On Occupancy Based Randomized Load Balancing for Large Systems with General Distributions

by

Thirupathaiah Vasantam

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Sem Borst Professor, Dept. of Mathematics and Computer Science Eindhoven University of Technology
Supervisor(s):	Ravi R. Mazumdar Professor, Dept. of ECE, University of Waterloo
Internal Members:	Andrew Heunis Professor, Dept. of ECE, University of Waterloo Patrick Mitran Associate Professor, Dept. of ECE, University of Waterloo
Internal-External Member:	Hossein Abouee Mehrizi Associate Professor, Dept. of Management Sciences University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Multi-server architectures are ubiquitous in today's information infrastructure whether for supporting cloud services, web servers, or for distributed storage. The performance of multi-server systems is highly dependent on the load distribution. This is affected by the use of load balancing strategies. Since both latency and blocking are important features, it is most reasonable to route an incoming job to a server that is lightly loaded. Hence a good load balancing policy should be dependent on the states of servers. Since obtaining information about the remaining workload of servers for every arrival is very hard, it is preferable to design load balancing policies that depend on occupancy or the number of progressing jobs of servers. Furthermore, if the system has a large number of servers, it is not practical to use the occupancy information of all the servers to dispatch or route an arrival due to high communication cost. In large-scale systems that have tens of thousands of servers, the policies which use the occupancy information of only a finite number of randomly selected servers to dispatch an arrival result in lower implementation cost than the policies which use the occupancy information of all the servers. Such policies are referred to as occupancy based randomized load balancing policies.

Motivated by cloud computing systems and web-server farms, we study two types of models. In the first model, each server is an Erlang loss server, and this model is an abstraction of Infrastructure-as-a-Service (IaaS) clouds. The second model we consider is one with processor sharing servers that is an abstraction of web-server farms which serve requests in a round-robin manner with small time granularity. The performance criterion for web-servers is the response time or the latency for the request to be processed. In most prior works, the analysis of these models was restricted to the case of exponential job length distributions and in this dissertation we study the case of general job length distributions.

To analyze the impact of a load balancing policy, we need to develop models for the system's dynamics. In this dissertation, we show that one can construct useful Markovian models. For occupancy based randomized routing policies, due to complex interdependencies between servers, an exact analysis is mostly intractable. However, we show that the multi-server systems that have an occupancy based randomized load balancing policy are examples of weakly interacting particle systems. In these systems, servers are interacting particles whose states lie in an uncountable state space. We develop a mean-field analysis to understand a server's behavior as the number of servers becomes large. We

show that under certain assumptions, as the number of servers increases, the sequence of empirical measure-valued Markov processes which model the systems' dynamics converges to a deterministic measure-valued process referred to as the mean-field limit. We observe that the mean-field equations correspond to the dynamics of the distribution of a non-linear Markov process. A consequence of having the mean-field limit is that under minor and natural assumptions on the initial states of servers, any finite set of servers can be shown to be independent of each other as the number of servers goes to infinity. Furthermore, the mean-field limit approximates each server's distribution in the transient regime when the number of servers is large.

A salient feature of loss and processor sharing systems in the setting where their time evolution can be modeled by reversible Markov processes is that their stationary occupancy distribution is insensitive to the type of job length distribution; it depends only on the average job length but not on the type of the distribution. This property does not hold when the number of servers is finite in our context due to lack of reversibility. We show however that the fixed-point of the mean-field is insensitive to the job length distributions for all occupancy based randomized load balancing policies when the fixed-point is unique for job lengths that have exponential distributions. We also provide some deeper insights into the relationship between the mean-field and the distributions of servers and the empirical measure in the stationary regime.

Finally, we address the accuracy of mean-field approximations in the case of loss models. To do so we establish a functional central limit theorem under the assumption that the job lengths have exponential distributions. We show that a suitably scaled fluctuation of the stochastic empirical process around the mean-field converges to an Ornstein-Uhlenbeck process. Our analysis is also valid for the Halfin-Whitt regime in which servers are critically loaded. We then exploit the functional central limit theorem to quantify the error between the actual blocking probability of the system with a large number of servers and the blocking probability obtained from the fixed-point of the mean-field. In the Halfin-Whitt regime, the error is of the order inverse square root of the number of servers. On the other hand, for a light load regime, the error is smaller than the inverse square root of the number of servers.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Ravi R. Mazumdar for his valuable personal and professional guidance and support throughout my doctoral studies. I am grateful to him for his valuable advice about academic career. His enthusiasm and positive attitude towards research is inspirational.

In addition, I would like to extend my gratitude to Dr. Arpan Mukhopadhyay of University of Warwick, UK for his constant encouragement and guidance that helped me to grow as a researcher. His collaboration led to the results of Chapters 2 and 3. I thoroughly enjoyed our collaboration.

I thank my thesis committee members: Prof. Andrew Heunis, Prof. Patrick Mitran, Prof. Hossein Abouee Mehrizi, and Prof. Sem Borst for serving on my thesis committee and carefully reading my dissertation. Their valuable comments have helped me to improve the quality of the presentation of this dissertation.

I extend my utmost gratitude to my friends for enriching my time in the beautiful city of Waterloo.

Finally, thanks to my mother and brother who supported and believed me throughout my education. I am deeply indebted to them for all their efforts to support my family. I am also thankful to my wife Shireesha for her consistent support throughout my doctoral studies. I am grateful to my son Yuvan for coming into our life and bringing us joy.

Dedication

This dissertation is dedicated to my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Computer systems such as web-server farms, high-traffic websites, and cloud computing systems often contain hundreds of thousands of servers for meeting the incoming demands [1–3]. All these systems are examples of large-scale multi-server systems. The main challenge is to distribute the incoming jobs (requests) that generate tasks to servers so that the system achieves good system performance. Multi-server systems use job dispatchers to dispatch an incoming job to one of the servers. A basic multi-server model with three servers and one job dispatcher is shown in Figure 1.1.

Load balancing in multi-server systems is the process of balancing the load on servers so that individual servers are not overburdened that can lead to performance degradation. The job dispatchers dispatch an incoming job according to a load balancing policy that can improve the system performance. The main focus of this dissertation is the design and analysis of load balancing policies for multi-server systems with one central job dispatcher.

When a job arrives at the system, the best method is to dispatch the arrival to the least loaded or best server in the entire system as the job would get processed more efficiently. Since the job sizes are unknown in practice, the estimation and exchange of the remaining workload information is impractical and hence the policies that require only the knowledge of the occupancy or number of jobs of servers are preferred. If an incoming job is dispatched to a server with the least occupancy among all the servers, then the load balancing policy is referred to as the Join-the-Shortest-Queue (JSQ) policy [4–6]. The implementation of this policy requires information about the occupancy of all the servers for every arrival. Since many cloud computing systems and web-server farms often contain hundreds of thousands of servers, the implementation of the JSQ policy for large-scale systems results

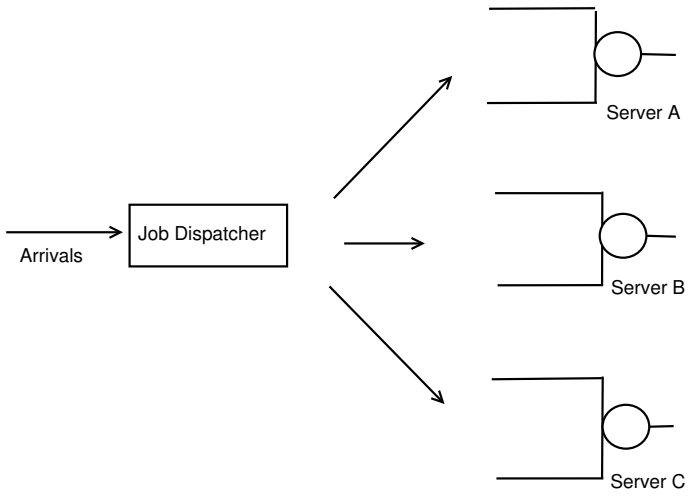

Figure 1.1: A multi-server system with three servers

in huge communication overhead and implementation cost. Hence, the JSQ policy is not preferred for large systems. An alternative type of policies that are easy to implement are randomized load balancing policies according to which the dispatcher first samples a few servers uniformly at random upon an arrival and one of them is chosen as the destination server for the arrival based on their occupancies. The criteria for selecting the destination server depends on the load balancing policy.

A popular randomized load balancing policy is the Shortest-Queue-Among d (SQ(d)) or Power-of-d policy according to which the dispatcher selects the server with the least occupancy among d randomly selected servers as the destination for an incoming job. This policy was introduced in [7] for multi-server systems with First-Come-First-Served (FCFS) servers when the job lengths have exponential distributions. The exact analysis of the system is not tractable due to complex interactions between servers as a result of the fact that the dispatcher uses the occupancy information of multiple servers to dispatch an arrival. Although servers are coupled, it was shown that the model is an example of a weakly interacting particle system. They used mean-field techniques to characterize the impact of the SQ(d) policy on the system performance. We give an introduction to mean-field techniques in Section 1.3.

The analysis of a randomized load balancing policy strongly depends on the type of servers that are present in the system. For example, servers could be FCFS, Erlang loss, or processor sharing (PS) servers. Each type of servers are used to study systems that serve jobs coming from different types of applications. For example, the FCFS model can be used to study supermarkets, hospitals, and path selection in networks. The loss models have

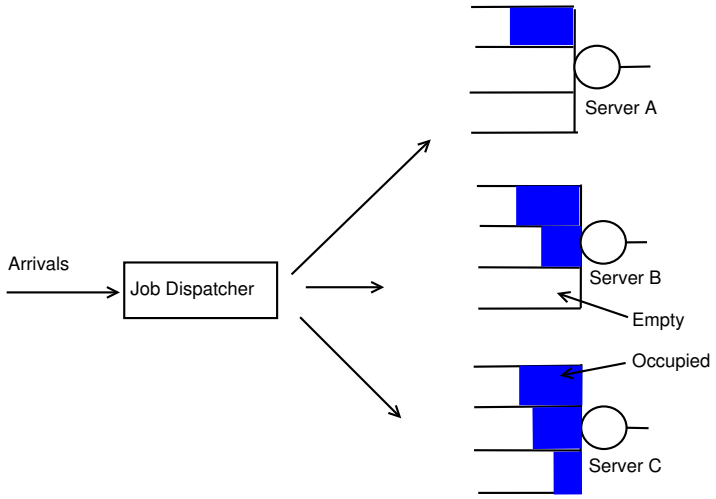

Figure 1.2: A multi-server Erlang loss model with $C = 3$.

applications in call centers and cloud computing systems. The PS servers have applications in web-server farms that provide service to delay-sensitive applications. Furthermore, the metric to measure the system performance also depends on the type of the system's servers. For FCFS, loss, and PS systems, the metrics of interest are the average delay, the average blocking probability, and the average response time (total execution time of a job) to measure the system performance, respectively. In this dissertation, we study occupancy based randomized load balancing for multi-server systems with Erlang loss servers and PS servers.

We now give a brief introduction to the architecture of the multi-server loss model that we study in this dissertation. Each loss server has capacity to serve only a finite number of jobs simultaneously, the maximum number of simultaneous jobs that a server can process is referred to as the server's capacity denoted by C. Furthermore, each job is processed at a constant unit rate irrespective of the occupancy of the server. A job arriving at a server will be accepted for service if its occupancy is less than the capacity C. Otherwise, the job is blocked from service and it is considered to be discarded from the system immediately. In Figure 1.2, we show a multi-server loss model with three servers each with capacity $C = 3$. In this figure, Server A, Server B, and Server C have occupancies one, two, and three, respectively. If an arrival is routed to Server C, then it will be blocked. On the other hand, if the destination for an arrival is either Server A or Server B, then it will be accepted. This example clearly shows the importance of a careful design of load balancing policies for efficient usage of resources so that the resulting average blocking probability is minimized.

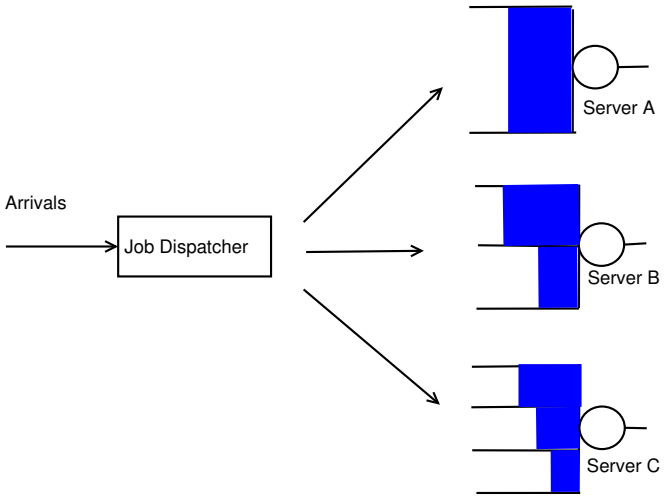

Figure 1.3: A multi-server PS model

Next, we give a brief introduction to the PS model. We assume that each server has capacity to serve jobs at a unit rate. A server with n jobs simultaneously processes all the n jobs at the rate of $\frac{1}{n}$. A job arriving at a server will be accepted for service and its processing begins immediately. In Figure 1.3, we show a PS model with three servers each having capacity to serve jobs at unit rate. The servers A, B, and C have occupancies one, two, and three, respectively. Hence, the processing speeds of jobs at severs A, B, and C are 1, $\frac{1}{2}$, and $\frac{1}{3}$, respectively. If an incoming job is routed to Server A, then it would get the highest possible instantaneous service rate in the system.

Our focus of interest in this dissertation is to study occupancy based randomized load balancing for the loss and PS models. In most prior works [8–12], the analysis was restricted to the case when job lengths have exponential distributions by using mean-field techniques. However, in many applications, the service time distributions are not exponential. For example, the service times follow log-normal distributions in call centers [13], and Gamma distributions in automatic teller machines (ATMs) [14] etc. Hence, it is of great interest to study the case of general job length distributions (JLDs). Our objective is to develop a mean-field analysis for this case.

1.1 Models

We now introduce the problems which we investigate in this dissertation. We study mean-field models in which servers are interacting particles, and the interactions between servers

take place when a job arrives at the system. The task of dispatching an incoming job to a server introduces dependence amongst the servers.

We study three problems. The first problem is on establishing the mean-field limit for large-scale multi-server loss models with general JLDs when the SQ(d) policy is used to dispatch the incoming jobs. This model is inspired by cloud computing systems which can be modeled as blocking models. We give an analysis of this problem in Chapter 2.

The second problem is on the mean-field analysis of randomized routing schemes for the PS model under the assumption of general JLDs. In this problem, we study not only the SQ(d) policy but also all occupancy based randomized routing schemes under a common framework. This model has applications in cloud computing systems and web-server farms that provide service to delay-sensitive applications. We provide an analysis of this problem in Chapter 3.

In Chapter 4, we investigate the third problem on the accuracy of mean-field approximations for multi-server loss models with the SQ(d) policy under the assumption of exponential JLDs. For this, we establish a suitable functional central limit theorem.

We next give a detailed description of all the three problems that we study in this dissertation. We also discuss the contributions of previous works and provide a summary of our contributions.

1.1.1 The SQ(d) Policy for Erlang Loss Systems

Due to tremendous growth in the trend to externalize storage and computing resources, cloud computing systems maintain a large number of servers to provide an efficient service to jobs. In particular, in Infrastructure-as-a-service (IaaS) clouds such as Microsoft's Azure [2] and Amazon EC2 [1], the incoming job requests are mapped into virtual machines (VMs) that request resources such as processor power, I/O bandwidth, disk etc. from a server that is picked from a large set of distributed servers. When a job arrives, the incoming request is routed to one of the servers where it is accepted for service if the requested amount of resources are available. Otherwise, it is blocked from service. Then the job is discarded from the system immediately. The resources allocated to a job will be released once the service of a job ends. Every incoming job requests a certain amount of resources for usage from a server in the system and clearly, since servers have finite amount of resources, we assume that servers can process only a limited number of jobs simultaneously. Therefore,

the available resources in the system should be used as efficiently as possible. The system's performance is measured in terms of the average blocking probability of a job. To serve jobs efficiently, the service provider of a cloud computing system should design a mechanism to quickly identify a server that has enough resources for an incoming arrival. Since the job requests arrive randomly and their job lengths are random too, the main challenge is to quickly choose the server that is the least loaded or has the smallest number of jobs in the system as the destination server for an arrival.

Cloud computing systems are distributed systems in practice. Hence, they have job dispatchers to distribute incoming jobs to servers. The job dispatchers follow a predefined load balancing policy to distribute the incoming job requests to servers. We focus on the homogeneous model in which all the servers have the same amount of resources and also the incoming jobs request same amount of resources and have identical JLDs. Each accepted job is assumed to be processed at a constant unit rate. Furthermore, we assume that the system has a single central job dispatcher to distribute the incoming jobs to servers. A server in IaaS clouds with capacity C indicating the maximum number of jobs that can be served simultaneously at the server can be viewed as an Erlang loss server with capacity C. In particular, an IaaS cloud with N servers each with capacity C can be modeled by a multi-server Erlang loss system with N servers each with capacity C.

Since cloud computing systems often contain hundreds of thousands of servers, the JSQ policy for such large-scale systems suffers from high implementation cost. On the other hand, if an arrival is routed to a server that is picked uniformly at random, then the resulting blocking probability is much larger than the resulting blocking probability under the JSQ policy although the implementation cost is less. In the literature [8,10,15], the SQ(d) (d 2) policy was studied for loss models. This policy was shown to improve the system performance significantly over the random routing ($d = 1$) policy even for the small value of $d = 2$. Although the SQ(d) policy performs worse than the JSQ policy, but it reduces the implementation cost drastically.

For systems with a finite number N of servers, characterizing the impact of the SQ(d) policy on each server thereby on the average blocking probability is a challenging task due to inter-dependencies between servers. However, under the assumption that the job arrival process is a Poisson process with rate N , due to weak interactions between servers, it was shown in previous works [8,10,15] by using mean-field techniques that a deterministic process referred to as the mean-field limit and its fixed-point approximate the distribution of a server's state of the system with large N in the transient and stationary regimes,

respectively. As a consequence, it can be shown that a function of the fixed-point of the mean-field approximates the average blocking probability.

Objective and Key Challenges

Our objective is to develop the mean-field analysis to study the impact of the SQ(d) routing policy for a multi-server Erlang loss system with N servers each with capacity C under the assumption of general JLDs. Furthermore, we consider the case when jobs arrive according to a Poisson process with rate N . If a routed job gets accepted at its destination, then its processing begins immediately at a unit rate until its service is completed. This model was investigated in previous works under the assumption of exponential JLDs [8, 10]. Since in practice jobs have general JLDs, it is important to characterize the impact of the SQ(d) policy on each server and more importantly, on the average blocking probability. Motivated by this, we aim to obtain the mean-field limit and understand whether this can approximate the distribution of a server under the assumption of general JLDs. Moreover, we would like to investigate the impact of the routing policy on the average blocking probability when job lengths have general distributions.

To analyze complex stochastic systems, the starting point is often the mathematical modeling of the time evolution of the system as the time evolution of a Markov process. For the mean-field analysis, the first step is also to obtain a mathematical formulation of the time evolution of the system starting from an initial state. For exponential JLDs, due to the memoryless property of exponential distributions, it is sufficient to keep track of only the occupancy information of a server and tracking the age indicating the amount of time for which the job is in service or the residual service time is not required to construct a Markov process that models the system dynamics. In this simple case, each server's state is its occupancy. Due to symmetry of the system to servers' identities (as servers are homogeneous and the SQ(d) policy never uses the identities of servers), the process $(\mathbf{X}^{(N)}(t); t \quad 0)$ where $\mathbf{X}^{(N)}(t) = (\mathbf{X}_i^{(N)}(t); 0 \quad i \quad C)$ and $\mathbf{X}_i^{(N)}(t)$ denotes the fraction of servers with at least i progressing jobs at time t out of N servers is a Markov process. Let $\mathbf{S}_i^{(N)}(t)$ be the random variable denoting the state of the i^{th} server at time t, then

$$\mathbf{X}_i^{(N)}(t) = \frac{1}{N} \sum_{k=1}^{N} I^{\mathsf{n}}{}_{\mathbf{S}_k^{(N)}(t) \quad i}{}^{\mathsf{o}};$$

The mean-field limit which is a deterministic process is the limit of the process $(\mathbf{X}^{(N)}(t); t$ $0)$ as $N \; ! \; 1$. The proof of the existence of a mean-field limit mainly uses techniques on

the convergence of a sequence of Markov processes. In this case, the results in [8,10] imply that the following interchange holds,

$$\lim_{N \to \infty} \lim_{t \to \infty} \mathbf{X}^{(N)}(t) = \lim_{t \to \infty} \lim_{N \to \infty} \mathbf{X}^{(N)}(t) \tag{1.1}$$

where we consider the limit of a random element as its weak convergence limit. From (1.1), we obtain the equivalence between the stationary distribution of the limiting system given by the left-hand side and the globally stable fixed-point or equilibrium of the mean-field given by the right-hand side under the SQ(d) load balancing policy. The key is that the mean-field evolves according to a set of ordinary differential equations (ODEs) referred to as the mean-field equations (MFEs). We can study these equations to understand the behavior of the mean-field in stationary regime. Moreover, (1.1) can be used to show that any finite set of servers become statistically independent of each other [10] in the limit as $N \to \infty$. As a result, if $= (\ _n; 0 \quad n \quad _C)$ is the unique fixed-point of the mean-field, then $_n$ is the probability that a server has at least n jobs and ^d_c is the blocking probability in the stationary regime as $N \to \infty$.

For general JLDs, due to the absence of the memoryless property, each progressing job's state must also be included in server's state. We can use either the age or residual service time of a job to track its status. In this dissertation, we use the ages of progressing jobs as their states. The state of a server must include the ages of all the jobs that are currently in progress at the server. At any time t, the state of a server for the case of general JLDs is of the form $(n; x_1; \quad ; x_n)$ where n denotes the number of progressing jobs and x_i denotes the age of the i^{th} progressing job at the server. Since the age of a job increases linearly with time and the state of a server lies in R^n_+ when it has n jobs, the classical methods used to study the exponential case cannot be used. We can overcome these challenges if we model the system evolution by a measure-valued Markov process, and then the mean-field limit can be established by using techniques on the convergence of a sequence of measure-valued Markov processes. We model the time evolution of the system as the time evolution of the measure-valued Markov process $(^{-(N)}_t; t \quad 0)$ where

$$^{-(N)}_t = \frac{1}{N} \sum_{k=1}^{N} \ \mathbf{s}_k^{(N)}(t)$$

and $_a$ is the Dirac measure with unit mass at a. The result that $(^{-(N)}_t; t \quad 0)$ is a Markov process follows from [16, Proposition 2.3.3]. Our main aim is to establish the limit of the process $(^{-(N)}_t; t \quad 0)$ as $N \to \infty$. If the limit exists, then the limiting process is

referred to as the mean-field limit. We then plan to show that the obtained mean-field limit approximates the distribution of a server's state in the transient regime.

To study the impact of the SQ(d) policy in the stationary regime, we first need to characterize the fixed-point of the mean-field, we then need to show that the stationary distribution of a server's state can be approximated by the fixed-point of the mean-field if it is unique.

An important feature of Erlang loss systems for the case $d = 1$ is that the system is insensitive. That is, the stationary distribution of a server's occupancy is independent of the type of the JLD, but it depends only on the average job length. This implies the robustness of the system to variations in the JLDs. Hence, one can dimension the system by assuming the exponential JLD for the given average job length. Motivated by this, we would like to understand whether the stationary distribution of a server's occupancy is insensitive or not in our model. In the system with N servers, the stationary distribution is not insensitive when d 2 due to violation of the Whittle balance condition [17] by the resulting routing probabilities which is a necessary condition to have insensitivity [18]. However, it was observed in simulation results of [10] that the stationary distribution of a server's occupancy for different classes of JLDs coincide with that of the exponential case as N ! 1. This implies that if the fixed-point of the mean-field coincides with the stationary distribution of a server's state as N ! 1, then the fixed-point of the mean-field should exhibit insensitivity. The proof of insensitivity of the fixed-point of the mean-field and the stationary distribution of a server's occupancy when N ! 1 are still open problems. We investigate these issues by characterizing the fixed-point of the mean-field. A major challenge for our model is that the resulting MFEs are partial differential equations (PDEs). As a result, understanding the behavior of the mean-field as t ! 1 is much more challenging than the exponential case since in that case the MFEs are ODEs.

Related Literature

The SQ(d) scheme was first introduced in [7] for multi-server systems with the First-Come-First-Served (FCFS) service discipline assuming $d = 2$ and exponential service times. When the number of servers N is finite, analyzing the impact of the SQ(d) policy is a difficult task due to dependence amongst the servers introduced by the SQ(d) policy. However, they obtained an elegant tractable way of characterizing the stationary distributions in the limit as N ! 1 by using mean-field techniques. Their results were then extended for the

case of $d > 2$ in [12] where it was argued that the case $d = 2$ provides most of the gains and hence, the term 'The power-of-2' came to be used.

