THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

Many Server Queueing Models with Heterogeneous Servers and Parameter Uncertainty with Customer Contact Centre Applications

Wenyi Qin

Doctor of Philosophy
University of Edinburgh
2018

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Wenyi Qin)

Acknowledgements

Abstract

In this thesis, we study the queueing systems with heterogeneous servers and service rate uncertainty under the Halfin-Whitt heavy traffic regime. First, we analyse many server queues with abandonments when service rates are i.i.d. random variables. We derive a diffusion approximation using a novel method. The diffusion has a random drift, and hence depending on the realisations of service rates, the system can be in *Quality Driven* (QD), *Efficiency Driven* (ED) or *Quality-Efficiency-Driven* (QED) regime. When the system is under QD or QED regime, the abandonments are negligible in the fluid limit, but when it is under ED regime, the probability of abandonment will converge to a non-zero value. We then analyse the optimal staffing levels to balance holding costs with staffing costs combining these three regimes. We also analyse how the variance of service rates influence abandonment rate.

Next, we focus on the state space collapse (SSC) phenomenon. We prove that under some assumptions, the system process will collapse to a lower dimensional process without losing essential information. We first formulate a general method to prove SSC results inside pools for heavy traffic systems using the hydrodynamic limit idea. Then we work on the SSC in multi-class queueing networks under the Halfin-Whitt heavy traffic when service rates are i.i.d. random variables within pools. For such systems, exact analysis provides limited insight on the general properties. Alternatively, asymptotic analysis by diffusion approximation proves to be effective. Further, limit theorems, which state the diffusively scaled system process weakly converges to a diffusion process, are usually the central part in such asymptotic analysis. The SSC result is key to proving such a limit. We conclude by giving examples on how SSC is applied to the analysis of systems.

Lay Summary

Contact centres have been playing a more and more important role in the society. Almost everyone has to interact with contact centres such as airline companies, banks, and utility companies. For managers, how to make decisions to balance the cost and service quality of the contact centre becomes a significant problem. Thus we need to analyse call centre properties such as the probability of waiting, staffing costs, holding costs (for delayed customers) and rate of customers abandoning the service without getting service. In reality, servers are unlikely to be identical in such systems. There are many reasons which will cause heterogeneity among servers, such as personal skills, health and weather. Hence we focus on systems where service rates are random variables. We develop approximations of the systems to analyse the basic properties and how they behave when the system becomes large. We then show that under some assumptions, the dimensions of the system process will reduce while still keeping the essential information. So called state space collapse (SSC) results will simplify the analysis of the system and help us gain key insights using approximations.

Contents

Chapter 1

Introduction and Literature Review

1.1 Introduction

Many server queues are widely seen nowadays. For example, in banks, customers come for service, then staffs process customers' requests, and after their services are completed, customers leave banks. This is a classical many server queue system with arrivals, services, and departures. It can also be observed in other scenarios such as emergency departments of hospitals, call centres, post offices, and computers. For these systems, servers are usually different from each other. They may possess different types of skills. Even when they have the same skill, their ability on this skill may be different. Such heterogeneous systems are not studied sufficiently in literature yet. We are going to investigate their behaviours and properties in this thesis. More specifically, call centres is a very important area where this heterogeneity can be applied. We are going to focus on the applications of heterogeneous systems in call centres.

A call centre is a service operation over the phone. It consists of groups of people, called agents or servers, who provide service to customers. Call centres are increasingly important in today's business world since they have become a preferred and prevalent means for companies to communicate with their customers. Call centres are data rich environments which triggers many interesting mathematical problems. For call centre managers, it is important to guide the system to achieve certain service levels while the costs remain reasonable. There are many factors that can influence service levels. For example, the routing scheme which leads arriving customers to specific servers, the deployment of servers, and the scheduling policy which guides servers to accept customers. Existing research provides fruitful results regarding these problems for call centres with identical servers. However, in reality, it is usually the case that servers will be heterogeneous. Individual servers can possess unique skills, and even when a group of servers have the same skills, their abilities to show the skills are subject to environment. In

this thesis, we are going to focus on call centres with heterogeneous servers and analyse their properties. Our proxy for heterogeneity is the servers' service time distribution.

A call centre can be seen as a queueing system. Customers arrive at the centre according to a stochastic process, then they are routed to available idle servers based on some routing policies. If there are no idle servers available upon their arrival, they will wait in the queue, and they may wait until they get service or abandon the queue before they get service. On the other hand, a prescribed scheduling policy is used to dispatch a server to serve a customer. Once a customer is routed to a server, s/he will be served with a specific rate based on the server, and s/he will leave the system when the service is finished.

The cost-service level trade-off has a central place in quantitative call centre management. When the cost of capacity is dominated by the waiting cost of customers, the decision maker concentrates on the waiting cost and sets the staffing levels so that the utilisation ends up being less than one. This is called the *Quality-Driven* (QD) regime. The other extreme is when staffing costs dominate the waiting costs. In such case, utilisation of servers is fixed and is equal to one. In the long run, such a large scale system will be unstable. Customers will accumulate and a significant portion of the customers will abandon. This is called the *Efficiency-Driven* (ED) regime. In between these two extremes is the *Quality-Efficiency-Driven* (QED) regime, under which quality and efficiency are balanced. Under this regime, utilisations will approach to one from below as the system size increases, and the proportion of customers who wait before service converges to a constant which is related to the staffing level.

In this thesis, we focus on the QED regime. [Halfin and Whitt, 1981] proposed the approach to analyse the system performance under the QED regime for homogeneous servers. This work is a milestone in queueing theory which initiated a lot of research. We are going to modify their assumptions and apply it to systems where each server is unique and different. In particular, we assume service rates are i.i.d. random variables, and remain the same once the system starts operating, i.e. they do not change with time.

Apart from the staffing decisions, how to route customers to different servers will also influence the system quality. We have the routing policy to ensure customers are routed to servers in a certain way. For homogeneous systems, routing policy does not play a major role as no matter to which server a customer is routed, there is no change in the system process. For heterogeneous systems, each routing policy will yield different performances. We will mainly focus on two policies : Longest Idle Server First (LISF) policy, which will route a customer to the server who has been idle for the longest time, and Faster Server First (FSF) policy, which will always route a customer to the fastest server among those idle servers. As for scheduling policies, in our study we let it be First-In-First-Out (FIFO), i.e. when customers are queueing, they will get service by the order of their arrival.

Routing policies will cause not only difference in service levels, but also different fairness among servers, which is another measure for system quality. We also analyse fairness for

heterogeneous servers under different routing policies.

For such large scale systems, exact analysis is usually intractable. Instead, we use asymptotic analysis. We use diffusion processes to approximate original queueing processes. By analysing properties of limiting diffusions, we can get insights for decision making in reality. Based on [Atar, 2008]'s result, first we analyse many server queues with random service rates under the QED regime. Then we use a similar framework as in [Borst et al., 2004] to obtain the optimal staffing levels while balancing waiting and staffing costs. Later in this thesis, we prove the diffusion limit result of [Atar, 2008] using a novel method.

Then, we analyse systems with random service rates and abandonments similarly as above. In addition to the same results for systems without abandonments, we also show that the influence of service rates variance on abandonment rate are different under different routing policies. For LISF, the abandonment rate is increasing with variance, while for FSF it may be decreasing.

Finally, we consider the state space collapse (SSC) phenomenon which implies that under some assumptions, the system process is asymptotically equivalent to a lower dimensional process. We first formulate a general method to prove SSC results for a single pool with random servers under the Halfin-Whitt regime, by which we can prove results presented in [Atar, 2008] in a different way. Then, adapting the results in [Dai and Tezcan, 2011], we show how SSC in multiclass queueing networks can be obtained under the Halfin-Whitt heavy traffic regime when service rates are i.i.d. random variables within pools.

Our main contributions can be summarised as follows.

- Use a martingale method to prove the diffusion limit for many server queues with random service rates and abandonments.

- Establish the optimal staffing problem for many server queues with random service rates. Provide a continuous approximation of this optimisation problem and validate it. Tightness of the steady state is proved in order to show the interchangeable limit.

- Prove SSC results for parallel random server systems. Use a coupling method to prove the almost Lipschitz condition for departure processes.

The thesis is organised as follows. In the rest of this chapter, we review the literature, comparing existing results with our new results. In Chapter 2, we first present [Atar, 2008]'s results, and show the diffusion limit using the method developed by Atar. Then we include abandonments to Atar's model, and also show its diffusion limit. In Chapter 3, we formulate an optimal staffing problem for models in [Atar, 2008]. Later we extend the problem to systems with abandonments. We also analyse how the variance of service rates influence abandonment rates. In Chapter 4, we talk about state space collapse results for many server queueing networks with server heterogeneity, and how it can be applied to system analysis. In Chapter 5, we conclude the thesis by summarising our contributions and point out future research directions.

1.2 Literature review

Queueing models are used broadly in many service systems such as call centres, healthcare and computer science. For systems with arrivals, service and departures, it is convenient to model them as queueing systems and analyse their performance. For example, for call centres, there are incoming calls, agents who answer calls and call departures, and such call centres can be analysed using queueing theory. As for queueing analysis used in other areas, [Mandelbaum et al., 2012] discuss fair routing between emergency departments and hospital wards under the QED regime. [Deo and Gurvich, 2011] use a game-theoretic queueing model and find an equilibrium on the accepted diverted ambulance from emergency departments of other hospitals. [Tezcan and Zhang, 2014] consider customer service chat systems where customers can receive real time service from agents using an instant messaging application over the internet. We will focus on call centres in this thesis using queueing modelling and analyse their performance.

The research on call centres can be viewed under different headings. [Gans et al., 2003] and [Aksin et al., 2007] review research on call centres, and provide a survey of literature on call centre operations management. They also identify some promising directions for future research. The most famous and basic queueing model for call centres is the Erlang-C or Erlang delay model, which deals with only one type of call and server without abandonments; thus every customer waits until s/he reaches a server. [Koole, 2007] gives a general idea of how the Erlang-C formula is used in call centre. More mathematical details on the Erlang-C formula can be found in [Cooper, 1981]. The Erlang-C formula is an important formula in the early stage of call centre research. It gives an explicit form of statistical-equilibrium distributions of the queueing process given the arrival rate, identical service rates, and number of servers. Thus, the probability of waiting can be calculated, to decide on the number of servers that are needed to make the system achieve a certain service level.

However, such direct analysis becomes impractical when the system grows large. Hence, the asymptotic analysis should be used. Diffusion approximations for stochastic processes in queueing models prove to be quite useful (see [Iglehart, 1965], [Stone, 1961], [Halfin and Whitt, 1981]); The work of [Halfin and Whitt, 1981] is the most relevant to our work. It considers a sequence of $GI/M/s$ systems in which the traffic intensities converge to one from below, which brings the Halfin-Whitt heavy traffic regime into the picture. With arrival rate λ, service rate μ, and the Halfin-Whitt regime, under a certain scaling, the probability of waiting converges to a constant α which is strictly greater than 0 and less than 1. α can be used to indicate the service level of the system. The offered load is a measure of traffic in a queue and is defined to be $\frac{\lambda}{\mu}$. The staffing level of such a system will be offered load $\frac{\lambda}{\mu}$ plus the square root of the offered load $\sqrt{\frac{\lambda}{\mu}}$ multiplied by a constant β, where the coefficient β depends on the service level α. The quantity $\beta\sqrt{\frac{\lambda}{\mu}}$ is called the "safety staffing" level against stochastic variability. Using this approach, staffing and waiting costs are well balanced. The Halfin-Whitt regime has

been extended in several directions. [Janssen et al., 2011] propose refinements of the celebrated square-root safety-staffing rule which have the appealing property that they are as simple as the conventional square-root safety-staffing rule. [Atar, 2008] introduces a new square root staffing policy for many servers systems with random service rates. [Puhalskii and Reiman, 2000] extend the results to a system with multiple customer classes, priorities, and phase-type service distributions. [Armony, 2005] establishes diffusion approximations and staffing levels for inverted-V systems. Our study also focuses on the Halfin-Whitt regime, and extends it to heterogeneous servers instead of identical ones.

Staffing has always been a central issue for call centre managers. There are many research papers on staffing problems for different models. See [Mandelbaum and Zeltyn, 2009], [Koçağa et al., 2015], [Whitt, 2006], [Mandelbaum and Zeltyn, 2009] and [Armony and Mandelbaum, 2011] for different discussions. Based on staffing level, systems can be in a QD, ED or QED regime. When the system is under a QD or QED regime, the abandonments are negligible in the limit, but when it is under an ED regime, the probability of abandonment will converge to a non-zero value. [Borst et al., 2004] determine the asymptotically optimal staffing level for $M/M/N$ queues under different regimes. [Whitt, 2004] investigates the ED many server heavy traffic regime for queues with abandonments. Our staffing model is based on the framework of [Borst et al., 2004].

For heterogeneous systems, how to route arriving customers to servers and how to schedule servers to serve customers are crucial decisions to the system performance. Routing and scheduling have been studied extensively in the literature. [Tezcan and Zhang, 2014] consider customer service chat systems where agents can serve multiple customers simultaneously. They propose routing policies for such system with impatient customers with the objective to minimise the probability of abandonment in steady state. [Gurvich et al., 2010] consider the staffing problem for call centres with multi-class customers and different agent types operating under QD constraints and arrival rate uncertainty. They propose a two-step solution which contains two actions: the number of agents of each type, and a dynamic routing policy. [Armony, 2005] shows that for the inverted-V model, the FSF policy is asymptotically optimal in the QED regime and no thresholds are needed. There is literature that carries out exact analysis and asymptotic analysis under conventional heavy traffic, such as [Rykov and Efrosinin, 2004] and [Kelly and Laws, 1993]. Routing policies also play an important role in staffing optimisation. Under different policies, the steady state behaviour of the system changes and fairness among servers is also different. FSF policy is commonly used. [Armony, 2005] establishes diffusion limits for the inverted-V systems under FSF policy and concludes they have a better performance than their corresponding homogeneous systems. [Atar, 2008] provides diffusion limits for many server queues with random servers under LISF and FSF policies. [Tezcan, 2008] develops limit theorems for inverted-V systems under minimum-expected-delay faster-server-first (MED-FSF) and minimum-expected-delay load-balancing (MED-LB) routing policies. Notice that LISF is

a blind policy, i.e. it only needs to track the state of the process in order to make routing decisions, and information about service rates is not needed in this case. We will mainly use this policy in our model since our service rates are random variables and thus their realisations are unknown before the systems start to operate.

State space collapse is an important phenomenon when we analyse the system behaviour. [Harrison and Van Mieghem, 1997] explain the dimension reduction in general terms, using an orthogonal decomposition. For some examples of SSC one can check [Reiman, 1984]. [Puhalskii and Reiman, 2000] prove SSC for a particular system which has phase type distributed service rates. [Bramson, 1998] uses the hydrodynamic scaling to build up the state space collapse results for multi-class queueing networks under the conventional heavy traffic regime. He shows that we can use a lower dimensional process, the workload processes of each service station, to represent the system because the original system process, which is the number of each type of customers in every station, can be obtained through the workload processes and some lifting functions. The paper by [Dai and Tezcan, 2011] uses the hydrodynamic scaling proposed by [Bramson, 1998] in a many server setting. Their contribution is the definition of a SSC function, which is used in that paper to show the dimension reduction for many server networks under the Halfin-Whitt heavy traffic regime. Our SSC result is based on their framework. [Tezcan, 2008] applies this method to a distributed parallel server system and does optimal control analysis.

The key point in our research is heterogeneity and uncertainty in parameters. The uncertainty in arrival rates is investigated in some prior work. For example, [Zan, 2012] analyses the staffing problems when the arrival rate is uncertain. However, for uncertainty in the service rates, there is still plenty of space for us to explore. For a general non-technical introduction to this topic, [Gans et al., 2010] is an excellent reference. The heterogeneity in the servers is modelled in various ways. A commonly used one is the inverted-V system, which contains a single customer class and multiple server types. [Armony, 2005] considers the asymptotic framework for such systems. She shows that the FSF policy is asymptotically optimal in the QED regime. Later in [Armony and Ward, 2010], an optimisation problem for inverted V systems is formulated. They minimise the steady-state expected customer waiting time subject to a "fairness" constraint and propose a threshold routing policy which is asymptotically optimal in the Halfin-Whitt regime. [Mandelbaum et al., 2012] introduce the randomised most-idle (RMI) routing policy for the inverted-V model and analyse it in the QED regime. [Atar, 2008]'s results about random servers set the cornerstones for our work. The diffusion limit in [Atar, 2008] contains a random drift, which comes from the heterogeneity of servers. From the diffusion, we derive its steady state distribution, then formulate the optimisation staffing problem using the distribution.

Chapter 2

Diffusion Limits for Single Server Pool Systems

2.1   Introduction

In this chapter we focus on many server queueing systems with random service rates under the Halfin-Whitt heavy traffic regime. Many server queueing models have been studied extensively. However, there are only a few papers about many server queues when service rates are random variables instead of identities. Models with identical servers are not sufficient when it comes to modelling human behaviours. Individual abilities are always influenced by environment thus they can not be constants. To capture the feature more accurately, we assume our model has N^r exponential servers with i.i.d. service rates $\mu_k, k = 1, \ldots, N^r$. N^r is also assumed to be a random variable. When customers arrive into the system they will either queue in a buffer with infinite room, or be routed to a server according to the LISF routing policy. Customers from the queue are routed to servers according to the FIFO rule. In this chapter, we first assume that the customers do not abandon and leave the system only after their service is completed. We relax this assumption later in the chapter. The routing policy is work conserving, in the sense that no server will be idle when there is at least one customer in the queue. The service policy is non-preemptive, i.e. once a customer is assigned to a server, it will continuously receive service until it is completed, i.e. the services will not be interrupted. This model is considered in [Atar, 2008]. [Atar, 2008] also analyses the same systems under the FSF routing policy. We will not focus on FSF policy because our decision is supposed to be made before the system starts running, and FSF policy requires the knowledge of each server's rate, which does not suit our case. This chapter is organised as follows: in Section 2.1, we give the detailed description of the mathematical model and notations used throughout this chapter; in Section 2.2, we rewrite the proof of the central theorem in [Atar, 2008], although it is already proven by Atar, we present it here for completeness of our discussion; in Section 2.3, we formulate the

optimal staffing problem and prove the validity of its asymptotic version.

2.2 Mathematical modelling and notation

First we introduce the notation used throughout the thesis. All of the random variables and stochastic processes are defined on a complete probability space $(\Omega, \mathscr{F}, \mathbb{P})$. For a positive integer d, we denote by $\mathbb{D}(\mathbb{R}^d)$ the space of functions from $\mathbb{R}^+$ to $\mathbb{R}^d$ that are right continuous and left limits exist (RCLL), endowed with the usual Skorohod J_1 topology. (See [Billingsley, 1999] for the definition.) We use $\Rightarrow$ to denote weak convergence. And for $X \in \mathbb{D}(\mathbb{R})$, we write $||X||_t := \sup_{0 \le s \le t} |X(s)|$.

The model is parameterised by $r \in \mathbb{R}^+$, where for each r, $N^r \in \mathbb{N}$ is a random variable, representing number of servers. Service times for customers served at server k are i.i.d. exponentially distributed with rate $\mu_k, k = 1, \dots, N^r$. The μ_ks are assumed to be nonnegative and lie in an interval $[p, q]$. The distribution of μ_k is denoted by m, and its expected value is

$$\bar{\mu} := \int_{[0,\infty)} x \, dm \in (0, \infty). \tag{2.2.1}$$

It is also assumed that N^r satisfies the following two assumptions

$$\mathbb{P}(N^r \le 2r) = 1,$$
$$\frac{N^r}{r} \Rightarrow 1, \text{ as } r \to \infty.$$

The arrivals are assumed to be renewal processes with finite second moments for the inter-arrival times. Let the arrival rate be λ^r such that $\lim_{r \to \infty} \frac{\lambda^r}{r} = \lambda > 0$, and a sequence of strictly positive i.i.d. random variables $\{\check{U}(l), l \in \mathbb{N}\}$, with mean $\mathbb{E}\check{U}(1) = 1$ and variance $C^2_{\check{U}} = Var(\check{U}(1)) \in [0, \infty)$. The Halfin-Whitt heavy traffic condition, which makes the system critically loaded, is assumed to be

$$\lim_{r \to \infty} \frac{1}{\sqrt{r}} (\lambda^r - r\lambda) = \hat{\lambda}, \tag{2.2.2}$$

where $\lambda = \bar{\mu}$, and $\hat{\lambda} < 0$.

For the rth system, let $A^r(t)$ be the total number of arrivals into the system up to time t, $X^r(t)$ be the total number of customers in the rth system at t (including customers both being served and waiting in the queue), $D^r_k(t)$ be the number of jobs completed by server k up to time t, and $T^r_k(t)$ be the accumulated busy time of server k by time t. Let $B^r_k(t) = 1$ if server k is busy at time t, and it equals to zero if the server is idle. Accordingly, let $I^r_k(t) = 1 - B^r_k(t)$, indicating that server k is idle if $I^r_k(t)$ equals 1.

2.3 Atar's results on many server queues with random servers

When there are no abandonments, it can be shown the systems satisfy the equation

$$X^r(t) = X^r(0) + A^r(t) - \sum_{k=1}^{N^r} D_k^r(t). \tag{2.3.1}$$

We let $\hat{X}^r(t)$ be a renormalized version of the process $X^r(t)$, which is defined as

$$\hat{X}^r(t) = \frac{1}{\sqrt{r}}(X^r(t) - N^r). \tag{2.3.2}$$

The initial value of $X(t)$ and the random variable N^r are assumed to satisfy

$$(\hat{X}^r(0), \hat{N}^r) := \left(\frac{1}{\sqrt{r}}(X^r(0) - N^r), \frac{1}{\sqrt{r}}(N^r - r)\right) \Rightarrow (\xi(0), \nu), \tag{2.3.3}$$

where $(\xi(0), \nu)$ is an $\mathbb{R}^2$-valued random variable.

[Atar, 2008]'s main result is the following theorem. It states that under the LISF policy, the sequence of the scaled processes of total number of customers weakly converges to a diffusion. What makes the diffusion distinctive from other diffusion limits is that it contains a random drift which arises from the randomness of service rates.

Theorem 2.3.1 ([Atar, 2008]). *Assume $\int x^2 dm < \infty$. Then, under the LISF policy, the processes $\hat{X}^r(t)$ weakly converge to the solution of the following SDE*

$$\xi(t) = \xi(0) + \sigma w(t) + \beta t + \gamma \int_0^t \xi(s)^- ds, t \geq 0, \tag{2.3.4}$$

where $\sigma^2 = \lambda C_{\bar{U}}^2 + \bar{\mu} = \bar{\mu}(C_{\bar{U}}^2 + 1)$, and $\beta = \hat{\lambda} - \zeta - \bar{\mu}\nu$, ζ is a normal random variable with parameters $(0, \int(x - \bar{\mu})^2 dm)$, $\gamma = \frac{\int x^2 dm}{\int x dm}$, $w(t)$ is a standard Brownian motion, and the three random elements $(\xi(0), \nu), \zeta$, and $w(t)$ are mutually independent.

[Atar, 2008] proves his result using a method particular to this model. We give an explanation of his proof here. The detailed mathematical proof is included in Appendix A.2 for completeness. Later we will use another approach inspired by state space collapse phenomenon to prove this theorem again. For more detailed discussion on the new proof, see Section 4.5. The process of scaled total number of customers are

$$\hat{X}^r(t) = \frac{1}{\sqrt{r}}(X^r(0) - N^r) + \frac{1}{\sqrt{r}}A^r(t) - \frac{1}{\sqrt{r}}\sum_{k=1}^{N^r} D_k^r(t)$$

$$= \hat{X}^r(0) + \frac{1}{\sqrt{r}}(A^r(t) - \lambda^r t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{k=1}^{N^r} D_k^r(t)$$

21

$$= \hat{X}^r(0) + \hat{A}^r(t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{k=1}^{N^r} D_k^r(t). \tag{2.3.5}$$

The convergence of arrival process $\hat{A}^r(t)$ follows from the basic functional central limit theorem for renewal processes. Analysing the convergence of departure process $D_k^r(t)$ is considerably harder. Each server needs to be considered individually because they have their unique service rate, and thus it is an N^r dimensional process. However, as the system grows large, the number of servers tends to infinity; hence analysing each $D_k^r(t)$ becomes intractable.

[Atar, 2008] solves this problem by partitioning the servers into finite I pools. Within each pool, the supremum of the difference between two service rates is bounded by some small positive number ϵ.

To see how this idea is used in the proof, first we need to obtain more insights about the departure processes. The departure process of each server is treated as a time changing Poisson process. Let $\{S_k(t), k \in \mathbb{N}\}$ be independent standard Poisson processes. Since $T_k^r(t)$ is the accumulated busy time for server k by time t, then by random time change, $D_k^r(t)$ satisfies

$$D_k^r(t) = S_k(\mu_k T_k^r(t)), \ k = 1, \dots, N^r. \tag{2.3.6}$$

where

$$T_k^r(t) = \int_0^t B_k^r(s)ds. \tag{2.3.7}$$

Denote $\Sigma^r = (A^r, X^r, Q^r, I^r, \{B_k^r, D_k^r\}_{k=1,\dots,N^r})$ which has a.s. piecewise constant and right-continuous sample paths. [Atar, 2008] partitions servers into pools as mentioned above. Pools are indexed by $\{i, i = 1, 2, \dots, I\}$. Using such configuration, the system can be regarded approximately as an inverted-V system, hence departure processes can be considered aggregately in each pool. Denote the total departure process of pool i as $D^{r,(i)}(t)$. He further defines the total service rate of pool i to be the sum of $\mu_k T_k^r(t)$ (the sum is over all of the servers in pool i). Then $D^{r,(i)}(t)$ is also a time changing Poisson process and

$$D^{r,(i)}(t) = S^{(i)}\left(\sum_{k \in \text{ pool }_i} \mu_k T_k^r(t)\right),$$

where $\{S^{(i)}(\cdot)\}$ are independent standard Poisson processes. For simplicity denote the total service rate in pool i to be

$$T^{r,(i)}(t) = \sum_{k \in \text{ pool }_i} \mu_k T_k^r(t).$$

He shows in Proposition 3.1 (see AppendixA.1) that with this pooling method, the total depar-

ture process of the original system $\sum_{k=1}^{N^r} D_k^r(t)$ and the total departure process of the approximate inverted-V system $\sum_{i=1}^{I} D^{r,(i)}(t)$ are equal in distribution.

Using this equivalence, system equation (2.3.5) is equal in distribution to

$$\hat{X}^r(t) = \hat{X}^r(0) + \hat{A}^r(t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{i=1}^{I} D^{r,(i)}(t).$$

With further manipulations, it can be expressed as

$$\hat{X}^r(t) = \hat{X}^r(0) + \hat{A}^r(t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{i=1}^{I}\left(S^{(i)}\left(T^{r,(i)}(t)\right) - T^{r,(i)}(t)\right) - \frac{1}{\sqrt{r}}\sum_{i=1}^{I} T^{r,(i)}(t)$$

$$= \hat{X}^r(0) + \hat{A}^r(t) + \frac{1}{\sqrt{r}}(\lambda^r t - r\bar{\mu}t) - \frac{1}{\sqrt{r}}\left(\sum_{k=1}^{N^r}\mu_k t - r\bar{\mu}t\right)$$

$$- \frac{1}{\sqrt{r}}\sum_{i=1}^{I}\left(S^{(i)}\left(T^{r,(i)}(t)\right) - T^{r,(i)}(t)\right) - \frac{1}{\sqrt{r}}\left(\sum_{i=1}^{I} T^{r,(i)}(t) - \sum_{k=1}^{N^r}\mu_k t\right).$$

Convergence of the first four items are easy. Let us pay attention to the latter two. First, it is proved that $\frac{1}{r}T^{r,(i)}(t)$ converges to its fluid limit $\rho_i t$, where ρ_i is the product of the expectation of μ_k and the weight probability of servers in pool i. $\frac{1}{\sqrt{r}}\left(S^{(i)}\left(T^{r,(i)}(t)\right) - T^{r,(i)}(t)\right)$ is a martingale, then [Atar, 2008] uses Functional Central Limit Theorem, and applies random time change to the fluid limit $\rho_i t$ to show that $\frac{1}{\sqrt{r}}\sum_{i=1}^{I}\left(S^{(i)}\left(T^{r,(i)}(t)\right) - T^{r,(i)}(t)\right)$ converges to some Brownian motion.

$\frac{1}{\sqrt{r}}\left(\sum_{i=1}^{I} T^{r,(i)}(t) - \sum_{k=1}^{N^r}\mu_k t\right)$ is the most concerning part in the proof. It is actually equal to

$$\frac{1}{\sqrt{r}}\sum_{k=1}^{N^r}(\mu_k T_k^r(t) - \mu_k t) = \frac{1}{\sqrt{r}}\sum_{k=1}^{N^r}\left(\mu_k \int_0^t B_k^r(s)ds - \mu_k t\right) = \frac{1}{\sqrt{r}}\sum_{k=1}^{N^r}\left(-\mu_k \int_0^t I_k^r(s)ds\right),$$

which is the total amount of unused service capacities due to idleness. Denote this lost capacity as $F^r(t)$.

$F^r(t)$ contains N^r different idleness processes. It is later proved that there is a state space collapse (SSC) in such systems in the sense that, in the limit, the total lost capacities $F^r(t)$ can be represented as the total accumulated idle time multiplied by some coefficient γ related to the service rate distribution and routing policy. Such a SSC result reduces the dimension of the original processes, which eventually helps to get a one dimensional diffusion limit.

To show such SSC result, [Atar, 2008] considers the difference between total lost capacities $F^r(t)$ and the product of total accumulated idle time and γ. The difference consists of four $e_i(t)$s: $e_1(t), e_2(t), e_3(t), e_4(t)$. He proves that each $e_i(t)$ tends to zero in the limit. These four $e_i(t)$s are unique to this system, thus cannot be directly extended to other models. In Chapter 4, we explain the SSC phenomenon in detail. Then we use a more generic method to prove this

SSC result again.

2.4 Extension of Atar's results to include abandonments

In this section, we assume customers may abandon the system prior to being served. Each customer has an associated patience time, and abandons the system without obtaining any service if the waiting time in the queue exceeds the customer's patience. Once his/her service starts, s/he cannot abandon the system. Assume each customer's patience time is exponentially distributed with rate ν. We will not deal with the abandonment processes directly due to complications of analysing each customer's patience individually. Instead, we use a "perturbed" abandonment processes similar to the one described in Section 2.1 of [Dai and Tezcan, 2011]. In perturbed systems, only the customer at the head of the queue will be able to abandon, and her/his abandonment rate is the sum of abandonment rates of all of the customers in the queue. Under the assumption of exponential service and patience time, the equivalence of systems with original abandonment processes and perturbed abandonment processes is proved in [Dai and Tezcan, 2011]. Note that [Dai and Tezcan, 2011] use perturbed system technique to analyse both abandonment processes and service processes, while we only use it for our abandonment processes since our service rates are no longer deterministic thus the equivalence to the perturbed systems is invalid.

Let $Q(t)$ be the queue length at time t, and let $M(t)$ denotes the number of customers who have abandoned queue by time t. The systems are assumed to be under LISF policy again.

Let $S_Q(t)$ be a standard Poisson process. We define

$$G^r(t) = \int_0^t Q^r(s)ds, \quad t \geq 0. \tag{2.4.1}$$

Then for the perturbed abandonment process

$$R^r(t) = S_Q(\nu G^r(t)). \tag{2.4.2}$$

Using the same notations as in the previous section, we can write the system dynamic equations

$$X^r(t) = X^r(0) + A^r(t) - \sum_{k=1}^{N^r} D_k^r(t) - R^r(t). \tag{2.4.3}$$

We can show similarly that diffusion limits exist in the presence of abandonments.

Theorem 2.4.1. *Assume $\int x^2 dm < \infty$. Then, under LISF policy, the diffusively scaled processes $\hat{X}^r(t)$ weakly converge to the solution of the following SDE*

$$\xi(t) = \xi(0) + \sigma w(t) + \beta t + \gamma \int_0^t \xi(s)^- ds - \nu \int_0^t \xi(s)^+ ds, t \geq 0, \tag{2.4.4}$$

24

where $\sigma, \beta, \zeta, \gamma$ and $w(t)$ are as in Theorem 2.3.1.

Proof. The proof is quite similar to Theorem 2.3.1. We will focus on abandonment processes here since the proofs of other parts are the same.

