \mu_j} > \frac{\tilde{Q}_{j'}(s)}{\beta_{j'} \mu_{j'}}$$

for all $s \in [t-\delta, t+\delta]$. Since ϵ is arbitrary, by (C.18) one can choose (r, m, ω) such that

$$\frac{Q_j^{r,m}(s)}{\beta_j \mu_j} > \frac{Q_{j'}^{r,m}(s)}{\beta_{j'} \mu_{j'}}$$

for all $s \in [t-\delta, t+\delta]$. Since $\beta_j = N_j^r/|N^r| - o(N^r)/|N^r|$, this implies, similar to (C.12), by (C.16) that $A_j^{q,r,m}(s)$ is flat on $[t-\delta, t+\delta]$. Hence $\tilde{A}^q(t)$ cannot increase on $[t-\delta, t+\delta]$. The second hydrodynamic equation (C.15) is proved similarly using (C.17). □

Lemma 7. *Let $\{\mathbb{X}^r\}$ be a sequence of MED-LB distributed server pool systems. Then, in addition to equations (4.9)-(4.14) in [22], every hydrodynamic limit $\tilde{\mathbb{X}}$ of $\{\mathbb{X}^r\}$ satisfies (C.14) and*

$$\dot{\tilde{A}}_j^s(t) = 0 \text{ when } \frac{\tilde{Z}_j(t)}{\beta_j} > \frac{\tilde{Z}_{j'}(t)}{\beta_{j'}} \text{ for some } j' \in \mathcal{J}. \tag{C.19}$$

The proof is similar to that of Lemma 6.

Next, we define the SSC function for the MED–FSF policy. Let $\hat{f}_j : \mathbb{R}^{2J} \to \mathbb{R}$ be defined by

$$\hat{f}_j(q, z) = \frac{q_j}{\mu_j \beta_j},$$

for $j \in \mathcal{J}$. The SSC function, $\hat{g} : \mathbb{R}^{2J} \to \mathbb{R}$, for the MED–FSF policy is defined as follows.

$$\hat{g}(q, z) = \max_{j \in \mathcal{J}} \{f_j(q, z)\} - \min_{j \in \mathcal{J}} \{f_j(q, z)\} + \sum_{j=2}^{J} |\tilde{Z}_j(t)|. \tag{C.20}$$

It is obvious that $\hat{g}$ is continuous and $\hat{g}(\alpha q, \alpha z) = \alpha \hat{g}(q, z)$ for all $(q, z) \in \mathbb{R}^{2J}$. Hence, $\hat{g}$ satisfies Assumption 3 in [22]. Also,

$$\hat{g}(\tilde{Q}(t), \tilde{Z}(t)) = 0$$

if and only if

$$\frac{\tilde{Q}_j(t)}{\beta_j \mu_j} = \frac{\tilde{Q}_{j'}(t)}{\beta_{j'} \mu_{j'}} \text{ and} \sum_{j=2}^{J} \tilde{Z}_j(t) = 0 \text{ for all } j, j' \in \mathcal{J}.$$

Therefore, $\hat{g}$ is the desired SSC function. Next we show that the hydrodynamic model solutions and $\hat{g}$ satisfy Assumption 4 in [22].

Proposition C.1. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Let $\hat{g}$ be defined as in (C.20). For any hydrodynamic model solution $\tilde{\mathbb{X}}$,*

$$\hat{g}(\tilde{Q}(t), \tilde{Z}(t)) \le H(t) \quad \text{for } t \ge 0$$

with $H(t) = (\hat{g}(\tilde{Q}(0), \tilde{Z}(0)) - (\mu_{\min}\beta_{\min} \wedge 1)t) \wedge 0$. Whenever $\hat{g}(\tilde{Q}(0), \tilde{Z}(0)) = 0$ then $\hat{g}(\tilde{Q}(t), \tilde{Z}(t)) = 0$ for $t \ge 0$. In particular, the hydrodynamic model solutions of $\{\mathbb{X}^r\}$ and $\hat{g}$ satisfy Assumption 4 in [22].

Proof. Let $\tilde{\mathbb{X}}$ be a hydrodynamic model solution and $t \ge 0$ a regular point of $\tilde{\mathbb{X}}$. Assume that $\hat{g}(\tilde{Q}(t), \tilde{Z}(t)) > 0$.

Let

$$\mathcal{U}_{\min}(t) = \{j \in \mathcal{J} : \hat{f}_j(\tilde{Q}(t), \tilde{Z}(t)) \le \hat{f}_{j'}(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J}\} \text{ and}$$
$$\mathcal{U}_{\max}(t) = \{j \in \mathcal{J} : \hat{f}_j(\tilde{Q}(t), \tilde{Z}(t)) \ge \hat{f}_{j'}(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J}\}.$$

Since t is a regular point of $\tilde{\mathbb{X}}$, by Lemma 2.8.6 of [19],

$$\dot{\hat{g}}(\tilde{Q}(t), \tilde{Z}(t)) = \dot{\hat{f}}_i(t) - \dot{\hat{f}}_j(t) - \sum_{j=2}^{J} \dot{\tilde{Z}}_j(t) \text{ for all } i \in \mathcal{U}_{\max}(t) \text{ and } j \in \mathcal{U}_{\min}(t).$$

First assume that $\sum_{j=2}^{J} \tilde{Z}_j(t) < 0$. Then,

$$\sum_{j=2}^{J} \dot{\tilde{Z}}_j(t) \ge \lambda - \sum_{j=2}^{J} \mu_j\beta_j \ge \mu_{\min}\beta_{\min}$$

by (C.15) and (4.5) in [22]. Also, by (C.14) and (4.8) in [22], $\dot{\hat{f}}_i(t) = -1$ or 0 and $\dot{\hat{f}}_j(t) = 0$ for all $i \in \mathcal{U}_{\max}(t)$ and $j \in \mathcal{U}_{\min}(t)$. Hence, $\dot{\hat{g}}(\tilde{Q}(t), \tilde{Z}(t)) \le -\mu_{\min}\beta_{\min}$.

Now assume that $\sum_{j=2}^{J} \tilde{Z}_j(t) = 0$. Then, $\sum_{j=2}^{J} \dot{\tilde{Z}}_j(t) = 0$ since $\sum_{j=2}^{J} \dot{\tilde{Z}}_j(t) > 0$ whenever $\sum_{j=2}^{J} \tilde{Z}_j(t) < 0$. Hence,

$$\dot{\hat{g}}(\tilde{Q}(t), \tilde{Z}(t)) = \dot{\hat{f}}_i(t) - \dot{\hat{f}}_j(t).$$

We get as in the proof of Proposition C.2 below that $\dot{\hat{g}}(\tilde{Q}(t), \tilde{Z}(t)) \le -(\mu_{\min} \wedge 1)$. This gives the first claim. Second claim is follows from the fact that $\dot{\hat{g}}(\tilde{Q}(t), \tilde{Z}(t)) < 0$ if $\hat{g}(\tilde{Q}(t), \tilde{Z}(t)) > 0$. □

Next, we define the SSC function for the MED–LB policy. Let $q = (q_1, \ldots, q_J) \in \mathbb{R}^J$, $z = (z_1, \ldots, z_J) \in \mathbb{R}^J$ and $f_j : \mathbb{R}^{2J} \to \mathbb{R}$ be defined by

$$f_j(q, z) = \frac{q_j}{\mu_j \beta_j} + \frac{z_j}{\beta_j},$$

for $j \in \mathcal{J}$. The SSC function, $g : \mathbb{R}^{2J} \to \mathbb{R}$, for MED–LB policy is defined by

$$g(q, z) = \max_{j \in \mathcal{J}} \{f_j(q, z)\} - \min_{j \in \mathcal{J}} \{f_j(q, z)\}. \tag{C.21}$$

It is easily checked that g is continuous and $g(\alpha q, \alpha z) = \alpha g(q, z)$ for all $(q, z) \in \mathbb{R}^{2J}$. Hence, g satisfies Assumption 3 in [22]. Also,

$$g(\tilde{Q}(t), \tilde{Z}(t)) = 0$$

if and only if

$$\frac{\tilde{Q}_j(t)}{\beta_j \mu_j} = \frac{\tilde{Q}_{j'}(t)}{\beta_{j'} \mu_{j'}} \text{ and } \frac{\tilde{Z}_j(t)}{\beta_j} = \frac{\tilde{Z}_{j'}(t)}{\beta_{j'}}, \text{ for all } j, j' \in \mathcal{J}.$$

Therefore, g is the desired SSC function. Next we show that g satisfies Assumption 4 in [22].

Proposition C.2. *Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. Let g be defined as in (C.21). For any hydrodynamic model solution $\tilde{\mathbb{X}}$*

$$g(\tilde{Q}(t), \tilde{Z}(t)) \le H(t) \quad \text{for } t \ge 0 \tag{C.22}$$

with $H(t) = (g(\tilde{Q}(0), \tilde{Z}(0)) - (\mu_{\min} \wedge 1)t) \wedge 0$. Whenever $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$ then $g(\tilde{Q}(t), \tilde{Z}(t)) = 0$ for $t \ge 0$. In particular, $\{\mathbb{X}^r\}$ with g satisfy Assumption 4 in [22].

Proof. Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit and $t \ge 0$ be a regular point of $\tilde{\mathbb{X}}$.