Multi-server loss models with randomized load balancing schemes were first studied in [15] when job lengths are exponentially distributed using a formal mean-field approach. However, the existence and uniqueness of the fixed-point of the mean-field were not shown. In [8,10,19], the existence and uniqueness of the fixed-point of the mean-field for the homogeneous loss model of [15] was addressed. For the heterogeneous loss model, the existence and uniqueness of the fixed-point of the mean-field were established in [8] under the assumption of asymptotic independence of servers in the stationary regime while [10] showed the asymptotic independence of servers and the interchange of limits (1.1). The propagation of chaos on path space implying that any finite set of random elements $f(\mathbf{S}_i^{(N)}(t); t \quad 0)g_{1 \ i \ N}$ become independent of each other as $N \ ! \ 1$ was established by [20, 21] in the context of alternate routing in circuit-switched networks. The notion of propagation of chaos is further explained in Definition 1.1.

The impact of the SQ(d) policy for the loss models similar to the one considered here was analyzed in [8,10,15,19] under the assumption of exponential JLDs. They also considered the more general heterogeneous case with an appropriate modification of the SQ(d) policy to account for server and job heterogeneity. It was shown in [19] that the SQ(d) load balancing scheme yields almost optimal blocking performance in that the average blocking is very close to the theoretical lower bound on the minimum average blocking achievable by any load balancing policy.

It is well known that the stationary distributions of single server loss systems even with state-dependent Poisson arrival rates are insensitive to the service time distribution, i.e., they only depend on the mean of the service times [22]. Hence, it is essential to investigate whether the insensitivity property carries over to systems with randomized load balancing such as the SQ(d) policy. When N is finite, randomized load balancing policies result in the individual servers being coupled. It can be shown as in [18, 23] that when N is finite, the system is not insensitive since the SQ(d) policy does not satisfy the necessary condition of state-dependent arrival rates to be balanced. However, the insensitivity of the fixed-point or equilibrium of the mean-field was observed for the limiting case (i.e., when $N \ ! \ 1$) via simulations in [8, 10] but no proofs were provided. This requires a proof in our context because as $N \ ! \ 1$, the resulting MFEs represent the time evolution of the distribution of a nonlinear Markov process that models a queueing system with a single server in which the arrival rates of jobs depend both on the instantaneous state and the

instantaneous distribution of the state of the server whereas in the classical models, the underlying Markov process is linear since the arrival rates of jobs are independent of the instantaneous distribution of the server's state. The main goal of the chapter is to address this issue.

For the case of general JLDs, Markovian modeling of the system requires us to keep track of the age or residual service time of each job that is in progress in the system. Therefore, the underlying space on which the Markov process is defined is not discrete, and hence the classical Markov chain techniques cannot be used. As a result, establishing the mean-field limit and characterizing the properties of its stationary behavior for general JLDs is a challenging task.

In the literature, some prior works discussed later in this section have studied queueing models in which either jobs or servers are interacting particles under the assumption of general JLDs by using mean-field techniques. Also, some works have investigated the randomized load balancing for multi-server systems with general JLDs but the analysis is not complete. There are still several open problems in the literature that need to be addressed. We now discuss contributions and drawbacks of some related works for the case of general JLDs below.

In [24], randomized load balancing schemes for queueing systems with general service time distributions were investigated when servers use the service disciplines FCFS, PS, and Last-In-First-Out (LIFO). The steady-state results were characterized by assuming the asymptotic independence of servers in the system when $N \to \infty$. However, the proof of asymptotic independence of servers remains an open problem and also, the mean-field limit and its fixed-point were not studied in any detail.

In [25], mean-field techniques were used to study a closed queueing network with M jobs and N queues under the FCFS service discipline. In this model, a job once served at a queue joins another queue chosen with probability $\frac{1}{N}$ from N queues. The mean-field limit was established for the regime when $\lim_{M,N \to \infty} \frac{M}{N} \to \tilde{}$. However, the equilibrium behavior of the system was not studied. Recently, the SQ(d) setting in a system of N servers with FCFS service discipline where jobs arrive according to a time-inhomogeneous Poisson process and general i.i.d. service times were studied in [26]. They studied only the transient behavior of the mean-field. The steady-state analysis of the model was not investigated.

The mean-field and fluid analysis of queues are closely related, the former usually in

the space of probability measures and the latter on the space of finite measures. The fluid limit analysis of the FCFS and the PS queues with general service time distributions has been studied using a measure-valued processes approach developed by Dawson [16] in [27–31]. In this chapter, we use the age of a progressing job, indicating the amount of time elapsed since its arrival to construct a measure-valued Markov process that models the system dynamics. We use this framework to establish the mean-field limit by studying the limit of the empirical measure-valued process when the number of servers $N \to \infty$, as in [16, 29]. Our proof techniques closely follow those in [29] in which a fluid limit and a functional central limit theorem for a sequence of M/GI/1 systems were established by using measure-valued processes.

In [32], the FCFS model was studied with exponential distributions under the SQ(d) policy by using measure-valued processes. In the exponential case, the set of states of servers is $\mathbb{Z}_+$. In [32], the law of large numbers on path space is established by studying the limit of the sequence of empirical measures with samples in $\mathcal{M}_1(\mathcal{D}_{\mathbb{Z}_+}([0, \infty)))$ where $\mathcal{D}_{\mathbb{Z}_+}([0, \infty))$ is the space of right continuous with left limits (RCLL) functions in $\mathbb{Z}_+$ and $\mathcal{M}_1(\mathcal{D}_{\mathbb{Z}_+}([0, \infty)))$ is the space of probability measures on $\mathcal{D}_{\mathbb{Z}_+}([0, \infty))$. For our analysis, we study stochastic processes with sample paths in $\mathcal{D}_{\mathcal{M}_1(\mathcal{U})}([0, \infty))$ where $\mathcal{U}$ is the set of states of severs which is uncountable.

Summary of Contributions

We obtained the following results for the two regimes of interest. Preliminary results were presented in [33] and the detailed results will appear in [34].

For the Transient Regime: We first establish the mean-field limit by using measure-valued processes. We then use this result to show the asymptotic independence of any finite set of servers at a given finite time t as in [26] under the exchangeability assumptions of initial servers' states. The MFEs represent dynamics of a classical single server system with Poisson job arrival process whose intensity depends both on the occupancy of the server and the distribution at any time t. From the initial form of the MFEs whose solution is a measure-valued process, we show that under certain assumptions, the mean-field satisfies a set of PDEs.

For the Stationary Regime: We show the existence of a unique fixed-point of the mean-

field. The fixed-point of the mean-field satisfies a product form given in (2.21) and the occupancy distribution obtained from the fixed-point of the mean-field coincides with the fixed-point in the exponential case having the same average job length. This concludes insensitivity of the fixed-point of the mean-field. We also provide numerical evidence suggesting that the fixed-point is globally asymptotically stable (GAS) as $t \to 1$ when job lengths have mixed-Erlang distributions.

1.1.2 Occupancy Based Randomized Routing for Processor Sharing Systems

Due to the emergence of the cloud computing paradigm and other applications, the server farms often contain large numbers of servers to process the incoming jobs. Unlike the model discussed in Section 1.1.1 which is a blocking model, we now consider a model in which servers provide service to delay-sensitive applications such as online-search (Google), Social networking (Facebook) etc. The main objective in this model is to minimize the average response time experienced by a job in the system. Furthermore, each server has an infinite buffer, and there is no blocking of jobs in the system. A server always accepts an incoming job.

The front end job dispatchers in web-server farms route an arriving job to one of the servers that provide minimal response times as the tasks in most cases are delay-sensitive. Therefore, the main challenge in these systems is to design low complexity load balancing algorithms that yield efficient usage of available resources, thereby good system performance. The system performance is measured in terms of the average response time experienced by a job in the system. In server farms, the resources are shared by processing requests in a round-robin manner with small time granularity. This model can be well approximated by a model with servers having the PS service discipline [4, 35, 36] in which the processing speed of a server is equally shared by all the progressing jobs. In practice, the service requirements of jobs are highly variable. Under the PS service discipline, as the processing speed of a server is shared by all the jobs, the short jobs will be processed quickly and the PS service discipline avoids the chance of getting waited for a long time by small jobs before their service begins. As a result, the average response time is minimized. Hence, the PS service discipline is preferred for delay-sensitive applications.

Traditionally, small-scale server farms such as Cisco Local Director, IBM Network Dispatcher, Microsoft Sharepoint use the classical Join-the-Shortest-Queue (JSQ) load balanc-

ing policy [4, 8]. It was shown in [4–6] that the JSQ policy is nearly optimal and further, it is robust to the JLDs since it is almost insensitive. For large scale systems that contain hundreds of thousands of servers, the JSQ policy requires information about the number of progressing jobs at all the servers. However, it was shown in [7, 9, 10, 12, 24, 37] that the SQ(d) policy can achieve almost the same gains at a much smaller sampling cost.

Objective and Key Challenges

We consider a multi-server system with N servers which employ the PS service discipline. Each server has capacity to process jobs at a unit rate and hence, a server with n jobs processes every job at rate $\frac{1}{n}$ because of the PS service discipline assumption. The job arrival process is a Poisson process with rate N and jobs have general JLDs with finite average job length. We study occupancy based randomized routing schemes which imply sampling of a finite number of servers upon an arrival and the dispatcher selects a server from this set as the destination server. Any occupancy based policy can be viewed as a mapping of occupancies of sampled servers to routing probabilities. For example, let us assume d servers are sampled upon an arrival and let n_i be the occupancy of the i^{th} sampled server. Furthermore, let $(n_1; \quad ; n_d)$ be the vector of occupancies of sampled servers. Then the routing decision according to a load balancing policy can be viewed as finding a probability vector $(p_1; \quad ; p_d)$ based on $(n_1; \quad ; n_d)$ where p_i denotes the probability with which the arrival should be routed to the i^{th} sampled server. Based on this logic, we aim to develop a mean-field analysis under a common framework to include all occupancy based routing policies which require sampling of a finite number of servers upon an arrival. The study of occupancy based routing policies for the loss model easily follows from the analysis developed for the PS model. This is because each server in the loss model can process at most C jobs simultaneously each with constant unit processing rate and hence, it can be seen that the analysis of the loss model follows easily from the analysis of the PS model.

In the literature, the complete mean-field analysis for the model of interest exists only for the SQ(d) policy [7, 11] and for some threshold based policies [12] under the assumption of exponential JLDs but the analysis for general JLDs remains an open problem. The mean-field analysis of the SQ(d) policy was developed in [7] under the assumption of Poisson arrival process of jobs with rate N and jobs have exponential JLDs for the case $d = 2$. Later in [12], the analysis was extended to the case $d > 2$. Unlike the Erlang loss model

considered in Section 1.1.1 that is stable iff the average job length is finite, if $\frac{1}{}$ is the average job length, then the PS model requires $<$ for the stability of the system [38]. In [7, 12], under the assumption of $<$, they used mean-field techniques to understand the impact of the SQ(d) policy on large-scale system behavior. In this approach, they first characterized the system's dynamics with time by studying the evolution of the empirical process $(\mathbf{X}^{(N)}(t); t \quad 0)$ where $\mathbf{X}^{(N)}(t) = (\mathbf{X}_i^{(N)}(t); i \quad 0)$ and $\mathbf{X}_i^{(N)}(t)$ denotes the fraction of servers with at least i progressing jobs at time t. Then they showed that under certain initial conditions, the empirical process $(\mathbf{X}^{(N)}(t); t \quad 0)$ converges to a deterministic process referred to as the mean-field. The mean-field is a unique solution of a set MFEs which are ODEs. Moreover, they showed that the interchange of limits (1.1) holds. As a result, the mean-field and its fixed-point can be used to approximate the transient and stationary distribution of a server when N is large. In fact, for $d \quad 2$, they found the fixed-point of the mean-field $= (\ _i; i \quad 0)$ given by $\ _i = \left(-\right)^{\frac{d^i \ 1}{d \ 1}}$. Then the average delay of a job in the system when the number of servers N is large is nearly $\sum_{i \ 1} \left(-\right)^{\frac{d^i \ 1}{d \ 1}}$. For the random routing case of $d = 1$, we have $\ _i = (-)^i$. This clearly shows a significant improvement in the system performance since $\ _i$ decreases double exponentially with i for $d \quad 2$ whereas $\ _i$ decreases exponentially with i for the case $d = 1$.

Since the JLDs are not exponential in practice, it is crucial to understand the impact of the SQ(d) policy for the PS model with general JLDs. In the literature, although there are some previous works [39], the complete analysis remains an open problem. There is another reason why the PS model is interesting: it is known to be insensitive if the arrival process to each individual server is a Poisson process, and so showing insensitivity of the mean-field under very general occupancy based randomized routing can then lead us to conclude that insensitivity is a generic property of the PS service discipline rather than the precise input to a queue that is controlled by the routing policy. Also, by using our analysis, one can obtain the mean-field limit immediately for the policies that fall into our framework and thereby, the considered policy's impact can be studied by using the obtained mean-field limit. It is known that the random routing SQ(d) with $d = 1$ is insensitive. We need to understand whether this holds for the SQ(d) policy with $d \quad 2$ and also, for other occupancy based routing policies. The system with finite N under the SQ(d) policy is not insensitive as the arrival rate of jobs do not satisfy the Whittle balance condition [17], which is a necessary condition [18, 23]. However, when $N \ ! \ 1$, it was observed through simulations that the insensitivity property holds for the SQ(d)

policy [11]. For other routing policies, insensitivity was not studied.

We now discuss why it is essential to study occupancy based randomized routing policies in addition to the SQ(d) policy. Upon an arrival, if the dispatcher samples a fixed number of d servers to decide the destination server, then the SQ(d) policy is optimal [5, 6]. Now let us consider the following example. For the case d 2, if the first server sampled upon an arrival is found to be idle, then the dispatcher can stop further sampling of servers to decide the destination. For some systems, limiting the amount of information exchanged to dispatch an arrival is important. Hence, while designing a routing policy, we need to consider both its performance and complexity. Furthermore, a routing policy should provide the system performance that is robust to variations in the system congestion level. For example, when the dispatcher uses the SQ(d) policy, if the system is heavily loaded, then it could result in bad system performance when d is fixed at a small value. On the other hand, if the system is lightly loaded, then the dispatcher might unnecessarily sample more servers when d is fixed at a large value. Since variations in the system congestion level are common in practice, the routing policy should be able to adapt to the system congestion level. Our framework should also include policies that are robust to variations in the system traffic.

Our objective is to study the mean-field of the empirical distributions for policies that fall into our framework introduced in Section 3.1.2. This includes policies that are adaptive to the system congestion level and threshold based policies. We provide results for four occupancy based policies defined in Section 3.4 that include the SQ(d) policy, and the remaining three are generalized threshold based policies. Such policies have been studied in other contexts. For example, in [12], a threshold based policy was studied in which for an incoming arrival, the dispatcher randomly selects a server as the destination server if its occupancy is less than or equal to the chosen threshold value. Otherwise, the dispatcher samples another server, and the same procedure is repeated until the dispatcher samples d servers. If all the d sampled servers have occupancy higher than the threshold value, then the shortest queue is chosen as the destination. This policy has lower complexity than the SQ(d) policy at the cost of a small drop in the performance. We investigate this policy in this dissertation, and we show that the fixed-point of the mean-field is insensitive, which is also confirmed by our simulation results. In [12], insensitivity of the threshold based policy was not studied.

In [24], a d-adaptive algorithm was studied using simulations to overcome the bad performance of the SQ(d) policy for small values of d for the case of Power-law distributions

in First-In-First-Out (FIFO) models. According to this policy, if the occupancy of the first selected server is k, then the dispatcher selects further $f(k)$ severs and then chooses the server with the least occupancy as the destination. They observed through simulations that for $f(k) = k$, the d-adaptive algorithm samples with high probability three servers per arrival providing the system performance that is similar to the SQ(4) scheme. This policy also belongs to the class of occupancy based policies that we study in this dissertation. Our simulation results show that the system exhibits insensitivity under this scheme when $N \; ! \; 1$. In [24], insensitivity of the d-adaptive policy was not discussed.

For queueing systems with general service times, we need to use the occupancy state as well as the age or residual service times of the progressing jobs in service to model the system's evolution by a Markov process. For the FCFS models, this results in a bi-dimensional Markov process defined on $\mathsf{Z}_+ \quad \mathsf{R}_+$. On the other hand, for the PS model, the Markov process is defined on $\bigcup_{n2\mathsf{Z}_+}(n; \mathsf{R}_+^n)$ as we need to track both occupancy and age of each progressing job. The analysis is much more difficult because the rate at which a job is served at a server depends on the server's occupancy. Similar to the loss model, we use measure-valued Markov processes to establish the mean-field limit under certain assumptions on initial conditions.

Related Literature

In [24], the SQ(d) policy was investigated for the PS, FCFS, and LIFO systems with general service time distributions under the assumption of asymptotic independence of any finite set of servers and the existence of a unique limiting stationary distribution as stated in an *ansatz*. However, the proofs of the assumptions considered in the *ansatz* are still unknown. The asymptotic independence of finite set of servers has been shown for the PS and the Erlang loss models in [9, 10, 20] for the case of exponential distributions.

The assumption of asymptotic independence of servers for the PS model for the case of general JLDs renders the arrival process to each server as a state-dependent Poisson process. Then the resulting model satisfies the Whittle's condition [17], which implies insensitivity. A proof of the *ansatz* remains an open problem except for the FCFS models when JLDs have decreasing hazard rate functions [40] that give rise to monotonicity in the model. From [40], the proof techniques cannot be extended to the PS model as we do not have ordering between the server states, unlike the FCFS systems.

A mean-field analysis to characterize the transient and stationary distributions has been

extended to heterogeneous systems in which servers are classified into different classes based on their capacity to process jobs but with exponential service distributions. This was done in a series of works [9, 11, 37].

Recently, the SQ(d) policy for FCFS systems with generally distributed service times was studied in [26] by using a mean-field approach. However, the analysis was restricted to the transient case, and no results were given on the stationary regime. Another important parameter of a policy is its complexity. Hence, it is of interest to generalize the mean-field analysis for occupancy based policies where the number of samples is adaptive to the system congestion level, which is what we also consider in this chapter.

In [27, 28], measure-valued processes have been used to study PS models with general JLDs in a different context. The fluid limit was obtained in [27] for the GI/GI/1 system with a single PS queue by using measure-valued processes. They used residual service times of jobs to construct measure-valued processes. The analysis was extended to the case when the customers are impatient in [28].

Summary of Contributions

Preliminary results were presented in [41, 42]. Our contributions are the following:

For the Transient Regime: The main contribution of our analysis is to show the existence of a mean-field limit for occupancy based randomized routing policies under the assumption of general JLDs. The MFEs are now PDEs, unlike the exponential case that results in ODEs. We also show asymptotic independence of servers for any given finite time. We observe that the MFEs represent dynamics of a classical single server system with Poisson job arrival process having intensity that depends on both the instantaneous occupancy of the server and the instantaneous distribution of the Markov process that models the system's dynamics. Any load balancing policy influences only the intensity of the job arrival process.

For the Stationary Regime: We show that every fixed-point of the PDEs that describe the mean-field limit corresponds to a probability measure on Z_+, is also a fixed point of the MFEs under exponential distributions having the same average job length. The fixed point is unique when the corresponding property can be shown in the exponential case. Therefore, insensitivity of the fixed-point holds if the exponential case has a unique fixed-

point. Finally, we present four occupancy based policies that have been discussed in the literature. We obtain the MFEs for these policies from our generalized mean-field analysis of occupancy based routing schemes. For the threshold based policies, Policy 2 and Policy 4 defined in Section 3.4, when job lengths are exponentially distributed, we show that the mean-field has a globally stable fixed-point. In these cases, it follows that the fixed point is the stationary distribution of the individual servers when $N \to \infty$. We also provide simulation results providing evidence for insensitivity of the stationary distribution of a server as $N \to \infty$ for all the considered policies.

For the SQ(d) policy, under the assumption of mixed-Erlang JLDs, we numerically solve the MFEs. We observe that the mean-field is not quasi-monotonic while it is in the exponential case. Furthermore, if $\pi^{(N)}$ is the stationary distribution of the empirical process and Z is a limit point of a subsequence $\{\pi^{(N_k)}\}_{k \geq 1}$, then in the limiting model as $k \to \infty$ we show that any finite set of servers are independent of each other if and only if $Z = \pi$, where π is the unique fixed-point of the mean-field. If Z is not a Dirac measure, then servers are coupled through the position of the mean-field where the initial point of the mean-field has distribution Z. This result is also true for other mean-field models.

1.1.3 A Functional Central Limit Theorem for Erlang Loss Systems Under the SQ(d) Policy

In the previous two problems stated in Sections 1.1.1 and 1.1.2, the objective was to study the mean-field and its fixed-point as an approximation of the transient and stationary distribution of a server's state when N is large. We need to characterize the resulting approximation error as a function of N. We focus on loss models with exponential JLDs because these results require a complete mean-field analysis that is lacking when general JLDs are assumed. The results that are obtained for the loss model case are suggestive of the more general type of results that can be obtained for other models.

Objective and Key Challenges

We consider large-scale multi-server loss systems with N servers where each server has capacity to process at most C jobs simultaneously at unit rate. We assume that the service times are exponentially distributed with unit mean. The central job dispatcher dispatches the incoming jobs according to the SQ(d) policy. For this model under the

assumption that the jobs arrive according to a Poisson process with rate N , from the previous works [8,10], the mean-field analysis shows that $\lim_{N \to \infty} \mathbf{X}^{(N)}(t) = \mathbf{x}(t;\mathbf{u})$; where $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ is the mean-field limit with initial point $\mathbf{u}$. Now the key question is, how do we characterize the stochastic fluctuation process $(\mathbf{X}^{(N)}(t) \quad \mathbf{x}(t;\mathbf{u}); t \quad 0)$? Moreover, if $P_{block}^{(N)}$ is the actual blocking probability in the system with N servers, can we quantify the gap between $P_{block}^{(N)} \quad {}_{C}^{d}$ as a function of N?

Similarly, we study the Halfin-Whitt regime by assuming that the arrival process is a Poisson process with rate $N \ ^{(N)}$ where $^{(N)}$ is a function of N with $\lim_{N \to \infty} \sqrt{N}(1 \quad \frac{^{(N)}}{C}) = b$ for some $b \neq 0$. In this case, how do we obtain approximations to $P_{block}^{(N)}$ when N is large? In this case, the system is heavily loaded when N is large. One way is first to obtain the asymptotic results when $N \to \infty$, and then estimate the error terms. The interest in the Halfin-Whitt regime is because, for $d = 1$, there is a phase change in the behavior of the blocking probabilities going from exponential (in C) decrease to a $\frac{1}{\sqrt{C}}$ scaling as C becomes large (see [43], [44]). As explained later in this section, we also have the similar result for the complete resource pooling case, i.e., a loss system with a single server having arrival rate of $N \ ^{(N)}$ with capacity NC, then the blocking probabilities satisfy $\frac{1}{\sqrt{N}}$ scaling as N becomes large. Although loss systems are stable for any finite average load, this phase change shows that the critically loaded regime is important for appropriate dimensioning of systems because of the change in the blocking sensitivity. Our objective is to investigate the existence of such a phenomenon for loss systems under the SQ(d) load balancing for $d \quad 2$. The main challenge is that we do not know $P_{block}^{(N)}$ while it is computed using the Erlang-B formula for the case $d = 1$.

We address the two cases $^{(N)} =$ and the Halfin-Whitt regime as special cases of our framework in which we assume that

$$^{(N)} = \quad \tilde{} \sqrt{N} \tag{1.2}$$

for $\in \mathbb{R}_+$ and $\tilde{} \in \mathbb{R}$. Here, if $\tilde{} = 0$, then we get the first case in which the arrival process rate is N . On the other hand, if $= C$ and $\tilde{} \neq 0$, then we get the Halfin-Whitt regime.

Related Literature

So far, in the literature [8,10,19], only the mean-field analysis was investigated for $^{(N)} =$. An important problem which was not studied in the literature is the quantification of the

error between the actual blocking probability for a system with large N and the asymptotic blocking probability, as a function of N.