Following arguments in the previous section we again omit the symbol r, and thus we have

$$\hat{X}(t) = \hat{X}(0) + \frac{1}{\sqrt{r}}A(t) - \frac{1}{\sqrt{r}}\sum_{i=1}^{I}S^{(i)}(T^{(i)}(t)) - \frac{1}{\sqrt{r}}R(t) \tag{2.4.5}$$

$$= \hat{X}(0) + W(t) + b^{r}t + F(t) - \hat{R}(t), \quad \left(\text{let } \hat{R}(t) = \frac{1}{\sqrt{r}}R(t)\right) \tag{2.4.6}$$

under the LISF policy,

$$\hat{X}(t) = \hat{X}(0) + W(t) + b^{r}t + \gamma\int_{0}^{t}\hat{X}(s)^{-}ds - \hat{R}(t) + e(t), \quad \left(\gamma = \frac{\int x^{2}dm}{\int xdm}\right). \tag{2.4.7}$$

We already showed in the previous section that $W(t) \Rightarrow \sigma w$, $b^{r} \Rightarrow \beta$, and $e(t) \to 0$ u.o.c. in probability. For the newly added term $\hat{R}(t)$, note that

$$\hat{R}(t) = \frac{1}{\sqrt{r}}S_{Q}(\nu G(t)) = \frac{1}{\sqrt{r}}S_{Q}\left(\nu\int_{0}^{t}Q(s)ds\right)$$

$$= \frac{1}{\sqrt{r}}\left(S_{Q}\left(\nu\int_{0}^{t}Q(s)ds\right) - \nu\int_{0}^{t}Q(s)ds + \nu\int_{0}^{t}Q(s)ds\right)$$

$$= \frac{1}{\sqrt{r}}\left(S_{Q}\left(\nu\int_{0}^{t}(X(s)-N)^{+}ds\right) - \nu\int_{0}^{t}(X(s)-N)^{+}ds + \nu\int_{0}^{t}(X(s)-N)^{+}ds\right)$$

$$= \frac{1}{\sqrt{r}}\left(S_{Q}\left(r\frac{\nu}{r}\int_{0}^{t}(X(s)-N)^{+}ds\right) - r\frac{\nu}{r}\int_{0}^{t}(X(s)-N^{r})^{+}ds\right) + \nu\int_{0}^{t}\hat{X}(s)^{+}ds.$$

Denote $\hat{M}(t) = \frac{1}{\sqrt{r}}\left(S_{Q}\left(\nu\int_{0}^{t}(X(s)-N)^{+}ds\right) - \nu\int_{0}^{t}(X(s)-N)^{+}ds\right)$. Then by Theorem 7.2 in [Pang et al., 2007], $\hat{M}(t)$ is a square-integrable martingale with respect to the filtrations $\mathbf{F}_{r} \equiv \{\mathcal{F}_{r,t} : t \geq 0\}$ defined by

$$\mathcal{F}_{r,t} \equiv \sigma\left(X(0), A(s), S^{(1)}(T^{(1)}(s)), \ldots, S^{(q)}(T^{(q)}(s)),\right.$$
$$\left. S_{Q}\left(\nu\int_{0}^{s}(X(u)-N)^{+}du\right) : 0 \leq s \leq t\right), t \geq 0, \tag{2.4.8}$$

augmented by including all null sets. Its predictable quadratic variation is

$$\langle\hat{M}\rangle(t) = \frac{\nu}{r}\int_{0}^{t}(X(s)-N)^{+}ds, t \geq 0. \tag{2.4.9}$$

By the same reasoning in Section 7.1 in [Pang et al., 2007], we obtain the deterministic limits

$$\langle\hat{M}\rangle(t) \Rightarrow 0. \tag{2.4.10}$$

We explain the proof briefly. For more details, see [Pang et al., 2007]. We have that

25

the sequence $\{\hat{X}^r\}$ is stochastically bounded in $\mathbb{D}(\mathbb{R})$. Then, by Lemma 5.9 and Section 6.1 in [Pang et al., 2007], we get the Functional Weak Law of Large Numbers (FWLLN) corresponding to Lemma 4.3, from which we can prove (2.4.10). Recall the basic Functional Central Limit Theorem (FCLT): $\hat{S}(t) = \frac{S(rt)-rt}{\sqrt{r}} \Rightarrow B$, where S is a standard Poisson process, and B is a standard Brownian motion. Then

$$\hat{R}(t) = \hat{S}\left(\langle \hat{M}\rangle(t)\right) + \nu \int_0^t \hat{X}(s)^+ ds, \tag{2.4.11}$$

and by the lemma on random change of time in [Billingsley, 1999, p. 150], we have $\hat{S}\left(\langle \hat{M}\rangle(t)\right) \Rightarrow 0$, in the uniform topology on the compact set $[0,T]$ for any $T > 0$. Thus,

$$\hat{R}(t) \Rightarrow \nu \int_0^t \xi(s)^+ ds, \ as \ r \to \infty. \tag{2.4.12}$$

By Skorohod Representation Theorem, we can assume without loss of generality that the random variables $\hat{X}(0), b, \xi(0)$, and β, and the processes $W, \hat{S}\left(\langle \hat{M}\rangle(t)\right)$, and w are realized in such a way that

$$\left(\hat{X}(0), b, W, \hat{S}\left(\nu \int_0^t \frac{1}{r}Q(s)ds\right)\right) \to (\xi(0), \beta, \sigma w, 0) \quad \text{in probability, as } r \to \infty. \tag{2.4.13}$$

Recall that $||X||_t := \sup_{0 \le s \le t}|X(s)|$. Combining (2.4.7), (2.4.11), and (2.4.4), the inequalities $|x^- - y^-| \le |x-y|$, $|x^+ - y^+| \le |x-y|$ and Gronwall's inequality ($u(t) \le \alpha(t)\exp(\int_a^t c(s)ds)$ if c is non-negative and u satisfies $u(t) \le \alpha(t) + \int_a^t c(s)u(s)ds$.) together show that

$$
\begin{aligned}
&||\hat{X}(t) - \xi(t)||_T \\
&= \left\| \hat{X}^r(0) - \xi(0) + W(t) - \sigma w(t) + b^r t - \beta t + e(t) \right. \\
&\quad \left. + \gamma \int_0^t \left(\hat{X}(s)^- - \xi(s)^-\right) ds - \nu \int_0^t \left(\hat{X}(s)^+ - \xi(s)^+\right) ds \right\|_T \\
&\le ||\hat{X}(0) - \xi(0)||_T + ||W(t) - \sigma w(t)||_T + ||b^r t - \beta t||_T + ||e(t)||_T \\
&\quad + \gamma \left\| \int_0^t \hat{X}(s)^- - \xi(s)^- ds \right\|_T + \nu \left\| \int_0^t \hat{X}(s)^+ - \xi(s)^+ ds \right\|_T \\
&\le ||\hat{X}(0) - \xi(0)||_T + ||W(t) - \sigma w(t)||_T + ||b^r t - \beta t||_T + ||e(t)||_T \\
&\quad + M \int_0^t \left\| \hat{X}(s)^- - \xi(s)^- \right\|_T ds + M \int_0^t \left\| \hat{X}(s)^+ - \xi(s)^+ \right\|_T ds \\
&\le ||\hat{X}(0) - \xi(0)||_T + ||W(t) - \sigma w(t)||_T + ||b^r t - \beta t||_T + ||e(t)||_T \\
&\quad + 2M \int_0^t \left\| \hat{X}(s) - \xi(s) \right\|_T ds,
\end{aligned}
$$

where $M = \max\{\gamma, \nu\}$. By (2.4.13) and the uniform convergence of e to zero, we have shown that $\hat{X}(t)$ converges to $X(t)$ in probability, uniformly on $[0,T]$. Since T is arbitrary, this shows that $\hat{X}(t) \Rightarrow X(t)$. $\qquad\square$

2.5   Summary

In this chapter, we show that diffusion limits are an effective way to approximate queuing processes because of its continuity feature. First we restate the limit theorem for many server queues with random service rates proved by [Atar, 2008]. The key part of this theorem is its random drift β, which comes from the Central Limit Theorem applied on random service rates. Another thing which needs our attention is the coefficient γ of the integral of negative part of the diffusion limit, i.e. $\int_0^t \xi(s)^- ds$. $\gamma \int_0^t \xi(s)^- ds$ approximates the capacities that are lost due to idleness. To some extent, γ reflects fairness of the routing policy. We will talk about this fairness issue more in the end of next chapter.

Then we extend the result of [Atar, 2008] to systems with abandonments. This is an important extension as it ensures the stability of the diffusion limits. We use a martingale central limit theorem to prove weak convergence. In the next chapter, we show how the diffusion limits are applied to our optimal staffing problems.

Chapter 3

Staffing and Routing for Single Server Pool Systems

3.1 Introduction

For call centre managers, how to decide on number of servers to be scheduled servers is one of the major problems. Overstaffing and understaffing will both cause immense unnecessary costs in the long term. Using queueing models, we can help managers make wise decisions on staffing levels. Particularly, the diffusion limits will be used in the analysis. In the last chapter, we proved diffusion limits for single server pool systems. Diffusion limits give us approximations of how the system processes behave. If steady states exist for diffusion limits, we can then have estimates for system steady states in the long run. In this chapter, first we formulate the optimisation problem for staffing single server pools without abandonments. Then in the second section, we extend this result to systems with abandonments. In the third section, we focus on the variance of service rates and show how the variance influences the abandonment rate, and thus the total costs. Finally in the fourth section, we analyse the coefficient γ in the diffusion limits and show how it reflects fairness among servers under different routing policies.

3.2 Staffing many server queues with random servers

We use a similar framework as in [Borst et al., 2004] for asymptotic optimisation of many-server queuing systems with random service rates. Consider the many-server queueing model without abandonments. In the rth system, arrival rate λ^r, service rates μ_k and its expectation $\bar{\mu}$ are as defined in Chapter 2.

Recall the second moment condition for arrival rate

$$\lim_{r \to \infty} \frac{1}{\sqrt{r}}(\lambda^r - r\lambda) = \hat{\lambda} < 0, \tag{3.2.1}$$

where $\lambda = \bar{\mu}$. It is easy to see that when $\sum_{k=1}^{N^r} \mu_k \leq \lambda^r$, the system is unstable, thus in the limit, every customer will have to wait before getting service. To this end, we assume there is a fixed waiting cost C_{un} for unstable systems.

In this work, we will mainly focus on the scenario where the service rates satisfy $\sum_{k=1}^{N^r} \mu_k > \lambda^r$. For the diffusion limit in Theorem 2.3.1, this condition corresponds to $\beta < 0$. i.e. when the system is stable. From [Cooper, 1981], we know that, given the realisation of $\sum_{k=1}^{N^r} \mu_k$, the waiting time distribution is given by

$$\mathbb{P}\left(\text{Wait}^r > t \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right) = \pi^r e^{-(H-\lambda^r)t}, \qquad (3.2.2)$$

where $\pi^r = \mathbb{P}\left(\text{Wait}^r > 0 \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right)$ is the probability of waiting. Notice that H has to be greater than λ^r for stability.

Let $F(N^r)$ be the staffing cost per unit time, and $D^r(t)$ be the waiting cost of a customer when s/he waits for t time units. Without loss of generality we may take $D^r(0) = 0$. Then the conditional expected total cost per unit of time is given by

$$C\left(N^r, \lambda^r \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right) = F(N^r) + \lambda^r \mathbb{E}\left(D^r(\text{Wait}) \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right)$$

$$= F(N^r) + \lambda^r \pi^r G\left(N^r, \lambda^r \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right),$$

where

$$G\left(N^r, \lambda^r \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right) = \mathbb{E}\left(D^r(\text{Wait}) \,\middle|\, \text{Wait} > 0, \sum_{k=1}^{N^r} \mu_k = H\right)$$

$$= (H - \lambda^r) \int_0^\infty D^r(t) e^{-(H-\lambda^r)t} dt. \qquad (3.2.3)$$

We are interested in determining the expected optimal staffing level

$$N_*^r := \underset{N^r > \lambda^r/\bar{\mu}}{\arg\min}\, C(N^r, \lambda^r), \qquad (3.2.4)$$

where $C(N^r, \lambda^r) = \int_{\lambda^r}^\infty C\left(N^r, \lambda^r \,\middle|\, \sum_{k=1}^{N^r} \mu_k = H\right) f_r(H) dH \frac{1}{\mathbb{P}(\sum_{k=1}^{N^r} \mu_k > \lambda^r)}$, and $f_r(\cdot)$ is the density function of $\sum_{k=1}^{N^r} \mu_k$.

3.2.1 Framework of the asymptotic optimisation problem

We use a similar framework as in [Borst et al., 2004], where, the cost function contains an expected waiting cost, which is the product of the arrival rate and the expected waiting time of a single customer. We use the same concept, but our expected waiting cost contains two recursive expectations instead of one. The first one is the normal expectation of the waiting cost

when the sum of the service rate is given, then we take the second expectation over the sum. As the first step, we translate the discrete optimisation problem (3.2.4) to a continuous one, and approximate the latter problem by a related continuous version, which is easier to solve. Finally, we prove that the optimal solution to the approximating continuous problem provides an asymptotically optimal solution to the original discrete problem. Our main contribution is in the last step. To show the validity of the continuous approximation, we need to prove that limits are interchangeable as shown in Figure 3.1. And to show this, we prove tightness of the sequence of steady state distributions.

We first translate the discrete problem into a continuous one. Let

$$
N^r(x) = \frac{\lambda^r}{\bar{\mu}} + x \sqrt{\frac{\lambda^r}{\bar{\mu}}}, \tag{3.2.5}
$$

so that the variable $x = (N^r - \frac{\lambda^r}{\bar{\mu}})/\sqrt{\frac{\lambda^r}{\bar{\mu}}}$ is the normalized number of servers in excess of the minimum number $\frac{\lambda^r}{\bar{\mu}}$ required for stability. In terms of x, we define

$$
F^r(x) := F(N^r(x)) - F\left(\frac{\lambda^r}{\bar{\mu}}\right),
$$

$$
G^r\left(x \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) := \lambda^r G\left(N^r(x), \lambda^r \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right),
$$

$$
C^r\left(x \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) := C\left(N^r(x), \lambda^r \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) - F\left(\frac{\lambda^r}{\bar{\mu}}\right),
$$

$$
\pi^r\left(x \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) := \text{The probability of waiting in the } r\text{th system}
$$

given the sum of service rates to be H.

Then the total cost per unit of time can be rewritten as

$$
C^r(x) = F^r(x) + \int_{\lambda^r}^{\infty} \pi^r\left(x \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) G^r\left(x \,\bigg|\, \sum_{k=1}^{\left|N^r(x)\right|} \mu_k = H\right) f_r(H) dH \frac{1}{\mathbb{P}(\sum_{k=1}^{N^r(x)} \mu_k > \lambda^r)}. \tag{3.2.6}
$$

Denote

$$
x_*^r := \underset{x>0}{\arg\min}\, C^r(x). \tag{3.2.7}
$$

Next, we a use simpler version of continuous function to approximate the function $C^r(x)$. If we can find simpler approximations for both π^r and G^r, we will have an approximation for the new cost function (3.2.6). In the next section, we will show the feasibility of such approximations.

3.2.2 Validity of the approximating model

In this section, we provide lemmas with proofs to show that we can use a simpler function to estimate the continuous cost function (3.2.6).

Approximation of π^r

First, we will find an approximating version for $\pi^r(x \mid \sum_{k=1}^{N^r(x)} \mu_k = H^r)$. We achieve this by calculating the steady state of the diffusion limit in Theorem 2.1 in [Atar, 2008]. As we discussed before, when $\beta > 0$, the diffusion $\xi(t)$ is not stable, so here we focus on the situation when $\beta < 0$, under which the limiting diffusion $\xi(t)$ has following expressions

$$\xi(t) = \begin{cases} \xi(0) + \sigma w(t) + \beta t & \xi(t) \geq 0 \\ \xi(0) + \sigma w(t) + \beta t - \gamma \int_0^t \xi(s)ds & \xi(t) < 0 \end{cases}, \qquad (3.2.8)$$

where

$$\sigma^2 = \bar{\mu}(C_{\check{U}}^2 + 1), \text{ and } \gamma = \frac{\int x^2 dm}{\int x dm}. \qquad (3.2.9)$$

Then by Section 4 of [Browne and Whitt, 1994], when $\xi(t) \geq 0$, it is a reflected Brownian motion. Thus its steady state density function conditional on $\xi(\infty) \geq 0$ is exponential, and will be $g_e(x) = \frac{-2\beta}{\sigma^2} e^{\frac{2\beta}{\sigma^2} x}$. Similarly, the process $\xi(t)$ restricted to the negative half-line is an O-U process , thus its steady state conditional on $\xi(\infty) < 0$ is normally distributed with density function $g_r(x) = \frac{\sqrt{2\gamma}/\sigma \phi(\frac{\sqrt{2\gamma}}{\sigma} x - \frac{\sqrt{2}\beta}{\sqrt{\gamma}\sigma})}{\Phi(\frac{-\sqrt{2}\beta}{\sqrt{\gamma}\sigma})}$. Hence, let ϱ be the probability of waiting when $\beta < 0$, the steady state distribution of $\xi(t)$ has density function

$$f(x) = \begin{cases} g_e(x)\varrho, & x \geq 0 \\ g_r(x)(1-\varrho), & x < 0 \end{cases}. \qquad (3.2.10)$$

Since the variance of the limiting diffusion is a constant on the real line, by [Browne and Whitt, 1994], the density function of its steady state should be continuous. ϱ may be solved by equating the right-handed limit of $g_e(\cdot)\varrho$ and the left-handed limit of $g_r(\cdot)(1-\varrho)$ at 0. This gives us

$$\varrho = P(y) = \mathbb{P}(\xi(\infty) \geq 0 \mid \beta = y) = \left(1 - \frac{\frac{\sqrt{2}y}{\sqrt{\gamma}\sigma} \Phi(-\frac{\sqrt{2}y}{\sqrt{\gamma}\sigma})}{\phi(\frac{\sqrt{2}y}{\sqrt{\gamma}\sigma})}\right)^{-1}. \qquad (3.2.11)$$

Thus we will have an analogue of Proposition 1 in [Halfin and Whitt, 1981] and Lemma 5.1 in [Borst et al., 2004], which gives us an approximation of π^r.

Lemma 3.2.1. *For any function x^r with $\limsup_{r \to \infty} x^r < \infty$,*

$$\lim_{r \to \infty} \frac{\pi^r \left(x^r \Big| \sum_{k=1}^{N^r(x^r)} \mu_k = H\right)}{P(y^r)} = 1, \ \forall H > N^r(x^r)\bar{\mu} - x\bar{\mu}\sqrt{r}, \qquad (3.2.12)$$

32

$$\hat{X}^r(t) \xrightarrow{\quad t \to \infty \quad} \hat{X}^r(\infty)$$

$$r \to \infty \Big\downarrow \qquad\qquad ? \Big\downarrow r \to \infty$$

$$\xi(t) \xrightarrow{\quad t \to \infty \quad} \xi(\infty)$$

Figure 3.1: The interchange-of-limit diagram

where $y^r = -\left(H - N^r(x^r)\bar{\mu}\right)\frac{1}{\sqrt{r}} - x^r\bar{\mu}$.

Lemma 5.1 in [Borst et al., 2004] is a direct application of Proposition 1 in [Halfin and Whitt, 1981], which says that, under the Halfin-Whitt heavy traffic condition, the probability of waiting in many server queue (identical servers) converges to a constant between zero and one. For our systems, service rates are no longer identical, and thus we cannot use this result directly. We come up with another approach to prove the convergence, which involves an associated sequence of homogeneous systems.

The idea of the proof lies in Figure 3.1. In order to show the convergence of the probability of waiting in (3.2.12), we need to show that the sequence of steady states $\hat{X}^r(\infty)$ weakly converges to the steady state $\xi(\infty)$ of the diffusion limit. To show weak convergence, the usual way is to first show tightness, then use the relations in Figure 3.1 to show that tightness of $\hat{X}^r(\infty)$ actually reflects the convergence on the right hand side.

The challenging part is to show the tightness. It is hard to come up with a direct way to show tightness if service rates are random and unknown. Instead, we compare our heterogeneous systems with a sequence of homogeneous systems which have identical service rates being less than $\bar{\mu}$, then use the properties of homogeneous systems to get tightness results.

More specifically, consider homogeneous systems with N^r servers and identical service rates $\mu^r < \bar{\mu}$. Assume there is a fixed number $M^r < N^r$ for each r, and denote Δ_{M^r} as the family of sets containing M^r numbers out of $\{1, 2, \ldots, N^r\}$. Now our problem can be considered in the following two cases:

- When $\inf_{\delta \in \Delta_{M^r}} \sum_\delta \mu_k < M^r \mu^r$, denote this scenario as A_r. (This A_r is independent from the arrival process $A^r(t)$.) There is a possibility that the heterogeneous systems serve faster than their corresponding homogeneous ones.

- When $\inf_{\delta \in \Delta_{M^r}} \sum_\delta \mu_k \geq M^r \mu^r$, this scenario is marked as A_r^c. The heterogeneous systems always serve faster than their corresponding homogeneous ones. Thus under such situations, $X^r(t) \leq_{st} X_{hom}^r(t)$. Therefore to show tightness of heterogeneous systems for this scenario, we only need to show the tightness of their corresponding homogeneous

systems, i.e.

$$\forall \epsilon > 0, \exists K^\epsilon > 0, \text{ s.t. } \forall r, \mathbb{P}(X^r_{hom}(t) \geq K^\epsilon) \leq \epsilon. \tag{3.2.13}$$

Combining these two scenarios, let $\bar{K}^\epsilon = K^\epsilon \vee M^r$. Then

$$\mathbb{P}(X^r(t) \geq \bar{K}^\epsilon) = \mathbb{P}(X^r(t) \geq \bar{K}^\epsilon, A_r) + \mathbb{P}(X^r(t) \geq \bar{K}^\epsilon, A^c_r)$$

$$\leq \mathbb{P}(A_r) + \mathbb{P}(X^r_{hom}(t) \geq \bar{K}^\epsilon). \tag{3.2.14}$$

If (3.2.13) is true, then after a reselection of ϵ, we can easily show $\mathbb{P}(X^r_{hom}(t) \geq \bar{K}^\epsilon) < \epsilon/2$. Furthermore, if we can also show $\mathbb{P}(A_r) \to 0$ as $r \to \infty$, then $\mathbb{P}(X^r(t) \geq \bar{K}^\epsilon) \to 0$ as $r \to \infty, \forall t > 0$, the tightness result will follow. Since the tightness result holds for every t, it should also hold as $t \to \infty$, i.e. $\mathbb{P}(X^r(\infty) \geq \bar{K}^\epsilon) < \epsilon$.

From the discussion above, it is important to define proper μ^r and M^r to get our convergence results. We have three lemmas for that.

Lemma 3.2.2. *Let $\mu^r = \bar{\mu} - \frac{1}{r^p}$ for any $p > \frac{1}{2}$, then the sequence $\{X^r_{hom}(\infty)\}$ is tight.*

Proof. In order to show $\{X^r_{hom}(\infty)\}$ is tight, first we need to specify the heavy traffic condition for the homogeneous systems and the existence of their steady states.

When

$$\mu^r = \bar{\mu} - \frac{1}{r^p}, p > \frac{1}{2}, \tag{3.2.15}$$

the heavy traffic condition becomes

$$\lim_{r \to \infty} \frac{1}{\sqrt{r}}(r\mu^r - \lambda^r) = \lim_{r \to \infty} \frac{1}{\sqrt{r}}\left(r\left(\bar{\mu} - \frac{1}{r^p}\right) - \lambda^r\right)$$

$$= \lim_{r \to \infty} \frac{1}{\sqrt{r}}\left(r\bar{\mu} - \lambda^r - r^{1-p}\right) = \lim_{r \to \infty}\left(\frac{1}{\sqrt{r}}\left(r\bar{\mu} - \lambda^r\right) - r^{\frac{1}{2}-p}\right),$$

$$= \lim_{r \to \infty} \frac{1}{\sqrt{r}}\left(r\bar{\mu} - \lambda^r\right) = \hat{\lambda} > 0.$$

This also means $r\bar{\mu} > \lambda^r, \forall r$.

To guarantee the existence of their steady states, we need

$$r\mu^r = r\left(\bar{\mu} - \frac{1}{r^p}\right) > \lambda^r,$$

i.e.

$$r\bar{\mu} - r^{1-p} > \lambda^r, \tag{3.2.16}$$

and $p > \frac{1}{2}$ makes the inequality above hold.

Since we already showed the existence of $X^r_{hom}(\infty)$, now we can prove the tightness of it, i.e. (3.2.13) is true when $t \to \infty$. Denote $\rho^r = \frac{\lambda^r}{r\mu^r}$, then $\rho^r < 1, \forall r$. Fix r. We want to show

$$\forall \epsilon > 0, \exists K^\epsilon_r > 0, \text{ s.t. } \mathbb{P}(X^r_{hom}(\infty) \geq K^\epsilon_r) \leq \epsilon. \tag{3.2.17}$$

34

Choose $K_r^\epsilon > N^r$. By (1.1) and (1.3) in [Halfin and Whitt, 1981], we have

$$\mathbb{P}(X_{hom}^r(\infty) \geq K_r^\epsilon) = \sum_{k=K_r^\epsilon}^{\infty} \frac{(N^r)^{N^r}(\rho^r)^k}{N^r!}\eta = \eta\frac{(N^r)^{N^r}}{N^r!}\sum_{k=K_r^\epsilon}^{\infty}(\rho^r)^k,$$

where $\eta = \left(\frac{(N^r\rho^r)^{N^r}}{N^r!(1-\rho^r)} + \sum_{k=0}^{N^r-1}\frac{(N^r\rho^r)^k}{k!}\right)^{-1}$. Substitute η into the equation above, and since $0 < \rho^r < 1$, it becomes

$$\mathbb{P}(X_{hom}^r(\infty) \geq K_r^\epsilon) = \eta\frac{(N^r)^{N^r}}{N^r!}\frac{(\rho^r)^{K_r^\epsilon}(1-(\rho^r)^x)}{1-\rho^r} \xrightarrow{x\to\infty} \eta\frac{(N^r)^{N^r}}{N^r!}\frac{(\rho^r)^{K_r^\epsilon}}{1-\rho^r}$$

$$= \frac{(N^r)^{N^r}}{N^r!}\left(\frac{(N^r\rho^r)^{N^r}}{N^r!(1-\rho^r)} + \sum_{k=0}^{N^r-1}\frac{(N^r\rho^r)^k}{k!}\right)^{-1}\frac{(\rho^r)^{K_r^\epsilon}}{1-\rho^r}$$

$$= \left(\frac{(\rho^r)^{N^r}}{1-\rho^r} + \frac{N^r!}{(N^r)^{N^r}}\sum_{k=0}^{N^r-1}\frac{(N^r)^k}{k!}(\rho^r)^k\right)^{-1}\frac{(\rho^r)^{K_r^\epsilon}}{1-\rho^r}$$

$$= \left((\rho^r)^{N^r} + (1-\rho^r)\frac{N^r!}{(N^r)^{N^r}}\sum_{k=0}^{N^r-1}\frac{(N^r)^k}{k!}(\rho^r)^k\right)^{-1}(\rho^r)^{K_r^\epsilon}.$$

$\forall\epsilon > 0$, as long as

$$K_r^\epsilon > \log\left(\epsilon\left(\frac{(\rho^r)^{N^r}}{1-\rho^r} + \frac{N^r!}{(N^r)^{N^r}}\sum_{k=0}^{N^r-1}\frac{(N^r)^k}{k!}(\rho^r)^k\right)\right),$$

$\mathbb{P}(X_{hom}^r(\infty) \geq K_r^\epsilon) < \epsilon$ holds. Here log has base ρ^r. Thus (3.2.17) is proved.

For a sequence of homogeneous systems, take $K^\epsilon = K_1^\epsilon \vee K_2^\epsilon \vee \cdots \vee K_n^\epsilon \vee \cdots$, then (3.2.13) holds when $t \to \infty$. $\qquad\square$

From Lemma 3.2.2, in order to guarantee heavy traffic condition and stability of the homogeneous systems, service rate of homogeneous systems is chosen with

$$p > \frac{1}{2} \tag{3.2.18}$$

in the rest of this chapter.

As for the choice of M^r, we will explain in the following two lemmas that when M^r grows slower than N^r, the convergence does not hold. Only when M^r grows with the same rate as N^r, the probability $\mathbb{P}(A_r)$ converges to zero.

Lemma 3.2.3. *Let $\mu^r = \bar{\mu} - \frac{1}{r^p}$. If M^r grows slower than N^r in the sense that $\lim_{r\to\infty}\frac{M^r}{N^r} = 0$, then $\mathbb{P}(A_r) \to \infty$.*

Proof. Fix r. Since δ is one element in Δ_{M^r}, denote $B_\delta = \{\sum_\delta \mu_k < M^r\mu^r\}$, we have

$$\mathbb{P}(A_r) = \mathbb{P}\left(\bigcup_\delta B_\delta\right) \leq \sum_{\delta\in\Delta_{M^r}}\mathbb{P}(B_\delta) = \binom{N^r}{M^r}\mathbb{P}(B_\delta). \tag{3.2.19}$$

Using stirling's approximation, we have a lower bound for the combination factor

$$\binom{N^r}{M^r} = \frac{N^r!}{M^r!\,(N^r - M^r)!} \geq \frac{\sqrt{2\pi}(N^r)^{N^r+\frac{1}{2}}e^{-N^r}}{e(M^r)^{M^r+\frac{1}{2}}e^{-M^r}e(N^r - M^r)^{N^r-M^r+\frac{1}{2}}e^{-(N^r-M^r)}}$$

$$= \frac{\sqrt{2\pi}}{e^2}\frac{(N^r)^{N^r+\frac{1}{2}}}{(M^r)^{M^r+\frac{1}{2}}(N^r - M^r)^{N^r-M^r+\frac{1}{2}}}. \tag{3.2.20}$$

Thus as $r \to \infty$,

$$\binom{N^r}{M^r}\mathbb{P}\left(\sum_\delta \frac{\mu_k}{M^r} < \mu^r\right) \to \binom{N^r}{M^r}\mathbb{P}(Y^r < \mu^r)$$

$$= \frac{N^r!}{M^r!\,(N^r - M^r)!}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\frac{\sqrt{M^r}}{\sigma r^p}}e^{-\frac{t^2}{2}}dt$$

$$\geq \frac{\sqrt{2\pi}}{e^2}\frac{(N^r)^{N^r+\frac{1}{2}}}{(M^r)^{M^r+\frac{1}{2}}(N^r - M^r)^{N^r-M^r+\frac{1}{2}}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\frac{\sqrt{M^r}}{\sigma r^p}}e^{-\frac{t^2}{2}}dt$$

$$= \frac{1}{e^2}\frac{(N^r)^{M^r}}{(M^r)^{M^r}}\frac{(N^r)^{N^r-M^r}}{(N^r - M^r)^{N^r-M^r}}\left(\frac{N^r}{M^r(N^r - M^r)}\right)^{\frac{1}{2}}\int_{-\infty}^{-\frac{\sqrt{M^r}}{\sigma r^p}}e^{-\frac{t^2}{2}}dt. \tag{3.2.21}$$

Notice that the second item $(\frac{N^r}{M^r})^{M^r} \to \infty$ as $r \to \infty$, the third item can be considered as follows:

$$\left(\frac{N^r}{N^r - M^r}\right)^{(N^r-M^r)} = \left(1 - 1 + \frac{N^r}{N^r - M^r}\right)^{(N^r-M^r)} = \left(1 + \frac{M^r}{N^r - M^r}\right)^{(N^r-M^r)}$$

$$= \left(1 + \frac{1}{\frac{N^r}{M^r} - 1}\right)^{(\frac{N^r}{M^r}-1)M^r} \to \exp(M^r) \text{ as } r \to \infty,$$

and the fourth item equals $\sqrt{\frac{1}{M^r(1-\frac{M^r}{N^r})}}$, where $\sqrt{\frac{1}{1-\frac{M^r}{N^r}}}$ converges to 1. As for the integral, the upper limit is equal to

$$-\frac{\sqrt{M^r}}{\sqrt{N^r}}\frac{\sqrt{N^r}}{\sqrt{r}}\frac{\sqrt{r}}{r^p} = -\frac{\sqrt{M^r}}{\sqrt{N^r}}\frac{\sqrt{N^r}}{\sqrt{r}}\frac{1}{r^{p-\frac{1}{2}}} \to 0$$

as $r \to \infty$, because $p > \frac{1}{2}$. So as $r \to \infty$, (3.2.21) is equivalent to

$$\frac{1}{e^2}\left(\frac{N^r}{M^r}\right)^{M^r}\left(1 + \frac{1}{\frac{N^r}{M^r} - 1}\right)^{(\frac{N^r}{M^r}-1)M^r}\sqrt{\frac{1}{M^r}}\sqrt{\frac{1}{1-\frac{M^r}{N^r}}}\int_{-\infty}^{-\frac{\sqrt{M^r}}{\sqrt{N^r}}\frac{\sqrt{N^r}}{\sqrt{r}}\frac{1}{r^{p-\frac{1}{2}}}}e^{-\frac{t^2}{2}}dt$$

$$\to \frac{1}{e^2}\left(\frac{N^r}{M^r}\right)^{M^r}\exp(M^r)\frac{1}{\sqrt{M^r}}\int_{-\infty}^{0}e^{-\frac{t^2}{2}}dt$$

$$\to \frac{1}{e^2}\cdot\infty\cdot\infty\cdot\frac{1}{2} \to \infty. \tag{3.2.22}$$

Thus the original $\binom{N^r}{M^r}\mathbb{P}\left(\sum_\delta \frac{\mu_k}{M^r} < \mu^r\right)$ will not converge to zero, and therefore we rule out this situation. $\qquad\square$

Now we conclude that there must be $M^r = C_3^r$ for some C_3^r such that $\mathbb{P}(A_r) \to 0$ holds.