Assume that

$$g(\tilde{Q}(t), \tilde{Z}(t)) > 0. \tag{C.23}$$

Let

$$\mathcal{U}_{\min}(t) = \{j \in \mathcal{J} : f_j(\tilde{Q}(t), \tilde{Z}(t)) \le f_{j'}(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J}\} \text{ and}$$
$$\mathcal{U}_{\max}(t) = \{j \in \mathcal{J} : f_j(\tilde{Q}(t), \tilde{Z}(t)) \ge f_{j'}(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J}\}.$$

Since t is a regular point of $\tilde{\mathbb{X}}$, by Lemma 2.8.6 of [19],

$$\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) = \dot{f}_i(t) - \dot{f}_j(t), \text{ for all } i \in \mathcal{U}_{\max}(t) \text{ and } j \in \mathcal{U}_{\min}(t).$$

Also, observe that $\mathcal{U}_{\max}(t) \cap \mathcal{U}_{\min}(t) = \emptyset$ since $g(\tilde{Q}(t), \tilde{Z}(t)) > 0$.

We first show that if $g(\tilde{Q}(t), \tilde{Z}(t)) > 0$, then

$$\dot{f}_i(\tilde{Q}(t), \tilde{Z}(t)) \leq -(\mu_{\min} \wedge 1) \quad \text{for all } i \in \mathcal{U}_{\max}(t). \tag{C.24}$$

First observe that if $g(\tilde{Q}(t), \tilde{Z}(t)) > 0$ and $i \in \mathcal{U}_{\max}(t)$

$$\frac{\tilde{Q}_i(t)}{\mu_i \beta_i} > \frac{\tilde{Q}_j(t)}{\mu_j \beta_j} \quad \text{or} \quad \frac{\tilde{Z}_i(t)}{\beta_i} > \frac{\tilde{Z}_j(t)}{\beta_j}$$

for some $j \in \mathcal{U}_{\min}(t)$. Therefore, by (C.14), (C.19), equations (4.2) and (4.5) in [22]

$$\dot{\tilde{Q}}_i(t) + \dot{\tilde{Z}}_i(t) \leq -\mu_{\min}\beta_{\min}.$$

In addition, either $\dot{\tilde{Q}}_i(t) = 0$ or $\dot{\tilde{Z}}_i(t) = 0$ by (4.7) in [22]. This gives (C.24).

Next we show that if $g(\tilde{Q}(t), \tilde{Z}(t)) > 0$, then

$$\dot{f}_i(\tilde{Q}(t), \tilde{Z}(t)) \geq 0 \quad \text{for all } i \in \mathcal{U}_{\min}(t). \tag{C.25}$$

By (4.1), (4.2) and (4.5) in [22]

$$\sum_{i \in \mathcal{U}_{\min}(t)} \left(\dot{\tilde{Q}}_i(t) + \dot{\tilde{Z}}_i(t) \right) = \lambda - \sum_{i \in \mathcal{U}_{\min}(t)} \mu_i \beta_i > 0,$$

where the last inequality follows from the fact that $\mathcal{U}_{\min}(t) \neq \mathcal{J}$. By (C.24) we get (C.25). Combining (C.24) with (C.25) gives (C.22). The second claim immediately follows from Lemma 2.8.6 of [19] and the fact that $\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) < 0$ whenever $g(\tilde{Q}(t), \tilde{Z}(t)) > 0$ as shown above. □

Next we show that under a non-idling routing policy the sequence of distributed systems satisfy condition (4.15) in [22].

Theorem C.3.1. *Let $\pi \in \Pi$ be a non-idling routing policy and assume that (4.20) holds. Then, for every $T > 0$*

$$\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}\left\{ \|\hat{Q}^r(t)\|_T \vee \|\hat{Z}^r(t)\|_T > C \right\} = 0. \tag{C.26}$$

Proof. Fix $\pi \in \Pi$, $T > 0$ and assume that (4.20) holds. Observe that (4.20) implies

$$\lim_{C\to\infty} \lim_{r\to\infty} \mathbb{P}\left\{ |\hat{Q}^r(0)| \vee |\hat{Z}^r(0)| > C \right\} = 0. \tag{C.27}$$

By (4.35), (4.38) and (4.39), for $C > 0$,

$$\begin{aligned}
\mathbb{P}\left\{ \|\hat{Z}^r(t)\|_T > C \right\} &\leq \mathbb{P}\Bigg\{ \|\hat{Z}^r(0)\| + \theta\sqrt{\lambda^r/|N^r|}T + |o(\sqrt{|N^r|})|/\sqrt{|N^r|} \\
&\quad + 2\sum_{j=1}^{J} \frac{\|S_j(|N^r|t) - |N^r|\mu_j t\|_T}{\sqrt{|N^r|}} + 2\frac{\|A^r(t) - \lambda^r t\|_T}{\sqrt{|N^r|}} > C \Bigg\} \\
&\leq \mathbb{P}\left\{ \|\hat{Z}^r(0)\| > C/4 \right\} + \mathbb{P}\left\{ \theta\sqrt{\lambda^r/|N^r|}T + |o(\sqrt{|N^r|})|/\sqrt{|N^r|} > C/4 \right\} \\
&\quad + \mathbb{P}\left\{ 2\sum_{j=1}^{J} \frac{\|S_j(|N^r|t) - |N^r|\mu_j t\|_T}{\sqrt{|N^r|}} > C/4 \right\} \\
&\quad + \mathbb{P}\left\{ 2\frac{\|A^r(t) - \lambda^r t\|_T}{\sqrt{|N^r|}} > C/4 \right\}.
\end{aligned}$$

For any $j \in \mathcal{J}$, $\frac{S_j(|N^r|t) - |N^r|\mu_j t}{\sqrt{|N^r|}}$ converges weakly to a Brownian motion with variance μ_j, and $\frac{A^r(t) - \lambda^r t}{\sqrt{|N^r|}}$ converges weakly to a Brownian motion with variance λ. Hence by the continuous mapping theorem

$$\begin{aligned}
&\lim_{C\to\infty} \lim_{r\to\infty} \mathbb{P}\left\{ 2\sum_{j=1}^{J} \frac{\|S_j(|N^r|t) - |N^r|\mu_j t\|_T}{\sqrt{|N^r|}} > C/4 \right\} \\
&\qquad = \lim_{C\to\infty} \lim_{r\to\infty} \mathbb{P}\left\{ 2\frac{\|A^r(t) - \lambda^r t\|_T}{\sqrt{|N^r|}} > C/4 \right\} = 0.
\end{aligned} \tag{C.28}$$

Thus, by (C.27) and (C.28)

$$\lim_{C\to\infty} \lim_{r\to\infty} \mathbb{P}\left\{ \|\hat{Z}^r(t)\|_T > C \right\} = 0.$$

By (4.35), (4.42) and (4.43), for $C > 0$,

$$\begin{aligned}\mathbb{P}\left\{\|\hat{Q}^r(t)\|_T > C\right\} \;\leq\; & \mathbb{P}\left\{\|\hat{Q}^r(0)\| > C/5\right\} \\ & + \mathbb{P}\left\{\theta\sqrt{\lambda^r/|N^r|}T + |o(\sqrt{|N^r|})|/\sqrt{|N^r|} > C/5\right\} \\ & + \mathbb{P}\left\{2T(\mu_{\max}\vee 1)\zeta^r(T)/\sqrt{|N^r|} > C/5\right\} \\ & + \mathbb{P}\left\{4\sum_{j=1}^{J}\frac{\|S_j(|N^r|t) - |N^r|\mu_j t\|_T}{\sqrt{|N^r|}} > C/5\right\} \\ & + \mathbb{P}\left\{2\frac{\|A^r(t) - \lambda^r t\|_T}{\sqrt{|N^r|}} > C/5\right\}.\end{aligned}$$

By (C.27) and (C.28) it is enough to show that for any $T > 0$

$$\lim_{C\to\infty}\lim_{r\to\infty}\mathbb{P}\left\{2T(\mu_{\max}\vee 1)\zeta^r(T)/\sqrt{|N^r|} > C/5\right\} = 0. \tag{C.29}$$

to complete the proof of (C.26). For notational simplicity assume that $\mu_{\max} > 1$ and choose r large enough so that $N^r_{\min}\mu_{\min} - \theta\sqrt{\lambda^r} > 2|N^r|B_3$ for some $B_3 > 0$. Observe that

$$\begin{aligned}& \mathbb{P}\left\{2\mu_{\max}T\frac{\zeta^r(T)}{\sqrt{|N^r|}} > C\right\} \\ & \leq \mathbb{P}\left\{2\mu_{\max}T\sup_{0\leq s_1\leq s_2\leq T}\left\{-\sqrt{|N^r|}B_3(s_2-s_1) + \frac{|\breve{A}^r(s_2) - \breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\} > C/2\right\} \\ & + \sum_{j=1}^{J}\mathbb{P}\Bigg\{2\mu_{\max}T\sup_{0\leq s_1^j\leq s_2^j\leq T}\Bigg\{-\frac{\sqrt{|N^r|}B_3}{J}(s_2^j - s_1^j) \\ & \qquad + \frac{\left|\breve{S}_j\left(|N^r|s_2^j\right) - \breve{S}_j\left(|N^r|s_1^j\right)\right|}{\sqrt{|N^r|}}\Bigg\} > C/(2J)\Bigg\} \\ & \leq \mathbb{P}\left\{4\mu_{\max}T\frac{\|A^r(t) - \lambda^r t\|_T}{\sqrt{|N^r|}} > C/2\right\} \\ & + \sum_{j=1}^{J}\mathbb{P}\left\{4\mu_{\max}T\frac{\|S_j\left(|N^r|t\right) - |N^r|\mu_j t\|_T}{\sqrt{|N^r|}} > C/(2J)\right\}\end{aligned}$$

We get (C.29), again, by virtue of the continuous mapping theorem. □

Proposition 4.7. Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that (4.2),(4.4), and (4.20) hold. We showed above that this sequence with $\hat{g}$ defined as in (C.20) satisfy Assumptions 1-4 in [22]. So we conclude by Theorem 4.2 in the same

paper that for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{ \frac{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \hat{g}(\hat{Q}^r(t), \hat{Z}^r(t)) \right|}{\sup_{L^r/\sqrt{|N^r|} \le t \le T} \left(\left| \hat{Q}^r(t) \right| \vee \left| \hat{Z}^r(t) \right| \vee 1 \right)} > \epsilon \right\} \to 0, \tag{C.30}$$

as $r \to \infty$. Combining (C.30) with Theorem C.3.1 and using Remark 4.4 in [22] yields

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \hat{g}(\hat{Q}^r(t), \hat{Z}^r(t)) \right| > \epsilon \right\} \to 0.$$

If in addition (4.46) and (4.47) hold we conclude similarly from Theorem 4.1 and Remark 4.4 in [22] and Theorem C.3.1 above that (4.48) holds. □

Proposition 4.9 is proved similarly by using g, defined in (C.21), instead of $\hat{g}$.