We now discuss related existing works and provide some insights into the system performance in the Halfin-Whitt regime. The average blocking probability depends on how efficiently we use the system resources. For example, let us consider the random routing case where an arrival is routed to a randomly selected server in the Halfin-Whitt regime. Then the average blocking probability experienced by an arrival is the same as in the single server loss system with capacity C where the jobs arrive at a Poisson process with intensity $^{(N)}$. In this case, since the arrival process to each server is a Poisson process, due to reversibility, the average blocking probability is equal to $Er(\ ^{(N)}; C)$, where $Er(a; n)$ denotes the Erlang-B formula for Poisson arrivals with rate a and server's capacity n. Then since C is fixed and $Er(\ ^{(N)}; C)$ is a continuous function of N, the average blocking probability converges to $Er(C; C)$ when $N \ ! \ 1$. On the other hand, if we consider the complete resource pooling case that corresponds to the single server system with Poisson job arrival process having rate $N\ ^{(N)}$ and server capacity NC, then the average blocking probability is given by $Er(N\ ^{(N)}; NC)$. Then from [45],

$$\lim_{N \ ! \ 1} \ ^{p}\overline{N}(Er(N\ ^{(N)}; NC)) = \ ^{p}\overline{C}\frac{_{(normal)}(\tilde{\ })}{\Phi_{(normal)}(\tilde{\ })}; \tag{1.3}$$

where $_{(normal)}(\ )$ and $\Phi_{(normal)}(\ )$ denote the density and distribution functions of the standard normal distribution, respectively.

It was shown in [46] that under the JSQ scheme, we obtain the same result (1.3) as in the case of complete resource pooling, i.e.,

$$\lim_{N \ ! \ 1} \ ^{p}\overline{N}P_{block}^{(N)} = \ ^{p}\overline{C}\frac{_{(normal)}(\tilde{\ })}{\Phi_{(normal)}(\tilde{\ })}: \tag{1.4}$$

This is expected since an arrival will not be blocked from service when there is an empty spot in the system similar to the case of complete resource pooling. As a result, the average blocking probability under the JSQ policy is equal to $Er(N\ ^{(N)}; NC)$. This result was not explicitly stated in [46] for the system with N servers. Under the influence of the SQ(d) policy, an arrival could be blocked from service even if there is an empty spot in the system. Therefore, we expect a decrease in the system utilization when we use the SQ(d) policy, but it has less computational cost than the JSQ policy. However, it was shown in [46] that if d is a function of N denoted by $d^{(N)}$, and if $\lim_{N \ ! \ 1} \ ^{p}\frac{d^{(N)}}{\overline{N}\log(N)} = 1$, then we still obtain (1.4) for the SQ($d^{(N)}$) scheme.

Thus, we observe that even if we scale up the system capacity NC with N when the arrival rate of jobs scales as N $^{(N)}$, if we do not assign available resources to job requests cleverly, the system performance may not be optimal. This is because for a fixed d, an arrival could be blocked from service even if there is an empty spot in the system under the SQ(d) scheme. We would like to understand the impact of the SQ(d) policy on blocking probabilities in the Halfin-Whitt regime.

The key challenge in our model with $^{(N)} =$ $\tilde{=}^{p}\overline{N}$ is that we cannot compute $P_{block}^{(N)}$ using the Erlang formula. Moreover, we cannot obtain an explicit expression for $P_{block}^{(N)}$ as the system is non-reversible and non-tractable.

In our analysis, we first establish a functional central limit theorem (FCLT) that characterizes the fluctuations of the empirical process around the mean-field of our model and then use this result to obtain efficient approximations for $P_{block}^{(N)}$ when N is large. The limiting diffusion scaled fluctuation process is an Ornstein-Uhlenbeck (OU) process that is characterized by the mean-field limit. We then exploit the obtained FCLT to quantify the error between $P_{block}^{(N)}$ and the asymptotic blocking probability $^{d}_{C}$, where $= (\ _{i};0 \quad i \quad C)$ is the fixed-point of the corresponding mean-field and $_{C}$ is the probability that a server is fully occupied when $N \ ! \ 1$. Our results conclude that $\lim_{N \ ! \ 1}$ $^{p}\overline{N}(P_{block}^{(N)} \quad ^{d}_{C})$ goes to a limit that can be explicitly characterized in terms of $\tilde{}; \ , \ ,$ and C. More importantly, we show that $\lim_{N \ ! \ 1}$ $^{p}\overline{N}(P_{block}^{(N)} \quad ^{d}_{C})$ is non-zero if and only if $\tilde{}$ is non-zero. Graham [47] established a similar FCLT for the FCFS queueing model under the SQ(d) policy but did not exploit the FCLT result further to characterize the system performance. An FCLT approach was also used by Hunt [48] to analyze large symmetric star loss networks where asymptotic independence between nodes was established and also, the error between the actual blocking probabilities and approximate blocking probabilities was quantified.

A by-product of our result is to show that the empirical occupancy process converges to the fixed point of the MFE at a rate of $O(\not{p}\frac{1}{\overline{N}})$, a result that was also shown by Ying [49] using Stein's method and more recently by Gast [50] where a refined $O(\frac{1}{N})$ approximation was also obtained by using Stein's method under the assumption that the mean-field is exponentially stable. More importantly, in [49,50], the focus was on quantifying the mean-square error between $\mathbf{X}^{(N)}(t)$ and $\mathbf{x}(t;\mathbf{u})$ for both the transient and stationary regimes but they did not study the fluctuation process $(\mathbf{X}^{(N)}(t) \quad \mathbf{x}(t;\mathbf{u});t \quad 0)$.

In [51], it was shown that even the the gap between $\mathsf{E}\left[f(\mathbf{X}^{(N)}(t))\right]$ and $f(\mathbf{x}(t;\mathbf{u}))$ is $O(\frac{1}{N})$ for both the transient and stationary regimes where f is a twice differentiable performance function. As a result, it was concluded that for the FCFS model, the gap

between the average waiting time of the system with N servers and the asymptotic average waiting time computed as function of the fixed-point of the mean-field is $O(\frac{1}{N})$. The result of [51] can be applied to loss models to conclude that $P_{block}^{(N)} \quad \overset{d}{C} = O(\frac{1}{N})$. However, the analysis of [51] assumes $\tilde{} = 0$.

Recently, Eschenfeldt and Gamarnik [52] studied the CLT scaling of the queue occupancy process for a system of $M/M/1$ queues with the JSQ policy in the Halfin-Whitt where they showed that asymptotically the distribution concentrates on queues having up to two jobs. They did not study the stationary regime. In [53], it was shown that for the system with $M/M/1$ queues, under the SQ($d^{(N)}$) policy where $\frac{d^{(N)}}{N \log(N)} \to 1$, the limiting diffusion process is the same as in the case of the JSQ policy in the Halfin-Whitt regime. Their analysis was also restricted to the transient regime.

In [54], for the system with $M/M/1$ queues and Poisson arrival process of jobs with rate $N^{(N)}$, the impact of the SQ(d) policy was studied under the assumption that there exists a sequence $\{g_N\}$ such that $\lim_{N\to\infty} g_N = \infty$ and $\lim_{N\to\infty} g_N(1 - {}^{(N)}) = \tilde{}$ where $\tilde{} > 0$. They showed that in the stationary regime, majority of queues have queue lengths at least $\log_d(1 - {}^{(N)}) + O(1)$ and the fraction of such queues approaches one as $N \to \infty$ with probability one. Therefore an incoming job experiences a delay of at least $\log_d(1 - {}^{(N)}) + O(1)$ with probability one as $N \to \infty$. For the case $= O(\sqrt{N})$, they showed that the behavior of queues that have queue lengths less than $\log_d(1 - {}^{(N)}) + O(1)$ can be studied by the time evolution of a deterministic process even under the diffusion scale. The complete diffusion limit analysis that characterizes queues of all sizes remains an open problem.

It is worth pointing out that the result we obtain is interesting: for the Halfin-Whitt regime, the effect of the randomized SQ(d) routing results in individual loss servers that are critically loaded but whose blocking cannot be obtained from the classical Halfin-Whitt blocking limit, instead, the blocking is obtained from the mean-field or a functional law of large numbers limit of the empirical occupancy distribution.

Summary of Contributions

We characterize the asymptotics of the gap between the stochastic empirical process $(\mathbf{X}^N(t); t \quad 0)$ and the corresponding mean-field limit $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ when $N \to \infty$ for both the transient and stationary regimes. To achieve this, we study the limit of the fluctuation process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ where $\mathbf{Z}^{(N)}(t) = \sqrt{N}(\mathbf{X}^{(N)}(t) \quad \mathbf{x}(t;\mathbf{u}))$ for both the

transient and stationary regimes when $N \to \infty$. We show that the limiting diffusion process is an OU process which depends on the mean-field limit in the transient regime and on the fixed-point of the mean-field in the stationary regime. We then use this result to show that $\lim_{N\to\infty} \sqrt{N}(P^{(N)}_{block} - \frac{d}{C})$ is a finite real number (Theorem 4.7). Since $\frac{d}{C}$ is a finite constant real number, we conclude that $\lim_{N\to\infty} P^{(N)}_{block} = 1$. Preliminary results for the lightly loaded case were presented in [55].

1.2 Common Notation

In this section, we introduce the notation and terminology that is used in the rest of the dissertation. Let $\mathbb{Z}$, $\mathbb{R}$ be the set of integers and real numbers, respectively. Further, let $\mathbb{Z}_+$, $\mathbb{R}_+$ be the set of non-negative integers and non-negative real numbers, respectively.

For any given metric space E, let $\mathbb{K}_b(E); \mathbb{C}_b(E); \mathbb{C}_s(E)$ be the space of bounded measurable real valued functions, the space of bounded continuous real valued functions, and the space of continuous real valued functions with compact support, that are defined on E, respectively. Furthermore, for a metric space E of interest in this dissertation, let $\mathbb{C}^1(E)$ be the space of continuously differentiable real valued functions defined on E and let the subspace of functions in $\mathbb{C}^1(E)$ which have compact support be denoted by $\mathbb{C}^1_s(E)$. The space of bounded functions in $\mathbb{C}^1(E)$ whose first derivatives are also bounded is denoted by $\mathbb{C}^1_b(E)$. For any function $f \in \mathbb{K}_b(E)$, $h \in \mathbb{C}^1(E)$, we define

$$\|f\| = \sup_{x\in E} |f(x)|; \quad \|h\|_1 = \|h\| + \|h'\|;$$

where h' denotes the first derivative of h. The space $\mathbb{C}_b(E)$ is equipped with the uniform topology, i.e., we say that a sequence of functions $\{f_n\}_{n\geq 1}$ in $\mathbb{C}_b(E)$ converges to a function $f \in \mathbb{C}_b(E)$ if $\|f_n - f\| \to 0$ as $n \to \infty$. The space $\mathbb{C}^1(E)$ is equipped with the topology induced by the norm $\|\cdot\|_1$. For two real valued functions $f(\cdot)$ and $h(\cdot)$, we write $f(x) = o(g(x))$ as $x \to l$ if $\lim_{x\to l} \frac{f(x)}{h(x)} = 0$. Similarly, we write $f(x) = O(g(x))$ as $x \to \infty$ if there exists $a > 0$ such that $|f(x)| \leq a|g(x)|$ for sufficiently large values of x. For a given set A, $|A|$ denotes the number of elements in A and $\overline{A}$ denotes the complement of A.

For a given metric space E, let the Borel σ-algebra be denoted by $\mathcal{B}(E)$. Let the space of finite non-negative measures on E be denoted by $\mathbb{M}_F(E)$. We use the notation $\mu(B)$ and $\mu\{y\}$ to denote the measure of a Borel set $B \in \mathcal{B}(E)$ and an element $y \in E$ with respect to the measure $\mu \in \mathbb{M}_F(E)$, respectively. The space of probability measures is denoted by

$\mathsf{M}_1(E)$. For a measure $\quad 2\ \mathsf{M}_1(E)$, we write the product measure obtained from the k-fold product of $\quad$ as $\quad^k$. Also, let $\mathsf{M}_1^{(N)}(E) \quad \mathsf{M}_1(E)$ be the subspace of probability measures defined as

$$\mathsf{M}_1^{(N)}(E) = f \quad 2\ \mathsf{M}_1(E) : N \quad (B)\ 2\ \mathsf{Z}_+;\ 8\, B\ 2\ \mathsf{B}(E)g:$$

For any $\quad 2\ \mathsf{K}_b(E), \quad 2\ \mathsf{M}_F(E)$, we define

$$h\ ;\ \ i = \int_E \ (y)\ \ (dy):$$

The space of measures $\mathsf{M}_F(E)$ is equipped with the weak topology induced by the weak convergence of measures. For a signed finite measure $\ $, the mapping $\quad 7! \ h\ ;\ \ i$ is a continuous linear operator on the space of functions $\quad 2\ \mathsf{C}_b(E)$ which induces a norm for $\ $ as follows,

$$k\ \ k = \sup_{2\mathsf{C}_b(E)} \frac{jh\ ;\ \ ij}{k\ \ k}: \tag{1.5}$$

For any Borel set $B\ 2\ \mathsf{B}(E)$, let us define the indicator function I_{fBg} of B as

$$I_{fBg}(u) = \begin{cases} 1 & \text{if } u\ 2\ B; \\ 0 & \text{otherwise}: \end{cases}$$

Let $\mathbf{1}$ be the function defined such that for all $u\ 2\ E$, we have

$$\mathbf{1}(u) = 1:$$

We use bold-faced Greek letters to write random measures in Chapters 2 and 3, and we use bold-faced capital letters to write other random variables. We also use small bold-faced letters to write vectors.

Consider a Polish space H and $T\ 2\ \mathsf{R}_+$. Let $\mathsf{D}_\mathsf{H}([0;T])$ and $\mathsf{D}_\mathsf{H}([0;\ 1\ ))$ be the space of the càdlàg functions (right continuous functions with left limits) that are defined on $[0;T]$ and $[0;\ 1\ )$ with values in H, respectively. The càdlàg functions are also referred to as RCLL functions. Similarly, let $\mathsf{C}_\mathsf{H}([0;T])$ and $\mathsf{C}_\mathsf{H}([0;\ 1\ ))$ be the space of continuous functions that take values in H defined on $[0;T]$ and $[0;\ 1\ ))$, respectively. The spaces $\mathsf{D}_\mathsf{H}([0;T])$ and $\mathsf{D}_\mathsf{H}([0;\ 1\ ))$ are equipped with the Skorohod J_1-topology and hence, they are Polish spaces [56, Thorem 5.6, p.121]. Let the covariation of two local martingales $(\mathbf{M}_t^1;t \quad 0)$ and $(\mathbf{M}_t^2;t \quad 0)$ in $\mathsf{D}_\mathsf{R}([0;T])$ be denoted by $(< \mathbf{M}^1;\mathbf{M}^2 >_t;t \quad 0)$ and the quadratic variation of $(\mathbf{M}_t^1;t \quad 0)$ be denoted by $(< \mathbf{M}^1 >_t;t \quad 0)$.

For our analysis, in Chapters 2 and 3, we define a set U that contains all the possible servers' states as its elements. In these two chapters, we study H-valued stochastic processes where $\mathsf{H} = \mathsf{M}_F(\mathsf{U})$. In Chapter 4, we define a set U that contains all the possible states of $\mathbf{X}^{(N)}(t)$. In this chapter, we study H-valued stochastic processes where $\mathsf{H} = \mathsf{U}$. The considered stochastic processes are random elements defined on $(\Omega; \mathsf{F}; \mathsf{P})$ with sample paths in $\mathsf{D}_{\mathsf{H}}([0; 1))$, and are equipped with the Borel algebra generated by the open sets under the Skorohod J_1-topology [57]. We say that a sequence of random elements $\{\mathbf{Y}_n\}_{n=1}$ defined on $(\Omega; \mathsf{F}; \mathsf{P})$ converges in distribution to $\mathbf{Y}$ defined on $(\Omega; \mathsf{F}; \mathsf{P})$, if for every bounded, continuous, and real valued functional f, we have $\lim_{n \to \infty} \mathsf{E}(f(\mathbf{Y}_n)) = \mathsf{E}(f(\mathbf{Y}))$. We denote the convergence of a random element $\{\mathbf{Y}_n\}_{n=1}$ in distribution to $\mathbf{Y}$ by $\mathbf{Y}_n \Rightarrow \mathbf{Y}$. For a random element $\mathbf{Y}$, let us write its law as $\mathcal{L}(\mathbf{Y})$.

1.3 An Introduction to Mean-field Techniques

In this section, we first provide a brief introduction to the mean-field techniques. We then summarize the mean-field analysis of the loss and PS models with the SQ(d) load balancing under the assumption of exponential JLDs.

1.3.1 An Overview of the Mean-field Techniques

Many models that arise in Computer, Information, and Societal systems are examples of complex stochastic systems can be modelled as interacting particle systems. A mathematical study of these systems to obtain some meaningful insights typically requires us to first model the time evolution of the system as the time evolution of a suitable Markov process. We can then understand the system's behavior by studying the Markov process. However, due to complex interactions between particles, the exact analysis of the Markov process is not tractable when the number of nodes or particles is finite. But as the number of particles becomes large, for some models, the influence of a particle on the rest of the system is negligible. In such a situation and under certain initial conditions, we can obtain a simpler limiting model referred to as the mean-field model by taking $N \to \infty$ which can be used to approximate the behavior of the system with a large number of particles.

Consider a system with N nodes. For the sake of simplicity, we assume that the state of each particle lies in a finite set $\mathsf{U} = \{1; 2; \ldots; n\}$. Let $\mathbf{S}_i^{(N)}(t)$ be the state of the i^{th}

particle at time t. Furthermore, assume that the process $(\hat{\mathbf{Y}}^{(N)}(t); t \quad 0)$ is a Markov process where $\hat{\mathbf{Y}}^{(N)}(t) = (\mathbf{S}_1^{(N)}(t); \quad ; \mathbf{S}_N^{(N)}(t))$. Then the state of $\hat{\mathbf{Y}}^{(N)}(t)$ lies in U^N where U^N is the N-fold product of U. Let $\hat{R}(\mathbf{x}; \mathbf{y})$ be the rate of transition of $\hat{\mathbf{Y}}^{(N)}(t)$ from the state $\mathbf{x}$ 2 U^N to $\mathbf{y}$ 2 U^N. Then the time evolution of $(\hat{\mathbf{Y}}^{(N)}(t); t \quad 0)$ is governed through a set of transition rates $(\hat{R}(\mathbf{x}; \mathbf{y}); \mathbf{x}; \mathbf{y}$ 2 $\mathsf{U}^N)$.

Recall that $\mathsf{M}_1(\mathsf{U})$ is the space of probability measures on U. For $\mathbf{x} = (x_1; \quad ; x_N)$ 2 U^N, let

$$^{(N;\mathbf{x})} = \frac{1}{N} \sum_{i=1}^{N} \; _{x_i}: \tag{1.6}$$

Let $\mathsf{M}_1^{(N)}(\mathsf{U}) \quad \mathsf{M}_1(\mathsf{U})$ be defined such that

$$\mathsf{M}_1^{(N)}(\mathsf{U}) \; , \; f \; ^{(N;\mathbf{x})} : \mathbf{x} \; 2 \; \mathsf{U}^N g: \tag{1.7}$$

At time t, let $-_t^{(N)}$ be the empirical measure defined as

$$-_t^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \; \mathbf{s}_i^{(N)}(t)}: \tag{1.8}$$

Then the empirical process $(-_t^{(N)}; t \quad 0)$ is a measure-valued Markov process with values in $\mathsf{M}_1^{(N)}(\mathsf{U})$ [16, Proposition 2.3.3].

We assume that with probability one, only one particle's state jumps at any time t. For p 2 $\mathsf{M}_1(\mathsf{U})$, let $\tilde{R}(a; b; p)$ be the rate of transition of a particle from state a 2 U to b 2 U given that the empirical measure of the system is p. Then the evolution of the process is governed through a set of rates $(\tilde{R}(p); p$ 2 $\mathsf{M}_1(\mathsf{U}))$ where $\tilde{R}(p) = (\tilde{R}(a; b; p); a; b$ 2 $\mathsf{U})$ and $\tilde{R}(a; b; p)$ is given by

$$\tilde{R}(a; b; p) = \lim_{h \; ! \; 0} \frac{\mathsf{P}\left(\mathbf{S}_i^{(N)}(t+h) = b j \mathbf{S}_i^{(N)}(t)) = a; \quad _t^{(N)} = p\right)}{h}: \tag{1.9}$$

Because of the above assumption, a particle's behaviour depends on its own state and the state of $-_t^{(N)}$. That is, for every measurable function $f : \mathsf{U} \; ! \; \mathsf{R}$, $1 \quad i \quad N$, and $0 \quad s \quad t$, we have

$$\mathsf{E}\left[f(\mathbf{S}_i^{(N)}(t)) j \hat{\mathbf{Y}}^{(N)}(s)\right] = \mathsf{E}\left[f(\mathbf{S}_i^{(N)}(t)) j \mathbf{S}_i^{(N)}(s); -_s^{(N)}\right]: \tag{1.10}$$

This means all the particles influence the state of $-_t^{(N)}$ but the evolution of each particle's state is influenced by its own state and the state of $-_t^{(N)}$ at any time t. Here, the interactions

between particles take place through the empirical measure $-_t^{(N)}$. The system is a weakly interacting particle system because the contribution of each particle is of order $\frac{1}{N}$ towards the empirical measure that captures the interactions between particles. As $N \to 1$, the influence of any finite number of particles on the rest of the system is negligible.

For many systems, it is sufficient to study the empirical measure-valued process $(-_t^{(N)}; t \geq 0)$. Let $A^{(N)}(\ )$ be the generator of the Markov process $(-_t^{(N)}; t \geq 0)$. Then for $f : \mathsf{M}_1(\mathsf{U}) \to \mathsf{R}$ and $p \in \mathsf{M}_1(\mathsf{U})$,

$$A^{(N)} f(p) = \sum_{x,y \in \mathsf{U}} N p_x \left(f\left(p + \frac{1}{N}\, y \quad \frac{1}{N}\, x\right) \quad f(p)\right) \tilde{R}(x; y; p): \tag{1.11}$$

Then although $(-_t^{(N)}; t \geq 0)$ is a stochastic process for each N, this process converges to a deterministic process referred to as the mean-field limit when $N \to 1$ under certain initial conditions. Let $(\mathbf{X}^{(N)}(t); t \geq 0)$ be the empirical process corresponding to the measure-valued process $(-_t^{(N)}; t \geq 0)$. From applying the Kurtz theorem [58, Theorem 2.11], we have

Theorem 1.1. *Let $\tilde{R}(x; y; \ )$ be Lipschitz continuous for all $x; y \in \mathsf{U}$. For a constant $\mathbf{u}$ and $T > 0$, if $\mathbf{X}^{(N)}(0) \to u$ in probability as $N \to 1$, then $(\mathbf{X}^{(N)}(t); 0 \quad t \quad T) \to (\mathbf{x}(t; \mathbf{u}); 0 \quad t \quad T)$ in probability as $N \to 1$ uniformly in t where $(\mathbf{x}(t; \mathbf{u}); t \geq 0)$ is the unique solution of*

$$\frac{d\mathbf{x}(t; \mathbf{u})}{dt} = \mathbf{h}(\mathbf{x}(t; \mathbf{u})) \qquad \mathbf{x}(0; \mathbf{u}) = \mathbf{u}; \tag{1.12}$$

where $\mathbf{h}(\ ) = (h_i(\ ); i \geq 0)$ is a function that depends on the system's dynamics.

The process $(\mathbf{x}(t; \mathbf{u}); t \geq 0)$ corresponds to the distribution of a non-linear Markov process with distribution $\mathbf{x}(t; \mathbf{u})$ and infinitesimal generator $\mathbf{h}(\mathbf{x}(t; \mathbf{u}))$ at time t. Existence of the mean-field limit for an interacting particle system implies that if there is a law of large numbers effect in the system at time $t = 0$, it guarantees the functional law of large numbers (FLLN) result due to weak interactions between particles in the transient regime. The initial law of large numbers effect can be achieved if we assume all the particles' states are i.i.d. random variables.

A consequence of having the mean-field limit for an interacting particle system is the result of propagation of chaos. We now recall some results on the propagation of chaos from [47, 59].

Definition 1.1. *For $Q \in \mathsf{M}_1(\mathsf{U})$, a sequence of random variables $(\tilde{\mathbf{W}}_i^{(N)})_{1 \le i \le N}$ on U^N is said to be Q-chaotic if for any $k \ge 1$,*

$$\lim_{N \to \infty} L(\tilde{\mathbf{W}}_1^{(N)}, \ldots, \tilde{\mathbf{W}}_k^{(N)}) = Q^{\otimes k}; \tag{1.13}$$

weakly in $\mathsf{M}_1(\mathsf{U}^k)$. Similarly, if a sequence of processes $(\tilde{\mathbf{Z}}_i^{(N)}(t); t \ge 0)_{1 \le i \le N}$ is chaotic whenever $(\tilde{\mathbf{Z}}_i^{(N)}(0))_{1 \le i \le N}$ is chaotic, then we say that there is a propagation of chaos.