Lemma 3.2.4. *Let $\mu^r = \bar{\mu} - \frac{1}{r^p}$, and $M^r = C_3^r N^r$ for some $C_3^r \in (0,1)$. Then $\mathbb{P}(A_r) \to 0$ as $r \to \infty$.*

Proof. Assume $M^r \to \infty$ as $r \to \infty$. Then $\frac{1}{M^r} \sum_\delta \mu_k$ weakly converges to a normal random variable Y^r with mean $\bar{\mu}$, variance $\frac{\sigma^2}{M^r}$ as $r \to \infty, \forall \delta$. Using the left bound for normal distribution,

$$\frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-\frac{x^2}{2}} \leq \mathbb{P}(X > x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{x^2}{2}}.$$

As $r \to \infty$, we have

$$\mathbb{P}(B_\delta) \to \mathbb{P}(Y^r < \mu^r) = \mathbb{P}\left(\frac{\sqrt{M^r}}{\sigma}(Y^r - \bar{\mu}) < \frac{\sqrt{M^r}}{\sigma}(\mu^r - \bar{\mu}) \right)$$

$$\leq 1 - \frac{1}{\sqrt{2\pi}} \frac{\frac{\sqrt{M^r}}{\sigma}(\mu^r - \bar{\mu})}{\frac{M^r}{\sigma^2}(\mu^r - \bar{\mu})^2 + 1} \exp\left(-\frac{M^r(\mu^r - \bar{\mu})^2}{2\sigma^2} \right), \tag{3.2.23}$$

then applying stirling's approximation again on the combination factor in (3.2.19) gives

$$\binom{N^r}{M^r} = \frac{N^r!}{M^r!(N^r - M^r)!} \leq \frac{e(N^r)^{(N^r + \frac{1}{2})}}{2\pi(M^r)^{(M^r + \frac{1}{2})}(N^r - M^r)^{(N^r - M^r + \frac{1}{2})}}. \tag{3.2.24}$$

Combine (3.2.23) and (3.2.24) and we have, as $r \to \infty$,

$$\mathbb{P}(A_r) = \binom{N^r}{M^r} \mathbb{P}\left(\sum_\gamma \frac{\mu_k}{M^r} < \mu^r \right) \to \binom{N^r}{M^r} \mathbb{P}(Y^r < \mu^r)$$

$$= \binom{N^r}{M^r} \mathbb{P}\left(\frac{\sqrt{M^r}}{\sigma}(Y^r - \bar{\mu}) < \frac{\sqrt{M^r}}{\sigma}(\mu^r - \bar{\mu}) \right)$$

$$\leq \frac{e(N^r)^{(N^r + \frac{1}{2})}}{2\pi(M^r)^{(M^r + \frac{1}{2})}(N^r - M^r)^{(N^r - M^r + \frac{1}{2})}}$$

$$\left(1 - \frac{1}{\sqrt{2\pi}} \frac{\frac{\sqrt{M^r}}{\sigma}(\mu^r - \bar{\mu})}{\frac{M^r}{\sigma^2}(\mu^r - \bar{\mu})^2 + 1} \exp\left(-\frac{M^r(\mu^r - \bar{\mu})^2}{2\sigma^2} \right) \right). \tag{3.2.25}$$

For notational simplicity let $C_2^r = \frac{\mu^r - \bar{\mu}}{\sigma} = -\frac{1}{\sigma r^p} < 0$. Since (3.2.25) is an upper bound of $\mathbb{P}(A_r)$, if we can demonstrate (3.2.25) converges to zero as $r \to \infty$, then $\mathbb{P}(A_r) \to 0$ as $r \to \infty$. Assume $C_3^r + C_4^r = 1$ for some $0 < C_4^r < 1$ and $M^r = C_3^r N^r$, then the right hand side of (3.2.25) can be rewritten as

$$\frac{e(N^r)^{(N^r + \frac{1}{2})}}{2\pi(C_3^r N^r)^{(C_3^r N^r + \frac{1}{2})}(N^r - C_3^r N^r)^{(N^r - C_3^r N^r + \frac{1}{2})}}$$

$$- \frac{e(N^r)^{(N^r + \frac{1}{2})}}{2\pi(C_3^r N^r)^{(C_3^r N^r + \frac{1}{2})}(N^r - C_3^r N^r)^{(N^r - C_3^r N^r + \frac{1}{2})}} \frac{1}{\sqrt{2\pi}} \frac{\sqrt{C_3^r N^r} C_2^r}{C_3^r N^r (C_2^r)^2 + 1} \exp\left(-\frac{1}{2} C_3^r N^r (C_2^r)^2 \right)$$

$$= \frac{e(N^r)^{N^r}(N^r)^{\frac{1}{2}}}{2\pi \left((C_3^r)^{C_3^r}\right)^{N^r} \left((N^r)^{N^r}\right)^{C_3^r} (C_3^r)^{\frac{1}{2}}(N^r)^{\frac{1}{2}} \left((C_4^r)^{C_4^r}\right)^{N^r} \left((N^r)^{N^r}\right)^{C_4^r} (N^r)^{\frac{1}{2}}}$$

$$-\frac{e}{2\pi\sqrt{2\pi}}\frac{(N^r)^{N^r}(N^r)^{\frac{1}{2}}}{((C_3^r)^{C_3^r})^{N^r}((N^r)^{N^r})^{C_3^r}(C_3^r)^{\frac{1}{2}}(N^r)^{\frac{1}{2}}((C_4^r)^{C_4^r})^{N^r}((N^r)^{N^r})^{C_4^r}(N^r)^{\frac{1}{2}}}$$

$$\frac{(C_3^rN^r)^{\frac{1}{2}}C_2^r}{C_3^rN^r(C_2^r)^2+1}\exp\left(-\frac{1}{2}C_3^rN^r(C_2^r)^2\right)$$

$$=\frac{e}{2\pi}\frac{1}{((C_3^r)^{C_3^r}(C_4^r)^{C_4^r})^{N^r}(C_3^rC_4^r)^{\frac{1}{2}}(N^r)^{\frac{1}{2}}}$$

$$-\frac{e}{2\pi\sqrt{2\pi}}\frac{1}{((C_3^r)^{C_3^r}(C_4^r)^{C_4^r})^{N^r}(C_3^rC_4^r)^{\frac{1}{2}}(N^r)^{\frac{1}{2}}}\frac{(C_3^rN^r)^{\frac{1}{2}}C_2^r}{C_3^rN^r(C_2^r)^2+1}\exp\left(-\frac{1}{2}C_3^rN^r(C_2^r)^2\right)$$

$$=\frac{e}{\sqrt{2\pi}}\frac{1}{((C_3^r)^{C_3^r}(C_4^r)^{C_4^r})^{N^r}(C_3^rC_4^r)^{\frac{1}{2}}}\left(\frac{1}{\sqrt{N^r}}-\frac{1}{\sqrt{2\pi}}\frac{(C_3^r)^{\frac{1}{2}}}{(C_3^rN^rC_2^r+\frac{1}{C_2^r})\exp(\frac{1}{2}C_3^rN^r(C_2^r)^2)}\right)$$

$$=\frac{e}{\sqrt{2\pi}}\frac{1}{((C_3^r)^{C_3^r}(C_4^r)^{C_4^r})^{N^r}(C_3^rC_4^r)^{\frac{1}{2}}}\frac{1}{\exp(\frac{1}{2}C_3^rN^r(C_2^r)^2)}$$

$$\left(\frac{\exp(\frac{1}{2}C_3^rN^r(C_2^r)^2)}{\sqrt{N^r}}-\frac{1}{\sqrt{2\pi}}\frac{(C_3^r)^{\frac{1}{2}}}{C_3^rN^rC_2^r+\frac{1}{C_2^r}}\right)$$

$$=\frac{e}{\sqrt{2\pi}}\frac{1}{((C_3^r)^{C_3^r}(C_4^r)^{C_4^r}e^{\frac{1}{2}C_3^r(C_2^r)^2})^{N^r}(C_3^rC_4^r)^{\frac{1}{2}}}\left(\frac{\exp(\frac{1}{2}C_3^rN^r(C_2^r)^2)}{\sqrt{N^r}}-\frac{1}{\sqrt{2\pi}}\frac{(C_3^r)^{\frac{1}{2}}}{C_3^rN^rC_2^r+\frac{1}{C_2^r}}\right).$$

$$(3.2.26)$$

First notice that two items in the parentheses both converge to zero, because of the following:

$$\frac{\exp(\frac{1}{2}C_3^rN^r(C_2^r)^2)}{\sqrt{N^r}}=\frac{\exp(\frac{1}{2}C_3^r\frac{N^r}{\sigma^2r^{2p}})}{\sqrt{N^r}}.\qquad(3.2.27)$$

Since $p>\frac{1}{2}$, $\frac{N^r}{\sigma^2r^{2p}}\to0$, (3.2.27) converges to zero as $r\to\infty$; similarly,

$$\frac{(C_3^r)^{\frac{1}{2}}}{C_3^rN^rC_2^r+\frac{1}{C_2^r}}=\frac{(C_3^r)^{\frac{1}{2}}}{C_3^rN^r(-\frac{1}{\sigma r^p})-\sigma r^p}.\qquad(3.2.28)$$

Again, since $p>\frac{1}{2}$, $\frac{N^r}{r^p}>0$, $\sigma r^p\to+\infty$ as $r\to\infty$, (3.2.28) converges to zero as $r\to\infty$. Therefore the difference in the parentheses converges to zero as $r\to\infty$, as long as we can show the left factor in (3.2.26) converges to zero as $r\to\infty$, we can claim (3.2.26) converges to zero as $r\to\infty$.

Obviously, if $0<(C_3^r)^{-C_3^r}(C_4^r)^{-C_4^r}e^{-\frac{1}{2}C_3^r(C_2^r)^2}<1$, then the factor in (3.2.26) converges to zero as $r\to\infty$. The first $<$ is trivial, As for the second $<$,

$$(C_3^r)^{-C_3^r}(C_4^r)^{-C_4^r}e^{-\frac{1}{2}C_3^r(C_2^r)^2}<1\Leftrightarrow\frac{1}{(e^{\frac{1}{2}(C_2^r)^2})^{C_3^r}}<(C_3^r)^{C_3^r}(C_4^r)^{C_4^r}$$

$$\Leftrightarrow\frac{1}{e^{\frac{1}{2}(C_2^r)^2}}<C_3^r(C_4^r)^{\frac{C_4^r}{C_3^r}}=C_3^r(1-C_3^r)^{\frac{1-C_3^r}{C_3^r}}$$

$$\Leftrightarrow-\frac{1}{2}(C_2^r)^2<\ln C_3^r+\frac{1-C_3^r}{C_3^r}\ln(1-C_3^r)$$

$$\Leftrightarrow-\frac{(\bar{\mu}-\mu^r)^2}{2\sigma^2}<\ln C_3^r+\frac{1-C_3^r}{C_3^r}\ln(1-C_3^r)$$

$$\Leftrightarrow-\frac{1}{2\sigma^2r^{2p}}<\ln C_3^r+\frac{1-C_3^r}{C_3^r}\ln(1-C_3^r).\qquad(3.2.29)$$
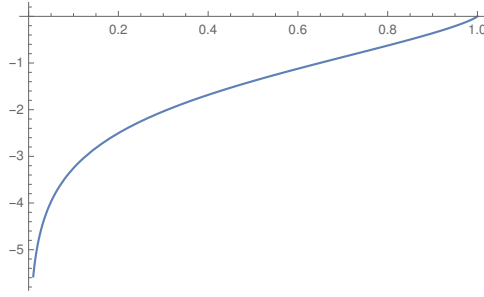
Figure 3.2: $f(C_3^r)$ on interval $(0, 1)$

The graph of function $f(C_3^r) = \ln C_3^r + \frac{1-C_3^r}{C_3^r} \ln(1 - C_3^r), C_3^r \in (0, 1)$ is shown in Figure 3.2. We can then choose C_3^r as

$$C_3^r = \min_{0<x<1} \left\{ x : -\frac{1}{2\sigma^2 r^{2p}} < \ln x + \frac{1-x}{x} \ln(1-x) \right\}. \tag{3.2.30}$$

If both (3.2.15) and (3.2.30) are satisfied, $\mathbb{P}(A_r) \to 0$ as $r \to \infty$ as required. $\qquad\square$

Now we can prove Lemma 3.2.1.

Proof. Since our system is under the Halfin-Whitt heavy traffic regime, assume the staffing is determined by

$$N^r = \frac{\lambda^r}{\bar{\mu}} + \theta \sqrt{\frac{\lambda^r}{\bar{\mu}}}. \tag{3.2.31}$$

Then from Theorem 2.1 in [Atar, 2008], the random drift is

$$\beta = \hat{\lambda} - \zeta - \bar{\mu}\nu. \tag{3.2.32}$$

Here we take $\hat{\mu} = 0$ since the purpose of the term $\hat{\mu}_k$ is just to make the random service rate more general. By substituting (2.3.3) and (3.2.31) into (3.2.32), we have

$$\begin{aligned}
\beta &= \hat{\lambda} - \zeta - \lim_{r\to\infty} \bar{\mu}\frac{1}{\sqrt{r}}(N^r - r) = \hat{\lambda} - \zeta - \lim_{r\to\infty} \bar{\mu}\frac{1}{\sqrt{r}}\left(\frac{\lambda^r}{\bar{\mu}} + \theta\sqrt{\frac{\lambda^r}{\bar{\mu}}} - r\right) \\
&= \hat{\lambda} - \zeta - \lim_{r\to\infty} \bar{\mu}\frac{1}{\sqrt{r}}\frac{\lambda^r + \theta\sqrt{\lambda^r\bar{\mu}} - r\bar{\mu}}{\bar{\mu}} \\
&= \hat{\lambda} - \zeta - \lim_{r\to\infty} \left(\frac{\lambda^r - r\bar{\mu}}{\sqrt{r}} + \theta\sqrt{\bar{\mu}}\sqrt{\frac{\lambda^r}{r}}\right) \\
&= \hat{\lambda} - \zeta - \hat{\lambda} - \theta\bar{\mu} = -\zeta - \theta\bar{\mu},
\end{aligned}$$

by (2.2.2) and the assumption $\lim_{r\to\infty} \frac{\lambda^r}{r} = \lambda = \bar{\mu}$. Denote

$$\beta(\theta) = -\zeta - \theta\bar{\mu}. \tag{3.2.33}$$

Recall that ζ is a normal random variable with parameters $(0, \int(x-\bar{\mu})^2 dm)$, and $(\sum_{k=1}^{N^r} \mu_k -$

39

$N^r\bar{\mu})\frac{1}{\sqrt{r}} \Rightarrow \zeta$. For the rth system, let $\theta = x^r$ be its extra staffing which is required for stability, and

$$\beta^r(x^r) = -\left(\sum_{k=1}^{N^r(x^r)} \mu_k - N^r(x^r)\bar{\mu}\right)\frac{1}{\sqrt{r}} - x^r\bar{\mu}. \qquad (3.2.34)$$

Keep x^r fixed for every r, then $\beta^r(x^r) \Rightarrow \beta(x^r)$ as $r \to \infty$. When $\sum_{k=1}^{N^r(x^r)} \mu_k = H$, denote $\beta^r(x^r)$ as $\beta^r_H(x^r)$.

Notice that

$$\pi^r\left(x^r \middle| \sum_{k=1}^{N^r(x^r)} \mu_k = H\right) = \mathbb{P}\left(\text{Wait}^r > 0 \middle| \beta^r(x^r) = -(H - N^r(x^r)\mu)\frac{1}{\sqrt{r}} - x^r\bar{\mu}\right). \qquad (3.2.35)$$

Choose ω_H such that

$$\zeta_H = \zeta(\omega_H) = (H - N^r(x^r)\bar{\mu})\frac{1}{\sqrt{r}}, \qquad (3.2.36)$$

then

$$P(y^r) = \mathbb{P}\left(\xi(\infty) > 0 \middle| \beta(x^r) = y^r = -\zeta_H - x^r\bar{\mu}\right). \qquad (3.2.37)$$

Then it remains to show that

$$\lim_{r\to\infty} \frac{\mathbb{P}\left(\text{Wait}^r > 0 \middle| \beta^r(x^r) = -(H - N^r(x^r)\bar{\mu})\frac{1}{\sqrt{r}} - x^r\bar{\mu}\right)}{\mathbb{P}\left(\xi(\infty) > 0 \middle| \beta(x^r) = -\zeta_H - x^r\bar{\mu}\right)} = 1, \ \forall H > N^r(x^r)\bar{\mu} - x\bar{\mu}\sqrt{r}. \qquad (3.2.38)$$

If we can show the family of random variables $\hat{X}^r(\infty)$ is tight, then we can claim $\hat{X}^r(\infty) \Rightarrow \xi(\infty)$ as $r \to \infty$ when $\beta^r < 0$, thus (3.2.38) is satisfied.

To show the tightness of $\{\hat{X}^r(\infty)\}$, we only need to show the tightness of $\{X^r(\infty)\}$. (If $\{X^r(\infty)\}$ is tight, then $\forall \epsilon > 0, \exists K^\epsilon > 0$, such that $\mathbb{P}(X^r(\infty) > K^\epsilon) < \epsilon, \forall r$. Then $\mathbb{P}(\hat{X}^r(\infty) > K^\epsilon) = \mathbb{P}((X^r(\infty) - r)/\sqrt{r} > K^\epsilon) = \mathbb{P}(X^r(\infty) > \sqrt{r}K^\epsilon + r) < \mathbb{P}(X^r(\infty) > K^\epsilon) < \epsilon, \forall r$.)

When μ^r and M^r are defined as in Lemma 3.2.2 and 3.2.4, $\{X^r_{hom}(\infty)\}$ are tight, and $\mathbb{P}(A_r) \to 0$ as $r \to \infty$. Thus the right side of (3.2.14) converges to zero. Let $t \to \infty$ on both sides of (3.2.14). We have that $\{X^r(\infty)\}$ is tight, hence $\{\hat{X}^r(\infty)\}$ is tight.

Now by the definition of tightness, for every sequence $\{\hat{X}^{r_i}(\infty)\}$ in $\{\hat{X}^r(\infty)\}$, there exists a subsequence $\{\hat{X}^{r_{i_j}}(\infty)\}$ weakly converging to a random variable Y. Figure 3.1 shows our goal. We already know the left, top and bottom arrows are true, now we need to prove the right one. Assume Y does not have the same distribution as $\xi(\infty)$ when $\beta < 0$. Take an infinite sequence $\{r_k\}$ and $\{t_l\}$, from the left and bottom arrows in Figure 3.1. The cumulative distribution functions (CDF) of random variables $\hat{X}^{r_k}(t_l)$ will converge to the CDF of $\xi(\infty)$ as $k \to \infty$ and $l \to \infty$. Next take a subsequence of $\{r_k\}$ and $\{t_l\}$, i.e. $\{r_{k_h}\}$ and $\{t_{l_g}\}$, then the CDFs of $\hat{X}^{r_{k_h}}(t_{l_g})$ will converge to the CDF of Y as $h \to \infty$ and $g \to \infty$. If $\xi(\infty)$ and Y are different random variables, then we find two subsequences of random variables' CDFs that converge to different limits, which means such a sequence does not converge, which furthermore contradicts

40

the fact that it is convergent (see left and bottom arrows in Figure 3.1). Thus Y should equal $\xi(\infty)$ in distribution, so we have proven that $\hat{X}^r(\infty) \Rightarrow \xi(\infty)$ as $r \to \infty$. Therefore (3.2.38) is proved, hence (3.2.12) is true. The proof is completed.

<div align="right">□</div>

Approximation of G^r

Next, we find an approximation for $G^r(\cdot)$, which will be more straightforward.

Lemma 3.2.5. *Denote $\beta_H(x^r) = -\zeta_H - x^r\mu \ \forall r$, where ζ_H is defined in (3.2.36), and let*

$$
\hat{G}\left(N^r, \lambda^r(x^r) \Bigg| \sum_{k=1}^{N^r} \mu_k = H \right) = -\beta_H(x^r)\sqrt{r} \int_0^\infty D^r(s) e^{\beta_H(x^r)\sqrt{r}s} ds, \tag{3.2.39}
$$

then $\hat{G}\left(N^r, \lambda^r(x^r) \Big| \sum_{k=1}^{N^r} \mu_k = H\right)$ is a valid approximation of (3.2.3), in the sense that

$$
\lim_{r \to \infty} \frac{\hat{G}\left(N^r, \lambda^r(x^r) \Big| \sum_{k=1}^{N^r} \mu_k = H\right)}{G\left(N^r, \lambda^r(x^r) \Big| \sum_{k=1}^{N^r} \mu_k = H\right)} = 1, \forall H > N^r(x^r)\bar{\mu} - x\bar{\mu}\sqrt{r}.
$$

Proof. Substituting $\beta^r(x^r)$ defined in (3.2.34) to (3.2.3), we have

$$
\begin{aligned}
G\left(N^r, \lambda^r(x^r) \Bigg| \sum_{k=1}^{N^r} \mu_k = H \right) &= (H - \lambda^r) \int_0^\infty D^r(s) e^{-(H-\lambda^r)s} ds \\
&= -\beta_H^r(x^r)\sqrt{r} \int_0^\infty D^r(s) e^{\beta_H^r(x^r)\sqrt{r}s} ds.
\end{aligned}
$$

Since $\beta^r(x^r) \Rightarrow \beta(x^r)$, the convergence is obtained. □

Verification of approximating function

Now apply the change of variables (3.2.34) into the original cost function (3.2.6), it can be rewritten as

$$
C^r(x) = F^r(x) + \lambda^r \frac{1}{\mathbb{P}(\beta^r(x) < 0)} L^r(x), \tag{3.2.40}
$$

where

$$
L^r(x) = \int_{-\infty}^0 \pi^r\left(x|\beta^r(x) = \beta_H^r(x)\right) \left(-\beta_H^r(x)\sqrt{r} \int_0^\infty D^r(s) e^{\beta_H^r s \sqrt{r}} ds\right) f_{\beta^r(x)}(\beta_H^r(x)) d\beta_H^r(x), \tag{3.2.41}
$$

and $f_{\beta^r(x)}(\cdot)$ is the probability density function of $\beta^r(x)$.

Now that we have approximations for both π^r and G^r, we will have an effective approximation for cost function (3.2.40). More specifically, with the new approximating function

$$
\hat{C}^r(x) = F^r(x) + \lambda^r \frac{1}{\mathbb{P}(\beta(x) < 0)} \hat{L}^r(x), \tag{3.2.42}
$$

<div align="center">41</div>

where

$$\hat{L}^r(x) = \int_{-\infty}^0 P\left(y^r | \beta(x) = \beta_H(x)\right) \left(-\beta_H(x)\sqrt{r}\int_0^\infty D^r(s)e^{\beta_H s \sqrt{r}}ds\right) f_{\beta(x)}(\beta_H(x))d\beta_H(x).$$

(3.2.43)

and $f_{\beta(x)}(\cdot)$ is the probability density function of $\beta(x)$, we have the following theorem

Theorem 3.2.1. $\hat{C}^r(x)$ *is a valid approximation for* $C^r(x)$ *in the sense that*

$$\lim_{r\to\infty} \frac{C^r(x)}{\hat{C}^r(x)} = 1, \text{ for any fixed } x.$$

(3.2.44)

Proof. Since we already have $\beta^r(x) \Rightarrow \beta(x)$ as $r \to \infty$ and $F^r(x)$ remains the same, we only need to show

$$\lim_{r\to\infty} \frac{L^r(x)}{\hat{L}^r(x)} = 1.$$

(3.2.45)

To make it more clear, let

$$Z(\beta^r(x)) = \pi^r(x)\left(-\beta^r(x)\sqrt{r}\int_0^\infty D^r(s)e^{\beta^r s \sqrt{r}}ds\right)$$

(3.2.46)

and

$$Z(\beta(x)) = P(y)\left(-\beta(x)\sqrt{r}\int_0^\infty D^r(s)e^{\beta s \sqrt{r}}ds\right),$$

(3.2.47)

where $y = -\left(\sum_{k=1}^{N^r(x)} \mu_k - N^r(x)\mu\right)\frac{1}{\sqrt{r}} - x\mu$. Then $\lim_{r\to\infty} \frac{L^r(x)}{\hat{L}^r(x)} = 1$ is actually

$$\lim_{r\to\infty} \frac{\mathbb{E}\left(Z(\beta^r(x))|\beta^r < 0\right)}{\mathbb{E}\left(Z(\beta(x))|\beta < 0\right)} = 1.$$

(3.2.48)

Since $\beta^r \Rightarrow \beta$ is proved, we only need to show $Z(\beta^r(x))$ is uniformly integrable when $\beta^r(x) < 0$, and $Z(\beta^r(x))$ is a continuous function of $\beta^r(x)$. Then by the Continuous Mapping Theorem and Theorem 3.5 in [Billingsley, 1999], one can conclude (3.2.48) is true.

To show $Z(\beta^r(x))$ is uniformly integrable, we only need to show

$$g^r(\beta^r(x)) := -\beta^r(x)\sqrt{r}\int_0^\infty D^r(s)e^{\beta^r s \sqrt{r}}ds$$

is integrable when $\beta^r(x) < 0$, since $\pi^r(x)$ are the probability of waiting and thus are bounded by 1.

It is easy to see the uniform integrability of $g^r(\beta^r(x))$. $D^r(s)$ is assumed to be a function such that $g^r(\beta^r(x))$ is finite. Since $\beta^r < 0$, given the realisation of $\beta^r(x)$, $-\beta^r s\sqrt{r} \to \infty$, and $e^{\beta^r s \sqrt{r}} \to 0$, as $r \to \infty$. Thus, $g^r \to 0$ as $r \to \infty$. Hence $\exists M > 0$, such that $g^r < M, \forall n$. Therefore using the Dominated Convergence Theorem,

$$\lim_{r\to\infty} \mathbb{E}\left(g^r(\beta^r(x))|\beta^r < 0\right) = \mathbb{E}\left(\lim_{r\to\infty}\left(g^r(\beta^r(x))|\beta^r < 0\right)\right) = 0,$$

42

and the uniform integrability is proved. The continuity of $g^r(\beta^r(x))$ is obvious. We show continuity of $\pi^r(\beta^r(x))$ when x is fixed, ignoring x in the proof for simplicity.

For every $a < 0$, we want to show, for every r, $\lim_{\beta^r \to a} \pi^r(\beta^r) = \pi^r(a)$. From the proof of Lemma 3.2.1 we know that for every realisation of β^r, there always exists a ω such that β has the same value as β^r. We choose β in such manner in the following proof. As $\beta^r \to a(\beta \to a)$,

$$
\begin{aligned}
|\pi^r(\beta^r) - \pi^r(a)| &= |\pi^r(\beta^r) - \mathbb{P}(\beta) + \mathbb{P}(\beta) - \mathbb{P}(a) + \mathbb{P}(a) - \pi^r(a)| \\
&\leq |\pi^r(\beta^r) - \mathbb{P}(\beta)| + |\mathbb{P}(\beta) - \mathbb{P}(a)| + |\mathbb{P}(a) - \pi^r(a)| \\
&\leq \epsilon + \epsilon + \epsilon \leq 3\epsilon.
\end{aligned}
$$

The first and third ϵs come from convergence of $\pi^r(\beta^r)$ to $\mathbb{P}(\beta)$ when β^r is given, the second ϵ is because of the continuity of $\mathbb{P}(\cdot)$. Now we proved continuity of $\pi^r(\beta^r)$, hence the function $Z(\beta^r(x))$ is continuous with respect to $\beta^r(x)$.

To this end we showed (3.2.48) and thus (3.2.42) is a valid approximation of the cost function (3.2.40). □

3.3 Staffing many server queues with random service rates and abandonments

Following the result we obtained in Section 3.1, we add abandonments to the model and design a cost function which we will optimise later.

All of the basic settings and notations are the same, except there are abandonments in the queue. Each customer has an associated patience time which are i.i.d. exponential random variables with rate ν. A customer abandons the system without getting any service if the waiting time in the queue exceeds the customer's patience. Once her/his service starts, s/he cannot abandon the system.

Adapting notations from Section 4.1, let $\hat{X}^r(t)$ be the scaled process of the number of customers in the rth system. We have the cost function

$$
\begin{aligned}
C^r(x) :&= F^r(x) + d\nu \mathbb{E}_{\beta^r} \left(\mathbb{E}_{\hat{X}^r(\infty)} \left(\hat{X}^r(\infty)^+ \Big| \hat{X}^r(\infty) \geq 0 \right) \mathbb{P} \left(\hat{X}^r(\infty) \geq 0 \right) \right) \\
&= F^r(x) + d\nu \mathbb{E}_{\beta^r} \left(\mathbb{E}_{\hat{X}^r(\infty)} \left(\hat{X}^r(\infty)^+, \hat{X}^r(\infty) \geq 0 \right) \right)
\end{aligned} \tag{3.3.1}
$$

where d is the cost of every customer abandonment, and $f_{\beta^r}(\cdot)$ is the density function of β^r. Notice that in this total cost function, we ignore the holding cost in the queue because abandonment and holding costs are both linear functions of expected queue length, and there is no need to consider expected queue length twice.

As proved in Section 3.3, $\hat{X}^r(t) \Rightarrow \xi(t)$ where $\xi(t)$ satisfies the equation

$$\xi(t) = \xi(0) + \sigma w(t) + \beta t + \gamma \int_0^t \xi(s)^- ds - \nu \int_0^t \xi(s)^+ ds. \tag{3.3.2}$$

More specifically, (3.3.2) can be seen as

$$\xi(t) = \begin{cases} \xi(0) + \sigma w(t) + \beta t - \nu \int_0^t \xi(s) ds, & \xi(t) \geq 0 \\ \xi(0) + \sigma w(t) + \beta t - \gamma \int_0^t \xi(s) ds, & \xi(t) < 0 \end{cases} . \tag{3.3.3}$$

Then by Section 4 of [Browne and Whitt, 1994], when $\xi(t) \geq 0$, it is an Ornstein-Uhlenbeck process, thus its steady state conditional on $\xi(\infty) \geq 0$ is normal distributed with density function $f_1(x) = \frac{\frac{\sqrt{2\nu}}{\sigma}\phi\left(\frac{\sqrt{2\nu}}{\sigma}\left(x - \frac{\beta}{\nu}\right)\right)}{\Phi\left(\frac{\sqrt{2\beta}}{\sqrt{\nu}\sigma}\right)}$. Similarly, when $\xi(t) < 0$, it is also an O-U process, thus its steady state conditional on $\xi(\infty) < 0$ is a normal random variable with density function $f_2(t) = \frac{\frac{\sqrt{2\gamma}}{\sigma}\phi\left(\frac{\sqrt{2\gamma}}{\sigma}\left(x - \frac{\beta}{\gamma}\right)\right)}{\Phi\left(-\frac{\sqrt{2\beta}}{\sqrt{\gamma}\sigma}\right)}$. Let $\varrho = \mathbb{P}(\xi(\infty) \geq 0)$. Then $\xi(\infty)$ has density function

$$f(x) = \begin{cases} f_1(x)\varrho, & x \geq 0 \\ f_2(x)(1 - \varrho), & x < 0 \end{cases} . \tag{3.3.4}$$

To find out ϱ, notice that $f(\cdot)$ is continuous because the infinitesimal variance of $\xi(t)$ is constant on the real line. Thus by equating the limits of $f(\cdot)$ from both left and right we get

$$\varrho = \mathbb{P}(\xi(\infty) \geq 0) = \left(1 + \sqrt{\frac{\nu}{\gamma}} \frac{\phi\left(-\frac{\sqrt{2\beta}}{\sqrt{\nu}\sigma}\right)}{\phi\left(-\frac{\sqrt{2\beta}}{\sqrt{\gamma}\sigma}\right)} \frac{\Phi\left(-\frac{\sqrt{2\beta}}{\sqrt{\gamma}\sigma}\right)}{\Phi\left(\frac{\sqrt{2\beta}}{\sqrt{\nu}\sigma}\right)}\right)^{-1} . \tag{3.3.5}$$

To this end, it is intuitive to use the following function as an approximation for the original cost function (3.3.1):

$$\hat{C}^r(x) = F^r(x) + d\nu \mathbb{E}_\beta \left(\mathbb{E}_{\xi(\infty)}\left(\xi(\infty)^+ \big| \xi(\infty) \geq 0\right) \mathbb{P}\left(\xi(\infty) \geq 0\right)\right)$$

$$= F^r(x) + d\nu \mathbb{E}_\beta \left(\mathbb{E}_{\xi(\infty)}\left(\xi(\infty)^+, \xi(\infty) \geq 0\right)\right). \tag{3.3.6}$$

We have a similar theorem as 3.2.1.

Theorem 3.3.1. (3.3.6) *is a valid approximation of* (3.3.1) *in the sense that*

$$\lim_{r \to \infty} \frac{C^r(x)}{\hat{C}^r(x)} = 1, \text{ for any fixed } x. \tag{3.3.7}$$

To prove this convergence, we will need uniform integrability of the steady state. To achieve this, for each heterogeneous system, we compare it with a homogeneous system with the same number of servers and the service rate being the lower bound of μ_k, i.e. p.

Lemma 3.3.1. *Let* $D_{hom}^{r,p}(t)$ *be the departure process of the rth homogeneous system described*

above where all the service rates take their minimum value p. Denote the departure process in the rth heterogeneous system as $D^r(t)$. Then $D^{r,p}_{hom}(t) \leq_{st} D^r(t)$, where 'st' means the inequality holds stochastically.

Proof. Denote the number of total customers in the rth homogeneous system as $X^{r,p}_{hom}(t)$. We have the system dynamic equations

$$X^r(t) = X^r(0) + A^r(t) - D^r(t) - R^r(t) \tag{3.3.8}$$

$$X^{r,p}_{hom}(t) = X^{r,p}_{hom}(0) + A^r(t) - D^{r,p}_{hom}(t) - R^r(t) \tag{3.3.9}$$

For simplicity let $X^r(0) = X^{r,p}_{hom}(0) = 0$. The arrival process and abandonment process are independent of the service rates and departure process, thus we can take them to be the same for these two processes. Let $S^{r,q}(t)$ be a generated poisson process with rate $N^r q$, where q is the upper bound of μ_k, and $0 < \tau_1 < \tau_2 < \cdots$ be the sequence of its occurrence times, i.e. $S^{r,q}(t) = \sum_{i=1}^{\infty} I(\tau_i < t)$. Let $\{U_l, l \in \mathbb{N}\}$ be a sequence of independent uniform$(0,1)$ random variables, and $I(A)$ be the indicator function for event A which takes the value 1 if A occurs and 0 otherwise. Assume all of the processes equal zero at $t = 0$. By splitting the process $S^{r,q}(t)$, we define the following processes

$$\breve{D}^{r,p,l}_{hom} = \sum_{i=1}^{l} I\left(U_i \leq \frac{(\breve{X}^{r,p}_{hom}(\tau_i-) \wedge N^r)p}{N^r q}\right), \tag{3.3.10}$$

$$\breve{D}^{r,l} = \sum_{i=1}^{l} I\left(U_i \leq \frac{\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_i-)}{N^r q}\right), \tag{3.3.11}$$

$$S^{r,q}(t) = l = \sum_{i=1}^{\infty} I(\tau_i < t), \tag{3.3.12}$$

$$\breve{D}^{r,p}_{hom}(t) = \breve{D}^{r,p,l}_{hom}, \breve{D}^r(t) = \breve{D}^{r,l}, \forall t \in [\tau_l, \tau_{l+1}), \tag{3.3.13}$$

$$\breve{X}^{r,p}_{hom}(t) = \breve{X}^{r,p}_{hom}(0) + A^r(t) - \breve{D}^{r,p}_{hom}(t) - R^r(t), \tag{3.3.14}$$

$$\breve{X}^r(t) = \breve{X}^r(0) + A^r(t) - \breve{D}^r(t) - R^r(t), \tag{3.3.15}$$

where $\breve{B}_k(t)$ is determined by the process $A^r(t)$, the sequence $\tau_1, \tau_2, \ldots$, and the selection scheme defined as follows: if, for some $i \in \{1, 2, \ldots, S^{r,q}(t)\}$, $I\left(U_i \leq \frac{\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_i-)}{N^r q}\right) = 1$, then we have that the potential departure occurring at time τ_i in process $S^{r,q}(t)$ is accepted as the real departure for process $\breve{D}^r(t)$. Assume there are m busy servers just before this departure occurs (time τ_i-). Then after it is accepted as the real departure, one of the m servers will be freed, which leads to our selection scheme. Let η_i be a uniformly distributed random variable on $(0,1)$. If $\frac{\sum_{a=0}^{j} \mu_k}{\sum_{k=1}^{m} \mu_k} \leq \eta_i < \frac{\sum_{k=0}^{j+1} \mu_k}{\sum_{k=1}^{m} \mu_k}$, $j = 0, 1, \ldots, m-1$, then server $j+1$ will be freed at time τ_i. Here we let $\mu_0 = 0$.

Under such definition, the process $(\mu^{(i)} D^{r,p}_{hom}(t), \breve{X}^{r,p}_{hom}(t))$ and $(\breve{D}^r(t), \breve{X}^r(t))$ are stochastically equivalent to $(D^{r,p}_{hom}(t), X^{r,p}_{hom}(t))$ and $(D^r(t), X^r(t))$ respectively.