C.3.2 Proofs of the results in Section 4.1.4.4

Proof of Proposition 4.11. Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that (4.2),(4.4), (4.20) and (4.48) hold.

By (4.48)

$$\left\| \frac{\hat{Q}^r_j(t)}{\beta_j \mu_j} - \frac{\hat{Q}^r_{j'}(t)}{\beta_{j'} \mu_{j'}} \right\|_T \vee \left\| \sum_{j=2}^{J} \hat{Z}^r_j(t) \right\| \le \epsilon(r), \tag{C.31}$$

for $j, j' \in \mathcal{J}$, where $\epsilon(r) \to 0$ as $r \to \infty$ in probability. This gives

$$\sum_{j=1}^{J} \mu_j \int_0^t \hat{Z}^r_j(s) \mathrm{d}s = \mu_1 \int_0^t \hat{Z}^r_1(s) \mathrm{d}s + \epsilon(r) \tag{C.32}$$

and

$$(\hat{X}^r(t))^- = -\hat{Z}^r_1(t) + J\epsilon(r) \tag{C.33}$$

Observe that

$$\begin{aligned} \hat{X}^r(t) &= \hat{X}^r(0) + \left(\frac{A^r(t) - \lambda^r t}{\sqrt{|N^r|}} \right) - \sum_{j=1}^{J} \frac{\left(S_j \left(|N^r| \int_0^t \hat{Z}^r(s) \mathrm{d}s \right) - |N^r| \mu_j \int_0^t \bar{Z}^r_j(s) \mathrm{d}s \right)}{\sqrt{|N^r|}} \\ &\quad - \sum_{j=1}^{J} \mu_j \int_0^t \hat{Z}^r_j(s) \mathrm{d}s + \frac{(\lambda^r - \sum_{j=1}^{J} \mu_j N^r_j)}{\sqrt{|N^r|}} t \end{aligned} \tag{C.34}$$

By (4.2), (4.4), (C.31)- (C.33)

$$\sum_{j=1}^{J} \mu_j \int_0^t \hat{Z}^r_j(s) \mathrm{d}s = \mu_1 \int_0^t \hat{Z}^r_1(s) \mathrm{d}s + J\epsilon(r) = -\mu_1 \int_0^t (\hat{X}^r(s))^- \mathrm{d}s + 2J\epsilon(r) \tag{C.35}$$

Also

$$\left(\frac{A^r(t)-\lambda^r t}{\sqrt{|N^r|}}\right) \Rightarrow W_a \text{ and } \sum_{j=1}^{J} \frac{\left(S_j\left(|N^r|\int_0^t \hat{Z}^r(s)\mathrm{d}s\right) - |N^r|\mu_j \int_0^t \bar{Z}_j^r(s)\mathrm{d}s\right)}{\sqrt{|N^r|}} \Rightarrow W_d, \quad \text{(C.36)}$$

where W_a and W_d are Brownian motions with variances equal to $\bar{\mu}$, by Lemma 1 and since $\lambda^r/|N^r| \to \bar{\mu}$, as $r \to \infty$. We combine (C.34)-(C.36) and appeal to the continuous mapping theorem to complete the proof.

□

Proof of Proposition 4.13. The proof is similar to the proof of Proposition 4.11 above. Let $\{\mathbb{X}^r\}$ be a sequence of MED–LB distributed server pool systems. Assume that (4.2),(4.4) and (4.20) hold. Let $\hat{Z}_\Sigma^r(t) = \sum_{j=1}^J \hat{Z}_j^r(t)$. By Proposition 4.9

$$\left\| \frac{\hat{Q}_j^r(t)}{\beta_j\mu_j} - \frac{\hat{Q}_{j'}^r(t)}{\beta_{j'}\mu_{j'}} \right\|_T \vee \left\| \frac{1}{\beta_j}\hat{Z}_j^r(t) - \hat{Z}_\Sigma^r(t) \right\|_T \leq \epsilon(r)$$

for all $j, j' \in \mathcal{J}$, where $\epsilon(r) \to 0$ as $r \to \infty$ in probability. This gives

$$(\hat{X}^r(t))^- = -\hat{Z}_\Sigma^r(t) + J\epsilon(r).$$

and

$$\sum_{j=1}^{J} \mu_j \int_0^t \hat{Z}_j(s)ds = \sum_{j=1}^{J} \mu_j\beta_j \int_0^t \hat{Z}_\Sigma^r(t) + J\epsilon(r) = \bar{\mu}\int_0^t \hat{Z}_\Sigma^r(t) + J\epsilon(r)$$

The other arguments in the previous proof can be repeated verbatim to conclude the proof.

□

Proof of Theorem 4.1.8. Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed server pool systems. Assume that (4.2),(4.4) and (4.20) hold. We prove the theorem for $J = 2$, the proof for an arbitrary J is similar. By Theorem 4.7 and Proposition 4.11, and Theorem 11.4.5 of [65],

$$\begin{aligned}\left(\hat{Q}_1^r(t), \hat{Z}_1^r(t), \hat{Q}_2^r(t), \hat{Z}_2^r(t)\right) \Rightarrow \Bigg(&\frac{\mu_1\beta_1}{\mu_1\beta_1+\mu_2\beta_2}(X(t))^+, \\ &-(X(t))^-, \frac{\mu_2\beta_2}{\mu_1\beta_1+\mu_2\beta_2}(X(t))^+, 0\Bigg), \end{aligned} \quad \text{(C.37)}$$

as $r \to \infty$ in $\mathbb{C}^4[0,\infty)$. Let

$$\hat{D}_j^r(t) = \frac{D_j^r(t) - \mu_j N_j^r t}{\sqrt{|N^r|}}.$$

Hence,

$$\hat{D}_j^r(t) = \sqrt{|N^r|}\left(\frac{S_j\left(|N^r|\int_0^t \bar{Z}_j^r(s)\mathrm{d}s\right)}{|N^r|} - \mu_j\int_0^t \bar{Z}_j^r(s)\mathrm{d}s\right) + \mu_j\int_0^t \hat{Z}_j^r(s)\mathrm{d}s$$

By virtue of Theorem 11.5.1 of [65] and the continuous mapping theorem

$$\left\{\mu_i\int_0^{\cdot} \hat{Z}_j^r(s)\mathrm{d}s\right\} \tag{C.38}$$

converges weakly to a continuous limit. Also, the sequence

$$\left\{\sqrt{|N^r|}\left(\frac{S_j\left(|N^r|\int_0^t \bar{Z}_j^r(s)\mathrm{d}s\right)}{|N^r|} - \mu_j\int_0^t \bar{Z}_j^r(s)\mathrm{d}s\right)\right\} \tag{C.39}$$

converges weakly by Lemma 1 and the convergence together theorem, hence it is tight. From (C.38) and (C.39) and Theorem 11.6.7 of [65] we have that $\left\{\hat{D}_j^r(\cdot)\right\}$ is tight in uniform topology. Thence, we have again from Theorem 11.6.7 of Whitt, (C.37), and the tightness of the scaled departure processes that the sequence

$$\left\{\left(\hat{Q}_1^r, \hat{Z}_1^r, \hat{D}_1^r, \hat{Q}_2^r, , \hat{Z}_1^r, \hat{D}_2^r\right)\right\}$$

is tight. Thus, there exists a subsequence r_k such that

$$\left(\hat{Q}_1^{r_k}, \hat{Z}_1^{r_k}, \hat{D}_1^{r_k}, \hat{Q}_2^{r_k}, , \hat{Z}_1^{r_k}, \hat{D}_2^{r_k}\right) \Rightarrow \left(\hat{Q}_1, \hat{Z}_1, \hat{D}_1, \hat{Q}_2, \hat{Z}_2, \hat{D}_2\right)$$

as $k \to \infty$ for some process $\left(\hat{Q}_1, \hat{Z}_1, \hat{D}_1, \hat{Q}_2, \hat{Z}_2\hat{D}_2\right)$. Let $a_j^r(t) = A_j^{q,r}(t) + A_j^{s,r}(t)$ be the total number of arrivals to the jth pool by time t. We define the diffusion scaled arrival process, $\hat{a}_j^r$, to queue j by

$$\hat{a}_j^r(t) = \sqrt{|N^r|}\left(\frac{a_j^r(t) - \mu_j N_j^r t}{|N^r|}\right).$$

Since

$$\hat{a}_j^r(t) = -\hat{Q}_j^r(0) - \hat{Z}_j^r(0) + \hat{Q}_j^r(t) + \hat{Z}_j^r(t) + \hat{D}_j^r(t),$$

we have by the continuous mapping theorem that

$$\left(\hat{a}_1^{r_k}, \hat{D}_1^{r_k}, \hat{a}_2^{r_k}, \hat{D}_2^{r_k}\right) \Rightarrow \left(\hat{a}_1, \hat{D}_1, \hat{a}_2, \hat{D}_2\right)$$

where $\hat{a}_i(t) = \hat{Q}_i(t) + \hat{Z}_i(t) + \hat{D}_i(t) - \hat{Q}_i(0) - \hat{Z}_i(0)$, for $i = 1, 2$.