The Q-chaoticity for random variables implies that asymptotically any finite number of random variables are independent of each other and furthermore, each random variable's law is Q. The propagation of chaos for processes implies that any finite collection of processes are asymptotically independent of each other whenever the initial states satisfy the chaoticity.

We define the notoin of exchangeability of random variables.

Definition 1.2. *Exchangeability of Random Variables: Let $\{\mathbf{Y}_k; 1 \le k \le N\}$ denote a collection of N random variables. Then the collection is called exchangeable if the joint law of the collection is invariant under any permutation of indices, $1 \le k \le N$, of random variables, i.e., if π is a permutation on indices $\{1, 2, \ldots, N\}$, then*

$$L(\mathbf{Y}_1, \ldots, \mathbf{Y}_N) = L(\mathbf{Y}_{\pi(1)}, \ldots, \mathbf{Y}_{\pi(N)}). \tag{1.14}$$

From [59, Proposition 2.2, p.177], we have

Proposition 1.1. *For $Q \in \mathsf{M}_1(\mathsf{U})$, let a collection of random variables $(\tilde{\mathbf{W}}_i^{(N)})_{1 \le i \le N}$ on U^N be Q-chaotic, then for $\mu^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{w}_i^{(N)}}$, we have $\lim_{N \to \infty} L(\mu^{(N)}) = \delta_Q$ weakly in $\mathsf{M}_1(\mathsf{M}_1(\mathsf{U}))$. If $(\tilde{\mathbf{W}}_i^{(N)})_{1 \le i \le N}$ is exchangeable for $N \ge 1$ and $\lim_{N \to \infty} L(\mu^{(N)}) = \delta_Q$ weakly, then $(\tilde{\mathbf{W}}_i^{(N)})_{1 \le i \le N}$ is Q-chaotic.*

Assuming that $(\mathbf{S}_i^{(N)}(0))_{1 \le i \le N}$ is exchangeable, we have the exchangeability of $(\mathbf{S}_i^{(N)}(t))_{1 \le i \le N}$ for any time $t \ge 0$. In this case, from Theorem 1.1 and Proposition 1.1, any finite set of particles are asymptotically independent of each other and furthermore, each particle's distribution is $\mathbf{x}(t; \mathbf{u})$. One can establish chaoticity not only for each t but also on the entire path space. For example, in [47, Theorem 3.4], chatoicity on path space was established for the FCFS model with exponential JLDs under the SQ(d) policy. Let $\mathbf{S}_i^{(N)} = (\mathbf{S}_i^{(N)}(t); t \ge 0)$. A random variable $\bar{\mu}^{(N)}$ was defined as follows

$$\bar{\mu}^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{S}_i^{(N)}}. \tag{1.15}$$

Here, $\bar{\mu}^{(N)}$ has samples in $\mathsf{M}_1(\mathsf{D}_{\mathsf{Z}_+}([0;\infty)))$. Then Theorem 3.4 of [47] concludes that if $(\mathbf{S}_i^{(N)}(0))_{1\le i\le N}$ is $\bar{\mu}_0$-chaotic, then $(\mathbf{S}_i^{(N)})_{1\le i\le N}$ is $\bar{\mu}$-chaotic where $\bar{\mu}=(\bar{\mu}_t;t\ge 0)$ is the mean-field with initial point $\bar{\mu}_0$.

So far, we have studied both the mean-field limit and the system with N particles in the transient regime. Often, we are interested in understanding the equilibrium or stationary behavior of complex stochastic systems. For many systems of interest, it is a challenging task to find the stationary distributions or their approximations. The key question is, can we use the mean-field limit to obtain the stationary distribution of a particle in the system with a large number of particles? Here, the main issue is that we have established the mean-field limit under a strong assumption on the initial measures, but we do not know whether such assumption holds in the stationary regime or not as it depends on the system's behavior. Suppose the mean-field has a unique fixed-point, then can we use the fixed-point of the mean-field to approximate the stationary distribution of a particle in the system with large N? In the literature, there are some sufficient conditions under which the fixed-point of the mean-field can be used to approximate the stationary distribution of a particle. The relationship between the mean-field and the stationary system was studied in [49–51, 60–65] for different contexts.

In [62], by working with empirical processes, the mean-field limit was established for a generic context. Let $\mathbf{x}(\cdot;\mathbf{u})$ be the mean-field limit with initial point $\mathbf{u}$. In this case, the mean-field limit is a solution to a set of MFEs. If the system with parameter N is stationary, let $\pi^{(N)}$ be the stationary distribution of the empirical process, then the support of every limit point of $\pi^{(N)}$ is a compact set included in the Birkhoff center of the mean-field where the Birkhoff center is the closure of all the recurrent points of the mean-field [62, 65]. The Birkhoff center contains all the limit cycles and fixed-points of the mean-field. As a result, it was argued that if the mean-field has a unique fixed-point $\quad$ and furthermore if it is GAS, then $\pi^{(N)}(\quad)\quad$. As a result, as $N\to\infty$, the distribution of a particle in the stationary regime is $\quad$. In [61], it was shown that every limit of $\pi^{(N)}$ is an invariant distribution of the mean-field. That is, let $\nu\in\mathsf{M}_1(\mathsf{V})$ be a limiting point of $\pi^{(N)}$ where V is the space in which the mean-field lies, if the initial point of the mean-field has distribution ν, then for every $f\in\mathsf{C}_b(\mathsf{M}_1(\mathsf{V}))\to\mathsf{R}$,

$$\int_{\mathsf{M}_1(\mathsf{V})} f(\mathbf{x}(t;\mathbf{u}))\,d\nu(\mathbf{u})=\int_{\mathsf{M}_1(\mathsf{V})} f(\mathbf{u})\,d\nu(\mathbf{u}). \tag{1.16}$$

In [63], it was shown that if the system with N particles is reversible system, then the support of every limit point concentrates on the fixed-points of the mean-field. As

a result, if the fixed-point of the mean-field has a unique fixed-point , then ( ) . In several contexts, the existence of multiple fixed-points and limit cycles was observed for the mean-field in the literature. In [60, 61], they give examples of models for which the mean-field limit has a unique fixed-point, but the mean-field limit starting from other than the fixed-point follows a cyclic path. It was also observed that the support of lies closer to the limit cycle as N becomes large. In [65], a mean-field model was studied for which the Birkhoff center of the mean-field is a limit cycle, and it was shown that the support of lies inside the limit cycle when N becomes large. In [66], a mean-field limit was found for which there exist two fixed-points of the mean-field, and a sample path of the empirical process oscillates between two regions with each region having a fixed-point supporting the meta-stability of the system. All these examples imply that although the mean-field limit can be used to approximate the distribution of a particle in the transient regime, it requires further investigation whether the obtained mean-field is useful or not to obtain approximations for the stationary distributions.

1.3.2 Mean-field Analysis of the SQ(d) Policy

We now provide a summary of the mean-field analysis of the SQ(d) policy for the loss and PS models under the assumption of exponential JLDs.

Analysis of the Loss System

In this section, we provide a summary of the analysis given in [8, 10]. Consider a system of N loss servers each with capacity C. Assume that jobs arrive according to a Poisson process with rate N and furthermore, jobs are dispatched according to the SQ(d) policy. The jobs have exponential JLDs with unit average job length. The first step is to obtain a mathematical formulation to the time evolution of the system. The empirical process $(\mathbf{X}^{(N)}(t); t \quad 0)$ is a Markov process where $\mathbf{X}^{(N)}(t) = (X_i^{(N)}(t); i \quad 0)$ and $X_i^{(N)}(t)$ denotes the fraction of servers with at least i progressing jobs at time t. Assume that a job arrives to the system when the state of $\mathbf{X}^{(N)}(t)$ is $\mathbf{b} = (b_0; \quad ; b_C)$ indicating that the fraction of servers with at least i progressing jobs is equal to b_i for $0 \quad i \quad C$. As servers are sampled with replacement uniformly at random, the probability that a sampled server has i jobs is equal to $b_i \quad b_{i+1}$. Since the dispatcher samples d servers with replacement to dispatch the arrival, the destination server of the job will have occupancy n with probability $b_n^d \quad b_{n+1}^d$. Then the probability that a tagged server with n jobs is selected as the destination server

is equal to $\frac{1}{N}\frac{b_n^d - b_{n+1}^d}{b_n - b_{n+1}}$. As a result, the arriving job joins a particular server with a certain probability that depends on the occupancy of that server and the empirical measure $\mathbf{b}$. Hence, the time evolution of each server depends on its state and the state of the empirical measure. Therefore each server influences $\mathbf{X}^{(N)}(t)$ and an arrival event to each server is influenced by the server's state and the state of $\mathbf{X}^{(N)}(t)$. Hence, the model is a weakly interacting particle system.

We now present results on the mean-field analysis of the model from [10]. The underlying space is

$$\mathsf{U} = f(b_i; 0 \quad i \quad C) : b_0 = 1 \quad b_1 \quad b_2 \qquad b_C \quad b_{C+1} = 0g: \tag{1.17}$$

The space U is equipped with the euclidean metric.

Theorem 1.2. *For* $\mathbf{u}$ *2* U*, if* $\mathbf{X}^{(N)}(0)$ *)* $\mathbf{u}$ *as* N *! 1 , then* $(\mathbf{X}^{(N)}(t); t \quad 0)$ *)* $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ *as* N *! 1 where* $(\mathbf{x}(t;\mathbf{u}); t \quad 0) = (x_n(t;\mathbf{u}); t \quad 0; 0 \quad n \quad C)$ *is the unique solution to the following equations:*
for $\mathbf{h}(\mathbf{x}(t;\mathbf{u})) = (h_n(\mathbf{x}(t;\mathbf{u})); 0 \quad n \quad C)$,

$$\mathbf{x}(0;\mathbf{u}) = \mathbf{u}; \quad \frac{dx_n(t;\mathbf{u})}{dt} = h_n(\mathbf{x}(t;\mathbf{u})); \tag{1.18}$$

where

$$h_0(\mathbf{x}(t;\mathbf{u})) = 0; \tag{1.19}$$

and for n 1,

$$h_n(\mathbf{x}(t;\mathbf{u})) = (x_{n\ 1}^d(t;\mathbf{u}) \quad x_n^d(t;\mathbf{u})) \quad n(x_n(t;\mathbf{u}) \quad x_{n+1}(t;\mathbf{u})); \tag{1.20}$$

with $x_0(t;\mathbf{u}) = 1$ *and* $x_{C+1}(t;\mathbf{u}) = 0$*. The deterministic process* $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ *is referred to as the mean- eld limit and the equations* (1.18)-(1.20) *are referred to as the mean- eld equations with the initial point* $\mathbf{u}$*.*

The mean-field $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ has a unique GAS fixed-point $= (\ _n; 0 \quad n \quad C)$ with $\ _0 = 1$. The proof that the fixed-point is GAS mainly uses the property that the mean-field is quasi-monotonic.

Definition 1.3. *Quasi-Monotonicity: Let* $\mathbf{r}(t;\mathbf{u})$ *be a deterministic process de ned on* R^n*,* n 1*, that is a solution to a set of ODEs, for* $\mathbf{f}(\ ) = (f_i(\ ); i \quad 1)$,

$$\frac{\mathbf{r}(t;\mathbf{u})}{dt} = \mathbf{f}(\mathbf{r}(t;\mathbf{u})); \quad \mathbf{r}(0;\mathbf{u}) = \mathbf{u}:$$

Then the process is said to satisfy quasi-monotonicity if $\mathbf{u}$ $\mathbf{v}$ *element-wise, then* $\mathbf{r}(t;\mathbf{u})$ $\mathbf{r}(t;\mathbf{v})$ *element-wise.*

It can be shown that the following exchange of limits holds

$$\lim_{N \to \infty} \lim_{t \to \infty} \mathbf{X}^{(N)}(t) = \lim_{t \to \infty} \lim_{N \to \infty} \mathbf{X}^{(N)}(t).$$

Using Theorem1.1, under the assumption of the exchangeability of initial servers' states, we can show the independence of any finite set of servers as $N \to \infty$. Also, for a server, $\mathbf{x}(t;\mathbf{u})$ denotes the distribution at time t and denotes the stationary distribution, as $N \to \infty$. As a result, denotes the stationary probability that a server is fully occupied when $N \to \infty$. Since the dispatcher samples d servers when a job arrives, the stationary average blocking probability of a job as $N \to \infty$ is equal to , where we use the fact that a sampled server will have C jobs with probability .

We now focus on finding the fixed-point . For loss models, a closed form for is not known while it is known for the case of the FCFS models [7,12]. We can find numerically as explained below. The fixed-point is the unique solution to the following equations

$$(\quad) = (\quad) \tag{1.21}$$

for 1 and $= 0$. Then from (1.21), we can also write

$$\frac{(\quad)}{(\quad)}(\quad) = (\quad) \tag{1.22}$$

for 1 and $= 0$. Let us define $\hat{} = \frac{(\quad)}{(\quad)}$. Then from (1.22), is the stationary distribution of the single server loss model with a Poisson arrival process of jobs having rate $\hat{}$ when there are progressing jobs, and is the probability that the server has at least progressing jobs. Let $\mathsf{M}_1(f0;1; \quad ;Cg)$ be the set of probability measures on $f0;1; \quad ;Cg$. Then from [10], the fixed-point can be computed using the formula for the stationary distribution of a single server loss system with state-dependent arrival rates. We first define two mappings, $\Theta : \mathsf{M}_1(f0;1; \quad ;Cg) \to \mathsf{R}_+^{C+1}$ and $\widehat{\Xi} : \mathsf{R}_+^{C+1} \to \mathsf{M}_1(f0;1; \quad ;Cg)$ that are used in computing . For every $(p_0; \quad ;p_C) \in \mathsf{M}_1(f0;1; \quad ;Cg)$, there exists $(r_0; \quad ;r_C) \in \mathsf{R}_+^{C+1}$ such that

$$\Theta((p_0; \quad ;p_C)) = (r_0; \quad ;r_C);$$

where

$$r_n = \frac{((\sum_{j=n}^{C} p_j)^d \quad (\sum_{i=n+1}^{C} p_i)^d)}{((\sum_{j=n}^{C} p_j) \quad (\sum_{i=n+1}^{C} p_i))}.$$

Similarly, for every $(b_0; \quad ;b_C) \in \mathsf{R}_+^{C+1}$, there exists $(a_0; \quad ;a_C) \in \mathsf{M}_1(f0;1; \quad ;Cg)$ such that

$$\widehat{\Xi}((b_0; \quad ;b_C)) = (a_0; \quad ;a_C);$$

where

$$a_n = \left(\prod_{i=1}^{n} \left\{ \frac{b_{i-1}}{i} \right\} \right) a_0$$

for $n \geq 1$ and $\sum_{i=0}^{C} a_i = 1$.

Lemma 1.1. *Since the mean-field is GAS, the mapping $\widehat{\Xi}(\Theta)$ has a unique fixed-point denoted by $\theta = (\theta_n; 0 \leq n \leq C)$. Furthermore, the unique fixed-point of the mean-field is given by, $n \geq 1$,*

$$\theta_n = \sum_{j=n}^{C} \theta_j.$$

Analysis of the PS System

In this section, we consider the mean-field analysis of the PS model with exponential JLDs from [7,9,11]. In [7], the FCFS model with exponential JLDs was studied. Heterogeneous PS model in which servers are categorized into different classes based on their capacities was studied in [9,11]. We provide a summary of the analysis given in [9,11] for our model. We first define the underlying spaces

$$U^{\emptyset} = \{(b_i; i \geq 0) : b_0 = 1; b_i \geq b_{i+1}; \forall i \in Z_+\} \tag{1.23}$$

$$U = \{(b_i; i \geq 0) : b_0 = 1; b_i \geq b_{i+1}; \forall i \in Z_+; \sum_{i \geq 0} b_i < \infty\}. \tag{1.24}$$

We equip the space $U^{\emptyset}$ with the metric $\rho(\cdot, \cdot)$ defined as

$$\rho(\mathbf{u}, \mathbf{v}) = \sup_{i \geq 1} \frac{|u_i - v_i|}{i+1}, \tag{1.25}$$

for all $\mathbf{u} = (u_i; i \geq 0)$, $\mathbf{v} = (v_i; i \geq 0)$ in U. The space $U^{\emptyset}$ equipped with the metric ρ is compact, complete, and separable.

A mean-field limit exists in this case also similar to the loss model but now the mean-field lies in a countable infinite dimensional space.

Theorem 1.3. *For $\mathbf{u} \in U^{\emptyset}$, if $\mathbf{X}^{(N)}(0) \Rightarrow \mathbf{u}$ as $N \to \infty$, then $(\mathbf{X}^{(N)}(t); t \geq 0) \Rightarrow (\mathbf{x}(t; \mathbf{u}); t \geq 0)$ as $N \to \infty$ where $(\mathbf{x}(t; \mathbf{u}); t \geq 0) = (x_n(t; \mathbf{u}); t \geq 0; n \geq 0)$ is the unique solution to the following equations:*
for $\mathbf{h}(\mathbf{x}(t; \mathbf{u})) = (h_n(\mathbf{x}(t; \mathbf{u})); n \geq 0)$,

$$\mathbf{x}(0; \mathbf{u}) = \mathbf{u}; \quad \frac{dx_n(t; \mathbf{u})}{dt} = h_n(\mathbf{x}(t; \mathbf{u})); \tag{1.26}$$

where

$$h_0(\mathbf{x}(t;\mathbf{u})) = 0; \tag{1.27}$$

and for $n \quad 1$,

$$h_n(\mathbf{x}(t;\mathbf{u})) = \quad (x_n^d \ _1(t;\mathbf{u}) \quad x_n^d(t;\mathbf{u})) \quad (x_n(t;\mathbf{u}) \quad x_{n+1}(t;\mathbf{u})); \tag{1.28}$$

with $x_0(t;\mathbf{u}) = 1$. *The deterministic process* $(\mathbf{x}(t;\mathbf{u}); t \quad 0)$ *is referred to as the mean- eld limit and the equations* (1.26)-(1.28) *are referred to as the mean- eld equations with the initial point* $\mathbf{u}$.

It was then shown that there exists a unique fixed-point of the mean-field $= (\ _i; i \quad 0)$ that belongs to the set U. Furthermore, the quasi-monotonicty of the mean-field was used to show $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ for all $\mathbf{u}\ 2\ \mathsf{U}$. After that the stochastic system was shown to be stable for < 1 and let $^{(N)}$ be the unique stationary distribution of the stochastic process $\mathbf{X}^{(N)}$. Since the space $\mathsf{U}^{\emptyset}$ is compact, from Prohorov's theorem [57], $f\ ^{(N)}g_{N\ 1}$ is tight. Therefore, it is sufficient to show that every limit point of $f\ ^{(N)}g_{N\ 1}$ coincides with . Let be the limit point of a converging subsequence $f\ ^{(N_k)}g_{k\ 1}$, then it was shown that $\mathsf{E}\ [\sum_{i\ 1} \mathbf{V}_i] < 1$ where $(\mathbf{V}_i; i \quad 0)$ is random quantity with distribution . As a consequence, we have $(\mathsf{U}) = 1$. Since is an invariant distribution of the mean-field and $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ for all $\mathbf{u}\ 2\ \mathsf{U}$, it implies $=$. Having $^{(N)}\ )$ implies that any finite set of servers are asymptotically independent in the stationary regime and furthermore, each server's distribution coincides with the fixed-point of the mean-field. More importantly, the fixed-point is given by $_i = \ ^{\frac{d^i\ 1}{d\ 1}}$, $i \quad 1$.

1.4 Outline

The rest of the dissertation is organized as follows. In Chapter 2, we provide the mean-field analysis of the first problem discussed in Section 1.1.1. We then study the second problem stated in Section 1.1.2 in Chapter 3. The analysis of the FCLT for the loss model introduced in Section 1.1.3 is given in Chapter 4. We conclude in Chapter 5 with a discussion on future work. We provide additional background material in the Appendix.

Chapter 2

Insensitivity of the Mean-field Limit of Loss Systems Under SQ(d) Load Balancing

In this chapter, we study a large multi-server loss model under a randomized dynamic load balancing scheme when the service time distributions are general with finite mean. In particular, we consider the SQ(d) load balancing scheme according to which an incoming job is dispatched to the server with the least number of progressing jobs among d randomly chosen servers. Previous works have addressed the exponential service time case when the number of servers goes to infinity, giving rise to a mean-field model. The fixed-point of the mean-field equations (MFEs) was seen to be insensitive to the service time distribution in simulations, but no proof was available. While insensitivity is well known for loss systems models even with state-dependent inputs such models belong to the class of linear Markov models. In the context of the SQ(d) load balancing, the resulting model belongs to the class of nonlinear Markov processes (processes whose generator itself depends on the distribution) for which traditional arguments do not directly apply. Showing insensitivity to the general service time distributions has thus remained an open problem. In this case, obtaining the MFEs poses a challenge due to the resulting Markov description of the system being in positive orthant as opposed to a finite chain in the exponential case. In this chapter, we first obtain the MFEs and then show that the MFEs have a unique fixed-point that corresponds to a distribution of occupancy coinciding with the fixed-point in the exponential case, thus establishing insensitivity. The approach is via a measure-valued Markov process representation and the martingale problem to establish the mean-field

limit.

Organization of the Chapter: The rest of the chapter is organized as follows: Section 2.1 describes the system model. In Section 2.2, we introduce the notation used in the rest of the chapter. In Section 2.3, we derive a measure-valued representation for the state of the system. The main results of the chapter are given in Section 2.4. We then establish the mean-field limit in Section 2.5. In Section 2.6, we prove the result on the uniqueness of the fixed-point of the MFEs and show that the fixed-point is insensitive to the distribution, i.e., it depends only on the mean service time. In Section 2.7, we provide numerical results that suggest the global asymptotic stability of the fixed-point of the MFEs and hence, the relation (1.1) indeed holds. In Section 2.8, we give proofs of the main results stated in Section 2.4. Finally, Section 2.9 concludes the chapter with some remarks and generalizations.

2.1 System Model

We consider a system consisting of a large number N of parallel servers. Jobs arrive according to a Poisson process with rate N , and the job lengths are assumed to be i.i.d. from a general distribution $G(\,)$ defined on R_+. Furthermore, we assume that the distribution $G(\,)$ has a continuous density function $g(\,)$. A central job dispatcher routes an incoming job to a server according to the SQ(d) policy described below. We assume that each server has capacity to process up to a number C of jobs simultaneously, and each accepted job is processed at a unit rate. At any time t, if a server is currently serving i jobs, then we say that this server has occupancy i and vacancy $C \quad i$ at time t. If an incoming job is routed to a server with occupancy C, then the routed job is blocked from service, and it is discarded from the system immediately. Otherwise, the processing of the job begins immediately, and it is processed at a unit rate.

Definition 2.1.1. *SQ(d) or Power-of-d load balancing policy: An incoming job is routed to the server with the minimum occupancy among d servers that are selected randomly with replacement. Ties among servers are broken by choosing a server uniformly at random. The randomly chosen d servers are referred to as the potential destination servers and the server to which a job is routed called the destination server.*

Remark 2.1. *In De nition 2.1.1, we assume sampling with replacement because of notational convenience, and it can be shown that the asymptotic results that are of interest in*

this chapter are not a ected whether we sample with or without replacement [32, 47].

We assume that the service times have finite mean $\frac{1}{}$. We make an assumption that $\overline{G}(\ )$ is supported on $[0; 1)$ where $\overline{G}(\ )$ denotes the complementary distribution. Otherwise, if M denotes the finite support of $\overline{G}(\ )$, then our analysis easily extends to this case by using the fact that the ages of jobs are at most M. The hazard rate function of $G(\ )$ is defined as

$$(x) = \frac{g(x)}{\overline{G}(x)} = \frac{g(x)}{1 \quad G(x)} \tag{2.1}$$

for $x \ 2 \ [0; 1)$. The hazard rate function indicates the instantaneous rate at which the service of a job ends. More precisely, a job with age y (where y denotes the time since its arrival) at time t exits the server in the interval $[t; t + dt)$ with probability $(y)dt + o(dt)$.

Assumption 2.1.1. *The hazard rate function satis es* $2 \ C_b(R_+)$*:*

Remark 2.2. *The Assumption 2.1.1 is true for several classes of distributions such as Phase-Type distributions, Gamma distributions, Log-Normal distributions, and any Pareto distribution with nite mean.*

2.2 Additional Notation and Terminology

In this section, we introduce the required additional notation and terminology that is specific to this chapter.