45

We prove this by contradiction. Define

$$l^* = \min\{l : \breve{D}_{hom}^{r,p,l} > \breve{D}^{r,l}\}. \tag{3.3.16}$$

Then we have

$$I\left(U_{l^*} \le \frac{\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-)}{N^r q}\right) = 0 \text{ and } I\left(U_{l^*} \le \frac{(\breve{X}_{hom}^{r,p}(\tau_{l^*}-) \wedge N^r)p}{N^r q}\right) = 1,$$

i.e.

$$\frac{\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-)}{N^r q} < U_{l^*} \le \frac{(\breve{X}_{hom}^{r,p}(\tau_{l^*}-) \wedge N^r)p}{N^r q}.$$

Thus

$$\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-) < (\breve{X}_{hom}^{r,p}(\tau_{l^*}-) \wedge N^r)p. \tag{3.3.17}$$

Notice that $\breve{D}_{hom}^{r,p,l^*} > \breve{D}^{r,l^*}$ implies $\breve{D}_{hom}^{r,p,l^*-1} = \breve{D}^{r,l^*-1}$. Since arrival processes and abandonment processes are identical, we have $\breve{X}_{hom}^{r,p}(\tau_{l^*-1}) = \breve{X}^r(\tau_{l^*-1})$. Also, during the time $[\tau_{l^*-1}, \tau_{l^*})$, there are no departures, thus, by equations (3.3.8) and (3.3.9),

$$\breve{X}_{hom}^{r,p}(\tau_{l^*}-) = \breve{X}_{hom}^{r,p}(\tau_{l^*-1}) + A^r(\tau_{l^*}-) - A^r(\tau_{l^*-1}) - (R^r(\tau_{l^*}-) - R^r(\tau_{l^*-1})), \text{ and}$$

$$\breve{X}^r(\tau_{l^*}-) = \breve{X}^r(\tau_{l^*-1}) + A^r(\tau_{l^*}-) - A^r(\tau_{l^*-1}) - (R^r(\tau_{l^*}-) - R^r(\tau_{l^*-1})).$$

This shows us that $\breve{X}_{hom}^{r,p}(\tau_{l^*}-) = \breve{X}^r(\tau_{l^*}-)$. Substituting this into (3.3.17) gives us

$$\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-) < (\breve{X}^r(\tau_{l^*}-) \wedge N^r)p. \tag{3.3.18}$$

If $\breve{X}^r(\tau_{l^*}-) < N^r$, then $\sum_{k=1}^{N^r} \breve{B}_k(\tau_{l^*}-) = \breve{X}^r(\tau_{l^*}-)$, and (3.3.18) becomes $\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-) < \breve{X}^r(\tau_{l^*}-)p$, but since $\mu_k \ge p$, this doesn't hold. If $\breve{X}^r(\tau_{l^*}-) \ge N^r$, then $\sum_{k=1}^{N^r} \breve{B}_k(\tau_{l^*}-) = N^r$, and (3.3.18) is $\sum_{k=1}^{N^r} \mu_k \breve{B}_k(\tau_{l^*}-) < N^r p$, which is also not true because of $\mu_k \ge p$.

Thereby we have found a contradiction to the assumption (3.3.16). Hence, we conclude that such an n^* does not exist and $\breve{D}_{hom}^{r,p}(t) \le \breve{D}^r(t)$ for every $t \ge 0$. That means $D_{hom}^{r,p}(t) \le_{st} D^r(t)$. $\square$

Now we are ready to prove the theorem.

Proof of Theorem 3.3.1. Since F^r does not change in both cost functions, we only need to show

$$\mathbb{E}_{\beta^r}\left(\mathbb{E}_{\hat{X}^r(\infty)}\left(\hat{X}^r(\infty)^+, \hat{X}^r(\infty) \ge 0\right)\right)$$

$$\to \mathbb{E}_\beta\left(\mathbb{E}_{\xi(\infty)}\left(\xi(\infty)^+, \xi(\infty) \ge 0\right)\right), \text{ as } r \to \infty. \tag{3.3.19}$$

Consider a sequence of homogeneous systems with abandonment and service rates being the lower bound of μ_k, i.e. p. All other settings are the same as in the heterogeneous systems. By

Lemma 3.3.1 and equations (3.3.8) and (3.3.9), we know that $X^r(t) \leq_{st} X^{r,p}_{hom}(t)$, $\forall t \geq 0$. Since $X^r(t) \Rightarrow X^r(\infty)$ and $X^{r,p}_{hom}(t) \Rightarrow X^{r,p}_{hom}(\infty)$ as $t \to \infty$, it is also true that $X^r(\infty) \leq_{st} X^{r,p}_{hom}(\infty)$. And by equations (3.3) and (3.4) in [Mandelbaum and Zeltyn, 2004],

$$\mathbb{P}(X^{r,p}_{hom}(\infty) = j) = \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \prod_{k=N^r+1}^{j} \left(\frac{\lambda^r}{N^r p + (k - N^r)\nu} \right), j \geq N^r + 1,$$

where

$$\pi_0 = \left(\sum_{j=0}^{N^r} \frac{(\lambda^r/p)^j}{j!} + \sum_{j=N^r+1}^{\infty} \prod_{k=N^r+1}^{j} \left(\frac{\lambda^r}{N^r p + (k - N^r)\nu} \right) \frac{(\lambda^r/p)^{N^r}}{N^r!} \right)^{-1}. \qquad (3.3.20)$$

Thus $\forall r$,

$$\mathbb{E}\left((X^{r,p}_{hom}(\infty) - N^r), X^r_{hom}(\infty) \geq N^r \right)$$

$$= \sum_{j=N^r+1}^{\infty} \left(\frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \prod_{k=N^r+1}^{j} \left(\frac{\lambda^r}{N^r p + (k - N^r)\nu} \right) \right) (j - N^r)$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{j=N^r+1}^{\infty} (j - N^r) \prod_{k=N^r+1}^{j} \left(\frac{\lambda^r}{N^r p + (k - N^r)\nu} \right)$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{j=N^r+1}^{\infty} (j - N^r) \frac{(\lambda^r)^{j-N^r}}{(N^r p + \nu)(N^r p + 2\nu) \cdots (N^r p + (j - N^r)\nu)}.$$

Let $m_1 = \min(p, \nu)$. Then the equation above can be enlarged

$$\mathbb{E}\left((X^{r,p}_{hom}(\infty) - N^r), X^r_{hom}(\infty) \geq N^r \right)$$

$$\leq \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{j=N^r+1}^{\infty} (j - N^r) \frac{(\lambda^r)^{j-N^r}}{m_1^{j-N^r}(N^r + 1)(N^r + 2) \cdots j}$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{i=1}^{\infty} \left(\frac{\lambda^r}{m_1} \right)^i \frac{i}{(N^r + 1)(N^r + 2) \cdots (N^r + i)}$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{i=1}^{\infty} \left(\frac{\lambda^r}{m_1} \right)^i \frac{1}{i!} \frac{i! N^r!}{(N^r + i)!} i \leq \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \sum_{i=1}^{\infty} \left(\frac{\lambda^r}{m_1} \right)^i \frac{1}{i!} i$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \frac{\lambda^r}{m_1} \sum_{i=1}^{\infty} \left(\frac{\lambda^r}{m_1} \right)^{i-1} \frac{1}{(i-1)!}$$

$$= \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \frac{\lambda^r}{m_1} \sum_{i=0}^{\infty} \left(\frac{\lambda^r}{m_1} \right)^i \frac{1}{i!} = \frac{(\lambda^r/p)^{N^r}}{N^r!} \pi_0 \frac{\lambda^r}{m_1} e^{-\frac{\lambda^r}{m_1}}. \qquad (3.3.21)$$

From [Mandelbaum and Zeltyn, 2004], we know π_0 converges, thus (3.3.21) $< \infty$, which means the expected queue length of the rth homogeneous system is bounded. Since $X^r(\infty) \leq_{st} X^{r,p}_{hom}(\infty)$, we have

$$\mathbb{E}((X^r(\infty) - N^r), X^r(\infty) \geq N^r) \leq \mathbb{E}((X^{r,p}_{hom}(\infty) - N^r), X^{r,p}_{hom}(\infty) \geq N^r) < \infty. \qquad (3.3.22)$$

Scaling the inequality on both sides, we have

$$\mathbb{E}(\frac{1}{\sqrt{r}}(X^r(\infty)-N^r), X^r(\infty) \geq N^r) \leq \mathbb{E}(\frac{1}{\sqrt{r}}(X_{hom}^{r,p}(\infty)-N^r), X_{hom}^{r,p}(\infty) \geq N^r) < \infty, \quad (3.3.23)$$

i.e.

$$\mathbb{E}(\hat{X}^r(\infty), \hat{X}^r(\infty) \geq 0) \leq \mathbb{E}(\hat{X}_{hom}^{r,p}(\infty), \hat{X}_{hom}^{r,p}(\infty) \geq 0) < \infty. \quad (3.3.24)$$

Notice that the expected value on the left is a function of the random variable β^r, thus it itself is also a random variable. (3.3.24) implies that in the rth heterogeneous systems, the (scaled) expected queue length is always bounded no matter what values the service rates take, which further implies that the (scaled) queue length is uniformly integrable.

Using the same reasoning process for Figure 3.1, but with abandonment in the systems, we can get $\hat{X}^r(\infty) \Rightarrow \xi(\infty)$. Hence, by Theorem 3.5 in [Billingsley, 1999], (3.3.19) is proved. $\quad\square$

To this end, we have proven the validity of the approximation (3.3.6) of the cost function (3.3.1).

3.4   Impact of service rate variation on abandonment rate

In this section we want to see how the variance of the service rate influence the queue length, and thus the abandonment cost. We analyse this by considering the expected (scaled) queue length. The analytical result seems rather intractable, so instead we show their numerical results and explain how it reflects such influence. We mainly focus on systems under the LISF policy. To have a better idea of how variance plays its role in a system, we also include the numerical results for systems under the FSF policy.

According to (3.3.6), the approximating cost function for systems with abandonment is

$$\hat{C}^r(x) = F^r(x) + d\nu\mathbb{E}_\beta \left(\mathbb{E}_{\xi(\infty)} \left(\xi(\infty)^+ \big| \xi(\infty) \geq 0 \right) \mathbb{P}\left(\xi(\infty) \geq 0 \right) \right)$$

$$= F^r(x) + d\nu\mathbb{E}_\beta \left(\mathbb{E}_{\xi(\infty)} \left(\xi(\infty)^+, \xi(\infty) \geq 0 \right) \right)$$

$$= F^r(x) + d\nu \int_\mathbb{R} \int_0^\infty x f_1(x) \varrho dx d\mathbb{P}(\beta)$$

$$= F^r(x) + d\nu \int_{-\infty}^\infty \int_0^\infty x \frac{\frac{\sqrt{2\nu}}{\sigma}\phi\left(\frac{\sqrt{2\nu}}{\sigma}\left(x - \frac{\beta}{\nu}\right)\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)} f_\beta(\beta)$$

$$\left(1 + \sqrt{\frac{\nu}{\gamma}} \frac{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\gamma}\sigma}\right)} \frac{\Phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\gamma}\sigma}\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)} \right)^{-1} dx d\beta. \quad (3.4.1)$$

We want to see how the variance of the service rate influences the steady state and, thus by (3.4.1), the abandonment cost. (3.4.1) contains a complicated integral which may not have a closed form. To simplify the problem, we consider a special distribution of service rates. Let

48

the random service rates be uniformly distributed on $(\bar{\mu} - \epsilon, \bar{\mu} + \epsilon)$, $\epsilon > 0$. Then ξ has a random drift β, where $\beta = -\zeta - \theta\bar{\mu}$, and $\zeta \sim N(0, \frac{\epsilon^2}{3})$, and $\gamma = \bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}$. Then, (3.4.1) becomes

$$F^r(x) + d\nu \int_{-\infty}^{\infty} \int_0^{\infty} x \frac{\frac{\sqrt{2\nu}}{\sigma} \phi\left(\frac{\sqrt{2\nu}}{\sigma}\left(x - \frac{\beta}{\nu}\right)\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)} \frac{\sqrt{3}}{\epsilon} \phi\left(\frac{\sqrt{3}(\beta + \theta\bar{\mu})}{\epsilon}\right)$$

$$\left(1 + \sqrt{\frac{\nu}{\gamma}} \frac{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\gamma}\sigma}\right)} \frac{\Phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\gamma}\sigma}\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}\right)^{-1} dx \, d\beta$$

$$= F^r(x) + d\nu \int_{-\infty}^{\infty} \int_0^{\infty} x \frac{\frac{\sqrt{2\nu}}{\sigma} \phi\left(\frac{\sqrt{2\nu}}{\sigma}\left(x - \frac{\beta}{\nu}\right)\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)} \frac{1}{\epsilon\sqrt{\frac{2\pi}{3}}} \exp\left(-\frac{3(\beta + \theta\bar{\mu})^2}{2\epsilon^2}\right)$$

$$\left(1 + \sqrt{\frac{\nu}{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}} \frac{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}\sigma}\right)} \frac{\Phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}\sigma}\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}\right)^{-1} dx \, d\beta. \qquad (3.4.2)$$

Since the only part depending on the service rate variance is the double integral in (3.4.2), which is actually the expected (scaled) queue length, we let

$$QL(\epsilon) = \int_{-\infty}^{\infty} \int_0^{\infty} x \frac{\frac{\sqrt{2\nu}}{\sigma} \phi\left(\frac{\sqrt{2\nu}}{\sigma}\left(x - \frac{\beta}{\nu}\right)\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)} \frac{1}{\epsilon\sqrt{\frac{2\pi}{3}}} \exp\left(-\frac{3(\beta + \theta\bar{\mu})^2}{2\epsilon^2}\right)$$

$$\left(1 + \sqrt{\frac{\nu}{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}} \frac{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}{\phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}\sigma}\right)} \frac{\Phi\left(-\frac{\sqrt{2}\beta}{\sqrt{\bar{\mu} + \frac{\epsilon^2}{3\bar{\mu}}}\sigma}\right)}{\Phi\left(\frac{\sqrt{2}\beta}{\sqrt{\nu}\sigma}\right)}\right)^{-1} dx \, d\beta. \qquad (3.4.3)$$

After simplification, (3.4.3) is still hard to tackle, thus we employ a numerical integral. In Figure 3.3, we show the function $QL(\epsilon)$ vs ϵ. From the graph, $QL(\epsilon)$ is increasing. This implies that when all the other conditions remain the same, the total cost will grow as the service rate variance grows.

In contrast, we consider the same function for the FSF policy. The only difference in the limiting diffusion for FSF is $\gamma = \mu_{\min}$, which is $\bar{\mu} - \epsilon$ in the above-mentioned uniform distribution. Keeping other parameters unchanged, we plot its graph in Figure 3.4. The result is surprisingly counter-intuitive. It shows that the expected (scaled) queue length will decrease as the variance grows. One can explain such a situation as follows: the FSF policy always routes customers to the fastest available servers, thus in the long run, only the slowest server will have the chance to be idle. When the service rate variance increase, the minimum service rate will decrease, thus when the arrival rate remains the same, the 'lost' capacities due to idleness will also decrease. This implies that the total service rates that are indeed utilised will increase, thus the queue length decreases.
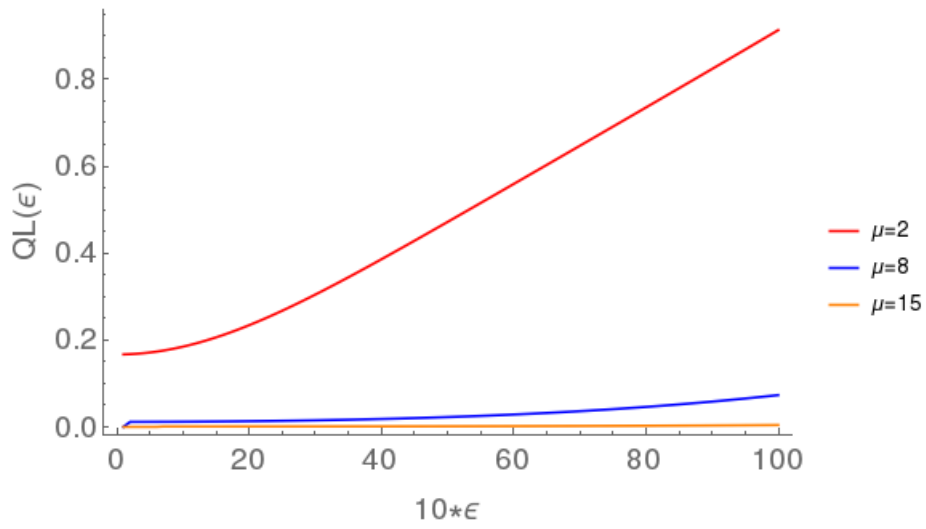
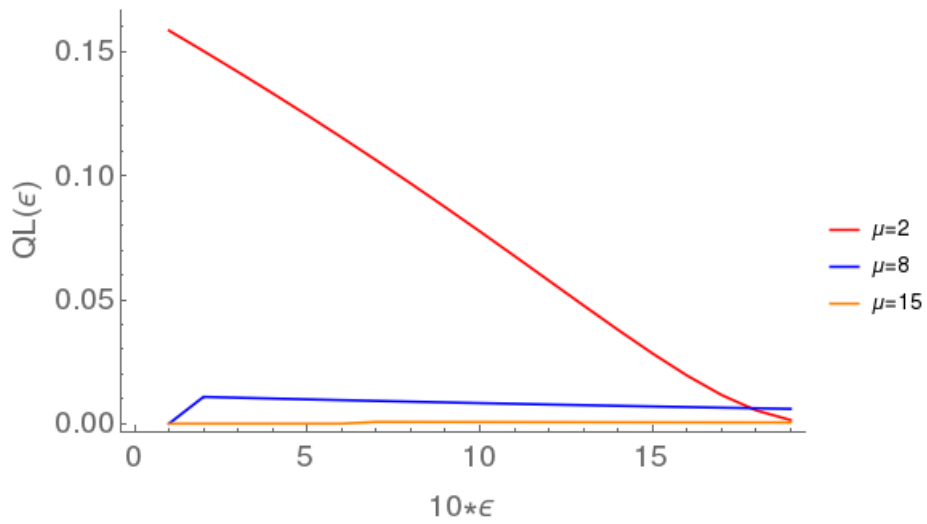Figure 3.3: $QL(\epsilon)$ vs ϵ for the LISF policy, when $\sigma = 4, \theta = 2, \nu = 2$.

Figure 3.4: $QL(\epsilon)$ vs ϵ for the FSF policy, when $\sigma = 4, \theta = 2, \nu = 2$.

3.5 Fairness among severs under different routing policies

From [Atar, 2008] and our analysis above, for a heterogeneous system, the longest idle server first (LISF) policy expresses a form of fairness, because when several servers are free, the one selected for the next incoming job is the one that has been idle for the longest time. On the other hand, policy faster server first (FSF) always routes customer to the fastest idle servers. Such difference in the routing scheme shows that LISF is more fair for servers than FSF, which provokes a question: can we quantify the fairness level for different routing policies? In this section we will answer this question, and demonstrate how it can be used in real systems.

3.5.1 Fairness measure

We introduce a new concept called "fairness measure". We still consider a sequence of many server queues with i.i.d. servers. For simplicity, we remove the randomness on the number of servers and assume the Nth system has exactly N servers. Denote Υ as the support of random variables $\{\mu_k\}$. To analyse fairness among servers, it is intuitive to consider their idle times. As before, denote $I_k(t)$ as the idleness process for server k, i.e. $I_k(t)$ equals 1 if server k is idle at t, and is equal to 0 if it is busy. For a routing policy π, its fairness measure is defined as $\eta_\pi : \Upsilon \to [0, 1]$, such that $\forall A \subset \Upsilon$, and $\forall\, T > 0$,

$$\sup_{0 \le t \le T} \frac{1}{\sqrt{N}} \left| \sum_{k=1}^{N} \mathbb{1}_{\mu_k}(A) I_k(t) - \eta_\pi(A) \sum_{k=1}^{N} I_k(t) \right| \to 0 \text{ in probability,} \qquad (3.5.1)$$

as $N \to \infty$, where $\mathbb{1}_{\mu_k}(\cdot)$ is the indicator function

$$\mathbb{1}_{\mu_k}(A) = \begin{cases} 1 & \mu_k \in A \\ 0 & \mu_k \notin A \end{cases}. \qquad (3.5.2)$$

From the definition, it is easy to see that η_π is a probability measure on Υ. One can understand (3.5.1) that the number of idle servers whose service rates are in A converges u.o.c to the product of the fairness measure of set A and the total number of idle servers.

We need to be aware that although (3.5.1) reflects fairness of a policy to some degree, it does not hold in every situation. There should be some limitations on the policies such that (3.5.1) is true, e.g. the policy should not depend on the total number of idle servers. Denote the set of all the eligible policies as Π. We have the following assumption,

Assumption 3.5.1. *For any $\pi \in \Pi$, (3.5.1) holds.*

3.5.2 Application of fairness measure on LISF and FSF policies

To see how Assumption 3.5.1 is used, we consider the diffusion limits proved in [Atar, 2008]. In Section 2 of Chapter 2, we showed that after some manipulations on the process $\hat{X}(t)$ (diffusively

scaled process of total number of customers in the system), it has expression

$$\hat{X}(t) = \hat{X}(0) + W(t) + bt + F(t), \tag{3.5.3}$$

with $F(t) = \frac{1}{\sqrt{N}} \int_0^t \sum_{k=1}^N \mu_k I_k(s) ds$ being the only item that is troublesome to deal with and also the only item that depends on policies.

Now we show the convergence of $F(t)$ under different policies. By Assumption 3.5.1,

$$F(t) = \frac{1}{\sqrt{N}} \int_0^t \sum_{k=1}^N \mu_k I_k(s) ds \to \frac{1}{\sqrt{N}} \int_0^t \sum_{k=1}^N \mu_k \mathbb{1}_{\mu_k}(\Upsilon) I_k(s) ds$$

$$\to \frac{1}{\sqrt{N}} \int_0^t \left(\int_\Upsilon \mu_1 d\eta_\pi(\mu_1) \right) \sum_{k=1}^N I_k(s) ds, \tag{3.5.4}$$

in probability u.o.c as $N \to \infty$.

- LISF

  For any $A \in \Upsilon$, define $\eta_{LISF}(A) = \frac{\int_A \mu_1 dm}{\int \mu_1 dm}$, then

  $$\int_\Upsilon \mu_1 d\eta_{LISF}(\mu_1) = \int_\Upsilon \mu_1 \frac{\mu_1 dm}{\int \mu_1 dm} = \frac{\int \mu_1^2 dm}{\int \mu_1 dm}, \tag{3.5.5}$$

  and (3.5.4) becomes

  $$F(t) \to \frac{\int \mu_1^2 dm}{\int \mu_1 dm} \frac{1}{\sqrt{N}} \int_0^t \sum_{k=1}^N I_k(s) ds, \tag{3.5.6}$$

  in probability u.o.c as $N \to \infty$. $\frac{\int \mu_1^2 dm}{\int \mu_1 dm}$ is actually γ in Theorem 2.1 in [Atar, 2008], and (3.5.6) matches with the result in that paper that $F(t) \to \gamma \int_0^t \hat{I}(s) ds$ in probability u.o.c as $N \to \infty$.

- FSF

  Similarly, for any $A \in \Upsilon$, define $\eta_{FSF}(A) = \mathbb{1}_{\mu_{\min}}(A)$. Then

  $$\int_\Upsilon \mu_1 d\eta_{FSF}(\mu_1) = \int_\Upsilon \mu_1 d\mathbb{1}_{\mu_{\min}}(\mu_1) = \mu_{\min}, \tag{3.5.7}$$

  and (3.5.4) becomes

  $$F(t) \to \mu_{\min} \frac{1}{\sqrt{N}} \int_0^t \sum_{k=1}^N I_k(s) ds, \tag{3.5.8}$$

  in probability u.o.c. as $N \to \infty$. Such a form also matches with the result of Theorem 2.2 in [Atar, 2008].

With fairness measure, we rephrase results in [Atar, 2008] in a more general way. The problem of proving diffusion limit of a particular system is reduced to finding the fairness

measure of the policy that is used in the system. Thus it is possible to invent a standard method for proving diffusion limits, which is more insightful than the proof in [Atar, 2008].

Chapter 4

State Space Collapse for Many Server Queues and Queueing Networks with Parameter Uncertainty

4.1 Introduction

In this chapter, we discuss the state space collapse phenomenon. Throughout this chapter, we adapt the framework developed by [Dai and Tezcan, 2011]. They consider a queueing network with multi-class customers and several server pools. In each pool, servers have the same capacities and capabilities. Customer arrivals are exogenous and independent from service processes. For such systems, exact analysis provides limited insight into the general properties of performances. One general way to overcome this is to use diffusion approximations. Similar to Chapter 3, the central part of the diffusion approximation is some heavy traffic limit theorems that state that a certain diffusively scaled performance processes converges to a diffusion in heavy traffic. Since the system processes in such networks are multidimensional, it becomes difficult to deal with when the system size grows large. Here is where state space collapse (SSC) plays a role. The SSC result reveals that under some conditions, the dimensions of system processes can be significantly reduced in the heavy traffic limit, while the essential information of the systems is still maintained. [Dai and Tezcan, 2011] gain the SSC result by using what they call an SSC function. They show that under some assumptions on the networks, the SSC function evaluated at diffusively scaled processes converges to zero as the system grows large.

For administrative and economic reasons, servers can be categorized and allocated such that, within pools, servers have the same capabilities, i.e. the set of customer classes that one server in

a pool is able to serve is the same as the set of customer class that any other server in the same pool can serve. This way of pooling reflects a kind of heterogeneity among servers. [Dai and Tezcan, 2011] also assume that within each pool, servers not only have the same capabilities, but also have the same capacities, i.e. when two servers in the same pool serve the same class of customers, their service rates are the same. However, in reality, it is more common that some differences are present among servers who are capable of doing the same tasks. Inside each pool, even though servers have the same capabilities, their skill levels can still be different, thus it will make more sense if we consider their rates to be different and random rather than identical.

In our work, service rates within pools are assumed to be i.i.d. random variables, and we also restrict our service times to being exponentially distributed. We analyse such networks, and demonstrate that with randomness within pools the SSC result still holds. We will show that the SSC function in such systems still converges to zero in the limit.

This chapter is organised as follows. In Section 4.2, we introduce our model and define *parallel random server systems*. In Section 4.3, we give our main result, the SSC for networks with random service rates. Then in Section 4.4, we provide essential proofs that are unique to our results. Proofs that are the same as in [Dai and Tezcan, 2011] are relegated to Appendix B.4. In Section 4.5, to show how the SSC can be used in queueing system analysis, we use the SSC method to show the diffusion limits proved in Chapter 3.

4.2   Notation and model descriptions

Our basic settings are similar to those in [Dai and Tezcan, 2011],but slightly different. Besides the randomness among servers, we do not include abandonments in our systems, and we also do not differentiate between the arrival streams and customer classes. Every arrival stream forms one class of customers. More specifically, we consider a system with parallel server pools and several customer classes. A server pool consists of several servers whose capabilities are the same, and their capacities are i.i.d. random variables (see more detailed definitions below). Customers of one class arrive in the system at a certain rate. Each class of customers have their own queue. Upon their arrival they will be routed to a capable server (idle and possesses the skill to serve this class of customer) if there is at least one; if all the capable servers are occupied, they will wait in their queues. Each customer is served by one of the servers. Once the service of a customer is completed by one of the servers, the customer leaves the system. And once a customer starts his service, he cannot abandon the service. For convenience, we refer to these systems as *parallel random server systems*.

4.2.1 Notation

We need to define some notation for convenience in presentation. Throughout this chapter, unless stated otherwise, for a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, its norm is defined as $|x| = \max_{\{i=1,\ldots,n\}} |x_i|$. For an $m \times n$ matrix M, its norm is $|M| = \max_{\{i=1,\ldots,m\}} |M_i|$, where $\{M_i, i = 1, \ldots, m\}$ are the row vectors of M.

In Chapter 2, we defined that for any function $x(t) \in \mathbb{D}^d$ and any $T > 0$, $||x(t)||_T = \sup_{0 \le t \le T} |x(t)|$. Now consider a sequence $x^r(t)$, we say $x^r \to x$ uniformly on a compact set (u.o.c) if $||x^r(t) - x(t)||_T \to 0$ as $r \to \infty$ for any $T > 0$.

4.2.2 Dynamics of the queueing networks

We use I to denote the number of server pools, and J to denote the number of customer classes. For notational convenience, we define $\mathcal{I} = \{1, \ldots, I\}, \mathcal{J} = \{1, \ldots, J\}$. And denote the number of servers in pool i by N_i for $i \in \mathcal{I}$ and set $N = (N_1, \ldots, N_I)$. The total number of servers in the system is denoted by $|N|$. Class j customers arrive the system according to a Poisson process with rate λ_j. We assume that the set of pools that can handle class j customers is fixed and denoted by $\mathcal{I}(j)$. Similarly, the set of customer classes that pool i can handle is fixed and denoted by $\mathcal{J}(i)$.

Upon arrival, each customer of class j is routed to a server if there is an available server in one of the pools in $\mathcal{I}(j)$. Otherwise, the customer joins the queue of class j, waiting to be served later. Assume the service time of a class j customer by the kth server in pool i is exponentially distributed with rate μ_{ijk}, where $j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i$. Then the service rates $\{\mu_{ijk}, k = 1, 2, \ldots, N_i\}$ are i.i.d. random variables. Denote their expectation value as $\bar{\mu}_{ij} = E\mu_{ijk}$. We also assume that $\mu_{ijk} \in [p_{ij}, q_{ij}], 0 < p_{ij} < q_{ij}$.

The object of study in this paper is a stochastic process $\mathbb{X} = (A, A_q, A_s, C, Q, Z, T, D)$. Assume all of the components are right continuous with left limits. We provide definitions of individual processes below.

- $A = (A_j; j \in \mathcal{J})$, $A_j(t)$ is the total number of class j arrivals by time t.

- $A_q = (A_{qj}; j \in \mathcal{J})$, $A_{qj}(t)$ denotes the total number of class j customers who are delayed and have to wait in the queue before their service starts.

- $A_s = (A_{sij}; i \in \mathcal{I}, j \in \mathcal{J})$, $A_{sij}(t)$ is the total number of class j customers who are routed to a server and start service in pool i immediately after their arrival by time t.

- $C = (C_{ij}; i \in \mathcal{I}, j \in \mathcal{J})$, $C_{ij}(t)$ is the total number of class j customers who are delayed in the queue and whose service started in pool i before time t.

- $Q = (Q_j; j \in \mathcal{J})$, $Q_j(t)$ is the total number of class j customers in queue at time t.

- $Z = (Z_{ij}; i \in \mathcal{I}, j \in \mathcal{J})$, $Z_{ij}(t)$ is the total number of servers in pool i who are busy with serving class j customers at time t.

- $T = (T_{ij}; i \in \mathcal{I}, j \in \mathcal{J})$, $T_{ij}(t)$ denotes the total time spent by servers in pool i in serving class j customers by time t.

- $D = (D_{ij}; i \in \mathcal{I}, j \in \mathcal{J})$, $D_{ij}(t)$ denotes the total number of class j customers whose service are completed by servers in pool i by time t.

Since there is heterogeneity among servers, we need to deal with each server individually, thus we also need the following notations:

- $B_{ijk}(t), i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i$: busy server indicator function. If the kth server in pool i is busy with a class j customer at time t, $B_{ijk}(t) = 1$, otherwise $B_{ijk}(t) = 0$.

- $T_{ijk}(t), i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i$: total time spent by the kth server in pool i in serving class j customers by time t.

- $D_{ijk}(t), i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i$: total number of class j customers whose service are completed by the kth server in pool i by time t.

Notice $B_{ijk}, Z_{ij}, T_{ijk}, T_{ij}, D_{ij}, D_{ijk}$ have such relation:

$$Z_{ij}(t) = \sum_{k=1}^{N_i} B_{ijk}(t), D_{ij}(t) = \sum_{k=1}^{N_i} D_{ijk}(t), \text{ and } T_{ij}(t) = \sum_{k=1}^{N_i} T_{ijk}(t)$$

for all $i \in \mathcal{I}, j \in \mathcal{J}$.

The main goal of this chapter is to study the SSC results of the above-mentioned queueing networks in the diffusion limit manner. Therefore, we analyse a sequence of systems indexed by r such that the arrival rate grows 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. $Q_j^r(t)$ is used to denote the number of class j customers in the queue in the rth system at time t. The arrival rate in the rth system is given by $\lambda^r = (\lambda_j^r, j \in \mathcal{J})$, and we assume that

$$\lambda_j^r \to \infty, \tag{4.2.1}$$

as $r \to \infty$.

Let $\{S_{ijk}, i \in \mathcal{I}, j \in \mathcal{J}, k = 1, 2, \ldots, N_i\}$ be i.i.d. standard Poisson processes, each having right-continuous sample paths. Combining the settings in [Atar, 2008] and [Dai and Tezcan, 2011], the processes D_{ijk} are assumed to satisfy

$$D_{ijk}^r(t) = S_{ijk}(\mu_{ijk} T_{ijk}^r(t)), \quad i \in \mathcal{I}, j \in \mathcal{J}, k = 1, 2, \ldots, N_i, \tag{4.2.2}$$

where

$$T^r_{ijk}(t) = \int_0^t B^r_{ijk}(s)ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, k = 1, 2, \ldots, N_i. \tag{4.2.3}$$

The process $\mathbb{X}^r$ depends on the control policy used in the system. To emphasize the dependence on the control policy π used, we use $\mathbb{X}_\pi$ to denote the process. Clearly, each element of $A^r, A^r_q, A^r_s, C^r, T^r, D^r$ is a nondecreasing process, and each element of Q^r and Z^r is nonnegative. Furthermore, the process $\mathbb{X}^r_\pi$ satisfies the following dynamic equations for all $t \geq 0$.

$$A^r_j(t) = A^r_{qj}(t) + \sum_{i \in \mathcal{I}(j)} A^r_{sij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.2.4}$$

$$Q^r_j(t) = Q^r_j(0) + A^r_{qj}(t) - \sum_{i \in \mathcal{I}(j)} C^r_{ij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.2.5}$$

$$Z^r_{ij}(t) = Z^r_{ij}(0) + A^r_{sij}(t) + C^r_{ij}(t) - \sum_{k=1}^{N^r_i} D^r_{ijk}(t), \text{ for all } i \in \mathcal{I}, j \in \mathcal{J}, \tag{4.2.6}$$

$$\sum_{j \in \mathcal{J}(i)} Z^r_{ij}(t) \leq N^r_i, \text{ for all } i \in \mathcal{I}, \tag{4.2.7}$$

$$Q^r_j(t) \left(\sum_{i \in \mathcal{I}(j)} \left(N^r_i - \sum_{j' \in \mathcal{J}(i)} Z^r_{ij'}(t) \right) \right) = 0, \text{ for all } j \in \mathcal{J}, \tag{4.2.8}$$

$$\int_0^t \sum_{i \in \mathcal{I}(j)} \left(N^r_i - \sum_{j' \in \mathcal{J}(i)} Z^r_{ij'}(s-) \right) dA^r_{qj}(s) = 0, \text{ for all } j \in \mathcal{J}, \tag{4.2.9}$$

Equations associated with the control policy π. $\tag{4.2.10}$

Equations (4.2.8) and (4.2.9) are based on the assumed non-idling property of a control policy. Equation (4.2.8) implies that there can be customers in the queue only when all of the servers that can serve that class of customers are busy. Equation (4.2.9) 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 (4.2.10) indicates the scheduling decisions to be made according to the selected scheduling policies.