Note that the processes $\hat{a}_i$ and $\hat{D}_i$, for $i = 1, 2$, are continuous a.s. since all the tightness results above hold in uniform topology and $\hat{a}_i(0) = 0$ and $\hat{D}_i(0) = 0$, $i = 1, 2$. By using the corollary in [55] we have that $(\sqrt{N^{r_k}}W_1^{r_k}, \sqrt{N^{r_k}}W_2^{r_k}) \Rightarrow (\hat{W}_1, \hat{W}_2)$ where

$$\hat{W}_i(t) = \frac{\left[\hat{X}\right]^+}{\mu} \text{ for } i = 1, 2 \tag{C.40}$$

and $\hat{X}(t) = \sum_{j=1}^{2} \hat{Q}_i(t))$. Since the limit $\hat{X}$ is independent of the subsequence chosen we have the convergence of the waiting time processes for each pool. To prove the convergence of W^r, we note that

$$(W_1^r(t) \wedge W_2^r(t)) \leq W^r(t) \leq (W_1^r(t) \vee W_2^r(t)) \text{ a.s.}$$

for all r and $t \geq 0$. Hence

$$\begin{aligned} 0 &\leq \sqrt{|N^r|}\,(W_1^r(t) \vee W_2^r(t)) - \sqrt{|N^r|}W^r(t) \\ &\leq \sqrt{|N^r|}\,(W_1^r(t) \vee W_2^r(t)) - \sqrt{|N^r|}\,(W_1^r(t) \wedge W_2^r(t)) \text{ a.s.} \end{aligned}$$

The last term converges to zero by continuous mapping theorem and from (C.40). But weak convergence to a deterministic limit implies convergence in probability, see, for example, [11]. Therefore, $\sqrt{|N^r|}\,(W_1^r(t) \vee W_2^r(t)) - \sqrt{|N^r|}W^r(t)$ converges to zero in probability. We have the convergence of W^r to $[X]^+/\mu$ by virtue of Theorem 3.1. of [11]. □

The proof of Theorem 4.1.8 is similar.

C.3.3 Proofs of the results in Section 4.1.4.4

C.3.3.1 Proof of Theorem 4.1.11

Fix a non-idling routing policy $\pi \in \Pi$ and let $\bar{\mathbb{X}}_\pi^{r,n}(\cdot) = \mathbb{X}_\pi^r(n\cdot)/n$. This scaling is known as the conventional fluid scaling. (These are not related to the fluid scaling that are discussed in Section 4.1.4.3 and will not be used elsewhere in this paper outside this section.) Similar

to [19], $\bar{\mathbb{X}}_\pi^r \in \mathbb{D}^{8J+1}$ is said to be a fluid limit of $\{\mathbb{X}_\pi^{r,n}\}$ if there exists a subsequence $\{n_k\}$ of $\{n\}$ and $\omega \in \Omega$ satisfying

$$\lim_{t\to\infty} E^r(\lambda^r t)/t = \lambda^r \text{ and } \lim_{t\to\infty} S_j(t)/t = \mu_j,$$

for all $j \in \mathcal{J}$, such that

$$\lim_{k\to\infty} \bar{\mathbb{X}}_\pi^{r,n_k}(\cdot,\omega) = \bar{\mathbb{X}}_\pi(\cdot)$$

u.o.c. It can be shown as in [18] that fluid limits for queueing systems with multiple servers exist and satisfy the following equations for all $t \geq 0$.

$$\begin{aligned}
&\lambda^r t = \sum_{j=1}^{J} (\bar{A}_j^{s,r}(t) + \bar{A}_j^{q,r}(t)),\\
&\bar{Q}_j^r(t) = \bar{Q}^r(0) + \bar{A}_j^{s,r}(t) + \bar{A}_j^{q,r}(t) - \mu_j \bar{T}_j(t), \text{for all } j \in \mathcal{J}\\
&\bar{T}_j^r(t) + \bar{Y}_j^r(t) = N_j^r \mu_j, \text{for all } j \in \mathcal{J}\\
&\bar{Y}_j^r(t) \text{ can only increase when } \bar{Q}_j^r(t) = 0, \text{for all } j \in \mathcal{J}\\
&\bar{A}_j^{s,r}, \bar{A}_j^{r,q}, \bar{T}_j^r, \text{ and } \bar{Y}_j^r \text{ are non-decreasing, for all } j \in \mathcal{J}
\end{aligned} \tag{C.41}$$

We note that $\bar{Z}_j^r(t) = 0$ for all $t \geq 0$, since $\bar{Z}_j^{r,n_k}(t) \leq |N^r|/n_k$ and so goes to zero as $k \to \infty$ for fixed r. It is clear from these equations that every fluid limit is absolutely continuous hence differentiable almost everywhere.

In this section we show that the fluid model of $\bar{\mathbb{X}}_\pi^r$ is stable (see Definition 4.1 of [18]) when $\pi \in \Pi$ and (4.59) holds. Then we appeal to Theorem 4.2 in [18]. This theorem is applicable only to single server systems but can be extended to cover the systems with multiple servers. We omit the proof since it follows straightforwardly from the analysis in [18].

Proof. Fix a routing policy $\pi \in \Pi$ and $r > 0$. Let $\bar{\mathbb{X}}^r$ be a fluid limit of $\{\mathbb{X}_\pi^{r,n}\}$. Fix a regular point $t > 0$.

We first show that fluid limits of $\{\mathbb{X}_\pi^{r,n}\}$ satisfy

$$\dot{\bar{A}}_j^{s,r}(t) + \dot{\bar{A}}_j^{q,r}(t) = 0 \text{ when } \bar{Q}_j^r(t) > 0 \text{ and } \bar{Q}_{j'}^r(t) = 0 \text{ for some } j' \in \mathcal{J} \tag{C.42}$$

for any $j \in \mathcal{J}$. To prove this, assume that $\bar{Q}^r_j(t) > 0$ and $\bar{Q}^r_{j'}(t) = 0$. By continuity of $\bar{Q}^r$, there exists a $\delta > 0$ such that

$$\bar{Q}^r_j(s) > 2\epsilon \text{ and } \bar{Q}^r_{j'}(s) < \epsilon/(2a^r_\pi) \text{ for all } s \in [t-\delta, t+\delta] \text{ and for some } \epsilon > 0.$$

Let $\mathbb{X}^{r,n_k}_\pi(\cdot,\omega) \to \bar{\mathbb{X}}^r_\pi(\cdot)$ u.o.c. as $k \to \infty$. Then, for k large enough,

$$\bar{Q}^{r,n_k}_j(s) > \epsilon \text{ and } \bar{Q}^{r,n_k}_{j'}(s) < \epsilon/a^r_\pi \text{ for all } s \in [t-\delta, t+\delta].$$

Hence, $Q^r_j(n_k s) > a^r_\pi Q^r_{j'}(n_k s)$ for all $s \in [t-\delta, t+\delta]$. Therefore, by (4.29), $A^{q,r}_j$ is flat on $s \in [n_k(t-\delta), n_k(t+\delta)]$. Note that $A^r_{s,j}$ is flat on $s \in [n_k(t-\delta), n_k(t+\delta)]$ since $Q^r_j(n_k s) > 0$. This gives (C.42).

Let $\bar{Q}^r_\Sigma(t) = \sum_{j=1}^J \bar{Q}^r_j(t)$ and assume that $\bar{Q}^r_\Sigma(t) > 0$. First assume that there exists $j' \in \mathcal{J}$ such that $\bar{Q}^r_{j'}(t) = 0$. Then, since $\bar{Q}^r_{j'}$ is absolutely continuous, and differentiable at time t and attains a minimum at $t > 0$, $\dot{\bar{Q}}^r_{j'}(t) = 0$. Hence

$$\dot{\bar{Q}}^r_\Sigma(t) = \sum_{j \in \mathcal{J}: \bar{Q}^r_j(t) > 0} \dot{\bar{Q}}^r_j(t) + \sum_{j \in \mathcal{J}: \bar{Q}^r_j(t) = 0} \dot{\bar{Q}}^r_j(t) \le -\mu_{\min} N^r_{\min}$$

by (C.41) and (C.42) .

If $\bar{Q}^r_j(t) > 0$ for all $j \in \mathcal{J}$, then

$$\dot{\bar{Q}}^r_\Sigma(t) = \lambda^r - \sum_{j \in \mathcal{J}} \mu_j N^r_j < -\epsilon$$

for some $\epsilon > 0$ by (4.59) and (C.41).

Hence, if $\bar{Q}^r_\Sigma(t) > 0$, and t is a regular point then $\dot{\bar{Q}}^r_\Sigma(t) < -(\epsilon \wedge \mu_{\min} N^r_{\min})$ so the fluid model of $\bar{\mathbb{X}}^r_\pi$ is stable by Lemma 5.2 of [18]. We conclude the existence of a stationary distribution of (Q^r, Z^r) by Theorem 4.2 of [18]. □

C.3.3.2 Proof of Theorem 4.1.12

Choose $t_0 > 0$ such that for large enough r

$$\mathbb{E}_x\left[\exp\left\{-\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\right\}\right] \mathbb{E}_x\left[\exp\left\{2\frac{\|A^r(t) - \lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right] \prod_{j=1}^J \mathbb{E}_x\left[\exp\left\{2\frac{\sum_{j=1}^J \|S_j(t) - \mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right] < 1/2. \tag{C.43}$$

Note that the existence of such t_0 and r is guaranteed by Lemma 10.