In this chapter, we define the term 'age' for a progressing job as the amount of time elapsed since its arrival. To model the dynamics of an Erlang loss system with capacity C for each server by a Markov process, we define the state of each server as $(n; a_1; a_2; \quad ; a_n)$ where n denotes the number of jobs that are in progress at the server and a_i denotes the age of the i^{th} progressing job. We now define a space U that was used in earlier works to study queueing models with general service time distributions by using the classical supplemental variable method [67, 68] such that it contains all the possible servers' states as its elements. The space U is defined as

$$U = [_{n=0}^{C} U_n;$$

where $U_0 = f0g$ and an element in U_n for $n \quad 1$ is of the form $(n; a_1; : : : ; a_n)$ where $1 \quad n \quad C$ and $a_i \ 2 \ R_+$. The state of a server that has n jobs belongs to the space

U_n. Here, one might omit the variable n and consider just $(a_1; \quad ; a_n)$ to denote a server's state, but such a representation does not account for idle servers while (0) is the state of idle servers in our representation. Furthermore, the variable n, directly gives us information about the number of progressing jobs at a server which changes upon every arrival and departure. Hence, it is convenient to work with the state representation that has a variable to denote the number of progressing jobs at a server.

It is also possible to define an element in U_n by $(n; a_1; \quad ; a_n; 0; \quad ; 0)$ of size $C+1$. This allows us to have constant size of $C+1$ for an element in U. Note that the zeros in the state $(n; a_1; \quad ; a_n; 0; \quad ; 0)$ act as dummy variables as there are only n jobs. Hence, to make it simple, we consider that an element in U_n is of the form $(n; a_1; \quad ; a_n)$ with size $n+1$. Without loss of generality, we refer to an element in the set U by $\mathbf{u}$ and an element in the set U_n by $\mathbf{u}_n$. Note that we have $\mathbf{u}_0 = 0$. For $\mathbf{y}_n = (n; y_1; \ldots; y_n); \mathbf{z}_m = (m; z_1; \ldots; z_m)$, we define the metric $d_{\mathsf{U}}(\mathbf{y}_n; \mathbf{z}_m)$ as

$$d_{\mathsf{U}}(\mathbf{y}_n; \mathbf{z}_m) = \begin{cases} \sum_{i=1}^{n} j y_i \quad z_i j & \text{if } n = m; \\ jm \quad nj & \text{otherwise.} \end{cases}$$

For $n \quad 1$, U_n is a complete, separable, and Polish space. Furthermore, U is a Polish space as it is a union of a set of disjoint Polish spaces. Also, U is separable and complete.

For $(n; u_1; \ldots; u_n)$ 2 U_n and $y \quad 0$, we use the following notation

$$\begin{aligned} \mathbf{u}_n &= (n; u_1; \quad ; u_n); \\ \mathbf{u}_n{}^j &= (n \quad 1; u_1; \quad ; u_{j \quad 1}; u_{j+1}; \quad ; u_n); \\ (\mathbf{u}_n^j; y) &= (n+1; u_1; \quad ; u_{j \quad 1}; y; u_j; \quad ; u_n); \\ (\mathbf{u}_n{}^j; y) &= (n; u_1; \quad ; u_{j \quad 1}; y; u_{j+1}; \quad ; u_n). \end{aligned}$$

A function $f : \mathsf{U} \not\to \mathsf{R}$ is said to be differentiable if for every $n \quad 1$, the function $\frac{@f(\mathbf{u}_n)}{@u_i}$ exists for all $1 \quad i \quad n$ at every $\mathbf{u}_n$ 2 U_n. As a result, from (1.6), the function $I_{f\mathsf{U}_n g}; n \quad 1$, is differentiable. For a differentiable function $f : \mathsf{U} \not\to \mathsf{R}$, we define

$$kf^{\prime}k = \max_{n \quad 1} \left(\sup_{\mathbf{u}_n 2 \mathsf{U}_n} \left(\max_{1 \quad i \quad n} \left| \frac{@f(\mathbf{u}_n)}{@u_i} \right| \right) \right).$$

Further, for a differentiable function $f : \mathsf{U} \not\to \mathsf{R}$, let the function $r_{\ 1} f$ be defined as

$$r_{\ 1} f(n; u_1; \quad ; u_n) = r f \quad \mathbf{1} = \sum_{i=1}^{n} \frac{@f(\mathbf{u}_n)}{@u_i}.$$

A measure $\in \mathcal{M}_F(\mathbb{U})$ when it is restricted to $\mathbb{U}_0$ is a Dirac measure at $\{0\}$ satisfying $(\mathbb{U}_0) = (\{0\})$. We say that a measure is absolutely continuous with respect to the Lebesgue measure if $(\{\mathbf{x}_n\}) = 0$ at every $\mathbf{x}_n \in \mathbb{U}_n$ for all $n \geq 1$. For any Borel measurable function f that is defined on $\mathbb{U}$, we define

$$\langle f, \rangle = f(0)\ (\{0\}) + \sum_{n=1}^{\infty} \int_{\mathbb{U}_n} f(\mathbf{z}_n)\ (d\mathbf{z}_n).$$

We now define a function $: \mathbb{U} \to \mathbb{R}$ as follows:

$$(\mathbf{x}_n) = \begin{cases} \sum_{i=1}^{n} x_i & \text{if } n \geq 1; \\ 0 & \text{otherwise.} \end{cases}$$

For $b \geq 0$, let $_b^+ : \mathbb{U} \to \mathbb{U}$ be the transition operator defined as

$$_b^+(\mathbf{x}_n) = \begin{cases} (n; x_1 + b; \quad ; x_n + b) & n \geq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for any $b \geq 0$ and $f \in \mathcal{K}_b(\mathbb{U})$, let the mapping $_b : \mathcal{K}_b(\mathbb{U}) \to \mathcal{K}_b(\mathbb{U})$ be defined as

$$_y f(\mathbf{u}) = f(\ _y^+ \mathbf{u}).$$

Also, for $b \geq 0$, let $_b \in \mathcal{M}_F(\mathbb{U})$ be the measure defined such that for any Borel set $B \in \mathcal{B}(\mathbb{U})$, we have

$$_b\ (B) = \ (\ _b^+(B)).$$

For $\in \mathcal{M}_F(\mathbb{U})$, the measure $_b \in \mathcal{M}_F(\mathbb{U})$ satisfies

$$\langle\ _b\ , f \rangle = \langle\ ,\ _b f \rangle \tag{2.2}$$

for all $f \in \mathcal{K}_b(\mathbb{U})$ and the existence of the unique measure $_b$ follows from the Riesz-Markov-Kakutani theorem [69, Theorem 2.14] stated in Theorem A.4. By using (2.2), we map a change in the system state in a given small time interval to an equivalent change in the function f. By working with the class of functions of the type $_t^{(N)} \to \langle\ _t^{(N)};\ \rangle$ for $\in \mathcal{C}_b^1(\mathbb{U})$, we obtain the generator of the Markov process $(_t^{(N)}; t \geq 0)$.

2.3 System Dynamics

In this section, we first define a system state descriptor. Using this, we then describe the time evolution of the system.

We index a sequence of systems by N that denotes the total numbers of servers. Incoming jobs arrive according to a Poisson process with rate N , and the job lengths are i.i.d. from a common distribution $G(\ )$ defined on R_+. The state of a server is written as $\mathbf{a}_n = (n; a_1; \quad ; a_n)\ 2\ \mathsf{U}$ when there are n progressing jobs and the i^{th} job has age a_i for $1 \quad i \quad n$. A server with a state $\mathbf{a}_n$ can be viewed as a particle with the given state. Therefore the system evolution can be considered as the evolution of a system with N particles where the interactions between particles take place while routing an arrival according to the SQ(d) load balancing policy.

The age of a progressing job indicating the time elapsed since its arrival increases linearly with time at a unit rate until its service is completed. We next describe the possible state of a server at time $t+h$ ($h > 0$) given that it has state $\mathbf{a}_n$ at time t. When h is small enough, in the interval $[t; t+h)$, the probability of having multiple events of arrivals or departures is negligible. In the interval $[t; t+h)$, if there is no arrival or departure at the given server, then the server's state will be equal to $\ _h^+(\mathbf{a}_n)$ at time $t+h$. On the other hand, if the i^{th} job expires in the interval $[t; t+h)$, then the server's state will be equal to $\ _h^+(\mathbf{a}_n{}^i)$ at time $t+h$. Considering arrivals, suppose there is an arrival that is accepted at the given server at time $t+r$ ($0 \quad r < h$), then the arriving job chooses its position uniformly at random out of $n+1$ possible positions. If the arrived job chooses the j^{th} position, then the server's state will be equal to $((\ _h^+(\mathbf{a}_n))^j; h \quad r)$ at time $t+h$.

Let $\mathbf{S}_i^{(N)}(t)\ 2\ \mathsf{U}$ be the random variable that indicates the state of the server i at time t. Although one can think of considering $(\mathbf{S}_1^{(N)}; \quad ; \mathbf{S}_N^{(N)}(t))$ to denote the system state at time t which is a Markovian representation of the system, the dimension of this state space increases with N as $N\ !\ 1$ which is inconvenient to work with since our focus of interest is to study the asymptotic behavior of the system as $N\ !\ 1$. Hence, we consider an alternative simple system state representation that can be used to describe the system evolution as the evolution of a Markov process. Note that the system is symmetric to the servers as they have the same parameters, and the servers' identities do not play any role in the time evolution of the system. Therefore, to model the system evolution by a Markov process, we will show that it is enough to keep track of the number of servers that lie in each state $\mathbf{u}\ 2\ \mathsf{U}$ in order to establish the mean-field limit. Measure-valued Markov processes have also been used to study other interacting particle systems as in [70–72] where the state of each particle lies in the space R^n, $n > 1$. Following these works, we consider the following system state descriptor.

Definition 2.3.1. *System state descriptor: At time t, the state descriptor of the system*

with index N is a random measure given by

$$ {}_t^{(N)} = \sum_{i=1}^{N} \ \mathbf{s}_i^{(N)}(t)}. \tag{2.3} $$

The interpretation of ${}_t^{(N)}$ is that for any measurable function f defined on U, we have

$$ h \ {}_t^{(N)}; f i = \sum_{i=1}^{N} f(\mathbf{S}_i^{(N)}(t)). $$

At time t, conditioned on servers' states say $\mathbf{S}_i^{(N)}(t) = \mathbf{s}_i(t)$, i 1, the state of the system can be represented by a measure defined as

$$ = \sum_{i=1}^{N} \ {}_{\mathbf{s}_i(t)}. \tag{2.4} $$

For ${}_t^{(N)} =$, an element $\mathbf{y}$ 2 U is an atom of if there exists at least one server with the state $\mathbf{y}$ at time t. The mass of an atom of is equal to the number of servers lying at that atom at time t. As a result, since the number of interacting particles in the system is equal to N, the measure defined on U contains a finite number of atoms which is bounded by N. If all the servers lie in different states, then the number of atoms is equal to N. Otherwise, the number of atoms is less than N. For , let $V(t)$ be the number of atoms at time t and let the i^{th} atom be denoted by $\mathbf{v}^{(i)}(t)$. Further, let the mass of the atom $\mathbf{v}^{(i)}(t)$ be denoted by $a^{(i)}(t)$. Here, $a^{(i)}(t)$ denotes the number of servers that lie in the state $\mathbf{v}^{(i)}(t)$ at time t where $a^{(i)}(t)$ 1. For given time t, from (2.4), we can also write as

$$ = \sum_{i=1}^{V(t)} a^{(i)}(t) \ {}_{\mathbf{v}^{(i)}(t)}. \tag{2.5} $$

For any Borel set B 2 $\mathsf{B}(\mathsf{U})$, the number of servers with ages lying in the set B is equal to ${}_t^{(N)}(B) = (B) = h \ ; I_{fBg} i$. We now define the measure of an element $\mathbf{y}_n = (n; y_1; \ ; y_n)$ as below. Let $B \ (\mathbf{y}_n) = f(n; r_1; \ ; r_n) : y_i \ r_i < y_i + \ ; 1 \ i \ ng$. Then as in [27], we define

$$ (f\mathbf{y}_n g) = \lim_{\prime 0} \ (B \ (\mathbf{y}_n)). \tag{2.6} $$

Essentially, $(f\mathbf{y}_n g)$ indicates the number of servers with state $\mathbf{y}_n$ at time t and can be viewed as an occupation count. The notation $d \ (\mathbf{y}_n)$ denotes the number of servers with state lying in the interval $[\mathbf{y}_n; \mathbf{y}_n + d\mathbf{y}_n)$, where $d\mathbf{y}_n = (dy_1; \ ; dy_n)$ and $\mathbf{y}_n + d\mathbf{y}_n$ is the

vector addition of $\mathbf{y}_n$ and $d\mathbf{y}_n$. If there is no server lying in the state $\mathbf{y}_n$ at time t, then $(f\mathbf{y}_n g) = 0$, otherwise $\mathbf{y}_n$ is an atom with mass $(f\mathbf{y}_n g)$. Note that the number of servers that have n progressing jobs at time t is given by $(U_n) = h\ ; I_{fU_n g} i$:

We now obtain the probability that the destination server of an arrival lies in a particular state.

Lemma 2.1. *At time t, given that the system state is , i.e., $^N_t =$, under the SQ(d) load balancing policy, the probability that the destination server of an arrival at time t lies in the state $\mathbf{z}_n = (n; z_1;\ ; z_n)$ where $\mathbf{z}_n$ is an atom of is given by*

$$p_r(\ : \mathbf{z}_n) = \left\{ \frac{(f\mathbf{z}_n g)}{N} \right\} \frac{f(\overline{R}_n(\frac{}{N}))^d \quad (\overline{R}_{n+1}(\frac{}{N}))^d g}{f\overline{R}_n(\frac{}{N}) \quad \overline{R}_{n+1}(\frac{}{N})g}; \tag{2.7}$$

where

$$\overline{R}_n\left(\frac{}{N}\right) = \sum_{j:j=n}^{C} \frac{(U_j)}{N} \tag{2.8}$$

denotes the fraction of the servers with at least n jobs.

Proof. When a potential destination server is chosen uniformly at random from N servers, it will have state $(n; z_1; : : : ; z_n)$ with probability $\frac{(f(n;z_1;\ ;z_n)g)}{N}$. Suppose out of the d potential destination servers, say j servers have occupancy n and the remaining $d \quad j$ servers have occupancy at least $n+1$. Further, out of the j $(j \quad 1)$ potential destination servers with occupancy n, assume r $(r \quad 1)$ servers lie in the state $\mathbf{z}_n$. Then the probability that the destination server is a server with state $\mathbf{z}_n$ is given by

$$\binom{d}{j}\binom{j}{r}\binom{r}{j}\left(\frac{(f\mathbf{z}_n g)}{N}\right)^r \left(\frac{(fU_n g) \quad (f\mathbf{z}_n g)}{N}\right)^{j \ r} \left(\sum_{i:\ i=n+1}^{C} \frac{(U_i)}{N}\right)^{d \ j} :$$

Finally, by summing over all the possible values of j $(j \quad 1)$ and r $(r \quad 1)$, we have

$$\sum_{j=1}^{d}\sum_{r=1}^{j}\binom{d}{j}\binom{j}{r}\binom{r}{j}\left(\frac{(f\mathbf{z}_n g)}{N}\right)^r \left(\frac{(fU_n g) \quad (f\mathbf{z}_n g)}{N}\right)^{j \ r} \left(\sum_{i:i=n+1}^{C} \frac{(U_i)}{N}\right)^{d \ j}$$

$$= \sum_{j=1}^{d}\binom{d}{j}\frac{1}{j}\left(\sum_{i:\ i=n+1}^{C} \frac{(U_i)}{N}\right)^{d \ j}\left(\frac{(fU_n g)}{N}\right)^j$$

$$\left[\sum_{r=1}^{j} r\binom{j}{r}\left(\frac{\frac{(f\mathbf{z}_n g)}{N}}{\frac{(fU_n g)}{N}}\right)^r \left(\frac{\frac{(fU_n g)}{N} \quad \frac{(f\mathbf{z}_n g)}{N}}{\frac{(fU_n g)}{N}}\right)^{j \ r}\right]:$$

The term inside the square bracket in the above equation is the average of a binomial random variable and hence, it is equal to $j \left\{ \frac{\left(\frac{(f\mathbf{z}_n g)}{N}\right)}{\left(\frac{(f\mathbb{U}_n g)}{N}\right)} \right\}$. As a result, the above expression simplifies to

$$p_r(\quad : \mathbf{z}_n) = \left\{ \frac{\left(\frac{(f\mathbf{z}_n g)}{N}\right)}{\left(\frac{(f\mathbb{U}_n g)}{N}\right)} \right\} \sum_{j=1}^{d} \binom{d}{j} \left(\sum_{i:\ i=n+1}^{C} \frac{(\mathsf{U}_i)}{N} \right)^{d\quad j} \left(\frac{(f\mathsf{U}_n g)}{N} \right)^{j} :$$

We can further write

$$p_r(\quad : \mathbf{z}_n) = \left\{ \frac{\left(\frac{(f\mathbf{z}_n g)}{N}\right)}{\left(\frac{(f\mathbb{U}_n g)}{N}\right)} \right\} \left[\left(\sum_{j=0}^{d} \binom{d}{j} \left(\sum_{i:\ i=n+1}^{C} \frac{(\mathsf{U}_i)}{N} \right)^{d\quad j} \left(\frac{(f\mathsf{U}_n g)}{N} \right)^{j} \right) \quad \left(\sum_{i:i=n+1}^{C} \frac{(\mathsf{U}_i)}{N} \right)^{d} \right] :$$

After simplifications, we get (2.7). □

Remark 2.3. *We can also interpret the expression of* $p_r(\quad : \mathbf{z}_n)$ *as follows: The probability that all the potential destination servers have occupancy at least* n *and there exists at least one potential destination server with occupancy* n*, is equal to* $(\overline{R}_n(\frac{\cdot}{N}))^d \quad (\overline{R}_{n+1}(\frac{\cdot}{N}))^d$*. From the SQ*$(d)$ *policy, the probability that the destination server has occupancy* n *is equal to* $(\overline{R}_n(\frac{\cdot}{N}))^d \quad (\overline{R}_{n+1}(\frac{\cdot}{N}))^d$*. From the list of the servers with occupancy* n*, the fraction of the servers with the state* $\mathbf{z}_n$ *is equal to* $\left\{ \frac{\left(\frac{(f\mathbf{z}_n g)}{N}\right)}{\left(\frac{(f\mathbb{U}_n g)}{N}\right)} \right\}$*. Therefore, the probability that the destination server lies in the state* $\mathbf{z}_n$ *is equal to* $\left\{ \frac{\left(\frac{(f\mathbf{z}_n g)}{N}\right)}{\left(\frac{(f\mathbb{U}_n g)}{N}\right)} \right\} ((\overline{R}_n(\frac{\cdot}{N}))^d \quad (\overline{R}_{n+1}(\frac{\cdot}{N}))^d)$*.*

For the case of exponential JLDs, $\mathsf{U}_n = f n g$ *and* $\mathbf{z}_n = n$*. Hence,* $p_r(\quad : \mathbf{z}_n) = (\overline{R}_n(\frac{\cdot}{N}))^d \quad (\overline{R}_{n+1}(\frac{\cdot}{N}))^d$ *coinciding with the analysis for the exponential case in [10, 12].*

As it is clear from (2.7), the routing decision depends only on the number of servers lying in each possible server state. Hence, we get the evolution of the process $(\ {}^{(N)}_{t}; t \quad 0)$ by tracking arrival events, routing decisions, and departure events.

2.4 Summary of Main Results

In this section, we present the main results of this chapter.

Our aim is to study the limit as $N \ ! \ 1$ of the empirical probability measures of the states of the servers, i.e., the limit of the sequence of processes $f(\frac{{}^{(N)}_{t}}{N}; t \quad 0) g_{N\ 1}$ as $N \ ! \ 1$. A system with index N has N servers that serve the incoming jobs arriving

according to a Poisson process with rate N , and all other system parameters remain the same for all N as given in Section 2.1. For given N, the process $(\ {}^{(N)}_{t}; t \quad 0)$ defined in (2.3) describes the dynamics of the system with index N. The goal is to characterize the limit of the normalized process $({}^{-(N)}_{t}; t \quad 0)$ as $N \ ! \ 1$ where

$$-^{(N)}_{t} = \frac{{}^{(N)}_{t}}{N}: \tag{2.9}$$

For a Borel set $B \ 2 \ \mathsf{B}(\mathsf{U})$, ${}^{-(N)}_{t}(B)$ is equal to fraction of the servers with state lying in the set B at time t.

2.4.1 An Overview of the Analysis

We now give a brief overview of the analysis of this chapter.

The mean-field limit corresponds to $\lim_{N \ ! \ 1}({}^{-(N)}_{t}; t \quad 0) = ({}^{-}_{t}; t \quad 0)$ that lies in $\mathsf{C}_{\mathsf{M}_1(\mathsf{U})}([0; \ 1 \ ))$ and it is a deterministic measure-valued process satisfying a set of evolution equations referred to as the MFEs. We then obtain an alternative form of the evolution equations satisfied by the process $(h^{-}_{t}; \ i; t \quad 0)$ for $2 \ \mathsf{C}_b(\mathsf{U})$. This is stated in Lemma 2.2. Using these equations, we show in Theorem 2.1 that there exists a unique solution to the MFEs for a given initial point.

We then show that the sequence of processes $f({}^{-(N)}_{t}; t \quad 0)g_{N \ 1}$ is tight. For this, we first study the Feller property of the Markov process $(\ {}^{(N)}_{t}; t \quad 0)$ and construct a martingale process in Theorem 2.2 by using the generator of the Markov process $(\ {}^{(N)}_{t}; t$ 0) by employing the Dynkin's formula [56, Proposition 1.7, p.162].

We prove the tightness of the sequence of processes $f({}^{-(N)}_{t}; t \quad 0)g_{N \ 1}$ using the result that the normalized version of the martingale process converges to the null process as $N \ ! \ 1$. Furthermore, we show that any limit point of the sequence $f({}^{-(N)}_{t}; t \quad 0)g_{N \ 1}$ coincides almost surely with the unique solution to the MFEs referred to as the mean-field limit. This is stated in Theorem 2.3.

Finally, we obtain a set of partial differential equations satisfied by the mean-field limit. We then prove the uniqueness of the fixed-point and its insensitivity in Theorem 2.5. The proofs of Theorems 2.3 and 2.5 are given in Sections 2.5 and 2.6, respectively. The remaining proofs are given in Section 2.8.

2.4.2 Transient Regime

In this section, we state the results that are related to the transient regime. We first study the proposed MFEs. For the given system parameters , C, d, and the probability density function $g(\ )$ of the JLDs, we state the MFEs in Proposition 2.1. The dynamics of a mean-field solution $(\bar{}_t; t \quad 0)$ are described by using a set of evolution equations of the real valued processes $(h\bar{}_t; \ i; t \quad 0)$ for all $\ 2\ C_b^1(\mathsf{U})$, referred to as the MFEs.

Proposition 2.1. *Mean- eld equations:*

For given system parameters $(\ ; C; d; g(\ ))$*, the process* $(\bar{}_t; t \quad 0)$ *satis es:*

1. *The mapping* $t \ 7! \ \bar{}_t$ *is a continuous.*

2. *For* $\ 2\ C_b^1(\mathsf{U})$*, the process* $(\bar{}_t; t \quad 0)$ *satis es*

$$h\bar{}_t; \ i = h\bar{}_0; \ i + \int_{s=0}^{t} h\bar{}_s; \ r_1 \ i \, ds$$
$$\int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{\mathsf{U}_n} (x_j) \left(\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n) \right) d\bar{}_s(\mathbf{x}_n) \right) ds$$
$$+ \int_{s=0}^{t} \left(\bar{}_s(f0g) \ \Phi_0(\bar{}_s) \left(\ (1;0) \quad (0) \right) + \sum_{n=1}^{C} \sum_{i=1}^{1} \sum_{n+1}^{n+1} \int \quad \int_{\mathsf{U}_n} \left\{ \frac{1}{n+1} \right\} \right.$$
$$\left. \Phi_n(\bar{}_s)(\ (\mathbf{x}_n^i; 0) \quad (\mathbf{x}_n)) \, d\bar{}_s(\mathbf{x}_n) \right) ds; \quad (2.10)$$

where the index j is used to denote the position of the departing job when there are n progressing jobs and the index i denotes the position of the arriving job when there are already n progressing jobs at the server. Further, $\Phi_n(\bar{}_s) = \frac{f(\overline{R}_n(\bar{}_s))^d \ (\overline{R}_{n+1}(\bar{}_s))^d g}{f\overline{R}_n(\bar{}_s) \ \overline{R}_{n+1}(\bar{}_s) g}$ *where* $\overline{R}_j(\bar{}_s) = \sum_{n:\ n=j}^{C} \bar{}_s(\mathsf{U}_n)$.

In (2.10), the second term on the right-hand side is due to the increase of the ages of the progressing jobs linearly with time at unit rate. The third and fourth terms on the right-hand side of (2.10) are due to departure and arrival of a job, respectively.