Let $\{S_{ijk}, i \in \mathcal{I}, j \in \mathcal{J}\}$ be the Poisson processes defined before, and $\{v_{ijk}(l); l = 1, 2, \ldots\}$ be the corresponding sequence of i.i.d. exponential random variables. Since S_{ijk} is a Poisson process, $v_{ijk}(l)$ has exponential distribution with rate μ_{ijk}. We define $V_{ijk} : \mathbb{N} \to \mathbb{R}$ by

$$V_{ijk}(m) = \sum_{l=1}^m \frac{v_{ijk}(l)}{\mu_{ijk}}, \quad m \in \mathbb{N}, \tag{4.2.11}$$

where, by convention, empty sums are set to be zero. The term $V_{ijk}(m)$ is the total service requirement of the first m class j customers who are served by the kth server in pool i, and

V_{ijk} is known as the cumulative service time process. By the duality of S_{ijk} and V_{ijk}, we have

$$S_{ijk}(\mu_{ijk}t) = \max\{m : V_{ijk}(m) \leq t\}, \quad t \geq 0. \tag{4.2.12}$$

It follows from (4.2.2) that

$$V_{ijk}(D_{ijk}^r(t)) \leq T_{ijk}^r(t) \leq V_{ijk}(D_{ijk}^r(t) + 1). \tag{4.2.13}$$

Next, we give the details of the arrival processes. Let $\chi(t)$ be a delayed renewal process with rate 1. Let

$$A^r(t) = \chi(\lambda^r t) \tag{4.2.14}$$

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

$$U(m) = \sum_{l=1}^r u(l), \quad m \in \mathbb{N}, \tag{4.2.15}$$

and so

$$\chi(t) = \max\{m : U(m) \leq t\}. \tag{4.2.16}$$

We require that the interarrival times of the arrival processes satisfy the following condition, which is similar to condition (3.4) in [Bramson, 1998]:

$$\mathbb{E}\left(u(2)^{2+\epsilon}\right) < \infty, \quad \text{for some } \epsilon > 0. \tag{4.2.17}$$

Condition (4.2.17) is automatically satisfied by the service times because they are assumed to be exponentially distributed. For the rest of the paper, we assume that the primitive processes of the system satisfy (4.2.17). We also assume that $Q^r(0), Z^r(0), \chi$ and S are independent.

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

$$\lim_{r \to \infty} \frac{|N^r|}{r} = 1, \tag{4.2.18}$$

$$\lim_{r \to \infty} \frac{N_i^r}{|N^r|} = \beta_i, \quad \text{for all } i \in \mathcal{I} \text{ and for some } \beta_i \in (0, 1), \tag{4.2.19}$$

$$\lim_{r \to \infty} \frac{\lambda_j^r}{|N^r|} = \lambda_j, \quad \text{for some } 0 < \lambda_j < \infty. \tag{4.2.20}$$

We shall denote $\lambda = (\lambda_j, j \in \mathcal{J})$.

4.3 Main results

In this section, we state our main result as Theorem 4.3.2, which is an SSC result for *parallel random server systems*. We extend the SSC result for systems from identical servers in each

pool which is proved by [Dai and Tezcan, 2011], to include random service rates in each pool. The proof framework of our results is similar to the one in [Dai and Tezcan, 2011]. Our main contribution is when showing almost Lipschitz condition for hydrodynamically scaled departure processes, the direct way in [Dai and Tezcan, 2011] is no longer valid, so instead we come up with a new coupling method. For more detailed differences between proofs of [Dai and Tezcan, 2011] and our results, see discussion at the beginning of Section 4.4.

Before stating the theorem, we need some preliminary definitions and assumptions.

We only consider systems under the heavy traffic condition. For a multi-class network with several server pools, it is not trivial to define the heavy traffic condition. Instead, we use the static planning problem (SPP) for the networks as a bridge to heavy traffic conditions.

4.3.1 The static planning problem

The SPP is introduced in [Dai and Tezcan, 2011]. We will modify the original problem such that it suits our models. The objective of an SPP is to minimise server utilisations in the network. [Dai and Tezcan, 2011] use identical service rates to define utilisations, while we consider their expected value $\bar{\mu}_{ij}$ instead.

Let $x = (x_{ij}, i \in \mathcal{I}, j \in \mathcal{J}(i))$, where x_{ij} is the long term proportion of pool i servers' working time in serving class j customers. We define the static planning problem

$$
\begin{aligned}
\min \quad & \rho \\
\text{s.t.} \quad & \sum_{i \in \mathcal{I}(j)} \beta_i \bar{\mu}_{ij} x_{ij} = \lambda_j, \text{ for all } j \in \mathcal{J}, \\
& \sum_{j \in \mathcal{J}(i)} x_{ij} \leq \rho, \text{ for all } i \in \mathcal{I}, \\
& x_{ij} \geq 0 \text{ for all } i \in \mathcal{I}, j \in \mathcal{J}.
\end{aligned} \tag{4.3.1}
$$

Denote (ρ^*, x^*) as the optimal solution of the above static planning problem. This means the average utilisation of the busiest pool will reach its minimum value ρ^* with the allocation x^*. If $\rho^* > 1$, it can easily be shown that the queue length grows without bound; thus in our analysis it is always assumed that $\rho^* \leq 1$.

Since we need a heavy traffic condition, we now consider a sequence of the following optimisation problem

$$
\begin{aligned}
\min \quad & \rho^r \\
\text{s.t.} \quad & \sum_{i \in \mathcal{I}(j)} N_i^r \bar{\mu}_{ij} x_{ij}^r = \lambda_j^r, \text{ for all } j \in \mathcal{J}, \\
& \sum_{j \in \mathcal{J}(i)} x_{ij}^r \leq \rho^r, \text{ for all } i \in \mathcal{I}, \\
& x_{ij}^r \geq 0 \text{ for all } i \in \mathcal{I}, j \in \mathcal{J}.
\end{aligned} \tag{4.3.2}
$$

Let $(\rho^r, x^{r,*})$ be an optimal solution of (4.3.2). These optimisation problems will be an impor-

tant part in the assumptions of heavy traffic below.

4.3.2 Assumptions

Before starting the analysis of the state space collapse phenomenon, we need to clarify that such phenomenon will not happen in all networks. We need some constraints on the systems to guarantee SSC. We have the following two assumptions for that reason.

Using the definitions in the static planning problem, we have the assumption about heavy traffic conditions and control policies.

Assumption 4.3.1. *For each static optimal solution (ρ^*, x^*) of the SPP (4.3.1), we have $\rho^* = 1$ and $\sum_{j \in \mathcal{J}(i)} x_{ij}^* = 1$. Moreover, for any sequence of optimal solutions $\{x^{r,*}\}$ of (4.3.2), we have*

$$x^{r,*} \to x^*,$$

as $r \to \infty$ for some optimal solution of (4.3.1).

In this work, we only consider control policies that will not cause the system to explode, i.e. the queue length does not grow to infinity. Thus we need the following assumption.

Assumption 4.3.2. *For a control policy π,*

$$\frac{Z_{ij}^r(t)}{|N^r|} \to z \quad u.o.c. \ a.s. \tag{4.3.3}$$

as $r \to \infty$ if $Z_{ij}^r(0)/|N^r| \to (0, z)$ a.s. as $r \to \infty$, where $z = (z_{ij}, i \in \mathcal{I}, j \in \mathcal{J})$, and $z_{ij} = \beta_i x_{ij}^*$ for an optimal solution (ρ^*, x^*) of the static planning problem (4.3.1).

We do not include the constraints on $Q^r(\cdot)$ because, unlike in [Dai and Tezcan, 2011], we do not have abandonments in our system.

4.3.3 Fluid limit and verification of the control policy assumption

Under a control policy, when Assumption 4.3.2 is satisfied, the fluid limits exist and do not explode, even though they are critically loaded. We assume that

$$Z^r(0)/|N^r| \to z \text{ a.s.} \tag{4.3.4}$$

as $r \to \infty$, where z is given as in Assumption 4.3.2. Under Assumption 4.3.2, condition (4.3.4) implies that

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

as $r \to \infty$, for $t \geq 0$.

To make Assumption 4.3.2 easier to check, we use the fluid limit concept. The fluid scaling

is defined as

$$\bar{\mathbb{X}}^r(t) = \frac{\mathbb{X}^r(t)}{|N^r|}. \tag{4.3.5}$$

The following definitions and notations are from [Dai and Tezcan, 2011], but we repeat them here for completeness.

Let $\mathscr{A} \in \Omega$ be such that $\{\bar{Q}^r(0)\}$ is bounded and the following Functional Strong Law of Large Numbers holds:

$$\frac{A_j^r(|N^r|\cdot)}{|N^r|} \to a_j(\cdot), \frac{\sum_{k=1}^{N_i^r} S_{ijk}(|N^r|\cdot)}{|N^r|} \to \alpha_{ij}(\cdot), \text{ u.o.c} \tag{4.3.6}$$

as $r \to \infty$, where $\alpha_{ij}(t) = \bar{\mu}_{ij}t, a_j(t) = t$. Note that we can take $\mathbb{P}(\mathscr{A}) = 1$.

We call $\bar{\mathbb{X}}^r(t)$ the fluid scaled process. $\bar{\mathbb{X}}$ is called a fluid limit of $\{\mathbb{X}^r\}$ if there exists an $\omega \in \mathscr{A}$ and a sequence $\{r_l\}$ with $r_l \to \infty$ as $l \to \infty$, such that $\bar{\mathbb{X}}^{r_l}(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}$ as $l \to \infty$, where $\mathscr{A}$ is taken from Theorem B.1 in [Dai and Tezcan, 2011]. The following theorem is analogous to Theorem B.1 in [Dai and Tezcan, 2011], with the main difference being in $\bar{T}$ and $\bar{I}$. These two processes are considered individually rather than aggregately due to the servers' heterogeneity.

Theorem 4.3.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Assume that (4.2.18) and (4.2.20) hold and $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Then $\{\bar{\mathbb{X}}_\pi^r\}$ is a.s. precompact(i.e. every sequence has a convergent subsequence) 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 $\{\bar{\mathbb{X}}_\pi^r\}$ satisfies the following equations for all $t \geq 0$:*

$$\lambda_j t = \bar{A}_{qj}(t) + \sum_{i \in \mathcal{I}(j)} \bar{A}_{sij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.3.7}$$

$$\bar{Q}_j(t) = \bar{Q}_j(0) + \bar{A}_{qj}(t) - \sum_{i \in \mathcal{I}(j)} \bar{C}_{ij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.3.8}$$

$$\bar{Z}_{ij}(t) = \bar{Z}_{ij}(0) + \bar{A}_{sij}(t) + \bar{C}_{ij}(t) - \bar{\mu}_{ij}\bar{T}_{ij}(t), \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{I}(i), \tag{4.3.9}$$

$$\bar{T}_{ij}(t) = \int_0^t \bar{Z}_{ij}(s)ds, \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{I}(i), \tag{4.3.10}$$

$$\bar{I}_i(t) = \beta_i t - \sum_{j \in \mathcal{J}(i)} \bar{T}_{ij}(t), \text{ for all } i \in \mathcal{I}, \tag{4.3.11}$$

$$\bar{Q}_j(t) \left(\beta_i - \sum_{j' \in \mathcal{J}(i)} \bar{Z}_{ij'}(t) \right) = 0, \text{ for all } j \in \mathcal{J}, \tag{4.3.12}$$

$$\int_0^t \sum_{j \in \mathcal{J}(i)} \bar{Q}_j(s)d\bar{I}_i(s) = 0, \text{ for all } i \in \mathcal{I}, \tag{4.3.13}$$

63

$$\int_0^t \sum_{i \in \mathcal{I}(j)} \left(\beta_i - \sum_{j' \in \mathcal{J}(i)} \bar{Z}_{ij'}(s) \right) d\bar{A}_{qj}(s) = 0, \text{ for all } j \in \mathcal{J}, \tag{4.3.14}$$

$$\bar{A}, \bar{A}_q, \bar{A}_s, \bar{T}, \text{ and } \bar{C} \text{ are nondecreasing,} \tag{4.3.15}$$

$$\bar{Q}(t) \geq 0, \ \bar{Z}_{ij}(t) \geq 0, \ and \sum_{j \in \mathcal{J}(i)} \bar{Z}_{ij}(t) \leq 1 \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i). \tag{4.3.16}$$

The proof of this theorem is in Appendix B.1.

The vector (q, z) is called a steady 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 $t \geq 0$.

We denote the set of all of the steady states of the fluid limits of $\{\mathbb{X}^r\}$ by $\mathcal{M}$. The following result is analogous to Lemma B.1 in [Dai and Tezcan, 2011], and it is an equivalent condition to Assumption 4.3.2.

Lemma 4.3.1. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel random server systems processes that satisfies conditions of Theorem 4.3.1 and Assumption 4.3.1. A control policy π satisfies Assumption 4.3.2 if $(0, z) \in \mathcal{M}$, where $z_{ij} = \beta_i x^*_{ij}$.*

The proof is the same as Lemma B.1 in [Dai and Tezcan, 2011] and is trivial so we omit it here.

In general, diffusion limits are introduced as refinements of the fluid limits. Under condition (4.3.4) and Assumption 4.3.2, we define the diffusive scaling as follows:

$$\hat{Q}^r(t) = \frac{Q^r(t)}{\sqrt{|N^r|}} \quad \text{and} \quad \hat{B}^r_{ijk}(t) = \frac{B^r_{ijk}(t)}{\sqrt{|N^r|}}, \quad \text{for} \quad t \geq 0. \tag{4.3.17}$$

and denote

$$\hat{Z}^r_{ij}(t) = \sum_{k=1}^{N^r_i} \left(\hat{B}^r_{ijk}(t) - \frac{x^*_{ij}}{\sqrt{N^r}} \right) = \frac{Z^r_{ij}(t) - x^*_{ij} N^r_i}{\sqrt{|N^r|}}. \tag{4.3.18}$$

Now we introduce the hydrodynamic model equations, and we borrow the SSC function from [Dai and Tezcan, 2011]. This function is the key point in our main theorem.

4.3.4 Hydrodynamic model equations

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

$$\lambda_j t = \tilde{A}_{qj}(t) + \sum_{i \in \mathcal{I}(j)} \tilde{A}_{sij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.3.19}$$

$$\tilde{Q}_j(t) = \tilde{Q}_j(0) + \tilde{A}_{qj}(t) - \sum_{i \in \mathcal{I}(j)} \tilde{C}_{ij}(t), \text{ for all } j \in \mathcal{J}, \tag{4.3.20}$$

$$\tilde{A}_{qj}, \tilde{A}_{sij}, \tilde{C}_{ij} \text{ are nondecreasing for all } i \in \mathcal{I}, j \in \mathcal{J}, \tag{4.3.21}$$

$$\tilde{Z}_{ij}(t) = \tilde{Z}_{ij}(0) + \tilde{A}_{sij}(t) + \tilde{C}_{ij}(t) - \bar{\mu}_{ij}\tilde{T}_{ij}(t), \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \quad (4.3.22)$$

$$\tilde{T}_{ij}(t) = \int_0^t z_{ij}ds = z_{ij}t, \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \quad (4.3.23)$$

$$\tilde{Q}_j(t) \geq 0, \text{ for all } j \in \mathcal{J}, \text{ and } \sum_{j \in \mathcal{J}(i)} \tilde{Z}_{ij}(t) \leq 0, \text{ for all } i \in \mathcal{I}, \quad (4.3.24)$$

$$\tilde{Q}_j(t) \left(\sum_{i \in \mathcal{I}(j)} \sum_{j' \in \mathcal{J}(i)} \tilde{Z}_{ij'}(t) \right) = 0, \text{ for all } j \in \mathcal{J}, \quad (4.3.25)$$

$$\int_0^t \left(\sum_{i \in \mathcal{I}(j)} \sum_{j' \in \mathcal{J}(i)} \tilde{Z}_{ij'}(s) \right) d\tilde{A}_{qj}(s) = 0, \text{ for all } j \in \mathcal{J}, \quad (4.3.26)$$

$$\text{Additional equations associated with the control policy } \pi, \quad (4.3.27)$$

where λ_j is defined as in (4.2.20). Equations (4.3.19)-(4.3.27) are called the *hydrodynamic model equations*, and they define the *hydrodynamic model* of the system under policy π. Any process $\tilde{\mathbb{X}}_\pi$ satisfying (4.3.19)-(4.3.27) for all $t \geq 0$ is called a hydrodynamic model solution.

Hydrodynamic model solutions are deterministic and absolutely continuous, hence almost everywhere differentiable. Absolute continuity follows from the following result.

Proposition 4.3.1. *Any process $\tilde{\mathbb{X}}_\pi$ satisfying* (4.3.19)-(4.3.27) *for all $t \geq 0$ is Lipschitz continuous.*

It will be proved in Proposition 4.4.3 that the hydrodynamic model equations are satisfied by hydrodynamic limits under certain general assumptions; these limits are obtained from the hydrodynamically scaled sequences.

4.3.5 SSC in the diffusion limits

Similar to Section 4.2 in [Dai and Tezcan, 2011], we define the state space collapse function. Let $g : \mathbb{R}^{J+d_z} \to \mathbb{R}^+$, where $d_z = \sum_{j \in \mathcal{J}} |\mathcal{I}(j)|$, be a nonnegative function that satisfies the following *homogeneity* condition:

$$g(\alpha x) = \alpha^c g(x), \quad (4.3.28)$$

for some $c > 0$, for all $x \in \mathbb{R}^{J+d_z}$, and for all $0 \leq \alpha \leq 1$. We call g a 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 4.3.3. *The function $g : \mathbb{R}^{J+d_z} \to \mathbb{R}^+$ satisfies* (4.3.28) *and is continuous on $\mathbb{R}^{J+d_z}$.*

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 [Dai and Tezcan, 2011, Assumption 4.2].

Assumption 4.3.4. *Let g be a function that satisfies Assumption 4.3.3. There exists a function $H(t)$ with $H(t) \to 0$ as $t \to \infty$ such that*

$$g(\tilde{Q}(t), \tilde{Z}(t)) \leq H(t) \ \text{for all } t \geq 0 \tag{4.3.29}$$

for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ satisfying $|(\tilde{Q}(0), \tilde{Z}(0))| \leq 1$. Furthermore, for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ with $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$ and $|(\tilde{Q}(0), \tilde{Z}(0))| \leq 1$, $g(\tilde{Q}(t), \tilde{Z}(t)) = 0$ for $t \geq 0$.

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

Theorem 4.3.2. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Suppose that Assumption 4.3.1 and Assumption 4.3.2 hold, g satisfies Assumption 4.3.3, the hydrodynamic model of the system satisfies Assumption 4.3.4, and*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \ \text{in probability} \tag{4.3.30}$$

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

$$\frac{\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_T}{(\|\hat{Z}^r(t)\|_T \vee 1)^c} \to 0 \ \text{in probability} \tag{4.3.31}$$

as $r \to \infty$, where $c > 0$ is given as in (4.3.28).

Theorem 4.3.2 tells us that, with a well-defined SSC function g for parallel random server systems under a specific control policy, one can show that function g evaluated at $\hat{Q}^r$ and $\hat{Z}^r$ over a compact set converges to zero in some way. This implies $\hat{Q}^r$ and $\hat{Z}^r$ can be represented by a lower dimensional process using the relation between $\hat{Q}^r$ and $\hat{Z}^r$ in function g, which means the process states collapse from a higher dimension to to lower dimension. This SSC result can be then applied to analysing diffusion limits of such systems, which is much simpler now since the dimensions of diffusions are reduced.

4.4   SSC framework

In this section we explain in detail how the SSC result is obtained. Since we use the framework developed by [Dai and Tezcan, 2011], some of the steps remain the same as in that paper. We will explain the process with an emphasis in our contributions.

The steps are as follows:

1. Define the *hydrodynamic scaling* $\mathbb{X}^{r,m}(\cdot)$ for the original system process $\mathbb{X}^r(\cdot)$.

2. Show that the SSC function $g(\cdot)$ evaluated at the hydrodynamically scaled process $\mathbb{X}^{r,m}(t)$ is bounded by some function $H(t)$ plus an arbitrary ϵ. The function $H(t)$ has the property

that $H(t) \to 0$ as $t \to \infty$. This step works as a bridge to prove the final SSC result in step 3.

3. Using the mathematical relation between the hydrodynamically scaled process $\mathbb{X}^{r,m}(\cdot)$ and diffusively scaled process $\hat{\mathbb{X}}$, together with the boundness result in step 2, we can show that the SSC function $g(\cdot)$ evaluated at the diffusively scaled process converges to zero in probability, i.e. (4.3.31) is true.

In Step 1, the hydrodynamic scaling is slightly different from [Dai and Tezcan, 2011]. They consider servers in each pool aggregately, thus they define the scaling for the process of the number of busy servers in each pool. We need to treat each server individually since their service rates are random, therefore, apart from the process scaled above, we also need to define hydrodynamic scaling for the process $B_{ijk}(t)$ of each server.

Step 3 is the same as in [Dai and Tezcan, 2011].

Our main contributions are in Step 2. To get a better idea of this, we need to look at Step 2 in more detail. Step 2 can be further decomposed into 5 smaller steps:

i) Show that hydrodynamically scaled process $\mathbb{X}^{r,m}(\cdot)$ is almost Lipschitz.

ii) Define the *hydrodynamic limit* $\tilde{\mathbb{X}}(\cdot)$ such that it is right continuous with left limits, and has the Lipschitz condition.

iii) Using i) and ii), one can show that $\mathbb{X}^{r,m}(\cdot)$ converges to $\tilde{\mathbb{X}}(\cdot)$ u.o.c. as $r \to \infty$.

iv) Show that $\tilde{\mathbb{X}}(\cdot)$ satisfies the hydrodynamic model equations (4.3.19) to (4.3.27).

v) Finally, by Assumptions 4.3.3 and 4.3.4, and using results in iii) and iv), show that the boundness result of step 2.

The difficult part is in Step i) - more specifically, the almost Lipschitz condition of departure processes. In [Dai and Tezcan, 2011], to show this they use Proposition 4.3 of [Bramson, 1998], which is an application of Chebyshev's inequality on renewal processes. They use this proposition on departure processes of each pool when the service rates are given. However, in our systems, service rates within pools are random and unknown, thus we can not use this proposition directly. We formulate a coupling method to achieve this. We assume the random service rates in each pool for all kind of customers are bounded, then generate a Poisson process with rate being the product of the number of servers in each pool and the upper bound of the service rates. Then we couple our real departure processes by splitting this Poisson process. We only need to show the almost Lipschitz condition for the generated Poisson process, then the same condition also holds for the coupled real departure processes since the time points when departures happen in the coupled processes are subsets of the generated Poisson process.

Another main difference in our proof is in Step iv). When we show that the hydrodynamic limits $\tilde{\mathbb{X}}(\cdot)$ satisfy the hydrodynamic model equations, we need to use the Law of Large Numbers because of the randomness among servers.

67

The rest of this section is organized as follows: in Section 4.4.1, we define the hydrodynamic scaling and prove i) of step 2. Then in Section 4.4.2, we use the hydrodynamic scaling to define the hydrodynamic limits, and show iii) and iv), hence provd step 2. In Section 4.4.3, we translate the boundness inequality in step 2 to a similar inequality for the diffusion-scaled processes. We finally show that this latter inequality implies the desired SSC result in the diffusion limit.

4.4.1 Hydrodynamic scaling and bounds

We begin by defining the *hydrodynamic scaling*. We divide the interval $[0, T]$ into $T\sqrt{|N^r|}$ intervals of length $\frac{1}{\sqrt{|N^r|}}$ and analyse the processes in each intervals. We index the intervals by m. For a nonnegative integer m, let

$$x_{r,m} = \left| Z^r \left(\frac{m}{\sqrt{|N^r|}} \right) - |N^r| \right|^2 \vee |N^r|, \tag{4.4.1}$$

Note that the square root of the first two terms of $x_{r,m}$ gives the deviations of these processes from their fluid limits.

We define the hydrodynamic scaling by shifting and scaling the processes of $\mathbb{X}^r$ as follows. For a process X^r associated with the rth process, we denote the hydrodynamic scaled version by $X^{r,m}$. For $A^r, A_s^r, A_q^r, C^r, D^r, T^r$, the hydrodynamic scaling is defined for $t \in [0, L]$, for some $L > 0$ by

$$X^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(X^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) - X^r \left(\frac{m}{\sqrt{|N^r|}} \right) \right). \tag{4.4.2}$$

The hydrodynamic scaled versions of Q^r and B^r and Z^r are defined as follows:

$$Q^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(Q^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) \right), \tag{4.4.3}$$

$$B_{ijk}^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(B_{ijk}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) - x_{ij}^* \right), \tag{4.4.4}$$

$$Z_{ij}^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(Z_{ij}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) - N_j^r x_{ij}^* \right). \tag{4.4.5}$$

Note that $Z_{ij}^{r,m}(t) = \sum_{k=1}^{N_i^r} B_{ijk}^{r,m}(t)$ and $D_{ij}^{r,m}(t) = \sum_{k=1}^{N_i^r} D_{ijk}^{r,m}(t)$.

Observe that $x_{r,m}$ must be in the order of $|N^r|$ for Q^r and Z_{ij}^r to have meaningful diffusion limits. Also, if $x_{r,m}$ is in the order of $|N^r|$, then $Q^{r,m}(\cdot)$ and $Z_{ij}^{r,m}(\cdot)$ are very similar to the diffusion scaling. This reveals the relationship between the hydrodynamic and diffusion scaling that will be used to translate a SSC result from hydrodynamic limits to diffusion limits.

For notational convenience, with a slight abuse of notation, we set

$$V_{ijk}^{r,m}(D_{ijk}^{r,m}(t), b) = \frac{1}{\sqrt{x_{r,m}}} \left(V_{ijk} \left(D_{ijk}^r \left(\frac{\sqrt{x_{r,m}} t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) + b_1 \right) \right.$$
$$\left. - V_{ijk} \left(D_{ijk}^r \left(\frac{m}{\sqrt{|N^r|}} \right) + b_2 \right) \right), \qquad (4.4.6)$$

and for $b = (b_1, b_2) \in \mathbb{R}^2$. By (4.2.13),

$$V_{ijk}^{r,m}(D_{ijk}^{r,m}(t), (0,1)) \le T_{ijk}^{r,m} \le V_{ijk}^{r,m}(D_{ijk}^{r,m}(t), (1,0)). \qquad (4.4.7)$$

Let $\mathbb{X}^{r,m} = (A^{r,m}, A_s^{r,m}, A_q^{r,m}, Q^{r,m}, B^{r,m}, Z^{r,m}, C^{r,m}, T^{r,m}, D^{r,m})$. We refer to $\mathbb{X}^{r,m}$ as the 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 $t \ge 0$:

$$A_j^{r,m}(t) = \sum_{i \in \mathcal{I}(j)} A_{sij}^{r,m}(t) + A_{qj}^{r,m}(t), \text{ for all } j \in \mathcal{J}, \qquad (4.4.8)$$

$$Q_j^{r,m}(t) = Q_j^{r,m}(0) + A_{qj}^{r,m}(t) - \sum_{i \in \mathcal{I}(j)} C_{ij}^{r,m}(t), \text{ for all } j \in \mathcal{J}, \qquad (4.4.9)$$

$$Z_{ij}^{r,m}(t) = Z_{ij}^{r,m}(0) + A_{sij}^{r,m}(t) + C_{ij}^{r,m}(t) - D_{ij}^{r,m}(t), \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \qquad (4.4.10)$$

$$D_{ij}^{r,m}(t) = \sum_{k=1}^{N_i^r} D_{ijk}^{r,m}(t), \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \qquad (4.4.11)$$

$$D_{ijk}^{r,m}(t) = \frac{S_{ijk}(\mu_{ijk}(\sqrt{x_{r,m}} T_{ijk}^{r,m}(t) + T_{ijk}^r(m/\sqrt{|N^r|}))) - S_{ijk}(\mu_{ijk}(T_{ijk}^r(m/\sqrt{|N^r|})))}{\sqrt{x_{r,m}}},$$

$$\text{for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \text{ and } k = 1, 2, \dots, N_i^r, \qquad (4.4.12)$$

$$T_{ijk}^{r,m}(t) = \frac{x_{ij}^*}{|N^r|} t + \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^t B_{ijk}^{r,m}(s) ds, \text{ for all } i \in \mathcal{I}, \text{ and } j \in \mathcal{J}(i), \text{ and } k = 1, 2, \dots, N_i^r, \qquad (4.4.13)$$

$$Q_j^{r,m}(t) \left(\sum_{i \in \mathcal{I}(j)} \sum_{j' \in \mathcal{J}(i)} Z_{ij'}^{r,m}(t) \right) = 0, \text{ for all } j \in \mathcal{J}, \qquad (4.4.14)$$

$$\int_0^t \left(\sum_{i \in \mathcal{I}(j)} \sum_{j' \in \mathcal{J}(i)} Z_{ij'}^{r,m}(s-) \right) dA_{qj}^{r,m}(s) = 0, \text{ for all } j \in \mathcal{J}. \qquad (4.4.15)$$

Now we have a similar result to Proposition 5.1 in [Dai and Tezcan, 2011] for random service rates systems.

Proposition 4.4.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Assume (4.2.19) and (4.2.20) hold, and π satisfies Assumption 4.3.2. Fix $\epsilon > 0, L > 0$ and $T > 0$. Then, for large enough r, there exists $N > 0$ such that*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|A^{r,m}(t) - \frac{\lambda^r}{|N^r|}t\right\|_L > \epsilon\right\} \le \epsilon, \tag{4.4.16}$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{0\le t_1,t_2\le L}|D^{r,m}(t_2) - D^{r,m}(t_1)| > N|t_2 - t_1| + \epsilon\right\} \le \epsilon, \ \ and \tag{4.4.17}$$

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

$$for \ all \ i \in \mathcal{I}, j \in \mathcal{J}(i), \ and \ k = 1, 2, \dots, N_i^r, \tag{4.4.18}$$

where $b = (1, 0)$ or $(0, 1)$.

The proof of (4.4.16) is the same as in [Dai and Tezcan, 2011], so we only need to show (4.4.17) and (4.4.18).

Proof of (4.4.17). For (4.4.17), [Dai and Tezcan, 2011] proved the analogous equation (77) in that paper when servers are identical inside each pool, thus they can show it directly using properties of Poisson processes. However, in our case, there is randomness among servers for each D_{ij}^r, which will cause troubles for a direct proof. To resolve this, we make a detour by comparing the departure processes of such systems to the ones of a homogeneous system. The departure processes of homogeneous systems can be proved to exhibit the almost Lipschitz condition in the direct way as shown in [Dai and Tezcan, 2011]. Then we use a coupling method to construct stochastically equivalent departure processes of the heterogeneous systems. Finally we compare the equivalent departure processes with the homogeneous departure processes, and obtain the almost Lipschitz condition for the former processes, and we conclude the original departure processes also have the almost Lipschitz property.