Let $x_i \in \mathbb{R}^J$ for $i = 1, 2$ and $x = (x_1, x_2)$. We define $\Phi_1^r(x) : \mathbb{R}^{2J} \to \mathbb{R}$ by

$$\Phi_1^r(x) = \exp\left\{(|N^r|t_0)^{-1/2}\varphi_1^r(x_2)\right\}, \tag{C.44}$$

where t_0 is as chosen in (C.43) and φ_1^r is defined as in (4.36). We show using Theorem 4.1.6 that Φ_1^r is a geometric Lyapunov function; see Definition 2 in Gamarnik and Zeevi [27], then we appeal to Theorem 5 in the same paper to complete the proof.

Recall that $\mathbb{P}_{\pi^r}$ denotes the stationary distribution of (Q^r, Z^r) under the routing policy π. We denote the expectation operator with respect to this distribution by $\mathbb{E}_{\pi^r}$. We set

$$\mathbb{E}_x[\cdot] = \mathbb{E}[\cdot|Q^r(0) = x_1, Z^r(0) = x_2]$$

for $x = (x_1, x_2)$, $x_i = (x_{i1}, \ldots, x_{iJ}) \in \mathbb{R}^J$, for $i = 1, 2$, with $x_1 \geq 0$, $0 \leq x_{2j} \leq N_j^r$ and $x_{1j}(N_j^r - x_{2j}) = 0$ for all $j \in \mathcal{J}$.

Proposition C.3. *Let Φ_1^r be defined as in (C.44). There exists $t_0 > 1$ and $0 < \gamma < 1$ such that for r large enough*

$$\sup_{x\in\mathbb{R}^{2J}:\Phi_1(x)>\kappa}\left\{\mathbb{E}_x\left[\Phi_1^r(Q^r(t_0), Z^r(t_0))/\Phi_1^r(x)\right]\right\} \leq \gamma \quad \textit{and} \tag{C.45}$$

$$\phi_1^r(t_0) \stackrel{\Delta}{=} \sup_{x\in\mathbb{R}^{2J}}\left\{\mathbb{E}_x\left[\Phi_1^r(Q^r(t_0), Z^r(t_0))/\Phi_1^r(x)\right]\right\} < \infty, \tag{C.46}$$

where $\kappa = \exp\left\{\frac{4\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}}{\mu_{\min}\wedge 1}\right\}$.

Proof. Fix a $t_0 > 1$ that satisfies (C.43). Note that if $\Phi_1^r(x) > \exp\left\{\frac{4\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}}{\mu_{\min}\wedge 1}\right\}$ then $\varphi_1^r(Z^r(0)) > \frac{4\theta\sqrt{\lambda^r}t_0}{\mu_{\min}\wedge 1}$. Hence, by (4.38), for r large enough

$$\sup_{x\in\mathbb{R}^{2J}:\Phi_1(x)>\kappa}\{\mathbb{E}_x\left[\Phi_1^r(Q^r(t_0), Z^r(t_0))/\Phi_1^r(x)\right]\} \leq 2\mathbb{E}_x\left[\exp\left\{-\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\right\}\right]$$
$$\mathbb{E}_x\left[\exp\left\{2\frac{\|A^r(t) - \lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\prod_{j=1}^J \mathbb{E}_x\left[\exp\left\{2\frac{\sum_{j=1}^J \|S_j(t) - \mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right].$$

This gives (C.45) by (C.43).

If $\Phi_1^r(x) \leq \exp\left\{\frac{4\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}}{\mu_{\min}\wedge 1}\right\}$ then $\phi_1^r(Q^r(0)) \leq \frac{4\theta\sqrt{\lambda^r}t_0}{\mu_{\min}\wedge 1}$. Hence, by (4.39),

$$\sup_{x\in\mathbb{R}^{2J}:\Phi_1(x)>\kappa}\{\mathbb{E}_x\left[\Phi_1^r(Q^r(t_0),Z^r(t_0))/\Phi_1^r(x)\right]\}\leq\mathbb{E}_x\left[\exp\left\{\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\right\}\right]$$
$$\mathbb{E}_x\left[\exp\left\{2\frac{\|A^r(t)-\lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\prod_{j=1}^{J}\mathbb{E}_x\left[\exp\left\{2\frac{\sum_{j=1}^{J}\|S_j(t)-\mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right].$$

We get (C.46) by virtue of Lemma 10.

□

In order to define the Lyapunov function for the queue length process we need the following result.

Lemma 8. *There exist $t_0>0$ and r_0 such that for $r>r_0$*

$$\exp\left\{-\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}+\sqrt{t_0}2(\mu_{\max}\vee 1)J/\sqrt{|N^r|}\right\}\left(\mathbb{E}_x\left[\exp\{4\sqrt{t_0}(\mu_{\max}\vee 1)\zeta^r(t_0)/\sqrt{|N^r|}\}\right]\right.$$
$$\left.+\mathbb{E}_x\left[\exp\left\{4\frac{\|A^r(t)-\lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\prod_{j=1}^{J}\mathbb{E}_x\left[\exp\left\{8\frac{\sum_{j=1}^{J}\|S_j(\mu_j t)-\mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\right)<1/2. \tag{C.47}$$

Proof. By Lemma 10 there exists r_1 such that for any $t_1>0$

$$\mathbb{E}_x\left[\exp\left\{4\frac{\|A^r(t)-\lambda^r t\|_{t_1}}{\sqrt{|N^r|t_1}}\right\}\right]\prod_{j=1}^{J}\mathbb{E}_x\left[\exp\left\{8\frac{\sum_{j=1}^{J}\|S_j(\mu_j t)-\mu_j t\|_{|N^r|t_1}}{\sqrt{|N^r|t_1}}\right\}\right]<B_1/2 \tag{C.48}$$

for some $J+2<B_1<\infty$ and all $r>r_1$. Now, for large enough $t_2>t_1$ and r_2, and for $r>r_2>r_1$

$$\exp\{-\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_2}+\sqrt{t_2}2(\mu_{max}\vee 1)J/\sqrt{|N^r|}\}<\frac{1}{4B_1} \tag{C.49}$$

By Lemma 11 we can choose r_3 large enough so that for $r>r_3>r_2$

$$\mathbb{E}_x\left[\exp\{4\sqrt{t_2}(\mu_{\max}\vee 1)\zeta^r(t_0')/\sqrt{|N^r|}\}\right]<B_1/2. \tag{C.50}$$

We get (C.47) by combining (C.48)-(C.50). □

For t_0 chosen as in Lemma 8, we define the function $\Phi_2^r:\mathbb{R}^{2J}\to\mathbb{R}$ by

$$\Phi_2^r(x)=\exp\left\{(|N^r|t_0)^{-1/2}\varphi_2^r(x_1)\right\} \tag{C.51}$$

for $x=(x_1,x_2)$ and $x_i\in\mathbb{R}^J$, $i=1,2$, where φ_2^r is defined in (4.1.7).

Proposition C.4. *Let Φ_2^r be defined as in (C.51). There exists $t_0 > 1$ and $0 < \gamma < 1$ such that for r large enough*

$$\sup_{x\in\mathbb{R}^{2J}:\Phi_2^r(x)>\kappa} \{\mathbb{E}_x [\Phi_2^r(Q^r(t_0), Z^r(t_0))/\Phi_2^r(x)]\} \leq \gamma \text{ and} \tag{C.52}$$

$$\phi_2^r(t_0) \stackrel{\Delta}{=} \sup_{x\in\mathbb{R}^{2J}} \{\mathbb{E}_x [\Phi_2^r(Q^r(t_0), Z^r(t_0))/\Phi_2^r(x)]\} < \infty, \tag{C.53}$$

where $\kappa = \exp\{\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\}$.

Proof. Choose t_0 and r_0 as in Lemma 8. Note that if $\Phi_2^r(x) > \exp\{\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\}$ then $\varphi_2^r(Q^r(0)) > \theta\sqrt{\lambda^r}t_0$. Hence, by (4.42),

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^{2J}:\Phi_2^r(x)>\kappa} \left\{\mathbb{E}_x [\Phi_2^r(Q^r(t_0), Z^r(t_0))/\Phi_2^r(x)]\right\} \leq \\
&\qquad \exp\left\{-\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0} + \sqrt{t_0}2(\mu_{max}\vee 1)J/\sqrt{|N^r|}\right\} \\
&\qquad \left(\mathbb{E}_x\left[\exp\{4\sqrt{t_0}(\mu_{\max}\vee 1)\zeta^r(t_0)/\sqrt{|N^r|}\}\right]\right. \\
&\qquad \left.+\mathbb{E}_x\left[\exp\left\{4\frac{\|A^r(t)-\lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\prod_{j=1}^J \mathbb{E}_x\left[\exp\left\{8\frac{\sum_{j=1}^J \|S_j(t)-\mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\right) \\
&\qquad <1,
\end{aligned}$$

where the last inequality follows from Lemma 8. This gives (C.52).