Remark 2.4. *The t-continuity of the mapping* $\bar{}_t$ *is equivalent to the continuity of the mapping* $t \ 7! \ h\bar{}_t; \ i$ *for all* $\ 2\ C_b^1(\mathsf{U})$ *since* $C_b^1(\mathsf{U})$ *is a separating class of* $\mathsf{M}_1(\mathsf{U})$ *[56, p. 111].*

Although the MFEs (2.10) are defined for the class of functions $\in C_b^1(U)$, it is more useful to obtain an approximation of the process $(h^{-(N)}_t; I_{fBg}i; t \geq 0)$ for an open set $B \in B(U)$. Therefore, we need to obtain the evolution equations of the real valued process $(h^{-}_t; I_{fBg}i; t \geq 0)$. In this direction, we first obtain the evolution equations of the real valued process $(h^{-}_t; \ i; t \geq 0)$ where $\in C_b(U)$. We then proceed to obtain the evolution equations of the process $(h^{-}_t; I_{fBg}i; t \geq 0)$ where B is an open set with the help of the monotone convergence theorem [69, Theorem 1.26] since there exists a sequence of functions in $C_b(U)$ that increase point wise to I_{fBg}.

Lemma 2.2. *A process $(\ _t \in M_1(U); t \geq 0)$ with continuity of the mapping $t \mapsto\ _t$ satisfies the MFEs (2.10) if it satisfies the following equation for all $\in C_b(U)$,*

$$h\ _t; \ i = h\ _0; \ _t\ i + \int_{r=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \ (x_j) \left(\ _t\ _r\ (\mathbf{x}_n{}^j) \quad \ _t\ _r\ (\mathbf{x}_n) \right) d\ _r(\mathbf{x}_n) \right.$$
$$+ \left[\ _r(f0g) \ \Phi_0(\ _r) \left(\ _t\ _r\ (1; 0) \quad \ _t\ _r\ (0) \right) \right.$$
$$\left. \left. + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{j=1}^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \ \Phi_n(\ _r)(\ _t\ _r\ (\mathbf{x}_n^j; 0) \quad \ _t\ _r\ (\mathbf{x}_n)) \, d\ _r(\mathbf{x}_n) \right] \right) dr: \quad (2.11)$$

Proof. See Section 2.8.2. □

Using (2.11), we now show that starting with an initial measure $\ _0$, for $t \geq 0$, there exists a unique measure $\ _t \in M_1(U)$ that satisfies (2.10).

Theorem 2.1. *There exists a unique solution in $C_{M_1(U)}([0; 1))$ to the MFEs. In particular, if $(\ ^1_t; t \geq 0)$ and $(\ ^2_t; t \geq 0)$ are two mean-field solutions starting at initial measures $\ ^1_0 \in M_1(U)$, $\ ^2_0 \in M_1(U)$, respectively, then*

$$k\ ^1_t \quad \ ^2_t k \quad k\ ^1_0 \quad \ ^2_0 k e^{(2Ck \ k+8d^2 \ )t}: \quad (2.12)$$

Proof. See Section 2.8.3. □

We next focus on the proof of the convergence of the sequence of the processes $f(^{-(N)}_t; t \geq 0)g_{N \ 1}$ as $N \to 1$. We first give a result on the Feller property of $(\ ^{(N)}_t; t \geq 0)$. We consider the filtration

$$F_t^{(N)} = \backslash_{>0}(\ (\ ^{(N)}_s : s \quad t + \ )): \quad (2.13)$$

Let $A^{(N)}(\ )$ be the generator of the Markov process $(\ ^{(N)}_t; t \geq 0)$. Using the Dynkin's formula [56], we have the following result.

Theorem 2.2. *The process* $(\ {}^{(N)}_t; t \quad 0)$ *is a Feller-Dynkin process in* $\mathsf{D}_{\mathsf{M}_F(\mathsf{U})}([0; 1))$. *Let* $2\ \mathsf{C}^1_b(\mathsf{U})$, *then the process* $(\mathbf{M}^{(N)}_t(\ ); t \quad 0)$ *de ned as*

$$\mathbf{M}^{(N)}_t(\ ) = h\ {}^{(N)}_t;\ \ i \quad h\ {}^{(N)}_0;\ \ i \quad \int_{s=0}^{t} A^{(N)} h\ {}^{(N)}_s;\ \ i\, ds \tag{2.14}$$

is a square integrable $F^{(N)}_t$*-martingale and it is an RCLL process.*

Proof. See Section 2.8.1. □

Our analysis requires the following assumptions.

Assumption 2.4.1. *The sequence of the initial random measures* $f^{-(N)}_0 g_{N\ 1}$ *satisfy*

$$(^{(N)}_0; h^{(N)}_0;\ \ i)\ \ )\ \ (\#; h\#;\ \ i); \tag{2.15}$$

where $\#\ 2\ \mathsf{M}_1(\mathsf{U})$ *is a probability measure that possesses a density (w.r.t. the Lebesgue measure) and* $h\#;\ \ i < 1$.

Here, we can interpret the condition $h\#;\ \ i < 1$ as follows. If # is the probability measure of the state of a server, then the average sum of ages of jobs is finite.

We use Theorem 2.2 and the Assumption 2.4.1 to establish the following main result.

Theorem 2.3. *Under the Assumptions 2.4.1 and 2.1.1, we show that* $(^{(N)}_t; t \quad 0)\ \ )$ $(^{-}_t; t \quad 0)$ *as* $N \ \ !\ \ 1$, *where* $(^{-}_t; t \quad 0)$ *is the unique solution to* (2.10) *with the initial point* #. *The process* $(^{-}_t; t \quad 0)$ *is referred to as the mean- eld limit.*

Proof. See Section 2.5. □

The existence of the mean-field limit allows us to show that any finite subset of servers become independent of each other in the limiting system.

Theorem 2.4. *If* $f\mathbf{S}^{(N)}_k(0); 1 \quad k \quad N g$ *are exchangeable and under the Assumptions 2.1.1 and 2.4.1, the following result holds:*

- *For each xed k and* $t\ 2\ [0; 1)$, $L(\mathbf{S}^{(N)}_k(t))\ \ )\ \ ^{-}_t$ *as* $N \ \ !\ \ 1$.
- *For any xed positive integer l and for each* $t\ 2\ [0; 1)$, *we have* $f\mathbf{S}^{(N)}_k(t); 1 \quad k \quad lg\ \ )\ \ f\mathbf{S}_k(t); 1 \quad k \quad lg$ *as* $N \ \ !\ \ 1$, *where* $\mathbf{S}_k(t)$, $1 \quad k \quad l$, *are independent random variables with* $L(\mathbf{S}_k(t))$ *equal to* $^{-}_t$ *for all* $1 \quad k \quad l$.

Proof. See Section 2.8.4. □

Remark 2.5. *For any time t, a consequence of Theorem 2.4 is that as $N \to \infty$, due to Poisson thinning, the arrival process to each server is a Poisson process with rate $\Phi_n(\bar{x}_t)$ when there are n ($n \geq 0$) progressing jobs.*

Lemma 2.3. *For any time t, the measure $\bar{x}_t$ has a density function w.r.t. the Lebesgue measure for almost all $\mathbf{u}_n \in \mathsf{U}_n$, $n \geq 1$.*

Proof. See Section 2.8.6. □

For a given subset $B \in \mathcal{B}(\mathsf{U})$, once we show that $(\bar{x}^{(N)}_t; t \geq 0) \Rightarrow (\bar{x}_t; t \geq 0)$ as $N \to \infty$, since $\bar{x}_t$ is absolutely continuous w.r.t. Lebesgue measure for every $t \geq 0$, the continuous mapping theorem [57, Theorem 2.7] implies that $(h\bar{x}^{(N)}_t; I_{\{B\}}i; t \geq 0) \Rightarrow (h\bar{x}_t; I_{\{B\}}i; t \geq 0)$. This shows that for large N, we can approximate $h\bar{x}^{(N)}_t; I_{\{B\}}i$ by $h\bar{x}_t; I_{\{B\}}i$.

2.4.3 Stationary Regime

We now present the results related to the stationary behavior of the mean-field.

We first demonstrate an analogy between the MFEs of the considered multi-server Erlang loss system under the SQ(d) load balancing policy and the dynamics of a single server Erlang loss system with state-dependent arrival rates. We then exploit this analogy to prove the uniqueness of the fixed-point of the mean-field and its insensitivity. We first recall the dynamics of the probability measure of the server's state of a single server Erlang loss system with capacity C, where jobs arrive according to a Poisson process with pre-specified state-dependent arrival rates.

Consider a single server system with capacity C where jobs arrive according to a Poisson process at rate λ_n when there are n progressing jobs in the system. The service times are generally distributed as stated in the system model of Section 2.1. Let $\nu^{(single)}_t$ be the probability measure of the server's state at time t defined on U. For $f \in \mathcal{C}^1_b(\mathsf{U})$, it can be verified that the Kolmogorov equations are given by,

$$h\nu^{(single)}_t; f i = h\nu^{(single)}_0; f i + \int_{s=0}^{t} h\nu^{(single)}_s; r_1 f i\, ds$$
$$\int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathsf{U}_n} \alpha(x_j) \left(f(\mathbf{x}_n^{\,j}) - f(\mathbf{x}_n) \right) d\nu^{(single)}_s(\mathbf{x}_n) \right.$$

$$+\left[\left(\ \ {}_0\ {}_s^{(single)}(f0g)\left(\ \ (1;0)\qquad(0)\right)\right)+\sum_{n=1}^{C}\sum_{j=1}^{1}\sum_{}^{n+1}\int\qquad\int_{U_n}\left\{\frac{1}{n+1}\right\}\right.$$
$$\left.{}_n\left(\ \ (\mathbf{x}_n^j;0)\qquad(\mathbf{x}_n)\right)d\ {}_s^{(single)}(\mathbf{x}_n)\right]\Bigg)\,ds: \quad (2.16)$$

On comparing the MFEs (2.10) with the Kolomogorov equation of a single-server system given by (2.16), it is clear that both the dynamics are similar except that $_n$ in (2.16) is replaced by $\Phi_n(^-_s)$ when the probability measure of the server's state is ^-_s at time s. Equation (2.10) differs from (2.16) in that the arrival rates at time t depend on ^-_t. This is an example of a non-linear Markov process which means that the generator of the Markov process at time t depends on the current distribution ^-_t of the Markov process [73] while the equation (2.16) for fixed $(\ _i;0\quad i\quad C)$ denotes a Markov process whose generator does not depend on the current distribution.

We now study the fixed-point of the mean-field. Let $P_t(0)$ be equal to $_t(f0g)$ and let $p_t(\mathbf{x}_n)$ be the probability density of $_t$ *w:r:t:* Lebesgue measure at $\mathbf{x}_n$. We obtain the differential equations satisfied by the process $(P_t;t\quad 0)$ with $P_t=(P_t(\mathbf{u});\mathbf{u}\ 2\ U)$ where

$$P_t(\mathbf{y}_n)=\int_{x_1=0}^{y_1}:::\int_{x_n=0}^{y_n}p_t(\mathbf{x}_n)\,dx_1\qquad dx_n: \quad (2.17)$$

Here, from the Remark 2.5, since ^-_t is the distribution of a server's state as $N\ !\ 1$, it implies that $P_t(\mathbf{y}_n)$ is the probability that a server has n jobs and the i^{th} job's age is at most y_i for $1\quad i\quad n$ as $N\ !\ 1$. Also, since $^{-(N)}_t(\ )\ )\ ^-(\ )$, for a large value of N, the fraction of servers with n jobs and the i^{th} job's age is at most y_i for $1\quad i\quad n$ can be approximated by $P_t(\mathbf{y}_n)$.

Lemma 2.4. *The process* $(P_t;t\quad 0)$ *satis es*

$$\frac{dP_t(0)}{dt}=\int_{y=0}^{1}(y)\left(\frac{@P_t(1;y)}{@y}\right)dy\qquad\Phi_0(P_t)P_t(0); \quad (2.18)$$

for $1\quad n\quad C\quad 1$,

$$\frac{dP_t(\mathbf{y}_n)}{dt}=\sum_{i=1}^{n}\frac{@P_t(\mathbf{y}_n)}{@y_i}+\sum_{j=1}^{n+1}\int_{x_j=0}^{1}(x_j)\left(\frac{@P_t(\mathbf{y}_n^j;x_j)}{@x_j}\right)dx_j$$
$$\sum_{j=1}^{n}\int_{x_j=0}^{y_j}(x_j)\left(\frac{@P_t(\mathbf{y}_n{}^j;x_j)}{@x_j}\right)dx_j$$
$$+\sum_{j=1}^{n}\frac{\Phi_{n\ 1}(P_t)}{n}P_t(\mathbf{y}_n{}^j)\qquad\Phi_n(P_t)P_t(\mathbf{y}_n); \quad (2.19)$$

and for $n = C$,

$$\frac{dP_t(\mathbf{y}_n)}{dt} = \sum_{i=1}^{n} \frac{@P_t(\mathbf{y}_n)}{@y_i} \sum_{j=1}^{n} \int_{x_j=0}^{y_j} (x_j) \left(\frac{@P_t(\mathbf{y}_n{}^j; x_j)}{@x_j} \right) dx_j + \sum_{j=1}^{n} \frac{\Phi_{n\ 1}(P_t)}{n} P_t(\mathbf{y}_n{}^j); \quad (2.20)$$

where $\Phi_n(P_t) = \frac{f(R_n(P_t))^d \quad (R_{n+1}(P_t))^d g}{f R_n(P_t) \quad R_{n+1}(P_t) g}$ *and* $R_n(P_t) = \sum_{j:\ j=n}^{C} \lim_{b\ !\ 1} P_t(j; b; \quad ; b)$.

Proof. See Section 2.8.5. □

Remark 2.6. *Specializing the results to the exponential case with mean* $\frac{1}{}$, $(x) = $, *and denoting* $q_n(t) = \lim_{b\ !\ 1} P_t(n; b; \quad ; b)$, *it can be veri ed that the process* $(\mathbf{q}(t); t \quad 0)$ *where* $\mathbf{q}(t) = (q_n(t); 0 \quad n \quad C)$ *is the unique solution of the MFEs given in [10] for the case of the exponential distributions with rate* $= 1$.

We next state the the principal result on the insensitivity of the fixed-point of the MFEs.

Theorem 2.5. *The process* $(P_t; t \quad 0) = (P_t(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U}; t \quad 0)$ *has a unique xed-point given by* $= (\ (\mathbf{y}); \mathbf{y}\ 2\ \mathsf{U})$ *where*

$$(\mathbf{y}_n) = \ {}^{(exp)}_{n} \ {}^{n} \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i; \quad (2.21)$$

and ${}^{(exp)} = (\ {}^{(exp)}_{n}; 0 \quad n \quad C)$ *denotes the unique xed-point of the mean- eld when the service times are exponentially distributed with the mean* $\frac{1}{}$ *and* ${}^{(exp)}_{n}$ *is the stationary probability that there are n jobs in the limiting system. Further, since* $\int_{x=0}^{1} \overline{G}(x)\, dx = \frac{1}{}$, *the xed-point of the mean- eld is insensitive, i.e.,*

$$\lim_{b\ !\ 1} \ (n; b; \quad ; b) = \ {}^{(exp)}_{n}: \quad (2.22)$$

Proof. See Section 2.6. □

2.5 Proof of Theorem 2.3

From Theorem 2.2, we now show that the normalized process $(\overline{\ }^{(N)}_{t}; t \quad 0)$ converges to the mean-field limit.

Let $(\overline{F}_t^{(N)}; t \quad 0)$ be the right continuous filtration associated with the process $(\overline{\ }_t^{(N)}; t \quad 0)$. Note that we have $(\overline{\ }_t^{(N)}; t \quad 0) \; 2 \; \mathsf{D}_{\mathsf{M}_1^{(N)}(\mathsf{U})}([0; 1))$. We first show that the sequence of processes $(\overline{\ }_t^{(N)}; t \quad 0)$ is relatively compact. We then prove that every limit point $(\ _t; t \quad 0)$ almost surely has sample paths that are continuous in t and furthermore, they coincide with the unique mean-field solution with the initial point #. For every limit point $(\ _t; t \quad 0)$, $\ _0$ almost surely coincides with the measure # from the Assumption 2.4.1. Further, we have that the mean-field solution is unique for the given initial measure. Hence, we conclude that for all the limit points, almost surely sample paths coincide with the unique mean-field solution $(\overline{\ }_t; t \quad 0)$ with the initial point #. The process $(\overline{\ }_t; t \quad 0)$ is referred to as the mean-field limit. As a result, $(\overline{\ }_t^{(N)}; t \quad 0)$ converges in distribution to the mean-field limit $(\overline{\ }_t; t \quad 0)$.

For $\quad 2 \; \mathsf{C}_b^1(\mathsf{U})$, from Proposition 2.3, the process $(\overline{\mathbf{M}}_t^{(N)}(\ ); t \quad 0)$ defined as follows is an RCLL square integrable $\overline{F}_t^{(N)}$-martingale

$$\overline{\mathbf{M}}_t^{(N)}(\ ) = h\overline{\ }_t^{(N)}; \ \ i \quad h\overline{\ }_0^{(N)}; \ \ i \quad \int_{s=0}^{t} h\overline{\ }_s^{(N)}; \ r_1 \ \ i \, ds \quad \int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{\mathsf{U}_n} \ (x_j) \right.$$
$$\left(\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n) \right) d\overline{\ }_s^{(N)}(\mathbf{x}_n)$$
$$+ \left[\overline{\ }_s^{(N)}(f0g) \ \ \Phi_0(\overline{\ }_s^{(N)}) \left(\ (1; 0) \quad (0) \right) + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{n+1}^{n+1} \int \quad \int_{\mathsf{U}_n} \left\{ \frac{1}{n+1} \right\} \right.$$
$$\left. \left. \Phi_n(\overline{\ }_s^{(N)}) \left(\ (\mathbf{x}_n^j; 0) \quad (\mathbf{x}_n) \right) d\overline{\ }_s^{(N)}(\mathbf{x}_n) \right] \right) ds: \quad (2.23)$$

We further have

$$< \overline{\mathbf{M}}^{(N)}(\ ) >_t = \frac{1}{N} \left[\int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{\mathsf{U}_n} \ (x_j) \left(\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n) \right)^2 d\overline{\ }_s^{(N)}(\mathbf{x}_n) \right. \right.$$
$$+ \left[\overline{\ }_s^{(N)}(f0g) \ \ \Phi_0(\overline{\ }_s^{(N)}) \left(\ (1; 0) \quad (0) \right)^2 \right.$$
$$\left. \left. + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{n+1}^{n+1} \int \quad \int_{\mathsf{U}_n} \left\{ \frac{1}{n+1} \right\} \ \Phi_n(\overline{\ }_s^{(N)}) \left(\ (\mathbf{x}_n^j; 0) \quad (\mathbf{x}_n) \right)^2 d\overline{\ }_s^{(N)}(\mathbf{x}_n) \right] \right) ds \Bigg]: \quad (2.24)$$

Since the space $\mathsf{D}_{\mathsf{M}_1(\mathsf{U})}([0; 1))$ endowed with the Skorohod topology is complete and separable, by using the Prohorov's theorem (Theorem A.2) [57], establishing the relative compactness of the sequence of the processes $f(\overline{\ }_t^{(N)}; t \quad 0)g_{N \ 1}$ is equivalent to proving the tightness of the processes $f(\overline{\ }_t^{(N)}; t \quad 0)g_{N \ 1}$. From Theorem 4:6 of [74], the Jakubowski's

criteria stated in Appendix A.3 can be used to establish the relative compactness of the sequence of the processes $f(\bar{\xi}_t^{(N)}; t \quad 0)g_{N\ 1}$. According to the Jakubowski's criteria, we need to show that the conditions J1 and J2 are true.

We next focus on the proof of the condition J2. For this, we prove the conditions C1 and C2 that are sufficient to prove the relative compactness of the sequence $f(h_t^{(N)};\ i; t \quad 0)g_{N\ 1}$ for $2\ C_b^1(\mathsf{U})$ in $D_{\mathsf{R}}([0; 1))$. For any $T > 0$, $t\ 2\ [0; T]$, we have $h_t^{(N)};\ i \quad k\ k_1 h_t^{(N)}; 1i$ and since $h_t^{(N)}; 1i = 1$, the condition C1 is trivially satisfied with $b = k\ k_1$.

We next prove that the condition C2 holds. For > 0, by using (2.24) and the Doob's inequality (Theorem A.8) [56, page 63], we have

$$\mathsf{P}\left(\sup_{t\ T}\left|\overline{\mathbf{M}}_t^{(N)}(\ )\right| \quad\right) \quad \frac{4}{2}\mathsf{E}\left[< \overline{\mathbf{M}}^{(N)}(\ ) >_T\right]$$
$$4Tk\ k^2\frac{1}{N}(k\ k + d\ );$$

and hence, $\mathsf{P}\left(\sup_{t\ T}\left|\overline{\mathbf{M}}_t^{(N)}(\ )\right| \quad\right) \ !\ 0$ as $N\ !\ 1$. Therefore, the sequence of processes $f(\overline{\mathbf{M}}_t^{(N)}(\ ); t \quad 0)g_{N\ 1}$ converges in distribution to the null process from the standard convergence criterion in $\mathsf{D}_{\mathsf{R}}([0; T])$ [56, Theorem 1.4, p.339]. Furthermore, the sequence of processes $f(\overline{\mathbf{M}}_t^{(N)}(\ ); t \quad 0)g_{N\ 1}$ is tight in $\mathsf{D}_{\mathsf{R}}([0; T])$ and hence, there exists $^{\emptyset} > 0$ and $N^{\emptyset} > 0$ such that for all $N \quad N^{\emptyset}$, we have

$$\mathsf{P}\left(\sup_{u;v\ T;ju\ vj\ {}^{\emptyset}}\left|\overline{\mathbf{M}}_v^{(N)}(\ ) \quad \overline{\mathbf{M}}_u^{(N)}(\ )\right| \quad \frac{}{2}\right) \quad \frac{}{2} \tag{2.25}$$

For any $u < v \quad T$, from (2.23), we have

$$\left|h_v^{(N)};\ i \quad h_u^{(N)};\ i\right| \quad \int_{s=u}^{v}\left|h_s^{(N)};\ r_1\ i\right| ds + 2k\ kk\ kC\, ju \quad vj + 2k\ k\ ju \quad vj$$
$$+ \left|\overline{\mathbf{M}}_v^{(N)}(\ ) \quad \overline{\mathbf{M}}_u^{(N)}(\ )\right|: \tag{2.26}$$

Further, we can write

$$\left|h_v^{(N)};\ i \quad h_u^{(N)};\ i\right| \quad jv \quad uj\, Ck\ k_1(1 + 2k\ k + 2d\ ) + \left|\overline{\mathbf{M}}_v^{(N)}(\ ) \quad \overline{\mathbf{M}}_u^{(N)}(\ )\right|: \tag{2.27}$$

Therefore, by using (2.25) and (2.27), there exists > 0 and $N_1 > 0$ such that for $N \quad N_1$, we have $\mathsf{P}\left(\sup_{u;v\ T;ju\ vj}\left|h_v^{(N)};\ i \quad h_u^{(N)};\ i\right| \quad\right) \quad :$ This proves the condition C2. Since the conditions C1 and C2 hold, the condition J2 also holds.