To this end, let $S_{ij}^{r;q}(t)$ be a Poisson process with rate $N_i q_{ij}$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}(i)$. First we show the almost Lipschitz condition for the sequence of processes $S_{ij}^{r;q}(t), i \in \mathcal{I}, j \in \mathcal{J}(i)$. Notice that we only need to investigate the process when $m = 0$, then multiply the error bounds. To see this, first define the hydrodynamic scaling for $S_{ij}^{r;q}(t)$:

$$S_{ij}^{r,q,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(S_{ij}^{r;q}\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - S_{ij}^{r;q}\left(\frac{m}{\sqrt{|N^r|}}\right)\right) \tag{4.4.19}$$

where

$$x_{r,m} = \left|Z^r\left(\frac{m}{\sqrt{|N^r|}}\right) - |N^r|\right|^2 \vee |N^r|. \tag{4.4.20}$$

Here $Q^r(\cdot)$, and $Z^r(\cdot)$ are the queue length and number of busy servers respectively. Then, similar to the proof in C.2.1 in [Dai and Tezcan, 2011], we have

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|S_{ij}^{r,q,m}(t)-\frac{N_i^r q_{ij}}{|N^r|}t\right\|_L>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\right\}$$

$$=P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\frac{1}{\sqrt{x_{r,m}}}\left(S_{ij}^{r,q}\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)-S_{ij}^{r,q}\left(\frac{m}{\sqrt{|N^r|}}\right)\right)\right.\right.$$
$$\left.\left.-\frac{N_i^r q_{ij}}{|N^r|}t\right\|_L>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\right\}$$

$$=P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\frac{1}{\sqrt{x_{r,m}}}\left(S\left(\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}t+\frac{N_i^r}{\sqrt{|N^r|}}q_{ij}m\right)-S\left(\frac{N_i^r}{\sqrt{|N^r|}}q_{ij}m\right)\right)\right.\right.$$
$$\left.\left.-\frac{N_i^r q_{ij}}{|N^r|}t\right\|_L>\frac{N_i^r q_{ij}}{|N^r|}\epsilon L\right\}$$

$$=P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\left(S\left(t+\frac{N_i^r}{|N^r|}q_{ij}\sqrt{|N^r|}m\right)-S\left(\frac{N_i^r}{|N^r|}q_{ij}\sqrt{|N^r|}m\right)\right)-t\right\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}L}\right.$$
$$\left.>\frac{N_i^r q_{ij}}{|N^r|}\epsilon L\sqrt{x_{r,m}}\right\}.$$

Since $S(t)$ is a homogeneous Poisson process, the probability above is equal to

$$P\left\{\max_{m<\sqrt{|N^r|}T}\|(S(t)-S(0))-t\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}L}>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\sqrt{x_{r,m}}\right\}$$

$$\leq\sum_{m<\sqrt{|N^r|}T}P\left\{\|S(t)-t\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}L}>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\sqrt{x_{r,m}}\right\},\qquad(4.4.21)$$

because $x_{r,m}\geq|N^r|$ by (4.4.20), and $\frac{N_i^r}{N^r}\to\beta_i\in(0,1)$, and by Proposition 4.3 of [Bramson, 1998], for given ϵ, any m and large enough r,

$$P\left\{\|S(t)-t\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}L}>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\sqrt{x_{r,m}}\right\}\leq\frac{\epsilon}{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,m}}L}\leq\frac{\epsilon}{2\beta_i q_{ij}\sqrt{|N^r|}L}.$$
$$(4.4.22)$$

Thus we only need to consider the process when $m=0$, then multiply by the error bound $\lceil\sqrt{|N^r|}T\rceil$. Using this result

$$P\left\{\sup_{0\leq t_1\leq t_2\leq L}\left|\left(S_{ij}^{r,q,0}(t_2)-\frac{N_i^r q_{ij}}{|N^r|}t_2\right)-\left(S_{ij}^{r,q,0}(t_1)-\frac{N_i^r q_{ij}}{|N^r|}t_1\right)\right|>\epsilon\right\}$$

$$\leq P\left\{\left\|S_{ij}^{r,q,0}(t_2)-\frac{N_i^r q_{ij}}{|N^r|}t_2\right\|_L+\left\|S_{ij}^{r,q,0}(t_1)-\frac{N_i^r q_{ij}}{|N^r|}t_1\right\|_L>\epsilon\right\}$$

$$\leq P\left\{\left\|S_{ij}^{r,q,0}(t_2)-\frac{N_i^r q_{ij}}{|N^r|}t_2\right\|_L>\frac{\epsilon}{2}\right\}+P\left\{\left\|S_{ij}^{r,q,0}(t_1)-\frac{N_i^r q_{ij}}{|N^r|}t_1\right\|_L>\frac{\epsilon}{2}\right\}$$

$$=P\left\{\|S(t_2)-t_2\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,0}}L}>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\sqrt{x_{r,0}}\right\}$$

$$+P\left\{\|S(t_1)-t_1\|_{\frac{N_i^r}{|N^r|}q_{ij}\sqrt{x_{r,0}}L}>\frac{N_i^r}{|N^r|}q_{ij}\epsilon L\sqrt{x_{r,0}}\right\}$$

$$\leq \frac{\epsilon}{\frac{N_i^r}{|N^r|} q_{ij} \sqrt{x_{r,0}} L} \leq \frac{\epsilon}{2\beta_i q_{ij} \sqrt{|N^r|} L}.$$

Therefore, reselecting ϵ, we get

$$P\left\{ \sup_{0 \leq t_1 \leq t_2 \leq L} \left| S_{ij}^{r,q,0}(t_2) - S_{ij}^{r,q,0}(t_1) \right| > \frac{N_i^r}{|N^r|} q_{ij}|t_2 - t_1| + \epsilon \right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{4.4.23}$$

Notice that $\frac{N_i^r}{|N^r|}$ depends on r in the Lipschitz coefficient. To get rid of this one can simply use the fact that, for large enough r, $\beta_i \in \left[\frac{N_i^r}{|N^r|} - \delta, \frac{N_i^r}{|N^r|} + \delta \right]$ for some really small $\delta > 0$, thus

$$P\left\{ \sup_{0 \leq t_1 \leq t_2 \leq L} \left| S_{ij}^{r,q,0}(t_2) - S_{ij}^{r,q,0}(t_1) \right| > \beta_i q_{ij}|t_2 - t_1| + \epsilon \right\}$$

$$\leq \left\{ \sup_{0 \leq t_1 \leq t_2 \leq L} \left| S_{ij}^{r,q,0}(t_2) - S_{ij}^{r,q,0}(t_1) \right| > \left(\frac{N_i^r}{|N^r|} - \delta \right) q_{ij}|t_2 - t_1| + \epsilon \right\}$$

$$\leq \left\{ \sup_{0 \leq t_1 \leq t_2 \leq L} \left| S_{ij}^{r,q,0}(t_2) - S_{ij}^{r,q,0}(t_1) \right| > \frac{N_i^r}{|N^r|} q_{ij}|t_2 - t_1| - \delta q_{ij}|t_2 - t_1| + \epsilon \right\}.$$

Since $|t_2 - t_1| < L$, the term $\delta q_{ij}|t_2 - t_1|$ can be chosen to be less than ϵ; therefore, after a reselection of ϵ, we have

$$P\left\{ \sup_{0 \leq t_1 \leq t_2 \leq L} \left| S_{ij}^{r,q,0}(t_2) - S_{ij}^{r,q,0}(t_1) \right| > \beta_i q_{ij}|t_2 - t_1| + \epsilon \right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{4.4.24}$$

Inequality (4.4.24) is sufficient for our comparison. Now we consider the departure processes with random service rates. We want to show (4.4.17). To overcome the difficulty in proving it directly, we compare D_{ij}^r with $S_{ij}^{r,q}$. In order to do this, we find a stochastically equivalent process $\breve{D}_{ij}^r$ of D_{ij}^r, which is the main difference between our proof and the one of [Dai and Tezcan, 2011]. We will do this in the following way.

For pool i, let $J(i)$ be the cardinality of $\mathcal{J}(i)$, i.e. total number of customer classes that servers in pool i are capable to serve. Then let $\{S_{ij_a}^{r,q}(t), a = 1, 2, \ldots, J(i)\}$ be a sequence of generated Poisson processes with rate $N_i^r q_{ij_a}$ respectively. And let $0 < \tau_1^a < \tau_2^a < \cdots$ be the sequence of occurrence times of the process $S_{ij_a}^{r,q}$, i.e. $S_{ij_a}^{r,q}(t) = \sum_{n=1}^{\infty} I(\tau_n^a < t)$. Let $\{U_n, n \in \mathbb{N}\}$ be a sequence of independent uniform $(0, 1)$ random variables, and $I(A)$ be the indicator function for event A which takes value 1 if A occurs and 0 otherwise. Assume all of the processes are equal to zero at $t = 0$. By splitting the processes $\{S_{ij_a}^{r,q}(t)\}$, we define the following processes

$$\breve{D}_{ij_1}^{r,n} = \sum_{l=1}^{n_1} I\left(U_l \leq \frac{\sum_{k=1}^{N_i^r} \mu_{ij_1k} \breve{B}_{ij_1k}(\tau_l^1-)}{\sum_{a=1}^{J(i)} N_i^r q_{ij_a}} \right), \tag{4.4.25}$$

$$\breve{D}_{ij_a}^{r,n} = \sum_{l=1}^{n_a} I\left(\frac{\sum_{c=1}^{a-1} N_i^r q_{ij_c}}{\sum_{a=1}^{J(i)} N_i^r q_{ij_a}} \leq U_l \leq \frac{\sum_{c=1}^{a-1} N_i^r q_{ij_c} + \sum_{k=1}^{N_i^r} \mu_{ij_ak} \breve{B}_{ij_ak}(\tau_l^a-)}{\sum_{a=1}^{J(i)} N_i^r q_{ij_a}} \right),$$

$$a = 2, 3, \ldots, J(i), \tag{4.4.26}$$

72

$$S_{ij_a}^{r,q}(t) = n_a = \sum_{l=1}^{\infty} I(\tau_l^a < t), \ a = 1, 2, \ldots, J(i), \tag{4.4.27}$$

$$\breve{D}_{ij_a}^r(t) = \breve{D}_{ij_a}^{r;n}, \forall t \in [\tau_n^a, \tau_{n+1}^a), \ a = 1, 2, \ldots, J(i), \tag{4.4.28}$$

where $\breve{B}_{ijk}(t)$ is determined by the arrival processes $A_i^r(t)$, the routing policies, the sequence $\tau_1^a, \tau_2^a, \ldots$, and the selection scheme defined as follows: we use class j_1 customer as an example, and other types of customer departure processes will be similar. If, for some $l \in \{1, 2, \ldots, S_{ij_1}^{r,q}(t)\}$, $I\left(U_l \leq \frac{\sum_{k=1}^{N_i^r} \mu_{ij_1 k} \breve{B}_{ij_1 k}(\tau_l^1-)}{N_i^r q_{ij_1}}\right) = 1$, then the potential departure occurring at time τ_l^1 in process $S_{ij_1}^{r,q}(t)$ is accepted as the real departure for process $\breve{D}_{ij_1}^r(t)$. Assume there are m busy servers just before this departure occurs (time τ_l^1-). Then after it is accepted as a real departure, one of the m servers will be freed, which leads to our selection scheme. Generate a uniformly distributed random variable ς on $(0, 1)$. If $\frac{\sum_{k=0}^{h} \mu_{ij_1 k}}{\sum_{k=1}^{m} \mu_{ij_1 k}} \leq \varsigma < \frac{\sum_{k=0}^{h+1} \mu_{ij_1 k}}{\sum_{k=1}^{m} \mu_{ij_1 k}}$, $h = 0, 1, \ldots, m - 1$, then the $h + 1$th server will be freed at time τ_l^1. Here we let $\mu_{ij_1 0} = 0$.

Under such a definition, the process $\breve{D}_{ij_a}^r(t)$ is stochastically equivalent to $D_{ij_a}^r(t)$ for all $a = 1, \ldots, J(i)$, thus in order to show

$$P\left\{\max_{m < \sqrt{|N^r|} T} \sup_{0 \leq t_1, t_2 \leq L} |D_{ij_a}^{r,m}(t_2) - D_{ij_a}^{r,m}(t_1)| > N|t_2 - t_1| + \epsilon\right\} \leq \epsilon,$$

it is equivalent to show

$$P\left\{\max_{m < \sqrt{|N^r|} T} \sup_{0 \leq t_1, t_2 \leq L} |\breve{D}_{ij_a}^{r,m}(t_2) - \breve{D}_{ij_a}^{r,m}(t_1)| > N|t_2 - t_1| + \epsilon\right\} \leq \epsilon. \tag{4.4.29}$$

From the construction of process $\breve{D}_{ij_a}^r$, it is easy to see that for every m,

$$\left|\breve{D}_{ij_a}^{r,m}(t_2) - \breve{D}_{ij_a}^{r,m}(t_1)\right|$$

$$= \frac{1}{\sqrt{x_{r,m}}} \left|\left(\breve{D}_{ij_a}^r\left(\frac{\sqrt{x_{r,m}} t_2}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - \breve{D}_{ij_a}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right)\right.$$

$$\left. - \left(\breve{D}_{ij_a}^r\left(\frac{\sqrt{x_{r,m}} t_1}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - \breve{D}_{ij_a}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right)\right|$$

$$= \frac{1}{\sqrt{x_{r,m}}} \left|\breve{D}_{ij_a}^r\left(\frac{\sqrt{x_{r,m}} t_2}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - \breve{D}_{ij_a}^r\left(\frac{\sqrt{x_{r,m}} t_1}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right)\right|$$

$$\leq \frac{1}{\sqrt{x_{r,m}}} \left|S_{ij_a}^{r,q}\left(\frac{\sqrt{x_{r,m}} t_2}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - S_{ij_a}^{r,q}\left(\frac{\sqrt{x_{r,m}} t_1}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right)\right|$$

$$= \left|S_{ij_a}^{r,q,m}(t_2) - S_{ij_a}^{r,q,m}(t_1)\right|. \tag{4.4.30}$$

The inequality is from the fact that the event points of $\breve{D}_{ij_a}^r(t)$ are chosen from the event points of $S_{ij_a}^{r,q}(t)$ (equations (4.4.26) and (4.4.27)), thus the departure difference between time t_2 and t_1 is a subset of the difference of $S_{ij_a}^{r,q}(t)$ during the same time. Then we also only need to consider

the process when $m = 0$. From (4.4.24) and (4.4.30), we have

$$P\left\{\sup_{0\leq t_1\leq t_2\leq L}\left|\breve{D}_{ij_a}^{r,0}(t_2) - \breve{D}_{ij_a}^{r,0}(t_1)\right| > \beta_i q_{ij_a}|t_2 - t_1|+\epsilon\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{4.4.31}$$

Note that we can conduct the same analysis for each pool, thus we can get rid of the a here.

Let $N = \max_{i\in\mathcal{I}, j\in\mathcal{J}}\{\beta_i q_{ij}\}$. Then, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$P\left\{\sup_{0\leq t_1\leq t_2\leq L}\left|\breve{D}_{ij}^{r,0}(t_2) - \breve{D}_{ij}^{r,0}(t_1)\right| > N|t_2 - t_1|+\epsilon\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{4.4.32}$$

Multiplying the error bound $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately we obtain (4.4.17). $\qquad\square$

Before proving (4.4.18), we need a lemma taken from [Bramson, 1998, Lemma 5.1].

Lemma 4.4.1. *Let $v_{ijk}^{r,T,\max} = \max\{|v_{ijk}(l)|:= V_{ijk}(l-1) \leq |N^r|T\}$ for all $i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i^r$. Then, for given T,*

$$v_{ijk}^{r,T,\max}/\sqrt{|N^r|} \to 0 \text{ in probability as } r \to \infty, \text{ for all } i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i^r. \tag{4.4.33}$$

The proof is in Appendix B.2.

Using this proposition, one can show that $\mathbb{X}^{r,m}$ is almost Lipschitz in probability, as described 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 4.4.2. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Assume that Assumption 4.3.1 and Assumption 4.3.2 hold. Fix $\epsilon > 0$, $L > 0$, and $T > 0$. Then for large enough r,*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{0\leq t_1\leq t_2\leq L}|\mathbb{X}^{r,m}(t_2) - \mathbb{X}^{r,m}(t_1)| > N|t_2 - t_1|+\epsilon\right\} \leq \epsilon, \tag{4.4.34}$$

where $N < \infty$ and only depends on λ.

The proof is similar to that of Proposition 5.3 in [Dai and Tezcan, 2011]. We include it into Appendix B.3.

For convenience, we assume for the rest of the paper that $N \geq 1$ and $L \geq 1$. Let

$$\mathscr{K}_0^r = \left\{\max_{m<\sqrt{|N^r|}T}\sup_{0\leq t_1\leq t_2\leq L}|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)|\leq N|t_1 - t_2|+\epsilon(r)\right\}, \tag{4.4.35}$$

where N, L, and T are fixed as before and $\epsilon(r)$ with $\epsilon \to 0$ as $r \to \infty$ is a sequence of real numbers. Similarly, we can replace ϵ in (4.4.16), (4.4.17), and (4.4.18) by $\epsilon(r)$. We denote

these new inequalities obtained from (4.4.16), (4.4.17), and (4.4.18) by (4.4.16)′, (4.4.17)′, and (4.4.18)′. Let $\mathscr{K}^r$ denote the intersection of $\mathscr{K}_0^r$ with the complements of the events in (4.4.16)′, (4.4.17)′, and (4.4.18)′. As in [Dai and Tezcan, 2011], when $\epsilon(r) \to 0$ sufficiently slowly as $r \to \infty$, one can show that $P(\mathscr{K}^r) \to 1$ as $r \to \infty$.

We summarize the above discussion in the following corollary for future reference, which is similar to [Dai and Tezcan, 2011, Corollary 5.1].

Corollary 4.4.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server system processes. Assume that Assumption 4.3.1 and Assumption 4.3.2 hold. Fix $L > 0$, and $T > 0$ and choose $\epsilon(r)$ as above. Then for $\mathscr{K}^r$ defined as above*

$$\lim_{r \to \infty} P(\mathscr{K}^r) = 1. \tag{4.4.36}$$

4.4.2 Hydrodynamic limits

In this section, we define the hydrodynamic limits. First, we define a set of functions that contains all of the hydrodynamic limits. The following definitions are similar to those in [Dai and Tezcan, 2011], 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 satisfy

$$|x(0)| \le 1 \tag{4.4.37}$$

and

$$|x(t_2) - x(t_1)| \le N|t_1 - t_2| \text{ for all } t_1, t_2 \in [0, L], \tag{4.4.38}$$

where the constant N is chosen as in Proposition 4.4.2. We set

$$E^r = \{\mathbb{X}^{r,m}, m < \sqrt{|N^r|}T, \omega \in \mathscr{K}^r\} \tag{4.4.39}$$

and

$$\mathscr{E} = \{E^r : r \in \mathbb{N}\} \tag{4.4.40}$$

where T is fixed, and $\mathscr{K}^r$ is defined as in the previous section.

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

Because

$$|\mathbb{X}^{r,m}(0)| \le 1 \tag{4.4.41}$$

for all $m < \sqrt{|N^r|}T$ and $r \in \mathbb{N}$, the following result is a corollary in [Dai and Tezcan, 2011], and is similar to Corollary 5.2 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. One can use the almost Lipschitz property of processes $\mathbb{X}^{r,m}$ to show this-see [Bramson, 1998, Lemma 4.2].

Corollary 4.4.2. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Assume that Assumption 4.3.1 holds and π satisfies 4.3.2. Let $\tilde{E}, E^r$, and $\mathscr{E}$ be as specified above. Fix $\epsilon > 0, L > 0$ and $T > 0$, and choose r large enough. Then, for $\omega \in \mathscr{K}^r$ and any $m < \sqrt{|N^r|}T$, there exists a hydrodynamic limit $\tilde{\mathbb{X}}(\cdot) \in E'$, such that*

$$\left\| \mathbb{X}^{r,m}(\cdot) - \tilde{\mathbb{X}}(\cdot) \right\|_L \leq \epsilon. \tag{4.4.42}$$

The next result is mainly needed to translate the condition on the hydrodynamic model solutions to hydrodynamic limits given in Assumption 4.3.4. It also reveals the origin of hydrodynamic model equations.

Proposition 4.4.3. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server system processes. Assume that Assumption 4.3.1 holds and π satisfies Assumption 4.3.2. Choose $L > 0$ and let $\tilde{\mathbb{X}}_\pi$ be a hydrodynamic limit of $\mathscr{E}$ over $[0, L]$. $\tilde{\mathbb{X}}_\pi$ satisfies the hydrodynamic model equations (4.3.19)-(4.3.27) on $[0, L]$.*

To prove 4.4.3, we need a lemma which appears to be the same as Lemma C.1 in [Dai and Tezcan, 2011], but it is for random rates systems.

Lemma 4.4.2. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel random server systems processes. Assume that Assumption 4.3.1 holds and π satisfies Assumption 4.3.2. Fix $\epsilon > 0, L > 0$, and $T > 0$. Then, for large enough r and $\omega \in \mathscr{A}$,*

$$\max_{m < \sqrt{|N^r|}T} \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^L |Z_{ij}^{r,m}(s)| ds < \epsilon, \forall\ i \in \mathcal{I},\ and\ j \in \mathcal{J}(i). \tag{4.4.43}$$

Proof. The proof of this lemma is the same as in [Dai and Tezcan, 2011], except we do not consider $Q^{r,m}$. Recalling that $Z_{ij}^{r,m}(t) = \sum_{k=1}^{N_i^r} B_{ijk}^{r,m}(t)$, the proof becomes obvious. □

Now we prove the proposition.

The proof of Proposition 4.4.3. Proof is similar to that in [Dai and Tezcan, 2011, Proposition 5.4]. Assume that Assumption 4.3.1 holds, and π satisfies Assumption 4.3.2. g satisfies Assumption 4.3.3. Fix $\omega \in \mathscr{K}^r$ and let $\mathbb{X}^{r,m}$ be given as in (4.4.2)-(4.4.4). By $(4.4.16)'$, we have, for large enough r, that

$$\left\| A^{r,m}(t) - \frac{\lambda^r}{|N^r|}t \right\|_L \leq \epsilon(r). \tag{4.4.44}$$

Using (4.4.13) and Lemma 4.4.2 gives

$$\left\| \sum_{k=1}^{N_i^r} \left(T_{ijk}^{r,m}(t) - \frac{x_{ij}^*}{|N^r|}t \right) \right\|_L \le \epsilon(r). \tag{4.4.45}$$

Now select any hydrodynamic limit $\tilde{\mathbb{X}}$ of $\mathscr{E}$. By Corollary 4.4.2, for given $\delta > 0$, choose (r,m) so that, $\epsilon(r) \le \delta$,

$$\left\| \tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t,\omega) \right\|_L \le \delta, \tag{4.4.46}$$

and

$$\left| \frac{\lambda^r}{|N^r|} - \lambda \right| \le \delta. \tag{4.4.47}$$

It follows from (4.4.44) and (4.4.47) that

$$\left\| \tilde{A}(t) - \lambda t \right\|_L \le (2+L)\delta. \tag{4.4.48}$$

Until now the proof remains the same as in [Dai and Tezcan, 2011, Proposition 5.4]. The next inequality is an analogy of inequality (C25) in [Dai and Tezcan, 2011]. However, we can not derive (C25) directly in our case because we do not have a constant μ_{jk} here. Since we have i.i.d. random service rates $\{\mu_{ijk}\}, i \in \mathcal{I}, j \in \mathcal{J}(i), k = 1, 2, \ldots, N_i$, using the Law of Large Numbers and properties of Poisson processes, we have

$$\left\| \sum_{k=1}^{N_i^r} \left(T_{ijk}^{r,m}(t) - \frac{D_{ijk}^{r,m}(t)}{\mu_{ijk}} \right) \right\|_L$$

$$= \frac{1}{\sqrt{x_{r,m}}} \left\| \sum_{k=1}^{N_i^r} \frac{1}{\mu_{ijk}} \left(S_{ijk} \left(\mu_{ijk} T_{ijk}^r \left(\frac{m}{\sqrt{|N^r|}} \right) \right) - S_{ijk} \left(\mu_{ijk} T_{ijk}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) \right) \right) \right.$$

$$\left. - \sum_{k=1}^{N_i^r} \frac{1}{\mu_{ijk}} \left(\mu_{ijk} T_{ijk}^r \left(\frac{m}{\sqrt{|N^r|}} \right) - \mu_{ijk} T_{ijk}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) \right) \right\|_L$$

$$\le \frac{1}{\sqrt{x_{r,m}}} \left\| \sum_{k=1}^{N_i^r} \left(S_{ijk} \left(\mu_{ijk} T_{ijk}^r \left(\frac{m}{\sqrt{|N^r|}} \right) \right) - S_{ijk} \left(\mu_{ijk} T_{ijk}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) \right) \right) \right.$$

$$\left. - \sum_{k=1}^{N_i^r} \left(\mu_{ijk} T_{ijk}^r \left(\frac{m}{\sqrt{|N^r|}} \right) - \mu_{ijk} T_{ijk}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) \right) \right\|_L$$

$$= \frac{1}{\sqrt{x_{r,m}}} \left\| \sum_{k=1}^{N_i^r} \left(S_{ijk} \left(\mu_{ijk} \int_0^{\frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds \right) - S_{ijk} \left(\mu_{ijk} \int_0^{\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds \right) \right) \right.$$

$$\left. - \sum_{k=1}^{N_i^r} \left(\mu_{ijk} \int_0^{\frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds - \mu_{ijk} \int_0^{\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds \right) \right\|_L$$

$$\overset{d}{=} \frac{1}{\sqrt{x_{r,m}}} \left\| \sum_{k=1}^{N_i^r} \left(S_{ijk} \left(\mu_{ijk} \int_{\frac{m}{\sqrt{|N^r|}}}^{\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds \right) \right) \right.$$

$$\left. - \sum_{k=1}^{N_i^r} \left(\mu_{ijk} \int_{\frac{m}{\sqrt{|N^r|}}}^{\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}} B_{ijk}(s)ds \right) \right\|_L$$

$$\leq \frac{1}{\sqrt{x_{r,m}}} \left\| \sum_{k=1}^{N_i^r} \left(S_{ijk} \left(\mu_{ijk} \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) \right) - \sum_{k=1}^{N_i^r} \left(\mu_{ijk} \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) \right\|_L$$

$$\overset{d}{=} \frac{1}{\sqrt{x_{r,m}}} \left\| S \left(\sum_{k=1}^{N_i^r} \mu_{ijk} \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) - \left(\sum_{k=1}^{N_i^r} \mu_{ijk} \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) \right\|_L$$

$$\leq \frac{1}{\sqrt{x_{r,m}}} \left\| S \left(\sum_{k=1}^{N_i^r} q \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) - \left(\sum_{k=1}^{N_i^r} q \frac{\sqrt{x_{r,m}}t}{|N^r|} \right) \right\|_L$$

$$= \frac{1}{\sqrt{x_{r,m}}} \left\| S \left(q\sqrt{x_{r,m}}t \right) - \left(q\sqrt{x_{r,m}}t \right) \right\|_L \tag{4.4.49}$$

where $\overset{d}{=}$ means equal in distribution and S is a standard Poisson process. The first $\overset{d}{=}$ is because every S_{ijk} is a stationary process. The second $\overset{d}{=}$ comes from the fact that the superposition of several Poisson processes is also a Poisson process with rate being the sum of the rates of the original processes. From (4.4.22), for r large enough, $(4.4.49) < \epsilon(r)$, thus

$$\left\| \sum_{k=1}^{N_i^r} \left(T_{ijk}^{r,m}(t) - \frac{D_{ijk}^{r,m}(t)}{\mu_{ijk}} \right) \right\|_L < \epsilon(r). \tag{4.4.50}$$

Therefore, from (4.4.50), (4.4.45), and (4.4.46),

$$\left\| \left(\tilde{D}_{ij}(t) - \sum_{k=1}^{N_i^r} \frac{\mu_{ijk}t}{|N^r|} \right) \right\|_L \leq \left\| \left(\tilde{D}_{ij}(t) - D_{ij}^{r,m}(t) \right) \right\|_L + \left\| \sum_{k=1}^{N_i^r} \left(D_{ijk}^{r,m}(t) - \mu_{ijk}T_{ijk}^{r,m}(t) \right) \right\|_L$$

$$+ \left\| \sum_{k=1}^{N_i^r} \left(\mu_{ijk}T_{ijk}^{r,m}(t) - \frac{\mu_{ijk}x_{ij}^*}{|N^r|}t \right) \right\|_L$$

$$\leq \delta(1 + 2q_{ij}), \tag{4.4.51}$$

and from the Law of Large Numbers

$$\left\| \sum_{k=1}^{N_i^r} \frac{\mu_{ijk}x_{ij}^*}{|N^r|}t - \bar{\mu}_{ij}z_{ij}t \right\|_L \leq \delta, \tag{4.4.52}$$

for large enough r. Thus, from (4.4.51) and (4.4.52), and reselecting δ, one has

$$\left\| \tilde{D}_{ij}(t) - \bar{\mu}_{ij}z_{ij}t \right\|_L \leq \delta. \tag{4.4.53}$$

78

By combining (4.4.46), (4.4.48), (4.4.8), and (4.4.9), we get

$$\left\| \lambda_j t - \tilde{A}_{qj}(t) - \sum_{i \in \mathcal{I}(j)} \tilde{A}_{sij}(t) \right\|_L \leq (2+L)\delta, \text{ and} \tag{4.4.54}$$

$$\left\| \tilde{Q}_j(t) - \tilde{Q}_j(0) - \tilde{A}_{aj}(t) + \sum_{i \in \mathcal{I}(j)} \tilde{C}_{ij}(t) \right\|_L \leq 4\delta. \tag{4.4.55}$$

By combining (4.4.46) with (4.4.53) and (4.4.10), we get

$$\left\| \tilde{Z}_{ij}(t) - \tilde{Z}_{ij}(0) - \tilde{A}_{sij}(t) - \tilde{C}_{ij}(t) + \bar{\mu}_{ij} z_{ij} t \right\|_L \leq 5\delta. \tag{4.4.56}$$

Equations (4.4.54)-(4.4.56) show that the hydrodynamic limits satisfy (4.3.19),(4.3.20), and (4.3.22).

That the hydrodynamic limits satisfy (4.3.25) and (4.3.26) is proved similarly to the fact that the fluid limits satisfy the fluid analogs of those equations. Hence, we only illustrate the proof of (4.3.25).

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 $\omega_l \in \mathcal{K}$ for all $l \geq 0$, such that

$$\mathbb{X}^{r_l, m_l}(\cdot, \omega_l) \to \tilde{\mathbb{X}}(\cdot) \text{ u.o.c. as } l \to \infty. \tag{4.4.57}$$

Fix $t > 0$. If for any $j \in \mathcal{J}$, $\tilde{Q}_j(t) = 0$, (4.3.25) holds trivially. Now we assume that $\tilde{Q}_j(t) > a$ for some $a > 0$. By (4.4.57), there exists an l_0 such that

$$Q_j^{r_l, m_l}(t, \omega_l) > a/2 \text{ for all } l > l_0.$$

This implies, by (4.4.14), that

$$\sum_{k=1}^{N_i^r} B_{ijk}^{r_l, m_l}(t, \omega_l) = 0,$$

hence,

$$Q_j^{r_l, m_l}(t, \omega_l) \sum_{k=1}^{N_i^r} B_{ijk}^{r_l, m_l}(t, \omega_l) = 0. \tag{4.4.58}$$

Convergence in (4.4.57) implies that

$$Q_j^{r_l, m_l}(t, \omega_l) \sum_{k=1}^{N_i^r} B_{ijk}^{r_l, m_l}(t, \omega_l) \to \tilde{Q}_j(t) \left(\sum_{k=1}^{N_i^r} \tilde{B}_{ijk}(t) \right) \text{ as } l \to \infty.$$

This gives (4.3.25) by (4.4.58). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Observe that, by (4.3.28) and the definitions of hydrodynamic and diffusion scalings,

$$|g(Q^{r,0}(0), Z^{r,0}(0))| \leq |g(\hat{Q}^r(0), \hat{Z}^r(0))|. \tag{4.4.59}$$

If condition (4.3.30) holds, then (4.4.59) implies that $g(Q^{r,0}(0), Z^{r,0}(0)) \to 0$ in probability as $r \to \infty$. Therefore, we can choose $\epsilon(r)$ with $\epsilon(r) \to 0$ as $r \to \infty$ such that, for $\mathscr{L}^r = \mathscr{K}^r \cap \mathscr{G}^r$, where

$$\mathscr{G}^r = \{|g(Q^{r,0}(0), Z^{r,0}(0))| \leq \epsilon(r)\}, \tag{4.4.60}$$

we have

$$\lim_{r \to \infty} P(\mathscr{L}^r) = 1. \tag{4.4.61}$$

We set

$$E_g^r = \{\mathbb{X}^{r,0}(\cdot, \omega), \omega \in \mathscr{L}^r\}, \tag{4.4.62}$$

and

$$\mathscr{E}_g = \{E_g^r, r \in \mathbb{N}\}. \tag{4.4.63}$$

The following proposition is similar to [Dai and Tezcan, 2011, Proposition 5.5]. It connects Assumption 4.3.4 with Corollary 4.4.2, and shows an inequality similar to (4.3.29) holds for the hydrodynamically scaled process $\mathbb{X}^{r,m}(\cdot)$.

Proposition 4.4.4. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel random server systems processes. Assume that Assumption 4.3.1 and 4.3.2 hold, g satisfies Assumption 4.3.3, and the hydrodynamic model of the system satisfies Assumption 4.3.4. Fix $\epsilon > 0, L > 0$, and $T > 0$, and assume that r is large. Then, for $\omega \in \mathscr{K}^r$,*

$$g(Q^{r,m}(t), Z^{r,m}(t)) \leq H(t) + \epsilon \tag{4.4.64}$$

for all $t \in [0, L]$, and $m < \sqrt{|N^r|}T$, with $H(\cdot)$ as given in Assumption 4.3.4. Furthermore, for $\omega \in \mathscr{L}^r$

$$||g(Q^{r,0}(t), Z^{r,0}(t))||_L \leq \epsilon. \tag{4.4.65}$$

If, in addition, condition (4.3.30) holds, then (4.4.61) holds.

The proof is the same as in [Dai and Tezcan, 2011]. We state the proof modified for our notation in Appendix B.4.

4.4.3 SSC in the diffusion limits

In this section we change the scaling from hydrodynamic to diffusion to prove Theorem 4.3.2. Once the scaling is changed, a few complications need to be dealt with regarding the change in the range of the time variable. In this process, all of the steps including lemmas and proofs

remain the same as in [Dai and Tezcan, 2011]. We include only final results in this section and leave the proofs in the Appendix.

We begin by changing the scaling. One can check by employing (4.3.17), (4.3.18), and (4.4.3), (4.4.4) that

$$
\begin{aligned}
Q^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}} \hat{Q}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) = \frac{1}{y_{r,m}} \hat{Q}^r \left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t + m) \right) \text{ and} \\
Z^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}} \hat{Z}^r \left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}} \right) = \frac{1}{y_{r,m}} \hat{Z}^r \left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t + m) \right),
\end{aligned}
\tag{4.4.66}
$$

where

$$
y_{r,m} = \sqrt{\frac{x_{r,m}}{|N^r|}} = \left| \hat{Z}^r \left(\frac{m}{\sqrt{|N^r|}} \right) \right| \vee 1. \tag{4.4.67}
$$

By changing the scaling in Proposition 4.4.4 as above, we can rephrase (4.4.64) and (4.4.65). Also, 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 (4.4.66) will correspond to

$$
\frac{m}{\sqrt{|N^r|}} \le t \le \frac{1}{\sqrt{|N^r|}}(y_{r,m}L + m) \tag{4.4.68}
$$

for the arguments on the right.

The next proposition uses the connection between hydrodynamic scaling $\mathbb{X}^{r,m}$ and diffusion scaling $\hat{\mathbb{X}}^r$ in (4.4.66) to translate the inequality (4.4.64) into another version for diffusively scaled processes.

Proposition 4.4.5. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel random server system processes. Assume that Assumption 4.3.1 and Assumption 4.3.2 hold, g satisfies Assumption 4.3.3, and the hydrodynamic model of the system satisfies Assumption 4.3.4. Fix $\epsilon > 0, L > 0,$ and $T > 0$, and assume that r is large. Then, for $\omega \in \mathscr{K}^r$ and for $H(\cdot)$ given as in Assumption 4.3.4,*

$$
g(\hat{Q}^r(t), \hat{Z}^r(t)) \le y^c_{r,m} H \left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t - m) \right) + \epsilon y^c_{r,m} \tag{4.4.69}
$$

for all $t \in [0, T]$ and m satisfying (4.4.68). Also

$$
\left\| g(\hat{Q}^r(t), \hat{Z}^r(t)) \right\|_{Ly r, 0/\sqrt{|N^r|}} \le \epsilon y^c_{r,0} \tag{4.4.70}
$$

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

At this point, if we can show that $(\sqrt{|N^r|}t - m)/y_{r,m}$ is large enough, we can conclude that the results in Theorem 4.3.2 hold by using the convergence property of $H(t)$, as given in Assumption 4.3.4. 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 $\mathscr{K}^r$ with the reselected L. Let H be given as in Assumption 4.3.4. Then $H(t) \to 0$ as $t \to \infty$, thus

for any fixed $\epsilon > 0$, there exists $s^*(\epsilon) > 1$ such that, for all $t > s^*(\epsilon), H(t) < \epsilon$. We assume for the rest of the paper that

$$L \geq 6Ns^*(\epsilon), \tag{4.4.71}$$

where N is chosen as in (4.4.35).

To make $\sqrt{|N^r|}t - m$ large, for a fixed $t \in [0, T]$, we take the smallest m that satisfies (4.4.68), which we denote by $m_r(t)$. We need the following lemmas to show that $\sqrt{|N^r|}t - m_r(t)$ is large.

Lemma 4.4.3. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel random server system processes. Assume that Assumption 4.3.1 and Assumption 4.3.2 hold. For fixed $L > 0$ and $T > 0$, and large enough r*

$$y_{r,m+1} \leq 3Ny_{r,m} \tag{4.4.72}$$

for $\omega \in \mathscr{K}^r$ and $m < \sqrt{|N^r|}T$, with the constant N chosen as in (4.4.35).

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

Lemma 4.4.4. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel random server systems processes. For fixed $L > 0$ and $T > 0$, and large enough r,*

$$\sqrt{|N^r|}t - m_r(t) \geq Ly_r(m_r(t))/6N \tag{4.4.73}$$

for $\omega \in \mathscr{K}^r$ and $t \in (Ly_{r,0}/\sqrt{|N^r|}, T]$, with the constant N chosen as in (4.4.35).