Now assume that $\Phi_2^r(x) \leq \exp\{\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}\}$. By (4.43),

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^{2J}:\Phi_2^r(x)>\kappa} \left\{\mathbb{E}_x [\Phi_2^r(Q^r(t_0), Z^r(t_0))/\Phi_2^r(x)]\right\} \leq \\
&\qquad \exp\left\{2J+\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0} + \sqrt{t_0}2(\mu_{max}\vee 1)J/\sqrt{|N^r|}\right\} \\
&\qquad \left(\mathbb{E}_x\left[\exp\{4\sqrt{t_0}(\mu_{\max}\vee 1)\zeta^r(t_0)/\sqrt{|N^r|}\}\right]\right. \\
&\qquad \left.+\mathbb{E}_x\left[\exp\left\{4\frac{\|A^r(t)-\lambda^r t\|_{t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\prod_{j=1}^J \mathbb{E}_x\left[\exp\left\{8\frac{\sum_{j=1}^J \|S_j(t)-\mu_j t\|_{|N^r|t_0}}{\sqrt{|N^r|t_0}}\right\}\right]\right)
\end{aligned}$$

This gives (C.53) by Lemmas 10 and 11. □

Theorem 4.1.12. Let $\pi \in \Pi$ be a non-idling routing policy. We claim that for r large

enough,

$$\mathbb{P}_{\pi^r}\left\{\sum_{j=1}^{J}\frac{Q_j^r(0)}{\sqrt{|N^r|}}>s\right\}\leq c_1\exp\{-c_2 s\}\text{ and} \tag{C.54}$$

$$\mathbb{P}_{\pi^r}\left\{\sum_{j=1}^{J}\frac{N_j^r-Z_j^r(0)}{\sqrt{|N^r|}}>s\right\}\leq c_1\exp\{-c_2 s\} \tag{C.55}$$

for some $c_1, c_2 > 0$.

Let t_0 be given as in Proposition C.3. By Theorem 5 of [27] and Theorem C.3,

$$\mathbb{E}_{\pi^r}\left[\Phi_1^r(Q^r(0),Z^r(0))\right]\leq\frac{\phi_1^r(t_0)\kappa}{1-\gamma}<c_0\exp\left\{\frac{4\theta\sqrt{\lambda^r/|N^r|}\sqrt{t_0}}{\mu_{\min}\wedge 1}\right\},$$

for some $c_0 > 0$. By Markov's inequality

$$\mathbb{P}_{\pi^r}\left\{\exp\left\{\frac{\varphi_1(Q^r(0))}{\sqrt{|N^r|}t_0}\right\}>\exp\{s\}\right\}\leq\exp\{-s\}\mathbb{E}_{\pi^r}\left[\Phi_1^r(Q^r(0))\right]<c_1\exp\{-s\}$$

for some $c_0 > 0$. This gives (C.54). The second inequality (C.55) is proved similarly using Proposition C.4.

Theorem 4.1.12 immediately follows from (C.54) and (C.55) since both $\frac{Q_j^r(0)}{\sqrt{|N^r|}}$ and $\frac{N_j^r-Z_j^r(0)}{\sqrt{|N^r|}}$ are nonnegative. □

C.3.3.3 Proofs of Theorems 4.1.13 and 4.1.14

Proof. The proof is similar to that of Theorem 8 in [27]. Assume that (4.2) and (4.4) hold.

By Theorem 4.1.11, $(Q^r(\infty), Z^r(\infty))$ exists for each r, and by Theorem 4.1.12 the sequence $\{(\hat{Q}^r(\infty)\,,\hat{Z}^r(\infty))\}$ is tight. Therefore, every subsequence of $\{(\hat{Q}^r(\infty),\hat{Z}^r(\infty))\}$ has a convergent subsequence. Hence, it is enough to show that every convergent subsequence of $\{(\hat{Q}^r(\infty),\hat{Z}^r(\infty))\}$ converges to the same limit [11, page 59] and this limit has the same distribution with the stationary distribution of $(\hat{Q},\hat{Z})$ that is given by

$$\begin{aligned}\hat{Q}_j(\infty) &= \frac{\mu_j\beta_j}{\sum_{j=1}^{J}\mu_j\beta_j}(\hat{X}(\infty))^+,\\ \hat{Z}_1(\infty) &= (\hat{X}(\infty))^-\text{ and }\hat{Z}_j(\infty)=0\text{ for }j\geq 2.\end{aligned}$$

by Remark 4.12, where $\hat{X}(\infty)$ has the density given by (4.13).

For notational simplicity let $\{(\hat{Q}^r(\infty), \hat{Z}^r(\infty))\}$ be a convergent subsequence and denote the weak limit by $(\hat{Q}^*(\infty), \hat{Z}^*(\infty))$. Let $\{\mathbb{X}^r\}$ be a sequence of MED–FSF distributed systems with

$$(\hat{Q}^r(0), \hat{Z}^r(0)) \sim (\hat{Q}^r(\infty), \hat{Z}^r(\infty)), \tag{C.56}$$

i.e., $\{(\hat{Q}^r(0), \hat{Z}^r(0))\}$ has the stationary distribution. Then, by Theorem 4.7, for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{ \sup_{L^r/\sqrt{|N^r|} \le t \le T} \left| \frac{\hat{Q}^r_j(t)}{\beta_j \mu_j} - \frac{\hat{Q}^r_{j'}(t)}{\beta_{j'} \mu_{j'}} \right| \vee \left| \sum_{j=2}^{J} \hat{Z}^r_j(t) \right| > \epsilon \right\} \to 0, \tag{C.57}$$

as $r \to \infty$.

Let

$$(q^r(\cdot), z^r(\cdot)) = (Q^r(\cdot + L^r/\sqrt{|N^r|}), Z^r(\cdot + L^r/\sqrt{|N^r|})),$$

By (C.56)

$$(q^r(0), z^r(0)) \sim (\hat{Q}^r(\infty), \hat{Z}^r(\infty)), \tag{C.58}$$

since $(\hat{Q}^r(\infty), \hat{Z}^r(\infty))$ is the unique stationary distribution. Therefore, $\{(q^r(0), z^r(0))\}$ satisfies the conditions of Proposition 4.11 by (C.57) and (C.58). Hence

$$(q^r, z^r) \Rightarrow (\hat{Q}, \hat{Z}) \tag{C.59}$$

where $\hat{Q}$ and $\hat{Z}$ are given by (4.52) and (4.53), respectively, and $(\hat{Q}, \hat{Z}(0)) \sim (\hat{Q}^*(\infty), \hat{Z}^*(\infty))$. Fix $t > 0$. Then $(q^r(t), z^r(t)) \sim (\hat{Q}^r(\infty), \hat{Z}^r(\infty))$, again by stationarity of $(\hat{Q}^r(\infty), \hat{Z}^r(\infty))$. Since $\{(q^r(t), z^r(t))\}$ converges weakly to $(\hat{Q}(t), \hat{Z}(t))$ by (C.59), $(\hat{Q}(t), \hat{Z}(t)) \sim (\hat{Q}^*(\infty), \hat{Z}^*(\infty))$. Hence $(\hat{Q}^*(\infty), \hat{Z}^*(\infty))$ is the unique stationary distribution of $(\hat{Q}, \hat{Z})$.

The weak convergence of $W^r(\infty)$ to $\hat{X}(\infty)/\mu$ can be proved similarly by starting the each process in its steady state and repeating the arguments in the proof of Theorem 4.1.8.

The proof of Theorem 4.1.14 is similar. □

C.4 Auxiliary Results

Lemma 9. *Let M be a renewal process with interarrival times given by the sequence of i.i.d. random variables $\{m(i) : i = 1, 2, \ldots\}$. Assume that $\mathbb{P}\{m(1) = 0\} = 0$. For $t_0 > 0$, let*

$$\mathcal{M} = \cap_{r=1}^{\infty} \{\|M(t) - M(t-)\|_{n^r t_0} \leq 1\},$$

where $\{n^r\}$ is a sequence of real numbers with $n^r = O(|N^r|)$. Then $\mathbb{P}(\mathcal{M}) = 1$.

Proof. Fix $r > 0$ and $t_0 > 0$. Then $\mathbb{P}\{M(n^r t) < \infty\} = 1$. Let $\mathcal{U} = \cup_{i=1}^{\infty}\{m(i) > 0\}$. By the assumption of the lemma $\mathbb{P}\{\mathcal{U}\} = 1$. Define for $k = 1, 2, \ldots,$

$$\mathcal{M}_k^r = \{M(n^r t_0) < k\} \text{ and}$$
$$\mathcal{M}^r = \{\|M(t) - M(t-)\|_{n^r t_0} \leq 1\}.$$

Observe that

$$\cup_{k=1}^{\infty} (\mathcal{M}_k^r \cap \mathcal{U}) \subset \mathcal{M}^r$$

Hence,

$$\mathbb{P}\{\mathcal{M}^r\} \geq \mathbb{P}\{\cup_{k=1}^{\infty} (\mathcal{M}_k^r) \cap \mathcal{U}\} = 1.$$

Since $\mathcal{M} = \cap_{r=1}^{\infty} \mathcal{M}^r$, $\mathbb{P}\{\mathcal{M}\} = 1$. □

Lemma 10. *Let M be a poisson process with rate $\gamma > 0$, $\{n^r\}$ be a sequence of nonnegative real numbers such that $n^r = O(|N^r|)$, and $\alpha > 0$. Then, there exists $B_1 < \infty$ such that for every $0 < t_0 < \infty$*

$$\limsup_{r \to \infty} \mathbb{E}\left[\exp\left\{\alpha \sup_{0 \leq t \leq n^r t_0} \frac{|M(t) - \gamma t|}{\sqrt{|N^r| t_0}}\right\}\right] < B_1. \tag{C.60}$$

Remark C.5. If M is a Poisson process with rate 1, then $M'(\cdot) = M(\gamma \cdot)$ is a Poisson process with rate γ, hence (C.60) also holds for the process $M(\gamma \cdot)$

Proof. Fix $\alpha > 0$ and $t_0 > 0$. Since $n^r = O(|N^r|)$, $n^r/|N^r| < a$ for r large enough and for some $a > 0$.