We next prove the compact containment condition J1. Let $(n_i(t); x_{i1}(t) : : : ; x_{in_i(t)}(t))$ be the state of the i^{th} server at time t where $x_{ij}(t)$ denotes the age of the j^{th} job at the i^{th} server. Clearly, we have $h_t^{(N)};\ i = \frac{1}{N}\sum_{i=1;\, n_i(t)>0}^{N}(x_{i1}(t) + \quad + x_{in_i(t)}(t))$:

We can classify the progressing jobs into two classes. The jobs that are in service from the beginning ($t = 0$) form the first class and the second class of jobs are the ones that entered the system in the interval $(0; t]$. At a server, the number of progressing jobs that belong to a class is upper bounded by C. Let $\mathbf{Y}_t$ be a random variable representing the age of a job belonging to the second class that is in progress at time t, and $\mathbf{Y}$ be a random variable with job length distribution G, then for any b 0, we have

$$\mathsf{P}(\mathbf{Y}_t \quad b) \quad \mathsf{P}(\mathbf{Y} \quad b): \tag{2.28}$$

Therefore, using (2.28), since each server has capacity C, for any time t 0, we can write

$$\mathsf{P}(h^{-(N)}_t; \ i \quad b) \quad \mathsf{P}\left(h^{-(N)}_0; \ i + Ct + \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Y}_{i1} + \quad + \mathbf{Y}_{iC}) \quad b\right); \tag{2.29}$$

where $(\mathbf{Y}_{ij}; 1 \quad i \quad N; 1 \quad j \quad C)$ are i.i.d. random variables with distribution G. Further, by the weak law of large numbers, we have $\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Y}_{i1} + \quad + \mathbf{Y}_{iC}) \ ) \ \underline{C}$ as $N \ ! \ 1$. Therefore, by choosing $Z_T = 2h\#; \ i + 2CT + \frac{2C}{}$, we have

$$\mathsf{P}\left(\sup_{t2[0;T]} h^{-(N)}_t; \ i > Z_T\right) \ ! \ 0 \tag{2.30}$$

as $N \ ! \ 1$. Let us define $\mathsf{L}_T \ , \ f \ 2 \ \mathsf{M}_1(\mathsf{U}) : h \ ; \ i \quad Z_T g:$ Since $h \ ; \ i \quad Z_T$ for $2 \ \mathsf{L}_T$, let $B_r = \mathsf{U}_0 \ [\ ([_n \ _1 f(n; y_1; \quad ; y_n) : 0 \quad y_i \quad r; 1 \quad i \quad ng)$ and $\overline{B}_r$ be the complement of B_r, then we have $(\overline{B}_r) \quad \frac{Z_T}{r}:$ Hence, $\lim_{r \ ! \ 1} \sup \ _{2\mathsf{L}_T} \ (\overline{B}_r) = 0:$ Therefore, from Lemma A7:5 of [75], L_T is relatively compact in $\mathsf{M}_1(\mathsf{U})$. Furthermore, from (2.30), we have $\liminf_{N \ ! \ 1} \mathsf{P}(-^{(N)}_t \ 2 \ \mathsf{L}_T; 8t \ 2 \ [0; T]) > 1 \quad :$ Let K_T be the closure of L_T, then we have a compact set K_T such that $\liminf_{N \ ! \ 1} \mathsf{P}(-^{(N)}_t \ 2 \ \mathsf{K}_T; 8t \ 2 \ [0; T]) > 1$ for all $0 < \ < 1$.

This establishes the condition J1 and hence, the proof of the tightness of the sequence of processes $f(-^{(N)}_t; t \quad 0)g_{N \ 1}$ is completed.

Let $(\ _t; t \quad 0)$ be a limit point of a converging subsequence $f(-^{(N_{i_k})}_t; t \quad 0)g_{k \ 1}$. From the condition C2, $\ _t$ is continuous in t, $P \quad a:s:$, where P is the probability law of $(\ _t; t \quad 0)$. Furthermore, from [74, Theorem 1.7] for $f \ 2 \ \mathsf{C}_b(\mathsf{U}); \ _t \ 2 \ \mathsf{M}_1(\mathsf{U})$, it follows that for any $T > 0$, we have $(\ _t; 0 \quad t \quad T) \ 7! \ (h \ _t; f i; 0 \quad t \quad T)$ is continuous in the Skorokhod topology. Then since the sequence of martingales $f(\overline{\mathbf{M}}^{(N_{i_k})}_t(\ ); t \quad 0)g_{k \ 1}$ converges to the null process, by the continuous mapping theorem [57, Theorem 2.7], we conclude

$$h \quad _t; \quad i = h \quad _0; \quad i + \int_{s=0}^{t} h \quad _s; r_1 \quad i\, ds$$
$$\int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \quad (x_j) \left(\quad (\mathbf{x}_n{}^j) \qquad (\mathbf{x}_n) \right) d \quad _s(\mathbf{x}_n) \right.$$
$$+ \left[\left(\quad _s(f0g) \quad \Phi_0(\quad _s) \left(\quad (1;0) \qquad (0) \right) \right) + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \right.$$
$$\left. \left. \Phi_n(\quad _s)(\quad (\mathbf{x}_n^j; 0) \qquad (\mathbf{x}_n))\, d \quad _s(\mathbf{x}_n) \right] \right) ds: \quad (2.31)$$

From the Assumption 2.4.1, $\quad _0 = \#$ almost surely and hence, the sample paths coincide almost surely with the unique mean-field solution with the initial point #. This argument holds for every limit point, and hence, the sample paths of every limit point are almost surely the same as the deterministic mean-field solution with the initial point #. This completes the proof.

2.6 Proof of Theorem 2.5

We now show that $\quad = (\quad (\mathbf{u}); \mathbf{u} \; 2 \; U)$ is the unique fixed-point of the mean-field. From [10], we first recall that under the assumption of exponential service time distributions, there exists a unique probability measure of occupancy $\quad ^{(exp)} = (\quad _n^{(exp)}; 0 \qquad n \qquad C)$ on $f0; 1; \qquad ; Cg$ to the stationary MFEs given below,

$$_n^{(exp)}(\quad ^{(exp)})\quad _n^{(exp)} = (n+1) \quad _{n+1}^{(exp)}; \quad (2.32)$$

where

$$_n^{(exp)}(\quad ^{(exp)}) = \frac{f(\sum_{j=n}^{C} \quad _j^{(exp)})^d \quad (\sum_{j=n+1}^{C} \quad _j^{(exp)})^d g}{f(\sum_{j=n}^{C} \quad _j^{(exp)}) \quad (\sum_{j=n+1}^{C} \quad _j^{(exp)}) g}. \quad (2.33)$$

Let $\quad = (\quad (\mathbf{u}); \mathbf{u} \; 2 \; U)$ be a fixed-point of the MFEs of the process $(P_t; t \qquad 0)$ under general service time distributions. Using $\quad$, let the corresponding probability measure of occupancy be $\mathbf{\Gamma} = (\Gamma_n; 0 \qquad n \qquad C)$ defined such that $\Gamma_n = \lim_{b \,!\, 1} \quad (n; b; \qquad ; b))$ and $\Gamma_0 = \quad (0)$. We now show that

$$(\mathbf{y}_n) = \frac{\left\{ \prod_{i=1}^{n} \frac{^{(exp)}_{i \quad 1}(\mathbf{\Gamma})}{i} \right\}}{1 + \sum_{m=1}^{C} \left\{ \prod_{i=1}^{m} \frac{^{(exp)}_{i \quad 1}(\mathbf{\Gamma})}{i} \right\}} \quad ^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i \quad (2.34)$$

and

$$(0) = \frac{1}{1 + \sum_{m=1}^{C} \left\{ \prod_{i=1}^{m} \frac{^{(exp)}_{i\ 1}(\mathbf{\Gamma})}{i} \right\}}: \tag{2.35}$$

Then it implies that $\mathbf{\Gamma}$ also satisfies (2.32)-(2.33), and hence, $\mathbf{\Gamma} = {}^{(exp)}$ concluding the insensitivity of the fixed-point. Furthermore, we have that $(\mathbf{y}_n) = {}^{(exp)}_{n}\ {}^{n} \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i$ concluding the uniqueness of the fixed-point of the mean-field under general service time distributions.

To complete the proof, it remains to show the validity of equations (2.34)-(2.35). We now recall the stationary distribution ${}^{(single)} = ({}^{(single)}(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ of a single server loss system with state-dependent Poisson arrival process with rate ${}_{n}$ $(0 \quad n \quad C)$ when there are n progressing jobs and the service time distributions are as in the system model of Section 2.1. Then from [22], the stationary probability that the server has n progressing jobs and the i^{th} job has age at most y_i $(1 \quad i \quad n)$ is given by

$${}^{(single)}(\mathbf{y}_n) = \frac{\left\{ \prod_{i=1}^{n} \frac{i\ 1}{i} \right\}}{1 + \sum_{m=1}^{C} \left\{ \prod_{i=1}^{m} \frac{i\ 1}{i} \right\}}\ {}^{n} \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i \tag{2.36}$$

and

$${}^{(single)}(0) = \frac{1}{1 + \sum_{m=1}^{C} \left\{ \prod_{i=1}^{m} \frac{i\ 1}{i} \right\}}: \tag{2.37}$$

For the given fixed-point of the mean-field and its corresponding occupancy probability measure $\mathbf{\Gamma}$, consider a single server system under the assumption of a Poisson arrival process with state-dependent rate ${}^{(exp)}_{n}(\mathbf{\Gamma})$ $(0 \quad n \quad C)$ when there are n progressing jobs. Then the unique stationary distribution is given by (2.36)-(2.37) with ${}_{n}$ replaced by ${}^{(exp)}_{n}(\mathbf{\Gamma})$ for all $0 \quad n \quad C$. But, from (2.10), (2.16), and Lemma 2.4, since $\overline{R}_n(\ ) = \sum_{j=n}^{C} \Gamma_j$, we have that is also another stationary distribution for the single server system with state dependent Poisson arrival process having rates ${}^{(exp)}_{n}(\mathbf{\Gamma})$ for all $0 \quad n \quad C$. Since the stationary distribution must be unique, equations (2.34)-(2.35) must hold. This completes the proof.

2.7 Numerical Results

In this section, we provide some numerical results to support that the mean-field is globally asymptotically stable (GAS) under the assumption of mixed-Erlang JLDs.

Showing that the fixed-point of the mean-field approximates the stationary distribution of the system with large N, remains an open problem. If one can establish that the equilibrium or fixed-point of the MFEs is GAS, then the conclusion of the interchange of limits would follow from the Prohorov's theorem (Theorem A.2) [57]. Proving that the fixed-point of the MFEs is GAS is a challenging problem because the joint distribution of the occupancy and ages does not possess any monotonicity properties unlike the case of exponential service time distributions [10]. In this section, we present numerical results on the validity of the GAS of the mean-field for the case in which the service time distributions are mixed-Erlang. In this case, the state of a server is also multi-dimensional and the mean-field is also non-monotonic unlike the exponential case. It is numerically easier to solve the MFEs for the case of mixed-Erlang distributions as they are systems of ODEs unlike the case of general service time distributions for which the MFEs are PDEs as we have shown. One more reason for using mixed-Erlang distributions is that such distributions are dense in the set of all distributions that have support on R_+, see [76]. Our numerical results show that the mean-field is GAS for the case of mixed-Erlang service time distributions.

We consider the system parameters as follows: The capacity of a server is assumed to be $C = 5$. The average job length is assumed to be equal to one, i.e. $= 1$. The service times have a Mixed-Erlang distribution given by sums of independent exponentially distributed random variables (known as an Erlang distribution) where the number of exponential phases is equal to $i \ 2 \ f1; 2; : : : ; Mg$ with probability p_i such that $\sum_{i=1}^{M} p_i = 1$. Each exponential phase is assumed to have rate $_p$. Therefore, we have,

$$\frac{1}{-} = \frac{\sum_{i=1}^{M} i p_i}{p}.$$

We choose $M = 3$, $p_1 = 0.3$, $p_2 = 0.3$, and $p_3 = 0.4$.

Under the mixed-Erlang service time distribution assumptions, let S be the set of all possible server states defined as $\mathsf{S} = [_{n=0}^{C} \mathsf{S}_n$ where $\mathsf{S}_0 = f(0)g$ and $\mathsf{S}_n = f(n; r_1; : : : ; r_n) : 1 \quad r_i \quad M; 1 \quad i \quad ng$. We refer to an element in the set S by $\mathbf{r}$ and an element in the set S_n by $\mathbf{r}_n$. The system dynamics can be modeled as a Markov process $\mathbf{Y}^{(N)}(t) = (\mathbf{Y}_{\mathbf{r}}^{(N)}(t); \mathbf{r} \ 2 \ \mathsf{S})$ where $\mathbf{Y}_{\mathbf{r}}^{(N)}(t)$ denotes the fraction of servers with n jobs such that the i^{th} job has r_i remaining phases at time t. Since the Markov process $(\mathbf{Y}^{(N)}(t); t \quad 0)$ is defined on a finite-dimensional space, we can establish the mean-field limit by using the same procedure as in the case of the exponential service times in [10]. Hence, we recall the following result without proof from [33].

Proposition 2.2. *If* $\{\mathbf{Y}^{(N)}(0)\}_{N\ 1}$ *converges in distribution to a state* $\mathbf{u}$ *as* $N\ !\ 1$*, then the sequence of processes* $\{(\mathbf{Y}^{(N)}(t); t\quad 0)\}_{N\ 1}$ *converges in distribution to a deterministic process* $\mathbf{x}(\ ;\mathbf{u})$ *as* $N\ !\ 1$ *called the mean- eld. The process* $\mathbf{x}(\ ;\mathbf{u})$ *is the unique solution to the following system of di erential equations*

$$\mathbf{x}(0;\mathbf{u}) = \mathbf{u}; \tag{2.38}$$

$$\dot{x}_{\mathbf{r}_n}(t;\mathbf{u}) = h_{\mathbf{r}_n}(\mathbf{x}(t;\mathbf{u})); \tag{2.39}$$

and $\mathbf{h}(\ ) = (h_{\mathbf{r}}(\ ); \mathbf{r}\ 2\ \mathsf{S})$ *with the mapping* $h_{\mathbf{r}_n}(\ )$ *given by*

$$h_{\mathbf{r}_n}(\mathbf{y}) = \sum_{b=1}^{n}\left\{\frac{p_{r_b}}{n}\right\} y_{(\mathbf{r}_n{}^{j})}\ {}^{(ME)}_{n\ 1}(\mathbf{y}) \quad y_{\mathbf{r}_n}\ {}^{(ME)}_{n}(\mathbf{y}) I_{\{n<C\}} + \sum_{b=1}^{n+1}\ {}_{p} I_{\{n<C\}} y_{(\mathbf{r}_n^b;1)} + \sum_{b=1}^{n}\ {}_{p} y_{(n;r_1;\quad ;r_{b\ 1};r_b+1;r_{b+1};\quad ;r_n)} \quad n\ {}_{p} y_{\mathbf{r}_n}; \tag{2.40}$$

where

$${}^{(ME)}_{n}(\mathbf{v}) = \left\{\frac{}{\sum_{\mathbf{r}_n} v_{\mathbf{r}_n}}\right\}\left[\left(\sum_{i=n}^{C}\sum_{\mathbf{b}_i} v_{\mathbf{b}_i}\right)^{d} \quad \left(\sum_{i=n+1}^{C}\sum_{\mathbf{b}_i} v_{\mathbf{b}_i}\right)^{d}\right]: \tag{2.41}$$

In Figure 2.1, we plot $d_E^2(\mathbf{x}(t;\mathbf{u});\ \ )$ as a function of t where d_E is the euclidean distance defined by

$$d_E(\mathbf{u};\mathbf{v}) = \sqrt{\sum_{\mathrm{l}2\mathsf{S}} ju_{\mathrm{l}} \quad v_{\mathrm{l}}j^2}:$$

It is observed that for $d = 2$, $\quad = 1$, and for four different initial points $\mathbf{u}_1;\mathbf{u}_2;\mathbf{u}_3$, and $\mathbf{u}_4$, the mean-field $\mathbf{x}(t;\mathbf{u})$ for the mixed-Erlang service time distribution converges to its unique fixed-point $\quad$. Note that the computed $\quad$ depends on the chosen value of d. This supports that $\quad$ is globally stable.

We conclude with some numerical results for the blocking probability of the above system showing closeness to the theoretical lower bound. Similar to the case of exponential distributions as in [10], under the asymptotic independence, any finite set of servers are independent, and the fixed-point of the mean-field implies that the fixed point is the stationary distribution of the state of a server. The average blocking probability is then given by Q_C^d where $Q_C = \sum_{\mathbf{v}_C\ 2\mathsf{S}_C}\ (\mathbf{v}_C)$. Let us recall the lower bound on the average blocking probability denoted by P_{block}^{avg} for any load balancing scheme shown in [19]. From

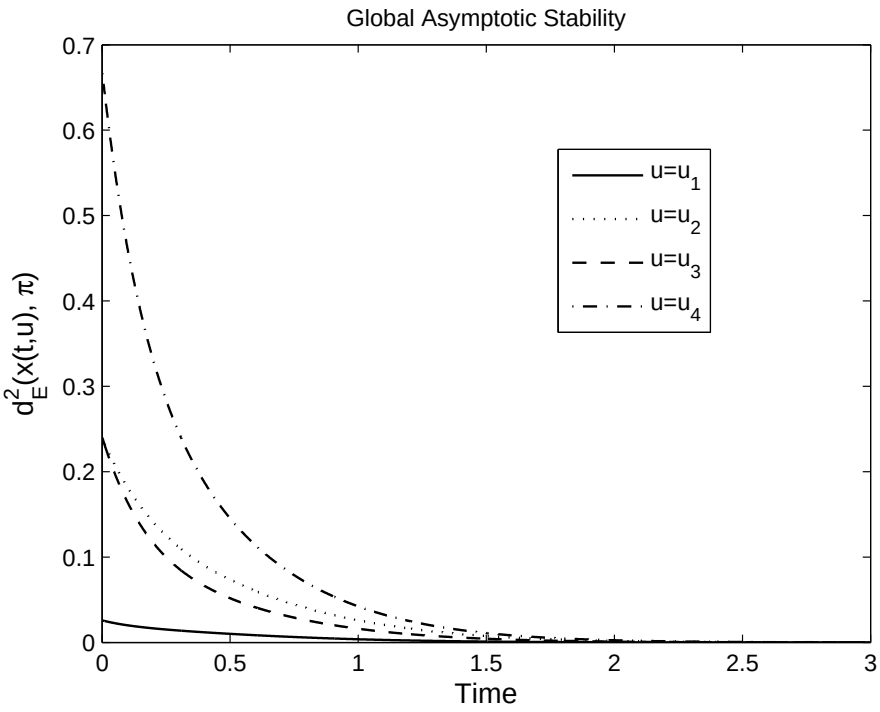

Figure 2.1: Convergence of the mean-field to the fixed-point

Little's law [76, Theorem 4.1], the average number of jobs in the system is equal to $(1 - P_{block}^{avg})N$ which is upper bounded by NC. Hence,

$$P_{block}^{avg} \quad \left(1 \quad \frac{C}{\quad}\right)_+ ;$$

where $(b)_+ = \max(b; 0)$. In Figure 2.2, we plot the lower bound $\left(1 \quad \frac{C}{\quad}\right)_+$ and the average blocking probability as a function of when the dispatcher uses the SQ(d) load balancing and the state-independent random routing where a destination server is chosen uniformly at random. It is clear that the resulting average blocking probability under the SQ(d) policy is much lower than the resulting average blocking probability when pure random routing is employed. Furthermore, the average blocking probability under the SQ(d) load balancing approaches the lower bound as d increases.

2.8 Proofs of Main Results

In this section, we provide proofs of the results stated in Section 2.4.

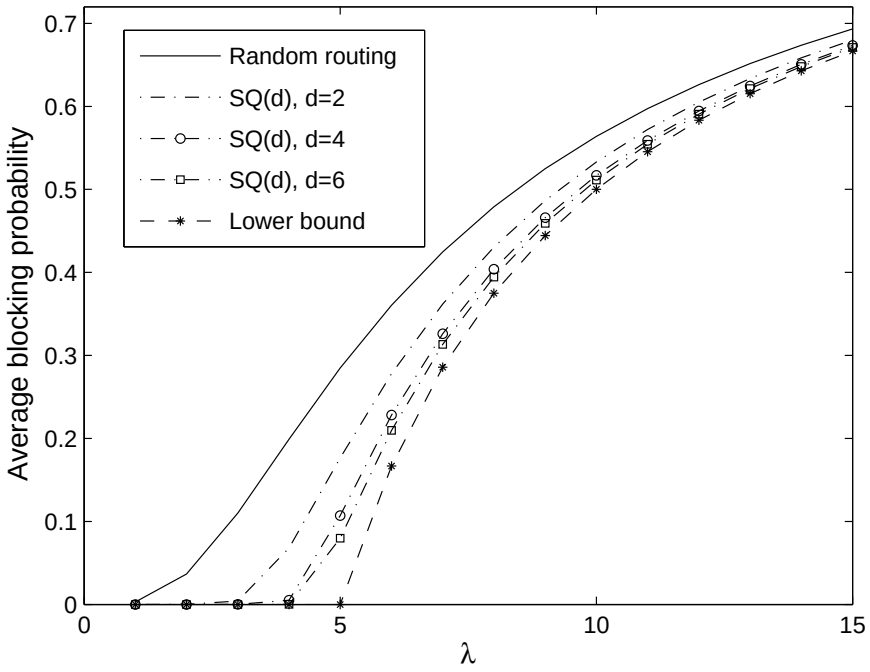

Figure 2.2: Comparison of the average blocking probability under SQ(d) with lower bound.

2.8.1 Proof Theorem 2.2

In this section, we first compute the semigroup of the Markov process ($_t^{(N)}; t \quad 0$) and then we show that the Markov process ($_t^{(N)}; t \quad 0$) is a Feller process. Finally, we construct a martingale using the generator of the Markov process ($_t^{(N)}; t \quad 0$) and the Dynkin's formula.

Conditioned on the initial state $_0^{(N)}$, let $\mathbf{A}_h$ and $\mathbf{D}_h$ be the number of arrivals and departures in the interval $[0; h]$, respectively. Note that a job with age x at time t departs from the system in the interval $[t; t+h]$ with the probability $\frac{G(x+h) \quad G(x)}{\overline{G}(x)}$. Further, from the definition of the hazard rate, we have that $\lim_{h \,!\, 0} \frac{1}{h} \frac{G(x+h) \quad G(x)}{\overline{G}(x)} = \quad (x)$ and hence,

$$\frac{G(x+h) \quad G(x)}{\overline{G}(x)} = \quad (x)h + o(h): \tag{2.42}$$

Let $T_h^{(N)}(\ )$ be the operator defined as

$$T_h^{(N)} f(\ ) = \mathsf{E}\left[f(\ _h^{(N)}) j \ _0^{(N)} = \ \right];$$

where f is a continuous bounded function $f : \mathsf{M}_F(\mathsf{U}) \ ! \ \mathsf{R}$ and the operator $T_h^{(N)}(\ )$ is a semigroup operator when ($_t^{(N)}; t \quad 0$) is a Markov process. Before computing the

expression for $T_h^{(N)} f(\ )$, we first introduce the following notation. Suppose the measure $_0^{(N)} =$ has mass at m atoms and let the i^{th} atom be $\mathbf{v}^{(i)} = (n_i; v_1^{(i)};\ ; v_{n_i}^{(i)})$ for $1 \quad i \quad m$ and let the number of servers with the state $\mathbf{v}^{(i)}$ be denoted by $(f\mathbf{v}^{(i)}g) = a^{(i)}$. If a server lies in the state $\mathbf{b}_n = (n; b_1; : : : ; b_n)$ at time t, let the probability that there is no departure in the interval $[t; t+h]$ be denoted by $p_{N\,D}(\mathbf{b}_n; h)$. We then have

$$p_{N\,D}(\mathbf{b}_n; h) = \prod_{i=1}^{n} \left\{ \frac{\overline{G}(b_i + h)}{\overline{G}(b_i)} \right\} : \tag{2.43}$$

Note that using (2.42), we can write

$$p_{N\,D}(\mathbf{b}_n; h) = \prod_{j=1}^{n} (1 \qquad (b_j)h) + o(h): \tag{2.44}$$

Lemma 2.5. *Let f be a real valued continuous bounded function de ned on $\mathsf{M}_F(\mathsf{U})$. Then the process $(\ _t^{(N)}; t \quad 0)$ is a Feller weak-homogeneous $\mathsf{M}_F(\mathsf{U})$-valued Markov process with semigroup operator $T_h^{(N)}(\ )$ given by*

$$\begin{aligned}
T_h^{(N)} f(\ ) = (1 \quad N \ \ h) & \left(\prod_{j=1;\, n_j>0}^{m} (p_{N\,D}(\mathbf{v}^{(j)}; h))^{a^{(j)}} \right) f(\ _h\ ) \\
& + (1 \quad N \ \ h) \sum_{j=1;\, n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)} \left\{ \frac{G(v_r^{(j)} + h) \quad G(v_r^{(j)})}{\overline{G}(v_r^{(j)})} \right\} \\
& \left(\prod_{w=1;\, w \in r}^{n_j} \left\{ \frac{\overline{G}(v_w^{(j)} + h)}{\overline{G}(v_w^{(j)})} \right\} \right) (p_{N\,D}(\mathbf{v}^{(j)}; h))^{(a^{(j)} \ \ 1)} \left(\prod_{i=1;\, n_i>0;\, i \in j}^{m} (p_{N\,D}(\mathbf{v}^{(i)}; h))^{a^{(i)}} \right) \\
& f(\ _h \ + \ ((\ _h^{+}\mathbf{v}^{(j)})\ _r) \qquad (\ _h^{+}\mathbf{v}^{(i)})) \\
& + (N \ \ h) \int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \frac{1}{n_i+1} p_r(\ _x\ : \mathbf{v}^{(i)}) \right. \\
& \left[I_{fn_i<Cg} f\left(\ _h \ + \ ((\ _h^{+}\mathbf{v}^{(i)})^j; h \ \ x) \qquad (\ _h^{+}\mathbf{v}^{(i)})\right) \prod_{k=1;\, n_k>0}^{m} (p_{N\,D}(\mathbf{v}^{(k)}; h))^{a^{(k)}} \overline{G}(h \quad x) \right. \\
& \left. \left. + I_{fn_i=Cg} f\left(\ _h\ \right) \prod_{k=1;\, n_k>0}^{m} (p_{N\,D}(\mathbf{v}^{(k)}; h))^{a^{(k)}} \right] \right) dx + \ (\ ; h); \qquad (2.45)
\end{aligned}$$

where $(\ ; h)$ is $o(h)$ for all . Moreover, the process $(\ _t^{(N)}; t \quad 0)$ is a Feller-Dynkin process.