Finally we can use the results from all of these lemmas and propositions to prove Theorem 4.3.2 now.

Proof of Theorem 4.3.2. Assume that Assumption 4.3.1 and Assumption 4.3.2 hold, g satisfies Assumption 4.3.3, the hydrodynamic model satisfies Assumption 4.3.4, and condition (4.3.30) holds.

Fix $\vartheta > 0$. By (4.4.36) and (4.4.61), there exists $r_0 > 0$ such that

$$P(\mathscr{K}^r) \geq P(\mathscr{L}^r) > 1 - \vartheta/2 \tag{4.4.74}$$

for all $r > r_0$. Fix $\epsilon > 0$ and take $L \geq 6Ns^*(\epsilon)$. Then, by (4.4.69) and Lemma 4.4.4, for $\omega \in \mathscr{K}^r, t \in (Ly_{r,0}/\sqrt{|N^r|}, T]$, and r large enough

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \leq 2\epsilon(y_r(m_r(t)))^c. \tag{4.4.75}$$

We have almost proven the result. Now recall by (4.4.67),

$$y_r(m_r(t)) = \left|\hat{Z}^r\left(\frac{m_r(t)}{\sqrt{|N^r|}}\right)\right| \vee 1 \leq ||\hat{Z}^r(\cdot)||_T \vee 1. \tag{4.4.76}$$

If we include the initial condition (4.3.30), we will have (4.4.70) as discussed above. Thus combine this and (4.4.76), for $t \in [0, Ly_{r,0}/\sqrt{|N^r|}]$ and $\omega \in \mathscr{L}^r$,

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \leq 2\epsilon(y_{r,0})^c \leq 2\epsilon(\|\hat{Z}^r(\cdot)\|_T \vee 1)^c. \tag{4.4.77}$$

Combining (4.4.75), (4.4.76), and (4.4.77) gives

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \leq 2\epsilon(y_{r,0})^c \leq 2\epsilon(\|\hat{Z}^r(\cdot)\|_T \vee 1)^c \tag{4.4.78}$$

for all $t \in [0, T]$ and $\omega \in \mathscr{L}^r$. Finally, by (4.4.74) and (4.4.78), for large enough r,

$$P\left\{ \frac{\|g(\hat{Q}^r(\cdot), \hat{Z}^r(\cdot))\|_T}{(\|\hat{Z}^r(\cdot)\|_T \vee 1)^c} > 2\epsilon \right\} < \vartheta. \tag{4.4.79}$$

This clearly implies (4.3.31) because $\epsilon > 0$ and $\vartheta > 0$ are arbitrary. $\qquad\square$

4.5 State space collapse in many server queue

Although [Dai and Tezcan, 2011]'s work is developed for multi-calss queueing networks, such SSC results can also be applied to the many server queue analysis. To gain more insight into how the SSC is related to many server queues with random servers, in this section, we provide a new approach using the SSC technique to prove diffusion limit result in [Atar, 2008, Theorem 2.1]. We state the theorem in Chapter 3. Our analysis is under the assumption of the LISF policy. By proving result of [Atar, 2008] using the state space collapse phenomenon, we provide a more generic method for showing diffusion limits for many server queues with random servers. Under other routing policies, one can also prove the diffusion limits using this approach by inventing their unique SSC functions.

Before the proof, we need some preliminary mathematics. First we have the shifted and scaled system process (the total number of customers in the system) as

$$\hat{X}^r(t) = \hat{X}^r(0) + W^r(t) + F^r(t) + b^r t, \tag{4.5.1}$$

where

$$b^r = \frac{\lambda^r t - r\lambda}{\sqrt{r}} + \sum_{k=1}^{N^r}\left(\frac{\mu_k - \bar{\mu}}{\sqrt{r}}\right) + \bar{\mu}\hat{N}, \tag{4.5.2}$$

and

$$F^r(t) = \sum_{i=1}^{I} F^{r,(i)}(t) = \sum_{i=1}^{I} \frac{\sum_{k \in K_i} \mu_k t - T^{r,(i)}(t)}{\sqrt{r}}$$

$$= \sum_{k=1}^{N^r} \frac{\mu_k t - T_k^r(t)}{\sqrt{r}} = \frac{\int_0^t \sum_{k=1}^{N^r} \mu_k I_k^r(s)ds}{\sqrt{r}} = \int_0^t \sum_{k=1}^{N^r} \mu_k \hat{I}_k^r(s)ds, \tag{4.5.3}$$

and all of the notations are as defined in Section 2 of Chapter 3.

We want to show that the process $\hat{X}^r(t)$ converges to a diffusion as the system size grows large. The convergence of $W^r(t)$, and $b^r t$ will be the same as in [Atar, 2008], and we also stated this in 3.2 of this thesis. We focus on showing that the item $F^r(t)$ weakly converges to $\gamma \int_0^t \xi(s)^- ds$. To achieve this, we use the SSC result from [Dai and Tezcan, 2011].

In order to show the convergence of $F(t)$, first we consider another sequence of systems - a sequence of inverted-V systems with I server pools. Each pool contains N_i^r identical servers with service rates $\mu^{(i)}, i = 1, 2, \ldots, I$. $\mu^{(i)}$ are defined in the proof of Theorem 2.3.1. When there are several servers idle at a moment, customers are routed to one of them according to the Longest Idle Server First(LISF) policy, which means that customers are routed to the server who has been idle for the longest time at the time of routing. Such a sequence of systems can be seen as a special case of [Atar, 2008]'s model in the sense that the random service rates have a discrete distribution with support $\{\mu^{(i)}, i = 1, 2, \ldots, I\}$, and servers are allocated according to different scenarios.

Assume the following conditions hold for the number of servers in the inverted-V systems:

$$\lim_{r \to \infty} \frac{N}{r} = 1, \tag{4.5.4}$$

$$\lim_{r \to \infty} \frac{N_i}{N} = \beta_i, \quad \text{for } i = 1, 2, \ldots, I, \tag{4.5.5}$$

where $\sum_{i=1}^I \beta_i = 1$.

The heavy traffic condition holds throughout the discussion:

$$\lim_{r \to \infty} \frac{1}{\sqrt{r}} \left(\lambda^r - \sum_{i=1}^I \mu^{(i)} N_i^r \right) = \hat{\lambda}, \tag{4.5.6}$$

and $\hat{\lambda}$ should be the same as in (2.2.2).

The idea of our new approach to Atar's results is to show the SSC result of the special model, then let the difference between $\mu^{(i-1)}$ and $\mu^{(i)}$ tends to zero, so that the discrete model will 'converge' to a continuous model, and thus the SSC result of the continuous model will also be proven.

The SSC result for such inverted-V systems is a simpler case of the results in [Dai and Tezcan, 2011], thus it is well established. For completeness, we repeat the notations here. Let $\mathbb{X}^r = (Q^r, Z^r)$ denote the system processes, where $Q^r(t)$ is number of customers waiting in the queue at time t, and $Z^r(t) = (Z_i^r(t), i = 1, 2, \ldots, I)$, where $Z_i^r(t)$ is the number of busy servers in pool i at time t. Since there are no abandonments in the queue, we will make a small modification on the hydrodynamic scaling factor $x_{r,m}$. Different to the $x_{r,m}$ defined in [Dai and Tezcan, 2011], let

$$x_{r,m} = \left| Z^r \left(\frac{m}{\sqrt{N^r}} \right) - N^r \right|^2 \vee |N^r|, \tag{4.5.7}$$

84

where $N^r = (N_1^r, N_2^r, \ldots, N_I^r)$ is an I dimensional vector, and $|N^r|$ is the total number of servers in the systems. Notice that this $x_{r,m}$ does not contain $Q^r(t)$.

We can show that the hydrodynamic limits results in [Dai and Tezcan, 2011] still hold with the new $x_{r,m}$ defined above. The proof is trivial thus we omit it here. We still use $\tilde{\mathbb{X}}$ to represent the hydrodynamic limit.

Finally the diffusive scaling is defined as

$$\hat{Q}^r(t) = \frac{Q^r(t)}{\sqrt{|N^r|}} \text{ and } \hat{Z}_i^r(t) = \frac{Z_i^r(t) - N_i^r}{\sqrt{|N^r|}}, \text{ for } t \geq 0.$$

We will need the following lemma in preparation for our SSC proof.

Lemma 4.5.1. *Let $\{\mathbb{X}^r\}$ be a sequence of inverted-V system processes as defined in the beginning of this section. Assume (4.5.6) and (4.5.19) hold, and $\tilde{\mathbb{X}}$ be any hydrodynamic limit of such system, then*

$$\lim_{t \to \infty} \left(\frac{\tilde{Z}_i(t)}{\tilde{Z}_l(t)} - \frac{\beta_i}{\beta_l} \right) = 0 \ \forall i \neq l, as \ t \to \infty. \tag{4.5.8}$$

Proof. From (4.5.4) and (4.5.19), we know that $\frac{Z_i^r(\cdot)}{N^r} \to \beta_i$ u.o.c. in probability as $r \to \infty$. Recall that $\bar{\mathbb{X}}$ is called a fluid limit of $\{\mathbb{X}^r\}$ if there exists an $\omega \in \mathscr{A}$ ($\mathscr{A}$ is taken from Appendix B. in [Dai and Tezcan, 2011]) and a sequence $\{r_n\}$ with $r_n \to \infty$ as $n \to \infty$ such that $\frac{\mathbb{X}^{r_n}(\cdot, \omega)}{N^{r_n}}$ converges u.o.c. to $\bar{\mathbb{X}}$ as $n \to \infty$. Therefore, $(\beta_1, \beta_2, \ldots, \beta_I)$ is the fluid limit of $\left(\frac{Z_1^r(\cdot)}{N^r}, \frac{Z_2^r(\cdot)}{N^r}, \ldots, \frac{Z_I^r(\cdot)}{N^r} \right)$. Obviously $(\beta_1, \beta_2, \ldots, \beta_I)$ is a time invariant state, thus it is also the steady state of the fluid limit.

Now we try to show (4.5.11). We know that $\tilde{Z}_i(t)$ is the hydrodynamic limit of $Z_i^{r,m}(t)$, i.e. if we fix $\epsilon > 0, L > 0$ and $T > 0$, and choose r large enough, then for $\omega \in \mathscr{K}^r$ ($\mathscr{K}^r$ as defined in [Dai and Tezcan, 2011, Corollary 5.1]) and any $m < \sqrt{N^r}T$,

$$\|Z_i^{r,m}(\cdot) - \tilde{Z}_l(\cdot)\|_L \leq \epsilon \tag{4.5.9}$$

for some hydrodynamic limit $\tilde{Z}_i(\cdot), i = 1, 2, \ldots, I$. Remember that

$$Z_i^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(Z_i^r \left(\frac{\sqrt{x_{r,m}} t}{N^r} + \frac{m}{\sqrt{N^r}} \right) - N_i^r \right). \tag{4.5.10}$$

Thus in order to show $\frac{\tilde{Z}_i(t)}{\tilde{Z}_l(t)} \to \frac{\beta_i}{\beta_l}$ as $t \to \infty$, we can first show

$$\frac{Z_i^{r,m}(t)}{Z_l^{r,m}(t)} = \frac{Z_i^r \left(\frac{\sqrt{x_{r,m}} t}{N^r} + \frac{m}{\sqrt{N^r}} \right) - N_i^r}{Z_l^r \left(\frac{\sqrt{x_{r,m}} t}{N^r} + \frac{m}{\sqrt{N^r}} \right) - N_l^r} \to \frac{\beta_i}{\beta_l}, \text{ as } r \to \infty \text{ and } t \to \infty. \tag{4.5.11}$$

Since $\frac{N_i^r}{N_l^r} \to \frac{\beta_i}{\beta_l}$ as $r \to \infty$ and of course as $t \to \infty$, we only need to show

$$Z_i^r \left(\frac{\sqrt{x_{r,m}t}}{N^r} + \frac{m}{\sqrt{N^r}} \right) \Big/ Z_l^r \left(\frac{\sqrt{x_{r,m}t}}{N^r} + \frac{m}{\sqrt{N^r}} \right) \to \frac{\beta_i}{\beta_l}, \text{ as } r \to \infty \text{ and } t \to \infty. \quad (4.5.12)$$

From the discussion about fluid limits in the beginning,

$$\frac{\bar{Z}_i^r(t)}{\bar{Z}_l^r(t)} = \frac{Z_i^r(t)/N^r}{Z_l^r(t)/N^r} \to \frac{\beta_i}{\beta_l} \text{ u.o.c. in probability as } r \to \infty. \quad (4.5.13)$$

This means that the limit of $\frac{\bar{Z}_i^r(\cdot)}{\bar{Z}_l^r(\cdot)}$ does not change with time t, so, as long as $r \to \infty$, we have

$$Z_i^r \left(\frac{\sqrt{x_{r,m}t}}{N^r} + \frac{m}{\sqrt{N^r}} \right) \Big/ Z_l^r \left(\frac{\sqrt{x_{r,m}t}}{N^r} + \frac{m}{\sqrt{N^r}} \right) - \frac{\beta_i}{\beta_l} \to 0, \forall\, t \geq 0. \quad (4.5.14)$$

Then we have shown (4.5.12), hence (4.5.11). Therefore, as $r \to \infty$,

$$\frac{\tilde{Z}_i(t)}{\tilde{Z}_l(t)} = \left(\frac{\tilde{Z}_i(t)}{\tilde{Z}_l(t)} - \frac{Z_i^{r,m}(t)}{Z_l^{r,m}(t)} \right) + \frac{Z_i^{r,m}(t)}{Z_l^{r,m}(t)} \to 0 + \frac{\beta_i}{\beta_l} = \frac{\beta_i}{\beta_l}, \forall\, t \geq 0,$$

and the lemma is proved as required. $\qquad \square$

Now we have the SSC result for inverted-V systems.

Lemma 4.5.2 (SSC for inverted-V). *For the sequence of inverted-V systems mentioned above, there exists a continuous function $g : \mathbb{R}^{I+1} \to \mathbb{R}^+$, such that it satisfies Assumption 4.1 and 4.2 in [Dai and Tezcan, 2011], and if*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \text{ in probability, as } r \to \infty, \quad (4.5.15)$$

then, for each $T > 0$,

$$||g(\hat{Q}^r(t), \hat{Z}^r(t))||_T \to 0 \text{ in probability,} \quad (4.5.16)$$

as $r \to \infty$.

Proof. To prove the SSC result in the inverted-V systems, first we need to check that four assumptions from [Dai and Tezcan, 2011] are satisfied.

By (4.5.6), [Dai and Tezcan, 2011, Assumption 3.1] is satisfied and the static planning problem in that paper has a unique optimal solution with $x_i^* = 1$ for all $i = 1, 2, \ldots, I$. Since such systems are special cases of the systems in [Atar, 2008], the condition in Lemma 3.1(ii) from that paper should also be satisfied by these systems. Thus, for fixed $T > 0$,

$$\left\| r^{-1} T^{(i)}(t) - \rho_i t \right\|_T \to 0 \text{ in probability as } r \to \infty, \quad (4.5.17)$$

86

which can be rewritten as

$$\left\| \int_0^t \left(\frac{1}{r} \sum_{k \in K_i} \mu_k B_k(s) - \beta_i \mu^{(i)} \right) ds \right\|_T \to 0 \text{ in probability as } r \to \infty. \qquad (4.5.18)$$

Recall that there is no randomness among servers in the same pool, i.e. $\mu_k = \mu^{(i)}, \forall k \in K_i$. Therefore the convergence above can be simplified to

$$\left\| \int_0^t \left(\frac{1}{r} Z_i^r(s) - \beta_i \right) ds \right\|_T \to 0 \text{ in probability as } r \to \infty. \qquad (4.5.19)$$

Since we do not consider abandonments in this model, the convergence of the (scaled) queue length does not have to be taken into account, thus equation (4.5.19) implies that the second part of [Dai and Tezcan, 2011, Assumption 3.2] is satisfied, which is adequate for our analysis.

Now we can define the SSC function $g : \mathbb{R}^{I+1} \to \mathbb{R}$:

$$g(q, z_1, z_2, \ldots, z_I) = \sum_{i=1}^I z_i \mu^{(i)} - \sum_{i=1}^I z_i \gamma(I), \qquad (4.5.20)$$

where $\gamma(I) = \frac{\sum_{l=1}^I \beta_l (\mu^{(l)})^2}{\sum_{l=1}^I \beta_l \mu^{(l)}}$.

It is easy to see that g is continuous and $g(\alpha q, \alpha z_1, \alpha z_2, \alpha z_I) = \alpha g(q, z_1, z_2, \ldots, z_I)$ for all $(q, z_1, z_2, \ldots, z_I) \in \mathbb{R}^{I+1}$ and for all $0 \le \alpha \le 1$. Hence g satisfies Assumption 4.1 in [Dai and Tezcan, 2011]. Next, we show that g satisfies Assumption 4.2 in [Dai and Tezcan, 2011].

From (4.5.20), we have

$$g(\tilde{Q}(t), \tilde{Z}_1(t), \tilde{Z}_2(t), \ldots, \tilde{Z}_I(t)) = \sum_{i=1}^I \tilde{Z}_i(t) \mu^{(i)} - \sum_{i=1}^I \tilde{Z}_i(t) \gamma(I)$$

$$= \sum_{i=1}^I \tilde{Z}_i(t) \left(\mu^{(i)} - \frac{\sum_{l=1}^I \beta_l (\mu^{(l)})^2}{\sum_{l=1}^I \beta_l \mu^{(l)}} \right) = \frac{1}{\sum_{l=1}^I \beta_l \mu^{(l)}} \sum_{i=1}^I \tilde{Z}_i(t) \left(\mu^{(i)} \sum_{l=1}^I \beta_l \mu^{(l)} - \sum_{l=1}^I \beta_l (\mu^{(l)})^2 \right)$$

$$= \frac{1}{\sum_{i=1}^I \beta_i \mu^{(i)}} \sum_{i=1}^I \tilde{Z}_i(t) \left(\sum_{l=1}^I \left(\mu^{(i)} \beta_l \mu^{(l)} - \beta_l (\mu^{(l)})^2 \right) \right)$$

$$= \frac{1}{\sum_{l=1}^I \beta_l \mu^{(l)}} \sum_{i=1}^I \frac{1}{\mu^{(i)}} \tilde{Z}_i(t) \mu^{(i)} \left(\sum_{l=1}^I \left(\mu^{(i)} \beta_l \mu^{(l)} - \beta_l (\mu^{(l)})^2 \right) \right)$$

$$= \frac{1}{\sum_{l=1}^I \beta_l \mu^{(l)}} \sum_{i=1}^I \sum_{l=1}^I \frac{1}{\mu^{(i)}} \left(\tilde{Z}_i(t) \mu^{(i)} \beta_l \mu^{(l)} \left(\mu^{(i)} - \mu^{(l)} \right) \right)$$

$$= \frac{1}{\sum_{l=1}^I \beta_l \mu^{(l)}} \sum_{i=1}^I \sum_{l>i}^I \frac{1}{\mu^{(i)}} \left(\tilde{Z}_i(t) \mu^{(i)} \beta_l \mu^{(l)} \left(\mu^{(i)} - \mu^{(l)} \right) + \tilde{Z}_l(t) \mu^{(l)} \beta_i \mu^{(i)} \left(\mu^{(l)} - \mu^{(i)} \right) \right)$$

$$= \frac{1}{\sum_{l=1}^I \beta_l \mu^{(l)}} \sum_{i=1}^I \sum_{l>i}^I \frac{1}{\mu^{(i)}} \left(\mu^{(i)} \mu^{(l)} \left(\mu^{(i)} - \mu^{(l)} \right) \left(\tilde{Z}_i(t) \beta_l - \tilde{Z}_l(t) \beta_i \right) \right). \qquad (4.5.21)$$

By Lemma 4.5.1, we know that (4.5.21) $\to 0$ as $t \to \infty$, thus (4.5.20) is a valid SSC function. So far all of the four assumptions in [Dai and Tezcan, 2011] are satisfied. However, we can

only derive the multiplicative SSC result of Theorem 4.1 under current conditions. What we really need is the strong SSC. Thus by Remark 4.2 in that paper, we need to check the so-called compact containment condition for $\hat{Z}^r$ (We do not need to worry about this condition for $\hat{Q}^r$ because the new hydrodynamic scaling factor (4.5.7) in this case does not include $\hat{Q}^r$).

The compact containment condition for $\{\hat{Z}^r(\cdot)\}$ is defined as

$$\lim_{K \to \infty} \limsup_{r \to \infty} \mathbb{P}(||\hat{Z}^r(t)||_T > K) = 0, \forall\, T > 0. \tag{4.5.22}$$

If $\{||\hat{Z}^r(t)||_T\}$ is tight, then obviously (4.5.22) holds, hence we only need to show tightness of $\{||\hat{Z}^r(t)||_T\}$. We will prove this by contradiction.

Fix $T > 0$. If $\{||\hat{Z}^r(t)||_T\}$ is not tight, then, $\forall r$, $\exists \epsilon > 0$ such that $\forall K > 0$, we always have

$$\mathbb{P}\left(||\hat{Z}^r(t)||_T > K\right) = C_K \geq \epsilon > 0, \tag{4.5.23}$$

i.e.

$$\mathbb{P}\left(\max_i \left\|\frac{Z_i^r(t) - N_i^r}{\sqrt{|N^r|}}\right\|_T > K\right) = C_K \geq \epsilon > 0. \tag{4.5.24}$$

However by (4.5.19), $\forall \delta > 0$,

$$\lim_{r \to \infty} \mathbb{P}\left(\left\|\frac{1}{r} Z_i^r(t) - \beta_i\right\|_T > \delta\right) = 0. \tag{4.5.25}$$

Using this and (4.5.4), and (4.5.5),

$$\lim_{r \to \infty} \mathbb{P}\left(\left\|\frac{Z_i^r(t)}{N_i^r} - 1\right\|_T > \delta\right) = 0, \tag{4.5.26}$$

which contradicts (4.5.24). Therefore $\{||\hat{Z}^r(t)||_T\}$ is indeed tight, so the condition (4.5.22) is satisfied. Thus by Theorem 4.1 and Remark 4.2 from [Dai and Tezcan, 2011], (4.5.16) holds, i.e. for any $T > 0$

$$\left\|\sum_{i=1}^{I} \hat{Z}_i^r(t)\left(\mu^{(i)} - \gamma(I)\right)\right\|_T \to 0 \text{ in probability as } r \to \infty. \tag{4.5.27}$$

$\square$

Now that we have shown SSC for the inverted-V systems, we can prove Atar's result, i.e. Theorem 2.3.1.

Proof of Theorem 2.3.1. All we need to show is

$$\left\|F^r(t) - \gamma \int_0^t \hat{I}^r(s)ds\right\|_T \to 0 \text{ as } r \to \infty \text{ for any } T > 0. \tag{4.5.28}$$

Since

$$F^r(t) - \gamma \int_0^t \hat{I}^r(s)ds = F^r(t) - \gamma(I) \int_0^t \hat{I}^r(s)ds + \gamma(I) \int_0^t \hat{I}^r(s)ds - \gamma \int_0^t \hat{I}^r(s). \quad (4.5.29)$$

The first difference from this equation is

$$
\begin{aligned}
F^r(t) - \gamma(I) \int_0^t \hat{I}^r(s)ds &= \int_0^t \sum_{k=1}^{N^r} \mu_k \hat{I}_k^r(s)ds - \gamma(I) \int_0^t \hat{I}^r(s)ds \\
&= \int_0^t \sum_{i=1}^{I} \sum_{k \in K_i} \left(\mu_k \hat{I}_k(s) - \gamma(I) \hat{I}_k^r(s) \right) ds \\
&= \int_0^t \sum_{i=1}^{I} \sum_{k \in K_i} \left(\mu_k - \mu^{(i)} + \mu^{(i)} - \gamma(I) \right) \hat{I}_k^r(s)ds \\
&= \int_0^t \sum_{i=1}^{I} \sum_{k \in K_i} \left(\mu_k - \mu^{(i)} \right) \hat{I}_k^r(s)ds + \int_0^t \sum_{i=1}^{I} \sum_{k \in K_i} \left(\mu^{(i)} - \gamma(I) \right) \hat{I}_k^r(s)ds.
\end{aligned}
$$
$$(4.5.30)$$

The first integral in (4.5.30) converges to zero u.o.c as $r \to \infty$, which is a result of [Atar, 2008, Lemma 3.1 (iii)] - actually being e_1 in the lemma. The second integral is

$$
\begin{aligned}
\int_0^t \sum_{i=1}^{I} \sum_{k \in K_i} \left(\mu^{(i)} - \gamma(I) \right) \hat{I}_k^r(s)ds &= \int_0^t \sum_{i=1}^{I} \left(\mu^{(i)} - \gamma(I) \right) \hat{I}^{r,(i)}(s)ds \\
&= \int_0^t \sum_{i=1}^{I} \left(\mu^{(i)} - \gamma(I) \right) \frac{1}{\sqrt{r}}(N_i^r - Z_i^r(s))ds = \int_0^t \sum_{i=1}^{I} \left(\mu^{(i)} - \gamma(I) \right) (-\hat{Z}_i^r(s))ds. \quad (4.5.31)
\end{aligned}
$$

By Lemma 4.5.2 and equation (4.5.27), (4.5.31) converges to zero u.o.c as $r \to \infty$. Hence (4.5.30) converges to zero u.o.c as $r \to \infty$.

For the latter half of equation (4.5.29), since $\{\mu^{(i)}\}$ are chosen in a way such that $0 < \mu^{(i)} - \mu^{(i-1)} \le \epsilon, \forall i = 1, 2, \ldots, I$, and ϵ is arbitrary, we can let $\epsilon \to 0$, and then $\gamma(I) \to \gamma$ (remember $\gamma = \frac{\int x^2 dm}{\int x dm}$). Thus the second half of (4.5.29) also converges to zero u.o.c. Therefore,

$$\left\| F^r(t) - \gamma \int_0^t \hat{I}^r(s)ds \right\|_T \to 0, \text{ as } r \to \infty, \quad (4.5.32)$$

and since $\hat{I}^r(t) = \frac{1}{\sqrt{r}}(X^r(t) - N^r)^- = \hat{X}^r(t)^-$, we have

$$\left\| F^r(t) - \gamma \int_0^t \hat{X}^r(s)^- ds \right\|_T \to 0, \text{ as } r \to \infty. \quad (4.5.33)$$

Combining with (4.5.1), we have $\hat{X}^r(t)$ weakly converge to the solution of SDE

$$\xi(t) = \xi(0) + \sigma w(t) + \beta t + \gamma \int_0^t \xi(s)^- ds, t \ge 0, \quad (4.5.34)$$

where all of the coefficients are as in Theorem 2.3.1. $\qquad\square$

Chapter 5

Conclusion and Future Work

In this work, we study queueing systems with heterogeneous servers and parameter uncertainty. In particular, we consider these systems under the *Quality-Efficiency Driven* regime and heavy traffic condition. Our work is motivated by call centres where agents' service speeds may be influenced by their environment, and thus it is uncertain and unknown prior to the operation of the system. We assume service rates to be i.i.d. random variables. Their realisations are given at time zero and are kept fixed during the running of the system. Although it is modelled on call centres, our results are generic and can be applied to other domains such as healthcare, computer science, instant messaging service, among others.

We start by deriving diffusion limits for many server queues under the LISF policy with random service rates and abandonments. [Atar, 2008] shows diffusion limits for such systems without abandonments. Then we obtain our results by extending his result to systems with abandonments, and use a martingale method to prove this. Unlike the constant drift of diffusion limits for identical server systems, our diffusion limits have a normal random drift which results from the randomness of the service rates. For systems without abandonments, the existence of such a random drift means that whether the diffusion has steady states depends on the value of the drift. From the perspective of the system, since the arrival rate is fixed, random service rates may cause the system to be unstable and the queue length may become unbounded. However, for systems with abandonments, even if it has random service rates it is always stable because of the abandonment process.

Then, we formulate an optimisation problem for the staffing for such systems. Staffing has always been of significance in call centre management since an unwise decision about it will cost companies immensely. For our systems, it is especially important because we need to take the randomness of servers into account. For systems without abandonments, we simplify the problem by assigning a fixed cost to unstable systems, and focus on the cost for stable systems. For stable systems, we consider both staffing costs and holding costs. The holding cost involves the expected queue length of steady state, thus we use the diffusion limits derived previously

to establish a continuous approximation of this cost. In order to show the validity of this estimation, we prove the tightness of the steady state, by which we can show interchangeability of the limits. For systems with abandonments, the systems are always stable and we also simplify the problem by neglecting the holding cost.

Finally, we show the state space collapse results for systems with random service rates. SSC is an important phenomenon because it reduces dimensions of the processes and significantly simplifies analysis. In particular, it is the central part in proving diffusion limits. Our work is based on the model developed by [Dai and Tezcan, 2011]. They show a generic SSC result for queueing networks with multi-class customers and skilled based parallel server pools. We generalise their results to systems with random service rates inside each pool. To show the SSC in such systems, the challenging part is to show the departure processes are almost Lipschitz. We cannot show it directly as in [Dai and Tezcan, 2011], thus we use a coupling method by splitting Poisson processes with the maximum rate. We later use the SSC result from [Dai and Tezcan, 2011] to show the diffusion limit in [Atar, 2008] again. By using this approach, we identify that such a SSC result can also be applied in systems with random service rates and thus gain more insight into this phenomenon. And since [Atar, 2008] uses a method particular to this model, this new approach also indicates that we can use a more general way to prove the limit theorem .

Queueing systems with random service rates exhibit many interesting properties, and we provide a few promising future research directions as follows. In our work, only the LISF routing policy is considered. One can consider other blind policies such as the random routing, where customers are routed to idle servers randomly, or the longest accumulated idle server first, where customers are routed to the idle server which has the longest accumulated idle time. Moreover, we can investigate the optimal routing and scheduling. In Section 3.3, we use a numerical method to analyse how the abandonment rate is influenced by the variation of service rates. We only consider the situation where random service rates are uniformly distributed. We can analyse other distributions as well as the theoretical proof of this result. In Section 3.4, we try to establish a fairness measure for general policies. We can continue this work and develop a general method for proving diffusion limits under different policies.

Appendix A

Proofs in Chapter 2

A.1   Proposition 3.1 of Atar

This proposition shows how Atar partitions servers such that the systems before and after pooling are equivalent in distribution. For simplicity, we will omit superscript r from the notation of all random variables and stochastic processes throughout this appendix. The deterministic parameters that depend on r will still keep the r in their notation.

Proposition A.1.1. *([Atar, 2008, Proposition 3.1]) Fix $r \in \mathbb{N}$. Let $(K_1, \ldots, K_I)$ be a partition of $1, \ldots, N$ that measurable on $\sigma\{N, \{\mu_k\}\}.Let\{S^{(1)}, \ldots, S^{(I)}\}$ be independent standard Poisson processes. For each $i = 1, \ldots, I$ and for each nonempty subset Θ of K_i, let $\{e(i, L, l), l \in \mathbb{N}\}$ be a sequence of i.i.d. random variables distributed uniformly on Θ, independent across i and Θ. Assume also that the four random objects $(N, \{\mu_k\}, X(0), \{B_k(0)\}), a, \{S^{(i)}\}$ and $\{e(i, \Theta, l)\}$ are mutually independent. Define*

$$D^{(i)}(t) = S^{(i)}(T^{(i)}(t)), i = 1, \ldots, I, \tag{A.1.1}$$

where

$$T^{(i)}(t) = \sum_{k \in K_i} T_k(t), i = 1, \ldots, I, \tag{A.1.2}$$

and consider

$$D_k(t) = \sum_{s \in (0,t]: \Delta D^{(i)}(s)=1} \mathbb{1}_{\{e(i, \{p \in K_i: B_p(s-)=1\}, D^{(i)}(s))=k\}}, \quad k \in K_i, i = 1, \ldots, I \tag{A.1.3}$$

as a substitute for equation (2.3.6). Then the process Σ', defined analogously to Σ, with (A.1.1)-(A.1.3) in place of (2.3.6), is equal in law to Σ.

A.2 Proof of Theorem2.3.1 by Atar

Proof of Theorem 2.3.1. We rephrase the method of [Atar, 2008]'s proof here.

From (2.3.1) and (2.3.2), we have

$$\hat{X}(t) = \frac{1}{\sqrt{r}}(X(0) - N) + \frac{1}{\sqrt{r}}A(t) - \frac{1}{\sqrt{r}}\sum_{k=1}^{N} D_k(t)$$

$$= \hat{X}(0) + \frac{1}{\sqrt{r}}(A(t) - \lambda^r t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{k=1}^{N} D_k(t). \tag{A.2.1}$$

Denote

$$\hat{A}(t) = \frac{1}{\sqrt{r}}(A(t) - \lambda^r t). \tag{A.2.2}$$

Proposition 3.1 in [Atar, 2008] provides a way to consider departure processes $D_k(t)$ aggregately instead of individually. In order to apply [Atar, 2008, Proposition 3.1] to $\sum_{k=1}^{N} D_k(t)$, first we need to define some notations.

Let ϵ be given and let $I \in \mathbb{N}$ and $\mu^{(i)} \in \mathbb{R}_+$, $i = 1, \ldots, I$ satisfying the following conditions:

- $\mu^{(1)} = 0, \mu^{(I)} \geq 1$,

- $0 < \mu^{(i)} - \mu^{(i-1)} \leq \epsilon, i = 1, \ldots, I$,

- $\int_{[\mu^{(I)}, \infty)} x^2 dm \leq \epsilon$,

- for $i = 2, \ldots, I$, $\mu^{(i)}$ is a continuity point of $x \mapsto m([0, x]) \equiv P(\mu_k^2 \leq x)$.