As in the proof of Lemma 1 in [27]

$$\limsup_{n\to\infty} \mathbb{E}\left[\sup_{0\le t\le n} \exp\left\{2\alpha\sqrt{a}n^{-1/2}|M(t)-\gamma t|\right\}\right] < B_2, \tag{C.61}$$

for some $B_2 < \infty$. We have

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\alpha \sup_{0\le t\le n^r t_0} \frac{|M(t)-\gamma t|}{\sqrt{|N^r|t_0}}\right\}\right] &\le \mathbb{E}\left[\exp\left\{\alpha\sqrt{a} \sup_{0\le t\le a|N^r|t_0} \frac{|M(t)-\gamma t|}{\sqrt{a|N^r|t_0}}\right\}\right] \\ &= \mathbb{E}\left[\sup_{0\le t\le a|N^r|t_0} \exp\left\{\alpha\sqrt{a}\frac{|M(t)-\gamma t|}{\sqrt{a|N^r|t_0}}\right\}\right]\end{aligned}$$

This together with (C.61) gives the desired result. □

Lemma 11. *Let ζ^r be defined as in (4.40). For every $t_0 > 0$ and $\alpha > 0$, there exists $r_{t_0} > 0$ such that for $r > r_{t_0}$*

$$\mathbb{E}\left[\exp\left\{4(\mu_{max}\vee 1)t_0\frac{\zeta^r(t_0)}{\sqrt{|N^r|}}\right\}\right] < (J+1) + 2\alpha.$$

Proof. For notational simplicity we assume that $\mu_{\max} > 1$. Choose r large enough so that $N^r_{\min}\mu_{\min} - \theta\sqrt{\lambda^r} > 2|N^r|B_3$ for some $B_3 > 0$. Then, for such r,

$$\begin{aligned}\mathbb{E}\Bigg[\exp\{4\mu_{\max}t_0\frac{\zeta^r(t_0)}{\sqrt{|N^r|}}\}\Bigg] \le& \\ &\mathbb{E}\left[\exp\left\{8\mu_{\max}t_0 \sup_{0\le s_1\le s_2\le t_0}\left\{-\sqrt{|N^r|}B_3(s_2-s_1) + \frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right\}\right] \\ &+\mathbb{E}\Bigg[\exp\Bigg\{8\mu_{max}t_0 \sup_{\substack{0\le s_1\le s_2\le t_0 \\ \nu_1,\ldots,\nu_J:\nu_j+(s_2-s_1)\le t_0}}\Bigg\{-\sqrt{|N^r|}B_3(s_2-s_1) \\ &+\frac{\sum_{j=1}^J\left|\breve{S}_j\left(\left(\upsilon_j+(s_2-s_1)\right)|N^r|\right)-\breve{S}_j\left(\upsilon_j|N^r|\right)\right|}{\sqrt{|N^r|}}\Bigg\}\Bigg\}\Bigg]\end{aligned} \tag{C.62}$$

We show that the first term on the right hand side (RHS) above is bounded by $1+\alpha$, it can similarly be shown that the second term is bounded by $J+\alpha$. First observe that for

any $\epsilon > 0$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{8\mu_{max}t_0 \sup_{0\le s_1\le s_2\le t_0}\left\{-\sqrt{|N^r|}B_3(s_2-s_1)+\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right\}\right] \\
&\le \mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(-\sqrt{|N^r|}B_3\epsilon+\sup_{0\le s_1\le s_2\le t_0}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right)\right\}\right] \\
&\quad+\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right)\right\}\right]
\end{aligned} \tag{C.63}$$

We next show that we can choose $\epsilon > 0$ so that the terms on the RHS of (C.63) are bounded by $1+\alpha$. Let $\hat{A}^r(t) = \breve{A}^r(t)/\sqrt{|N^r|} = (A^r(t)-\lambda^r t)/\sqrt{|N^r|}$. Then, $A^r \Rightarrow W_a$, as $r\to\infty$, where W_a is a Brownian motion with variance λ. By the continuous mapping theorem

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\} > \frac{\log u}{8\mu_{max}t_0}\right\} \\
&\qquad\to \mathbb{P}\left\{\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\} > \frac{\log u}{8\mu_{\max}t_0}\right\}.
\end{aligned}$$

Hence, by virtue of the dominated convergence theorem we have that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right)\right\}\right] \\
&\quad\to \mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\right]
\end{aligned} \tag{C.64}$$

as $r\to\infty$. By a.s. continuity of a Brownian motion,

$$\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\to 1$$

a.s. as $\epsilon\to 0$. Since for every $\epsilon > 0$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\right] \\
&\quad\le \mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{0\le s_1\le s_2\le t_0}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\right] < \infty
\end{aligned}$$

Another application of the dominated convergence theorem yields that

$$\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\right]\to 1$$

as $\epsilon\to 0$. Thus, for every $\alpha>0$ we can find $\epsilon>0$ such that

$$\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\{|W_a(s_2)-W_a(s_1)|\}\right)\right\}\right]<1+\alpha/4.$$

Hence, from (C.64), we can find $r_0>0$ such that for all $r>r_0$

$$\mathbb{E}\left[\exp\left\{8\mu_{max}t_0\left(\sup_{\substack{0\le s_1\le s_2\le t_0\\ |s_1-s_2|<\epsilon}}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right)\right\}\right]<1+\alpha/3. \tag{C.65}$$

Next, observe that

$$\begin{aligned}\mathbb{E}\Bigg[\exp\Bigg\{8\mu_{max}t_0\sup_{0\le s_1\le s_2\le t_0}\left\{\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\Bigg\}\Bigg]&\\ \le\mathbb{E}\left[\exp\left\{16\mu_{max}t_0\sup_{0\le t\le t_0}\left\{\frac{|A^r(t)-\lambda^r t|}{\sqrt{|N^r|}}\right\}\right\}\right].&\end{aligned} \tag{C.66}$$

Note that, for large enough r, the term on the RHS of (C.66) is bounded by Lemma 10. Hence, by selecting r large enough we can make the first term on the RHS of (C.63) arbitrarily small for any $\epsilon>0$.

Fix $\epsilon>0$ and choose r_1 large enough so that (C.65) holds for every $r>r_1$. Now, for this choice of $\epsilon>0$ choose r_2 large enough so that for $r>r_2$ the first term on the RHS of (C.63) is bounded by $\alpha/4$. Therefore, for every $r>r_2$

$$\mathbb{E}\quad\left[\exp\left\{8\mu_{\max}t_0\sup_{0\le s_1\le s_2\le t_0}\left\{-\sqrt{|N^r|}B_3(s_2-s_1)+\frac{|\breve{A}^r(s_2)-\breve{A}^r(s_1)|}{\sqrt{|N^r|}}\right\}\right\}\right]<1+\alpha$$

by (C.63).

Next we outline the details how the second term on the RHS of (C.62) is handled. Observe that

$$\begin{aligned}&\mathbb{E}\Bigg[\exp\Bigg\{8\mu_{\max}t_0\sup_{\substack{0\le s_1\le s_2\le t_0\\ \nu_1,\ldots,\nu_J:\nu_j+(s_2-s_1)\le t_0}}\Bigg\{-\sqrt{|N^r|}B_3(s_2-s_1)\\ &\qquad+\frac{\sum_{j=1}^J\left|\breve{S}_j\left(\left(\upsilon_j+(s_2-s_1)\right)|N^r|\right)-\breve{S}_j\left(\upsilon_j|N^r|\right)\right|}{\sqrt{|N^r|}}\Bigg\}\Bigg\}\Bigg]\\ &\le\sum_{j=1}^J\mathbb{E}\left[\exp\left\{8J\mu_{\max}t_0\sup_{0\le s_1^j\le s_2^j\le t_0}\left\{-\frac{\sqrt{|N^r|}B_3}{J}(s_2^j-s_1^j)+\frac{\left|\breve{S}_j\left(|N^r|s_2^j\right)-\breve{S}_j\left(|N^r|s_1^j\right)\right|}{\sqrt{|N^r|}}\right\}\right\}\right].\end{aligned}$$

It can be shown as above that for large enough r

$$\mathbb{E}\left[\exp\left\{ \quad 8J\mu_{\max}t_0 \sup_{0\le s_1^j\le s_2^j\le t_0}\left\{-\frac{\sqrt{|N^r|}B_3}{J}(s_2^j-s_1^j)+\frac{\left|\breve{S}_j\left(|N^r|s_2^j\right)-\breve{S}_j\left(|N^r|s_1^j\right)\right|}{\sqrt{|N^r|}}\right\}\right\}\right] < 1+\alpha/J,$$

for each $j \in \mathcal{J}$. □

BIBLIOGRAPHY

[1] Armony, M., "Dynamic routing in large-scale service systems with heterogenous servers," *Queueing Systems*, vol. 51, pp. 287–329, 2005.

[2] Armony, M. and Maglaras, C., "Contact centers with a call-back option and real-time delay information," *Operations Research*, vol. 52, pp. 527–545, 2004.

[3] Armony, M. and Maglaras, C., "On customer contact centers with a call-back option: Customer decisions, routing rules and system design," *Operations Research*, vol. 52, pp. 271–292, 2004.

[4] Ata, B. and Kumar, S., "Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies," *Ann. Appl. Probab.*, vol. 15, pp. 331–391, 2005.

[5] Atar, R., "A diffusion model of scheduling control in queueing systems with many servers," *The Annals of Applied Probability*, vol. 15, pp. 820–852, 2005.

[6] Atar, R., "Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic," *(working paper)*, 2005.

[7] Atar, R., Mandelbaum, A., and Reiman, M., "Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic," *The Annals of Applied Probability*, vol. 14, pp. 1084–1134, 2004.