Set $\mu^{(I+1)} = \infty$, and

$$K_i = \{k \in \{1, 2, \ldots, N^r\} : \mu_k \in [\mu^{(i)}, \mu^{(i+1)})\}, i = 1, 2, \ldots, I. \tag{A.2.3}$$

Let $\{S^{(1)}, \ldots, S^{(q)}\}$ be independent standard Poisson processes. Now we are ready to use Proposition 3.1.

$$\sum_{k=1}^{N} D_k(t) = \sum_{i=1}^{I} D^{(i)}(t) = \sum_{i=1}^{I} S^{(i)}(T^{(i)}(t)), \tag{A.2.4}$$

where

$$T^{(i)}(t) = \sum_{k \in K_i} \mu_k T_k(t). \tag{A.2.5}$$

Hence, by (A.2.2) and (A.2.4), (A.2.1) can be rewritten as

$$\hat{X}(t) = \hat{X}(0) + \hat{A}(t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{i=1}^{I} D^{(i)}(t)$$

$$= \hat{X}(0) + \hat{A}(t) + \frac{1}{\sqrt{r}}\lambda^r t - \frac{1}{\sqrt{r}}\sum_{i=1}^{I} \left(S^{(i)}(T^{(i)}(t)) - T^{(i)}(t) + T^{(i)}(t)\right). \tag{A.2.6}$$

Define $W^{(i)}(t) = S^{(i)}\left(T^{(i)}\right) - T^{(i)}(t)$. Then (A.2.6) becomes

$$\hat{X}(t) = \hat{X}(0) + \hat{A}(t) + \frac{\lambda^r t}{\sqrt{r}} - \sum_{i=1}^{I} \left(W^{(i)}(t) - \frac{T^{(i)}(t)}{\sqrt{r}} \right)$$

$$= \hat{X}_0 + \hat{A}(t) + \frac{\lambda^r t}{\sqrt{r}} - \sum_{i=1}^{I} \left(W^{(i)}(t) - \frac{T^{(i)}(t) - \sum_{k \in K_i} \mu_k t}{\sqrt{r}} \right) + \sum_{k=1}^{N^r} \frac{\mu_k}{\sqrt{r}} t. \qquad \text{(A.2.7)}$$

Let $F^{(i)}(t) = -\frac{T^{(i)}(t) - \sum_{k \in K_i} \mu_k t}{\sqrt{r}}$. Then (A.2.6) is

$$\hat{X}(t) = \hat{X}(0) + \hat{A}(t) + \frac{\lambda^r t}{\sqrt{r}} - \sum_{i=1}^{I} \left(W^{(i)}(t) + F^{(i)}(t) \right) + \sum_{k=1}^{N} \frac{\mu_k}{\sqrt{r}} t. \qquad \text{(A.2.8)}$$

To see this more clearly, let $W(t) = \hat{A}(t) - \sum_{i=1}^{I} W^{(i)}(t)$ so (A.2.8) is

$$\hat{X}(t) = \hat{X}(0) + W(t) + \frac{\lambda^r t}{\sqrt{r}} + \sum_{i=1}^{I} F^{(i)}(t) + \sum_{k=1}^{N} \frac{\mu_k}{\sqrt{r}} t$$

$$= \hat{X}(0) + W(t) + \frac{\lambda^r t}{\sqrt{r}} + \sum_{i=1}^{I} F^{(i)}(t) + \sum_{k=1}^{N} \frac{\mu_k}{\sqrt{r}} t. \qquad \text{(A.2.9)}$$

Recall that $\lambda = \bar{\mu}$, thus

$$\hat{X}(t) = \hat{X}(0) + W(t) + \frac{\lambda^r t - N\lambda t - r\lambda t + r\lambda t}{\sqrt{r}} + \sum_{i=1}^{I} F^{(i)}(t) + \sum_{k=1}^{N} \frac{\mu_k - \bar{\mu}}{\sqrt{r}} t$$

$$= \hat{X}(0) + W(t) + \frac{\lambda^r t - r\lambda t}{\sqrt{r}} + \sum_{i=1}^{I} F^{(i)}(t) + \sum_{k=1}^{N} \left(\frac{\mu_k - \bar{\mu}}{\sqrt{r}} t \right) + \bar{\mu} \frac{N - r}{\sqrt{r}} t$$

$$= \hat{X}(0) + W(t) + \frac{\lambda^r t - r\lambda t}{\sqrt{r}} + \sum_{i=1}^{I} F^{(i)}(t) + \sum_{k=1}^{N} \left(\frac{\mu_k - \mu}{\sqrt{r}} t \right) + \bar{\mu} \hat{N} t.$$

Finally define $b^r = \frac{\lambda^r t - r\lambda}{\sqrt{r}} + \sum_{k=1}^{N} \left(\frac{\mu_k - \bar{\mu}}{\sqrt{r}} \right) + \bar{\mu} \hat{N}$. Thus

$$\hat{X}(t) = \hat{X}(0) + W(t) + \sum_{i=1}^{I} F^{(i)}(t) + b^r t. \qquad \text{(A.2.10)}$$

$\sum_{i=1}^{I} F^{(i)}(t)$ is a troublesome component which needs our particular attention. For this reason, we define

$$F(t) = \sum_{i=1}^{I} F^{(i)}(t) = \sum_{i=1}^{I} \frac{\sum_{k \in K_i} \mu_k t - T^{(i)}(t)}{\sqrt{r}}$$

$$= \sum_{k=1}^{N} \frac{\mu_k t - T_k(t)}{\sqrt{r}} = \frac{\int_0^t \sum_{k=1}^{N} \mu_k I_k(s) ds}{\sqrt{r}}.$$

Now define $I^{(i)}(t)$ to be the number of idle servers in pool i. We can write

$$I^{(i)}(t-) = D^{(i)}(t-) - D^{(i)}(H(t)-) + e_0^{(i)}(t),$$

where $H(t)$ is defined in (2.16) in [Atar, 2008] to be the time at which the longest idle server at t became idle. So, the difference of $D_k(t-) - D_k(H(t)-)$ can be 1 or 0 based on whether server k is idle at time $t-$. Also, e_0 accounts for the fact that servers can be idle from 0 to t. This uses the fact that we are routing customers to the longest idle server first. If we use another policy, this will be the key point in proving SSC.

The inequality below follows because some of the servers idle at 0 can start serving:

$$\sum_{i=1}^{I} \hat{e}_0^i(t) = \frac{\sum_{i=1}^{I} e_0^i(t)}{\sqrt{r}} \leq \frac{I(0)I_{\{H(t)=0\}}}{\sqrt{r}} = \hat{X}(0)^-(t)I_{\{H(t)=0\}}.$$

Now, we will do the following manipulation:

$$
\begin{aligned}
\hat{I}^{(i)}(t-) &= \frac{I^{(i)}(t-)}{\sqrt{r}} \\
&= \frac{D^{(i)}(t-) - D^{(i)}(H(t)-) + e_0^{(i)}(t)}{\sqrt{r}} \\
&= \frac{S^{(i)}(T^{(i)}(t-)) - S^{(i)}(T^{(i)}(H(t)-)) + e_0^{(i)}(t)}{\sqrt{r}} \\
&= \frac{S^{(i)}(T^{(i)}(t-)) - T^{(i)}(t-) + T^{(i)}(t-) - S^{(i)}(T^{(i)}(H(t)-)) - T^{(i)}(H(t)-) + T^{(i)}(H(t)-)}{\sqrt{r}} \\
&\quad + \hat{e}_0^{(i)}(t) \\
&= W^{(i)}(t-) - W^{(i)}(H(t)-) + \frac{T^{(i)}(t-) - T^{(i)}(H(t)-)}{\sqrt{r}} + \hat{e}_0^{(i)}(t).
\end{aligned}
$$

Recall $h(t) = t - H(t)$. Hence

$$\hat{I}^{(i)}(t-) = W^{(i)}(t-) - W^{(i)}(H(t)-) + \frac{T^{(i)}(t-) - T^{(i)}(H(t)-)}{\sqrt{r}} + \frac{\mu^{(i)}N^{(i)}h(t)}{\sqrt{r}} - \frac{\mu^{(i)}N^{(i)}h(t)}{\sqrt{r}} + \hat{e}_0^{(i)}(t).$$

$$(\text{A.2.11})$$

To manipulate this even further, we need to realize

$$T^{(i)}(t-) - T^{(i)}(H(t)-) = \sum_{k \in K_i} \mu_k \int_{H(t)}^{t} B(s)ds,$$

$$N^{(i)}h(t) = \int_{H(t)}^{t} (I(s) + B(s))ds.$$

Now, substituting these into (A.2.11),

$$\hat{I}^{(i)}(t-) = W^{(i)}(t-) - W^{(i)}(H(t)-) + \sum_{k \in K_i} \left(\left(\mu_k - \mu^{(i)}\right) \int_{H(t)}^{t} \hat{B}(s)ds - \mu^{(i)} \int_{H(t)}^{t} \hat{I}(s)ds \right)$$

$$+ \frac{\mu^{(i)} N^{(i)} h(t)}{\sqrt{r}} + \hat{e}_0^{(i)}(t),$$

and aggregating terms gives us

$$\hat{I}^{(i)}(t-) = E^{(i)}(t) + \frac{\mu^{(i)} N^{(i)} h(t)}{\sqrt{r}}. \qquad (A.2.12)$$

Now, we note the following relation

$$h(t) = \sqrt{r} \frac{\sum_{i=1}^{I} \hat{I}^{(i)}(t-) - \sum_{i=1}^{I} E^{(i)}(t)}{\sum_{i=1}^{I} \mu^{i} N^{(i)}},$$

then use this to write

$$\hat{I}^{(i)}(t) = E^{(i)}(t) + \frac{\mu^{(i)} N^{(i)}}{\sum_{i=1}^{I} \mu^{(i)} N^{(i)}} \left(\sum_{i=1}^{I} \hat{I}^{(i)}(t-) - \sum_{i=1}^{I} E^{(i)}(t) \right). \qquad (A.2.13)$$

Now, let us try to rewrite $F(t)$:

$$\begin{aligned}
F(t) &= \int_0^t \sum_{k=1}^{N} \mu_k \hat{I}_k(s) ds \\
&= \int_0^t \left(\sum_{k=1}^{N} \mu_k \hat{I}_k(s) - \sum_{i=1}^{I} \mu^{(i)} \hat{I}^{(i)}(s) + \sum_{i=1}^{I} \mu^{(i)} \hat{I}^{(i)}(s) \right) ds \\
&= \int_0^t \left(\sum_{i=1}^{I} \sum_{k \in K_i} (\mu_k - \mu^{(i)}) \hat{I}_k(s) + \sum_{i=1}^{I} \mu^{(i)} \hat{I}^{(i)}(s) \right) ds. \qquad (A.2.14)
\end{aligned}$$

We define the four e_is of [Atar, 2008, Lemma 3.1 (iii)] as follows.

Let $e_1(t) = \sum_{i=1}^{I} \sum_{k \in K^i} (\mu_k - \mu^{(i)}) \int_0^t \hat{I}_k(s) ds$. Then (A.2.14) is equal to

$$F(t) = e_1(t) + \sum_{i=1}^{I} \mu^{(i)} \int_0^t \hat{I}^{(i)}(s) ds.$$

Use (A.2.13), we have

$$F(t) = e_1(t) + \sum_{i=1}^{I} \mu^{(i)} \int_0^t \left(E^{(i)}(s) + \frac{\mu^{(i)} N^{(i)}}{\sum_{i=1}^{I} \mu^{(i)} N^{(i)}} \left(\sum_{i=1}^{I} \hat{I}^{(i)}(s-) - \sum_{i=1}^{I} E^{(i)}(s) \right) \right) ds.$$

Define $e_2(t) = \sum_{i=1}^{I} \mu^{(i)} \int_0^t E^{(i)}(s) ds$. Then

$$F(t) = e_1(t) + e_2(t) + \sum_{i=1}^{I} \int_0^t \left(\frac{(\mu^{(i)})^2 N^{(i)}}{\sum_{i=1}^{I} \mu^{(i)} N^{(i)}} \left(\sum_{i=1}^{I} \hat{I}^{(i)}(s-) - \sum_{i=1}^{I} E^{(i)}(s) \right) \right) ds.$$

Further define $e_3(t) = \frac{\sum_{i=1}^{I}(\mu^{(i)})^2 N^{(i)}}{\sum_{i=1}^{I}\mu^{(i)}N^{(i)}} \sum_{i=1}^{I}\int_0^t E^{(i)}(s)ds$, so

$$F(t) = e_1(t) + e_2(t) + e_3(t) + \sum_{i=1}^{I}\int_0^t \left(\frac{(\mu^{(i)})^2 N^{(i)}}{\sum_{i=1}^{I}\mu^i N^{(i)}} \left(\sum_{i=1}^{I}\hat{I}^{(i)}(s-) \right) \right) ds.$$

Remember that $\hat{I} = \sum \hat{I}^{(i)}$ and add and subtract $\gamma \int_0^t \hat{I}(s)ds$ to get

$$F(t) = e_1(t) + e_2(t) + e_3(t) + e_4(t) + \gamma \int_0^t \hat{I}(s)ds.$$

Notice that $e_4(t) = \left(\frac{\sum_{i=1}^{I}(\mu^{(i)})^2 N^{(i)}}{\sum_{i=1}^{I}\mu^{(i)}N^{(i)}} - \gamma \right) \int_0^t \hat{I}(s)ds$, where $\gamma = E(\mu_1^2)/E(\mu_1)$.

Summarising all of these gives

$$\hat{X}(t) = \hat{X}(0) + W(t) + bt + \gamma \int_0^t \hat{X}(s)^- ds + \sum_{i=1}^{4} e_i(t).$$

First, we use Gronwall's inequality:

$$||\hat{X} - \xi||_t \leq \left(|\hat{X}(0) - \xi(0)| + |b - \beta| + ||W - \sigma w||_t + ||e||_t \right) \exp(\gamma t). \tag{A.2.15}$$

Now we need to analyse these terms separately:

1. $\hat{X}(0)$ is simply the usual Central Limit Theorem applied to the initial random variables.

2. $W(t) = \hat{A}(t) - \sum_{i=1}^{I} W^{(i)}(t)$

   (a) $\hat{A}(t) \Rightarrow B_1(t)$, where $B_1(t)$ is a Brownian Motion with 0 drift and diffusion coefficient $\sqrt{\lambda}C_{\check{U}}$ using the Functional Central Limit Theorem for renewal processes.

   (b) $W^{(i)}(t) = S^{(i)}\left(T^{(i)}(t)\right) - T^{(i)}(t)$. By [Atar, 2008, Lemma 3.1(ii)], we have $r^{-1}T^{(i)} \to \rho_i t$, where $\rho_i = \int_{\mu^{(i)}}^{\mu^{(i+1)}} x dm$. Again using the FCLT for renewal processes, $S^{(i)}(t) - t$ converges to a standard Brownian motion. Then replace $T^{(i)}(t)$ with t. By the random change of time, we show that $W^{(i)}(t)$ weakly converges to Brownian motion $B_2(t)$ with zero mean and diffusion coefficient $\sqrt{\rho_i}$.

   Summarising (a) and (b), we show that $W(t)$ converges weakly to σw, where w is a standard Brownian motion and $\sigma^2 = \lambda C_{\check{U}}^2 + \mu$.

3. For $j = 1, 2, 3, 4$, $||e_j||_\theta \to 0$ in probability as $r \to \infty$. This is shown in [Atar, 2008, Lemma 3.1 (iii)].

By steps 1, 2 and 3, the right hand side of equation (A.2.15) converges to zero uniformly on a compact set, thus $\hat{X}(t) \Rightarrow \xi(t)$ as required.

$\square$

Appendix B

Proofs in Chapter 4

B.1 Proof of fluid limits in Theorem 4.3.1

Proof of Theorem 4.3.1. The proof is similar to that in Theorem B.1 in [Dai and Tezcan, 2011]. Specifically, the proof of the precompactness of $\bar{A}, \bar{A}_q, \bar{A}_s, \bar{Q}$, and $\bar{C}$ is the same as in Theorem B.1 in [Dai and Tezcan, 2011], and equations (4.3.7), (4.3.8), (4.3.12), (4.3.13), and (4.3.14) are also proved in the same paper, so in the rest of the proof we use these results directly and focus on the precompactness of $\bar{Z}, \bar{T}, \bar{I}$ and equations (4.3.9), (4.3.10), and (4.3.11) here.

First we prove the precompactness. Assume (4.2.18) and (4.2.20) hold. With a slight abuse of notation, consider a sequence of numbers that is denoted as $\{r\}$. We show that $\{\bar{\mathbb{X}}^r(\cdot, \omega)\}$ has a convergent subsequence, for all $\omega \in \mathscr{A}$. Fix ω in the rest of the proof. One can observe that

$$\left| \frac{T^r(t_2, \omega)}{N^r} - \frac{T^r(t_1, \omega)}{N^r} \right| \le |t_2 - t_1|,$$

for all $0 \le t_1 \le t_2$. Hence $\{\bar{T}^r(\cdot, \omega)\}$ is uniformly bounded and uniformly continuous, which means, by [Billingsley, 1999, Theorem 12.3], that there exists a subsequence $\{r_l\}$ such that $\bar{T}^{r_l}(\cdot, \omega)$ converges u.o.c. to some continuous function $\bar{T}$.

We define the fluid scaled total idle process for the ith server pool by

$$\bar{I}_i^r(t) = \frac{N_i^r}{|N^r|}t - \sum_{j \in \mathcal{J}(i)} \bar{T}_{ij}^r(t). \tag{B.1.1}$$

Then obviously $\bar{I}_i^{r_l}(\cdot)$ is precompact.

For departure processes, we need to treat each server individually. Since

$$\bar{D}_{ijk}^{r_l}(t) = \frac{1}{|N^r|} S_{ijk}(\mu_{ijk}|N^r|\bar{T}_{ijk}^{r_l}(t)),$$

99

we have

$$\bar{D}_{ij}^{r_l}(t) = \sum_{k=1}^{N_i^r} \bar{D}_{ijk}^{r_l}(t) = \frac{1}{|N^r|} \sum_{k=1}^{N_i^r} S_{ijk}(|N^r|\mu_{ijk}\bar{T}_{ijk}^{r_l}(t)) \stackrel{d}{=} \frac{1}{|N^r|} S\left(|N^r|\sum_{k=1}^{N_i^r} \mu_{ijk}\bar{T}_{ijk}^{r_l}(t)\right)$$

where S is a standard Poisson process. $\stackrel{d}{=}$ means equal in distribution and it is a basic statement of the superposition of Poisson processes.

To see the convergence of $\bar{D}_{ij}^{r_l}(t)$, fix $L > 0$, and consider

$$\left| \sum_{k=1}^{N_i^r} \mu_{ijk}\bar{T}_{ijk}^{r_l}(t) - \bar{\mu}_{ij}\bar{T}_{ij}(t) \right|$$

$$= \left| \sum_{k=1}^{N_i^r} \left(\mu_{ijk}\bar{T}_{ijk}^{r_l}(t) - \bar{\mu}_{ij}\bar{T}_{ijk}^{r_l}(t) \right) + \sum_{k=1}^{N_i^r} \left(\bar{\mu}_{ij}\bar{T}_{ijk}^{r_l}(t) - \bar{\mu}_{ij}\bar{T}_{ij}(t)\frac{1}{N_i^r} \right) \right|$$

$$= \left| \frac{1}{|N^r|} \sum_{k=1}^{N_i^r} (\mu_{ijk} - \bar{\mu}_{ij})\,T_{ijk}^{r_l}(t) + \bar{\mu}_{ij} \sum_{k=1}^{N_i^r} \left(\bar{T}_{ijk}^{r_l}(t) - \bar{T}_{ij}(t)\frac{1}{N_i^r} \right) \right|$$

$$\leq \left| \frac{N_i^r}{|N^r|} \frac{1}{N_i^r} \sum_{k=1}^{N_i^r} (\mu_{ijk} - \bar{\mu}_{ij})\,t \right| + \left| \bar{\mu}_{ij} \left(\bar{T}_{ij}^{r_l}(t) - \bar{T}_{ij}(t) \right) \right|. \tag{B.1.2}$$

By the Law of Large Numbers, the first term in (B.1.2) converges u.o.c. to 0, and the second term also converges u.o.c. to 0 as proved already. This means $\sum_{k=1}^{N_i^r} \mu_{ijk}\bar{T}_{ijk}^{r_l}(\cdot)$ converges u.o.c. to $\bar{\mu}_{ij}\bar{T}_{ij}(\cdot)$. Then using [Ata et al., 2005, Lemma 11] and the Functional Strong Law of Large Numbers we can get

$$\bar{D}_{ij}^{r_l}(\cdot) \text{ converges u.o.c. to } \bar{D}_{ij}(\cdot), \tag{B.1.3}$$

where $\bar{D}_{ij}(t) = \bar{\mu}_{ij}\bar{T}(t)$. By the precompactness of $\bar{A}_s^{r_l}, \bar{C}^{r_l}$ and (B.1.3), and the process equation (4.2.6), $\bar{Z}^{r_l}(\cdot, \omega)$ is precompact.

Next we show that every fluid limit satisfies (4.3.9)-(4.3.11). (4.3.10)-(4.3.11) are trivial. For (4.3.10), let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that

$$\bar{\mathbb{X}}^r(\cdot, \omega) \to \bar{\mathbb{X}} \quad \text{u.o.c. as } r \to \infty \text{ for some } \omega \in \mathscr{A}. \tag{B.1.4}$$

Then equation (4.3.9) follows from (B.1.3), the convergence of $\bar{Z}^r(0, \omega)$, $\bar{A}_s^r(\cdot, \omega)$, and $\bar{C}^r(\cdot, \omega)$.

$\square$

B.2 Proof of (4.4.18)

Proof of (4.4.18). By setting $\epsilon = 1, t_2 = L$ and $t_1 = 0$ in (4.4.32), and since $D_{ij}^r(t)$ has the same probability distribution as $\check{D}_{ij}^r(t)$, we have that

$$P\left\{D_{ij}^r\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right) \geq 2NL\sqrt{x_{r,0}}\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{B.2.1}$$

Then for r large enough,

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

Also notice that $D_{ijk}^r(t) \leq D_{ij}^r(t)$ for all $t \geq 0$ and all $a = 1, 2, \ldots, |N^r|$, we have that

$$D_{ijk}^r\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right) + 1 \leq 3NL\sqrt{x_{r,0}}. \tag{B.2.2}$$

Let $e = 0$ or 1. It follows from [Bramson, 1998, Proposition 4.2] that, for large enough n,

$$P\left\{\left\|V_{ijk}(l) - \frac{l}{\mu_{ijk}}\right\|_n \geq \epsilon n\right\} \leq \frac{\epsilon}{n}. \tag{B.2.3}$$

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

$$P\left\{\left\|V_{ijk}(D_{ijk}^r(t) + e) - \frac{D_{ijk}^r(t)}{\mu_{ijk}}\right\|_{(\sqrt{x_{r,0}}/|N^r|)L} \geq 3NL\sqrt{x_{r,0}}\epsilon\right\} \leq B_2\frac{\epsilon}{\sqrt{|N^r|}}, \tag{B.2.4}$$

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

$$P\left\{\left\|V_{ijk}^{r,0}(D_{ijk}^{r,0}(t), \tilde{b}) - \frac{D_{ijk}^{r,0}(t)}{\mu_{ijk}}\right\|_L \geq \epsilon\right\} \leq \frac{\epsilon}{\sqrt{|N^r|}}. \tag{B.2.5}$$

Multiplying the error bound $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately, we obtain

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V_{ijk}^{r,m}(D_{ijk}^{r,m}(t), \tilde{b}) - \frac{D_{ijk}^{r,0}(t)}{\mu_{ijk}}\right\|_L \geq \epsilon\right\} \leq \epsilon. \tag{B.2.6}$$

For $b = (0, 1)$ and $\tilde{b} = (0, 0)$, by (4.4.6),

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V_{ijk}^{r,m}(D_{ijk}^{r,m}, \tilde{b}) - V_{ijk}^{r,m}(D_{ijk}^{r,m}(t), b)\right\|_L \geq \epsilon\right\}$$

$$= P\left\{\max_{m<\sqrt{|N^r|}T}\left|V_{ijk}\left(D_{ijk}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right) - V_{ijk}\left(D_{ijk}^r\left(\frac{m}{\sqrt{|N^r|}}\right) + 1\right)\right| \geq \sqrt{x_{r,m}}\epsilon\right\}. \tag{B.2.7}$$

Observe that, by (4.2.13), $V_{ijk}(D_{ijk}^r(m/\sqrt{|N^r|})) \leq |N^r|T$ and, by Lemma 4.4.1,

$$P\{v_{ijk}^{r,T,\max} \geq \sqrt{x_{r,m}}\epsilon\} \leq \epsilon \tag{B.2.8}$$

for large enough r. Thus, we get (4.4.18) by combining (4.2.13) with (B.2.6)-(B.2.8). □

B.3 Proof of Proposition 4.4.2

Proof. We use the bounds established in Proposition 4.4.1. Fix L, T and $\epsilon > 0$. Let $\mathscr{V}^r$ be the intersection of the complements of the events given in (4.4.16)-(4.4.18), so $P\{\mathscr{V}^r > 1 - \epsilon\}$. We show that for r large enough and all $\omega \in \mathscr{V}^r$

$$\max_{m < \sqrt{|N^r|T}} \sup_{0 \le t_1 \le t_2 \le L} |\mathbb{X}^{r,m}(t_2) - \mathbb{X}^{r,m}(t_1)| \le \tilde{N}|t_2 - t_1| + \epsilon \tag{B.3.1}$$

for some $\tilde{N}$ that depends only on λ. We fix $\omega \in \mathscr{V}^r$ for the rest of the proof and so omit it from the notation. Let $t_1, t_2 \in [0, T]$ and $m \ge 0$. We first show that

$$|C^{r,m}(t_2) - C^{r,m}(t_1)| \le N_0|t_2 - t_1| + \epsilon \tag{B.3.2}$$

for some $N_0 > 0$. Because $C_{ij}^{r,m}$ is nondecreasing, we have, by (4.4.2) and (4.4.10), that

$$0 \le C_{ij}^{r,m}(t_2) - C_{ij}^{r,m}(t_1) \le D_{ij}^{r,m}(t_2) - D_{ij}^{r,m}(t_1). \tag{B.3.3}$$

Then (B.3.2) is immediately satisfied from (4.4.17), and $N_0 = N$. Combining (B.3.3) with (4.4.10) yields

$$|Z_{ij}^{r,m}(t_2) - Z_{ij}^{r,m}(t_1)| \le 2\left|D_{ij}^{r,m}(t_2) - D_{ij}^{r,m}(t_1)\right| + |A_{sij}^{r,m}(t_2) - A_{sij}^{r,m}(t_1)|. \tag{B.3.4}$$

By (4.4.16), $|A_i^{r,m}(t_2) - A_i^{r,m}(t_1)| < 2|\lambda||t_2 - t_1| + \epsilon$ for r large enough. By setting $N_1 = 2N_0 + 2|\lambda|$, and using (4.4.17), we get

$$|Z^{r,m}(t_2) - Z^{r,m}(t_1)| \le N_1|t_2 - t_1| + \epsilon. \tag{B.3.5}$$

Combining the results above with (4.4.9) gives

$$|Q^{r,m}(t_2) - Q^{r,m}(t_1)| \le N_2|t_2 - t_1| + \epsilon, \tag{B.3.6}$$

for $N_2 = N_0 + 2|\lambda|$. Note that $N_1 \ge N_2$. Also, for r large enough, by (4.2.3), (4.4.2), and the fact that $T_{ij}^r(t) = \sum_{k=1}^{N_i^r} T_{ijk}^r(t)$,

$$|T_{ij}^{r,m}(t_2) - T_{ij}^{r,m}(t_1)| \le 2\beta_i|t_2 - t_1|. \tag{B.3.7}$$

Note that, by definition of $\mathscr{V}^r$, the inequalities above hold for all $m < \sqrt{|N^r|T}$. This shows that (B.3.1) holds for r large enough with $\tilde{N} = N_1 \vee 2$.

Summarise the discussion above, processes

$$\{\mathbb{X}^{r,m} = (A^{r,m}, A_s^{r,m}, A_q^{r,m}, Q^{r,m}, Z^{r,m}, C^{r,m}, T^{r,m}, D^{r,m})\}$$

are almost Lipschitz. □

B.4   Proof of Proposition 4.4.4

Proof. The proof is similar to that of Proposition 5.5 in [Dai and Tezcan, 2011]. Fix $L > 0$ and let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathscr{E}$. By the definition of hydrodynamic limit, $|\tilde{\mathbb{X}}(0)| \leq 1$. By Proposition 4.4.3, $\tilde{\mathbb{X}}$ satisfies the hydrodynamic model equations (4.3.19)-(4.3.27) on $[0, L]$, thus by (4.3.19), (4.3.20), and (4.3.22), one can easily find a R_L such that

$$||\tilde{\mathbb{X}}(t)||_L \leq R_L. \tag{B.4.1}$$

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

$$|g(x) - g(y)| < \epsilon \tag{B.4.2}$$

if $|x - y| < \delta$ and $x, y \in [-2R_L, 2R_L]$.

By Corollary 4.4.2, fix $T > 0$, and also fix $0 < \delta < R_L$. For $\omega \in \mathscr{K}^r$, any $m < \sqrt{|N^r|}T$, and choosing r large enough, there exists a hydrodynamic limit $\tilde{\mathbb{X}}$ such that

$$||\mathbb{X}^{r,m}(t) - \tilde{\mathbb{X}}(t)||_L \leq \delta. \tag{B.4.3}$$

Together with (B.4.1), we have

$$||\mathbb{X}^{r,m}(t)||_L \leq ||\mathbb{X}^{r,m}(t) - \tilde{\mathbb{X}}(t)||_L + ||\tilde{\mathbb{X}}(t)||_L \leq 2R_L. \tag{B.4.4}$$

Thus by (B.4.1), (B.4.2), (B.4.4), and Assumption 4.3.4, $\forall t \in [0, L]$,

$$g(Q^{r,m}(t), Z^{r,m}(t)) = g(Q^{r,m}(t), Z^{r,m}(t)) - g(\tilde{Q}(t), \tilde{Z}(t)) + g(\tilde{Q}(t), \tilde{Z}(t))$$
$$\leq \epsilon + H(t),$$

which is (4.4.64).

(4.4.65) is obtained similarly. Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathscr{E}_g$. Then there exists a subsequence r_k of r such that

$$||\mathbb{X}^{r_k,0}(t) - \tilde{\mathbb{X}}(t)||_L \to 0, \tag{B.4.5}$$

as $k \to \infty$. By the definition of $\mathscr{E}_g$, and the fact that $g(Q^{r_k,0}(0), Z^{r_k,0}(0)) \to 0$, combined with (B.4.5) and the continuity of g, we have $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$, thus, by the last statement of

Assumption 4.3.4,

$$g(\tilde{Q}(t), \tilde{Z}(t)) = 0 \text{ for any } t \geq 0. \tag{B.4.6}$$

Similar to (B.4.3), for $\omega \in \mathscr{L}^r$, and r large enough, we can find a hydrodynamic limit $\tilde{\mathbb{X}}$ in $\mathscr{E}_g$ such that

$$||\mathbb{X}^{r,m}(t) - \tilde{\mathbb{X}}(t)||_L \leq \delta,$$

and by (B.4.6) and (B.4.1)

$$g(Q^{r,0}(0), Z^{r,0}(0)) \leq \epsilon.$$

$\square$

Bibliography

[Aksin et al., 2007] Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

[Armony, 2005] Armony, M. (2005). Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3):287–329.

[Armony and Mandelbaum, 2011] Armony, M. and Mandelbaum, A. (2011). Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*, 59(1):50–65.

[Armony and Ward, 2010] Armony, M. and Ward, A. R. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Operational Research*, 58(3):624–637.

[Ata et al., 2005] Ata, B., Kumar, S., et al. (2005). Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. *The Annals of Applied Probability*, 15(1A):331–391.

[Atar, 2008] Atar, R. (2008). Central limit theorem for a many-server queue with random service rates. *The Annals of Applied Probability*, 18(4):1548–1568.

[Billingsley, 1999] Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley.

[Borst et al., 2004] Borst, S., Mandelbaum, A., and Reiman, M. I. (2004). Dimensioning large call centers. *Operations research*, 52(1):17–34.

[Bramson, 1998] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148.

[Browne and Whitt, 1994] Browne, S. and Whitt, W. (1994). Piecewise-linear diffusion processes. *preprint*.

[Cooper, 1981] Cooper, R. B. (1981). *Introduction to Queueing Theory*. Edward Arnold.

[Dai and Tezcan, 2011] Dai, J. and Tezcan, T. (2011). State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research*, 36(2):271–320.

[Deo and Gurvich, 2011] Deo, S. and Gurvich, I. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319.

[Gans et al., 2003] Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Operations Management*, 5(2):79–141.

[Gans et al., 2010] Gans, N., Liu, N., Mandelbaum, A., Shen, H., Ye, H., et al. (2010). Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 99–123. Institute of Mathematical Statistics.

[Gurvich et al., 2010] Gurvich, I., Luedtke, J., and Tezcan, T. (2010). Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115.

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

[Harrison and Van Mieghem, 1997] Harrison, J. M. and Van Mieghem, J. A. (1997). Dynamic control of brownian networks: state space collapse and equivalent workload formulations. *The Annals of Applied Probability*, pages 747–771.

[Iglehart, 1965] Iglehart, D. L. (1965). Limiting diffusion approximations for the many server queue and the repairman problem. *Applied Probability Trust*, 2(2).

[Janssen et al., 2011] Janssen, A., Van Leeuwaarden, J., and Zwart, B. (2011). Refining square root safety staffing by expanding erlang c. *Operations Research*, 59(6):1512–1522.

[Kelly and Laws, 1993] Kelly, F. and Laws, C. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing systems*, 13(1-3):47–86.

[Koçağa et al., 2015] Koçağa, Y. L., Armony, M., and Ward, A. R. (2015). Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, 24(7):1101–1117.

[Koole, 2007] Koole, G. (2007). *Call Center Mathematics-A Scientific Method for Understanding and Improving Contact Centers*.

[Mandelbaum et al., 2012] Mandelbaum, A., Močilvić, and Tseytlin, Y. (2012). On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science*, 58(7):1273–1291.

[Mandelbaum and Zeltyn, 2004] Mandelbaum, A. and Zeltyn, S. (2004). The palm/erlang-a queue, with applications to call centers. In *Technical Report*. Service Engineering Lecture Notes.

[Mandelbaum and Zeltyn, 2009] Mandelbaum, A. and Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205.

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

[Puhalskii and Reiman, 2000] Puhalskii, A. and Reiman, M. (2000). The multiclass gi/ph/n queue in the halfin-whitt regime. *Advances in Applied Probability*, 32(2):564–595.

[Reiman, 1984] Reiman, M. I. (1984). Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology*, pages 207–240. Springer.

[Rykov and Efrosinin, 2004] Rykov, V. and Efrosinin, D. (2004). Optimal control of queueing systems with heterogeneous servers. *Queueing Systems*, 46(3-4):389–407.

[Stone, 1961] Stone, C. (1961). *Limit theorems for birth and death processes, and diffusion process*. PhD thesis, Stanford University.

[Tezcan, 2008] Tezcan, T. (2008). Optimal control of distributed parallel server systems under the halfin and whitt regime. *Mathematics of operations research*, 33(1):51–90.

[Tezcan and Zhang, 2014] Tezcan, T. and Zhang, J. (2014). Routing and staffing in customer service chat systems with impatient customers. *Operations research*, 62(4):943–956.

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

[Whitt, 2006] Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and operations management*, 15(1):88.

[Zan, 2012] Zan, J. (2012). *Staffing Service Centers Under Arrival-rate Uncertainty*. PhD thesis, The University of Texas at Austin.