[8] Atar, R., Mandelbaum, A., and Shaikhet, G., "Queueing systems with many servers: null controllability in heavy traffic," *(working paper)*, 2005.

[9] Bell, S. L. and Williams, R. J., "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy," *Annals of Applied Probability*, vol. 11, pp. 608–649, 2001.

[10] Bell, S. L. and Williams, R. J., "Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy," *Electronic J. of Probability*, vol. 10, pp. 1044–1115, 2005.

[11] Billingsley, P., *Convergence of probability measures*. New York: Wiley, 1999.

[12] Borovkov, A., "On limit laws for service processes in multi-channel systems," *Siberian Math.J.*, vol. 8, pp. 983–1004, 1967.

[13] Bramson, M., "State space collapse with application to heavy traffic limits for multiclass queueing networks," *Queueing Systems: Theory and Applications*, vol. 30, pp. 89–148, 1998.

[14] Bramson, M. and Dai, J. G., "Heavy traffic limits for some queueing networks," *Annals of Applied Probability*, vol. 11, pp. 49–90, 2001.

[15] Chen, H. and Yao, D., *Fundamentals of queueing networks : Performance, asymptotics, and optimization*. New York: Springer, 2001.

[16] Chung, K. L., *A Course in Probability Theory*. Academic Press, 3 ed., 2001.

[17] Cisco, "ICM enterprise edition." http://www.cisco.com/en/US/products/sw/custcosw/ps1001/, 2005.

[18] Dai, J. G., "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Annals of Applied Probability*, vol. 5, pp. 49–77, 1995.

[19] Dai, J. G., *Stability of fluid and Stochastic Processing Networks*. MaPhySto, 1999.

[20] Dai, J. G. and Lin, W., "Maximum pressure policies in stochastic processing networks," *Operations Research*, vol. 53, pp. 197–218, 2005.

[21] Dai, J. G. and Lin, W., "Asymptotic optimality of maximum pressure policies in stochastic processing networks," *Unpublished manuscript*, 2005.

[22] Dai, J. G. and Tezcan, T., "State space collapse in many server diffusion limits of parallel server systems," tech. rep., School of Industrial and Systems Engineering, Georgia Insitute of Technology, 2005.

[23] Davis, M. H. A., "Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models," *J. Roy. Statist. Soc. Ser. B*, vol. 46, no. 3, pp. 353–388, 1984.

[24] Ethier, S. and Kurtz, T., *Markov Processes: Characterization and Convergence*. New York: John Wiley and Sons, 1986.

[25] Fleming, P., Stolyar, A., and Simon, B., "Heavy traffic limit for a mobile system model," in *Second International Conference on Telecommunication Systems, Modeling and Analysis*, (Nashville, TN), pp. 158–176, March 23-26 1994.

[26] Foschini, G. and Salz, J., "A basic dynamic routing problem and diffusion," *IEEE Transactions on Communications*, vol. 26, pp. 320–327, 1978.

[27] Gamarnik, D. and Zeevi, A., "Validity of heavy traffic steady-state approximations in open queueing networks," *Working paper*, 2004.

[28] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: Tutorial, review and research prospects," *Manufacturing and Service Operations Management*, vol. 5, pp. 79–141, 2003.

[29] Garnett, O., Mandelbaum, A., and Reiman, M., "Designing a call center with impatient customers," *Manufacturing and Service Operations Management*, vol. 48, pp. 566–583, 2002.

[30] Gross, D. and Harris, C. M., *Fundamentals of Queueing Theory*. Probability and Statistics, Wiley, 3 ed., 1998.

[31] Gurvich, I., Armony, M., and Mandelbaum, A., "Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers," *(working paper)*, 2004.

[32] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Operations Research*, vol. 29, pp. 567–588, 1981.

[33] Harrison, J. M., *Brownian Motion and Stochastic Flow Systems*. New York: John-Wiley and Sons, 1985.

[34] Harrison, J. M., "Brownian models of queueing networks with heterogeneous customer populations," in *Stochastic Differential Systems, Stochastic Control Theory and Their Applications* (Fleming, W. and Lions, P. L., eds.), vol. 10 of *The IMA Volumes in Mathematics and Its Applications*, (New York), pp. 147–186, Springer-Verlag, 1988.

[35] Harrison, J. M., "Brownian models of open processing networks: Canonical representation of workload," *Annals of Applied Probability*, vol. 10, pp. 75–103, 2000.

[36] Harrison, J. M. and Van Mieghem, J. A. *Annals of Applied Probability*, pp. 747–771.

[37] Harrison, J. M. and Williams, R. J., "Workload reduction of a generalized Brownian network," *Ann. Appl. Probab.*, vol. 15, no. 4, pp. 2255–2295, 2005.

[38] Harrison, J. M. and Zeevi, A., "Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime," *Operations Research*, vol. 52, pp. 243–257, 2004.

[39] Harrison, J. M., "Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies," *Ann. Appl. Probab.*, vol. 8, no. 3, pp. 822–848, 1998.

[40] Harrison, J. M., "A broader view of Brownian networks," *Ann. Appl. Probab.*, vol. 13, no. 3, pp. 1119–1150, 2003.

[41] Harrison, J. M. and López, M. J., "Heavy traffic resource pooling in parallel-server systems," *Queueing Systems Theory Appl.*, vol. 33, no. 4, pp. 339–368, 1999.

[42] Iglehart, D., "Weak convergence in queueing theory," *Advances in Applied Probability*, vol. 5, pp. 570–594, 1973.

[43] Karatzas, I. and Shreve, S., *Brownian motion and stochastic calculus*. New York: Springer-Verlag, 1991.

[44] Kogan, Y., Levy, Y., and Milito, R., "Call routing to distributed queues: Is FIFO really better than MED?," *Telecommunication Systems - Modeling, Analysis, Design and Management*, vol. 7, pp. 299–312, 1997.

[45] Law, A. M. and Kelton, W. D., *Simulation Modeling and Analysis*. McGraw Hill, 3 ed., 2000.

[46] Laws, C., "Resource pooling in queueing networks with dynamic routing," *Adv. Appl. Prob.*, vol. 24, pp. 699–724, 1992.

[47] Maglaras, C. and Zeevi, A., "Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations," *Management Science*, vol. 49, p. 10181038, 2003.

[48] Maglaras, C. and Zeevi, A., "Diffusion approximations for a Markovian multiclass service system with "guaranteed" and "best-effort" service levels," *to appear in Mathematics of Operations Research*, 2004.

[49] Maglaras, C. and Zeevi, A., "Pricing and design of differentiated services: Approximate analysis and structural insights," *to appear in Operations Research*, 2005.

[50] Maglaras, C., "Discrete-review policies for scheduling stochastic networks: trajectory tracking and fluid-scale asymptotic optimality," *Ann. Appl. Probab.*, vol. 10, no. 3, pp. 897–929, 2000.

[51] Mandelbaum, A., Massey, W., and Reiman, M., "Strong approximations for Markovian service networks," *Queueing Systems: Theory and Applications*, vol. 30, pp. 149–201, 1998.

[52] Mandelbaum, A. and Stolyar, A., "Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule," *Operations Research*, vol. 52, pp. 836–855, 2004.

[53] Milner, J. and Olsen, T. L., "Service level agreements in call centers: Perils and prescriptions," *Working Paper*, 2005.

[54] Pats, G., *State-Dependent Queueing Networks: Approximations and Applications*. PhD thesis, Technion, 1995.

[55] Puhalskii, A., "On the invariance principle for the first passage time," *Mathematics of Operations Research*, vol. 19, pp. 946–954, 1994.

[56] Puhalskii, A. and Reiman, M., "The multiclass GI/PH/N queue in the Halfin-Whitt regime," *Advances in Applied Probability*, vol. 32, pp. 564–595, 2000.

[57] Randhawa, R. and Kumar, S., "Multi-server loss systems with subscribers," *Working paper*.

[58] Reiman, M., "Some diffusion approximations with state space collapse," in *Proceedings International Seminar On Modeling And Performance Evaluation Methodology*, (Berlin), pp. 209–240, Springer-Verlag, 1983.

[59] Sheldon, R., *Stochastic Processses*. John Wiley and Sons, 1996.

[60] Stolyar, A., "Optimal routing in output-queued flexible server systems," *Probability in the Engineering and Informational Sciences*, vol. 19, pp. 141–189, 2005.

[61] Stolyar, L., "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. 14, pp. 1–53, 2004.

[62] Tezcan, T., "Optimal control of distributed call centers," *(working paper)*, 2005.

[63] Weber, R. R., "On the optimal assignment of customers to parallel servers," *J. Appl. Probability*, vol. 15, no. 2, pp. 406–413, 1978.

[64] Whitt, W., "On the heavy-traffic limit theorem for GI/G/∞ queues," *Adv. Appl. Prob.*, vol. 14, pp. 171–190, 1982.

[65] Whitt, W., *Stochastic-Process Limits*. New York: Springer, 2002.

[66] Whitt, W., "A diffusion approximation for the G/GI/n/m queue," *Operations Research*, vol. 52, pp. 922–941, 2004.

[67] Whitt, W., "Heavy-traffic limits for the G/H_2^*/n/m queue," *Mathematics of Operations Research*, vol. 30, pp. 1–27, 2005.

[68] Whitt, W., "Deciding which queue to join: some counterexamples," *Operations Research*, vol. 34, no. 1, pp. 55–62, 1986.

[69] Williams, R. J., "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse," *Queueing Systems: Theory and Applications*, vol. 30, pp. 27–88, 1998.

[70] Winston, W., "Optimality of the shortest line discipline," *J. Appl. Probability*, vol. 14, no. 1, pp. 181–189, 1977.

[71] Wolff, R. W., "Poisson arrivals see time averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, 1982.