Proof. We can write

$$T_h^{(N)} f(\ ) = \mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=0;\,\mathbf{D}_h=0g} j\ {}_0^{(N)} = \ \right] + \mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=0;\,\mathbf{D}_h=1g} j\ {}_0^{(N)} = \ \right]$$
$$+ \mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=1;\,\mathbf{D}_h=0g} j\ {}_0^{(N)} = \ \right] + \sum_{i\ 1;\,j\ 1} \mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=i;\,\mathbf{D}_h=jg} j\ {}_0^{(N)} = \ \right]: \quad (2.46)$$

The proof mainly depends on the simplification of the four terms on the right side of (2.46).

We first simplify the first term on the right side of equation (2.46). In this case, since there are no arrivals or departures, we have ${}_h^{(N)} = \ {}_h$. As a consequence, we have

$$\mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=0;\,\mathbf{D}_h=0g} j\ {}_0^{(N)} = \ \right] = f(\ {}_h\ )\mathsf{P}\left((f\mathbf{A}_h = 0;\ \mathbf{D}_h = 0g) j\ {}_0^{(N)} = \ \right): \quad (2.47)$$

Further, we can write

$$\mathsf{P}\left((f\mathbf{A}_h = 0;\ \mathbf{D}_h = 0g) j\ {}_0^{(N)} = \ \right) = \mathsf{P}\left(f\mathbf{A}_h = 0gj\ {}_0^{(N)} = \ \right) \mathsf{P}\left((fD_h = 0g) j\mathbf{A}_h = 0;\ {}_0^{(N)} = \ \right):$$

Since the arrival process is a Poisson process with rate N , it is independent of the state . Therefore, we have $\mathsf{P}\left(f\mathbf{A}_h = 0gj\ {}_0^{(N)} = \ \right) = \mathsf{P}(f\mathbf{A}_h = 0g) = e^{\ (N\ h)}$: On the other hand, the number of departures $\mathbf{D}_h$ is influenced by the number of arrivals $\mathbf{A}_h$. Hence, we need to compute the expression for $\mathsf{P}\left((f\mathbf{D}_h = jg) j\mathbf{A}_h = i;\ {}_0^{(N)} = \ \right)$ that gives the probability that there are j departures in the interval $[0; h]$ conditioned on the event that there are i arrivals in the interval $[0; h]$. If the arrival process is a Poisson process, conditioned on the number of arrivals $\mathbf{A}_h$, the arrival instants are independent random variables with uniform distribution in the interval $[0; h]$ [77, p. 325]. It can be seen that

$$\mathsf{P}\left((f\mathbf{D}_h = 0g) j\mathbf{A}_h = 0;\ {}_0^{(N)} = \ \right) = \prod_{j=1;\,n_j>0}^{m} (p_{N\,D}(\mathbf{v}^{(j)}; h))^{a^{(j)}}:$$

We can write

$$\mathsf{E}\left[f(\ {}_h^{(N)}) I_{f\mathbf{A}_h=0;\,\mathbf{D}_h=0g} j\ {}_0^{(N)} = \ \right] = f1 \quad N\ \ hg \left(\prod_{j=1;\,n_j>0}^{m} (p_{N\,D}(\mathbf{v}^{(j)}; h))^{a^{(j)}} \right) f(\ {}_h\ )$$
$$+\ {}_1(\ ; h);$$

where

$${}_1(\ ; h) = f\mathsf{P}(f\mathbf{A}_h = 0g) \quad (1 \quad N\ \ h)g \prod_{j=1;\,n_j>0}^{m} (p_{N\,D}(\mathbf{v}^{(j)}; h))^{a^{(j)}} f(\ {}_h\ )$$

is $o(h)$ for all .

Similarly, we can write the second term of the right side of (2.46) as

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{\{\mathbf{A}_h=0;\,\mathbf{D}_h=1\}} j\ {}^{(N)}_{0} = \ \right] = (1 \quad N \quad h) \sum_{j=1;\,n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)} \left\{ \frac{G(v_r^{(j)}+h) \quad G(v_r^{(j)})}{\overline{G}(v_r^{(j)})} \right\}$$
$$\left(\prod_{w=1;\,w\in r}^{n_j} \left\{ \frac{\overline{G}(v_w^{(j)}+h)}{\overline{G}(v_w^{(j)})} \right\} \right) (p_{N\,D}(\mathbf{v}^{(j)};h))^{(a^{(j)} \ 1)} \left(\prod_{i=1;n_i>0;i\in j}^{m} (p_{N\,D}(\mathbf{v}^{(i)};h))^{a^{(i)}} \right)$$
$$f(\ h \ + \ (({}^{+}_{h}\mathbf{v}^{(j)}) \ r) \quad ({}^{+}_{h}\mathbf{v}^{(i)})) + \ {}_2(\ ;h);$$

where we use r to denote the index of the departing job at a server with the state $\mathbf{v}^{(j)}$ and ${}_2(\ ;h)$ is $o(h)$ for all $\quad$ given by

$${}_2(\ ;h) = f\mathsf{P}(f\mathbf{A}_h = 0g) \quad (1 \quad N \quad h)g \sum_{j=1;\,n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)}$$
$$\left\{ \frac{G(v_r^{(j)}+h) \quad G(v_r^{(j)})}{\overline{G}(v_r^{(j)})} \right\} \left(\prod_{w=1;\,w\in r}^{n_j} \left\{ \frac{\overline{G}(v_w^{(j)}+h)}{\overline{G}(v_w^{(j)})} \right\} \right) (p_{N\,D}(\mathbf{v}^{(j)};h))^{(a^{(j)} \ 1)}$$
$$\left(\prod_{i=1;\,n_i>0;i\in j}^{m} (p_{N\,D}(\mathbf{v}^{(i)};h))^{a^{(i)}} \right) f(\ h \ + \ (({}^{+}_{h}\mathbf{v}^{(j)}) \ r) \quad ({}^{+}_{h}\mathbf{v}^{(i)})):$$

We next compute the third term on the right side of (2.46). We can write

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{\{\mathbf{A}_h=1;\,\mathbf{D}_h=0\}} j\ {}^{(N)}_{0} = \ \right] = f\mathsf{P}(f\mathbf{A}_h = 1g)g$$
$$\int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \left\{ \frac{1}{n_i+1} \right\} p_r(\ x \ : \ {}^{+}_{x}\mathbf{v}^{(i)}) \left[I_{\{fn_i<Cg\}} f\left(\ h \ + \ (({}^{+}_{h}\mathbf{v}^{(i)})^{j};h \ \ x) \quad ({}^{+}_{h}\mathbf{v}^{(i)}) \right) \right.\right.$$
$$\left.\left. \prod_{k=1;n_k>0}^{m} (p_{N\,D}(\mathbf{v}^{(k)};h))^{a^{(k)}} \overline{G}(h \quad x) + I_{\{fn_i=Cg\}} f(\ h \ ) \prod_{k=1;\,n_k>0}^{m} (p_{N\,D}(\mathbf{v}^{(k)};h))^{a^{(k)}} \right] \right) dx;$$

where the arrival instant x is chosen uniformly in $[0;h]$ given $\mathbf{A}_h = 1$, i denotes the index of the atom corresponding to the state of the destination server, and j is the position of the routed job at the destination server chosen uniformly at random from $n_i + 1$ positions. Further, we write

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{\{\mathbf{A}_h=1;\,\mathbf{D}_h=0\}} j\ {}^{(N)}_{0} = \ \right] = (N \quad h) \int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \left\{ \frac{1}{n_i+1} \right\} p_r(\ x \ : \ {}^{+}_{x}\mathbf{v}^{(i)}) \right.$$
$$\left. \left[I_{\{fn_i<Cg\}} f\left(\ h \ + \ (({}^{+}_{h}\mathbf{v}^{(i)})^{j};h \ \ x) \quad ({}^{+}_{h}\mathbf{v}^{(i)}) \right) \prod_{k=1;n_k>0}^{m} (p_{N\,D}(\mathbf{v}^{(k)};h))^{a^{(k)}} \overline{G}(h \quad x) \right.\right.$$

$$+I_{fn_i=Cg} f(\quad h\quad)\prod_{k=1;n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)};h))^{a^{(k)}}\Bigg]\Bigg)\,dx+\quad{}_3(\quad;h);$$

where

$$\begin{aligned}{}_3(\quad;h)=f\mathsf{P}(f\mathbf{A}_h=1g)\quad N\quad hg\int_{x=0}^{h}\frac{1}{h}\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\left\{\frac{1}{n_i+1}\right\}p_r(\quad x\quad:\quad {}^{+}_{x}\mathbf{v}^{(i)})\\ \Bigg[I_{fn_i<Cg} f\Bigg(\quad h\quad+\quad((\,{}^{+}_{h}\mathbf{v}^{(i)})^{j};h\quad x)\qquad({}^{+}_{h}\mathbf{v}^{(i)})\Bigg)\prod_{k=1;n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)};h))^{a^{(k)}}\overline{G}(h\quad x)\\ +I_{fn_i=Cg} f(\quad h\quad)\prod_{k=1;n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)};h))^{a^{(k)}}\Bigg]\Bigg)\,dx:\end{aligned}$$

We now show that ${}_3(\quad;h)$ is a $o(h)$ term for all . For this, we apply the method of change of variables by replacing x with hy. As a consequence, we have

$$\begin{aligned}{}_3(\quad;h)=f\mathsf{P}(f\mathbf{A}_h=1g)\quad N\quad hg\,h\int_{y=0}^{1}\frac{1}{h}\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\left\{\frac{1}{n_i+1}\right\}p_r(\quad hy\quad:\quad {}^{+}_{hy}\mathbf{v}^{(i)})\\ \Bigg[I_{fn_i<Cg} f\Bigg(\quad h\quad+\quad((\,{}^{+}_{h}\mathbf{v}^{(i)})^{j};h\quad hy)\qquad({}^{+}_{h}\mathbf{v}^{(i)})\Bigg)\prod_{k=1;n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)};h))^{a^{(k)}}\overline{G}(h\quad hy)\\ +I_{fn_i=Cg} f(\quad h\quad)\prod_{k=1;n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)};h))^{a^{(k)}}\Bigg]\Bigg)\,dy:\end{aligned}$$

By using the dominated convergence theorem [69, Theorem 1.34], we have $\lim_{h\,!\,0}\frac{{}_3(\quad;h)}{h}=0$ for all .

Finally, by using the fact that f is a bounded function, we now prove that the fourth term on the right side of (2.46) is a $o(h)$ term denoted by ${}_4(\quad;h)$. Since $f\ 2\ \mathsf{C}_b(\mathsf{M}_F^N(\mathsf{U}))$, it is enough to prove that $\sum_{i\quad 1;j\quad 1}\mathsf{P}(f\mathbf{A}_h=i;\,\mathbf{D}_h=jgj\ {}^{(N)}_{0}=\quad)$ is a $o(h)$ term. Due to the result that $f\sum_{i\quad 2;j\quad 1}\mathsf{P}(f\mathbf{A}_h=i;\,\mathbf{D}_h=jgj\ {}^{(N)}_{0}=\quad)g\quad\mathsf{P}(f\mathbf{A}_h\quad 2g)$ and $\mathsf{P}(f\mathbf{A}_h\quad 2g)$ is a $o(h)$ term, we conclude that $\sum_{i\quad 2;j\quad 1}\mathsf{P}(f\mathbf{A}_h=i;\,\mathbf{D}_h=jgj\ {}^{(N)}_{0}=\quad)$ is a $o(h)$ term for all . We now show that $f\sum_{j\quad 1}\mathsf{P}(f\mathbf{A}_h=1;\,\mathbf{D}_h=jgj\ {}^{(N)}_{0}=\quad)g$ is a $o(h)$ term. We can write

$$\begin{aligned}\Big(\mathsf{P}(f\mathbf{A}_h=1;\,\mathbf{D}_h\quad 1gj\ {}^{(N)}_{0}=\quad)\Big)&=\mathsf{P}(f\mathbf{A}_h=1gj\ {}^{(N)}_{0}=\quad)\quad\mathsf{P}(f\mathbf{A}_h=1;\,\mathbf{D}_h=0gj\ {}^{(N)}_{0}=\quad)\\ &=\mathsf{P}(f\mathbf{A}_h=1g)(1\quad\mathsf{P}(f\mathbf{D}_h=0gj\mathbf{A}_h=1;\ {}^{(N)}_{0}=\quad)):\end{aligned}$$

Again, by using the method of change of variables and the dominated convergence theorem as in the proof of the result that shows ${}_3(\quad;h)$ is a $o(h)$ term, we get that $\lim_{h\,!\,0}\mathsf{P}(f\mathbf{D}_h=$

$0gj\mathbf{A}_h = 1;\ {}^{(N)}_0 = \ ) = 1$ for all . Since $\lim_{h\,!\,0} \frac{\mathsf{P}(f\mathbf{A}_h = 1g)}{h} = N$, for all , we have that $\mathsf{P}(f\mathbf{A}_h = 1;\ \mathbf{D}_h \quad 1gj\ {}^{(N)}_0 = \ )$ is $o(h)$. Therefore, ${}_4(\ ;h)$ is $o(h)$ for all .

By combining the expressions for all the four terms on the right side of (2.46), and by defining $(\ ;h) = {}_1(\ ;h) + {}_2(\ ;h) + {}_3(\ ;h) + {}_4(\ ;h)$; we get the expression for $T_h^{(N)} f(\ )$ as in (2.45). Finally, from [16, p.18], $(\ {}^{(N)}_t; t \quad 0)$ is a weak homogeneous Markov process.

We next give the proof of the Feller-Dynkin property of $(\ {}^{(N)}_t; t \quad 0)$. From Lemma 3.5.1 and Corollary 3.5.2 of [16], we need to prove the following conditions to show the Feller-Dynkin property of $(\ {}^{(N)}_t; t \quad 0)$. For $f\ 2\ \mathsf{C}_s^1(\mathsf{U})$, $\ 2\ \mathsf{M}_F(\mathsf{U})$, let $Q_f : \mathsf{M}_F(\mathsf{U})\ 7\ !\ \mathsf{R}$ be defined as $Q_f(\ ) = e^{\ h\ ;f\,i}$, then we must have

1. For all $f\ 2\ \mathsf{C}_s^1(\mathsf{U})$ and $h > 0$, the mapping $\ 7\ !\ \mathsf{E}\left[Q_f(\ {}^{(N)}_h)j\ {}^{(N)}_0 = \ \right]$ is continuous.

2. For all $h > 0$, we have
$$\mathsf{E}\left[Q_1(\ {}^{(N)}_h)j\ {}^{(N)}_0 = \ \right]\ !\ 0 \tag{2.48}$$
as $(\mathsf{U})\ !\ 1$.

3. For all $\ 2\ \mathsf{M}_F(\mathsf{U})$ and $f\ 2\ \mathsf{C}_s^1(\mathsf{U})$, we have
$$\mathsf{E}\left[Q_f(\ {}^{(N)}_h)j\ {}^{(N)}_0 = \ \right]\ !\ Q_f(\ ) \tag{2.49}$$
as $h\ !\ 0$.

From (2.45), we obtain

$$\begin{aligned}
\mathsf{E}\left[Q_f(\ {}^{(N)}_h)j\ {}^{(N)}_0 = \ \right] = e^{\ h\ h\ ;f\,i}\Bigg(& (1 \quad N\ \ h)\left(\prod_{j=1;\,n_j>0}^{m}(p_{N\,D}(\mathbf{v}^{(j)};h))^{a^{(j)}}\right) \\
& + (1 \quad N\ \ h)\sum_{j=1;\,n_j>0}^{m}\sum_{r=1}^{n_j} a^{(j)}\left\{\frac{G(v_r^{(j)}+h) \quad G(v_r^{(j)})}{\overline{G}(v_r^{(j)})}\right\} \\
& \left(\prod_{w=1;\,w6r}^{n_j}\left\{\frac{\overline{G}(v_w^{(j)}+h)}{\overline{G}(v_w^{(j)})}\right\}\right)(p_{N\,D}(\mathbf{v}^{(j)};h))^{(a^{(j)}\ \ 1)}\left(\prod_{i=1;\,n_i>0;i6j}^{m}(p_{N\,D}(\mathbf{v}^{(i)};h))^{a^{(i)}}\right) \\
& Q_f(\ (({}^{+}_h\mathbf{v}^{(j)})\ {}^{r}) \qquad ({}^{+}_h\mathbf{v}^{(i)}))\Bigg) \\
& + (N\ \ h)\int_{x=0}^{h}\frac{1}{h}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\left\{\frac{1}{n_i+1}\right\}p_r(\ {}_x\ :\mathbf{v}^{(i)})\right.
\end{aligned}$$

$$\left[I_{fn_i<Cg} Q_f \left(\left({}^{+}_{h}\mathbf{v}^{(i)} \right)^{j}; h \quad x \right) \quad \left({}^{+}_{h}\mathbf{v}^{(i)} \right) \right) \prod_{k=1;\, n_k>0}^{m} \left(p_{ND}(\mathbf{v}^{(k)}; h) \right)^{a^{(k)}} \overline{G}(h \quad x) \right.$$
$$\left. + I_{fn_i=Cg} \prod_{k=1;\, n_k>0}^{m} \left(p_{ND}(\mathbf{v}^{(k)}; h) \right)^{a^{(k)}} \right] \Bigg) \, dx + \quad {}_f(\quad ; h) \Bigg) \quad (2.50)$$

where

$${}_f(\quad ; h) = \quad {}_{1f}(\quad ; h) + \quad {}_{2f}(\quad ; h) + \quad {}_{3f}(\quad ; h) + \quad {}_{4f}(\quad ; h)$$

such that

$${}_{1f}(\quad ; h) = (\mathrm{P}(f\mathbf{A}_h = 0g) \quad (1 \quad N \quad h)) \left(\prod_{j=1;\, n_j>0}^{m} \left(p_{ND}(\mathbf{v}^{(j)}; h) \right)^{a^{(j)}} \right);$$

$${}_{2f}(\quad ; h) = (\mathrm{P}(f\mathbf{A}_h = 0g) \quad (1 \quad N \quad h)) \sum_{j=1;\, n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)} \left\{ \frac{G(v_r^{(j)} + h) \quad G(v_r^{(j)})}{\overline{G}(v_r^{(j)})} \right\}$$
$$\left(\prod_{w=1;\, w \in r}^{n_j} \left\{ \frac{\overline{G}(v_w^{(j)} + h)}{\overline{G}(v_w^{(j)})} \right\} \right) \left(p_{ND}(\mathbf{v}^{(j)}; h) \right)^{(a^{(j)} \quad 1)} \left(\prod_{i=1;\, n_i>0;\, i \in j}^{m} \left(p_{ND}(\mathbf{v}^{(i)}; h) \right)^{a^{(i)}} \right)$$
$$Q_f(\quad (({}^{+}_{h}\mathbf{v}^{(j)}) \quad {}_r) \quad ({}^{+}_{h}\mathbf{v}^{(i)}));$$

$${}_{3f}(\quad ; h) = (\mathrm{P}(f\mathbf{A}_h = 1g) \quad N \quad h) \int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \left\{ \frac{1}{n_i + 1} \right\} p_r(\quad x \quad : \mathbf{v}^{(i)}) \right.$$
$$\left[I_{fn_i<Cg} Q_f \left(\left({}^{+}_{h}\mathbf{v}^{(i)} \right)^{j}; h \quad x \right) \quad \left({}^{+}_{h}\mathbf{v}^{(i)} \right) \right) \prod_{k=1;\, n_k>0}^{m} \left(p_{ND}(\mathbf{v}^{(k)}; h) \right)^{a^{(k)}} \overline{G}(h \quad x)$$
$$\left. \left. + I_{fn_i=Cg} \prod_{k=1;\, n_k>0}^{m} \left(p_{ND}(\mathbf{v}^{(k)}; h) \right)^{a^{(k)}} \right] \right) dx;$$

and

$${}_{4f}(\quad ; h) = \sum_{i \quad 1;\, j \quad 1} \mathrm{E} \Bigg[Q_f \left(\sum_{r=1}^{i} \left({}_{(S((M_r; Z_{1r}; \quad ; Z_{M_r r}); L_r; h; T_r))} \quad {}_{(M_r; Z_{1r}+h \quad T_r; \quad ; Z_{M_r r}+h \quad T_r)} \right) \right.$$
$$\left. + \sum_{l=1}^{j} \left({}_{(Y((n_l; X_{1l}; \quad ; X_{n_l l}); K_l; h; W_l))} \quad {}_{(n_l; X_{1l}+h \quad W_l; \quad ; X_{n_l l}+h \quad W_l)} \right) \right)$$
$$I_{f\mathbf{A}_h = i; \mathbf{D}_h = jg} j \quad {}^{(N)}_{0} = \quad \Bigg]; \quad (2.51)$$

where in equation (2.51), T_r is the arrival time of the r^{th} arrival and this job is routed to a server with state $(M_r; Z_{1r};\ \ ; Z_{M_r r})$ at time T_r and L_r is its position at the destination server. Further, $S((M_r; Z_{1r};\ \ ; Z_{M_r r}); L_r; h; T_r)$ denotes the possible state of the r^{th} arrival's destination at time $t = h$. Similarly, for the l^{th} departure, W_l is the departure time and this job departs from a server with state $(n_l; X_{1l};\ \ ; X_{n_l l})$ at time W_l and its position is K_l. Further, $Y((n_l; X_{1l};\ \ ; X_{n_l l}); K_l; h; W_l)$ is the possible state of the server where the l^{th} departure occurs at time $t = h$. It can be checked that $_f(\ ; h)$ is also $o(h)$ and this follows from the same arguments as that of $(\ ; h)$ in (2.45).

We now prove the first condition required to establish the Feller property. For this, we write (2.50) as

$$\mathsf{E}\left[Q_f(\ {}^{(N)}_{h})j\ {}^{(N)}_{0} = \ \right] = (e^{\ h\ h\ ;f\,i})V(\ ; h):$$

The mapping $e^{\ h\ h\ ;f\,i}$ is a continuous mapping of $\ $. To prove the first condition, we need to show $V(\ ; h)$ is a continuous mapping of $\ $. Since $\ $ is a point measure at finite N, the continuity of $V(\ ; h)$ *w:r:t:* $\ $ follows from the continuity of the routing probabilities and the departure probabilities. The second condition follows due to the fact that $_h\ (\mathsf{U}) = \ (\mathsf{U}) = N$. Finally, the third condition follows from the relation $h\ _h\ ; f\,i = h\ ;\ _h f\,i$ and the dominated convergence theorem.

□

We next compute the infinitesimal generator of the Markov process $(\ {}^{(N)}_{t}; t\ \ 0)$, and then we construct a martingale $(\mathbb{M}^{(N)}_t(\ ); t\ \ 0)\ \ 2\ \ D_{\mathsf{R}}([0;\ 1))$ where $\ \ 2\ \ \mathsf{C}^1_b(\mathsf{U})$ by applying the Dynkin's formula.

Since the set of linear combinations of $Q_f : \mathsf{M}_F(\mathsf{U})\ 7\ !\ \ \mathsf{R}$ for $f\ \ 2\ \ \mathsf{C}^1_s(\mathsf{U})$ defined by $Q_f(\ ) = e^{\ h\ ;f\,i}$ is dense in the set $\mathsf{C}(\mathsf{M}_F(\mathsf{U}))$ [78, proposition 7:10], by using $A^{(N)}Q_f(\ )$, for any continuous function $F\ \ 2\ \ \mathsf{C}(\mathsf{M}_F(\mathsf{U}))$ such that the infinitesimal generator $A^{(N)}F(\ ) = \lim_{h\ !\ 0} \frac{\mathsf{E}[F(\ {}^{(N)}_{h})j\ {}^{(N)}_{0} = \ ]\ \ F(\ )}{h}$ is well-defined for all $\ $, we have

$$A^N F(\ ) = \lim_{h\ !\ 0} \frac{F(\ _h\ )\ \ F(\ )}{h}\ \ N\ \ F(\ )\ \ \ F(\ )\sum_{n=1}^{C}\sum_{j=1}^{n}\int\ \ \int_{\mathsf{U}_n}\ (x_j)\,d\ (\mathbf{x}_n)$$

$$+\sum_{n=1}^{C}\sum_{j=1}^{n}\int\ \ \int_{\mathsf{U}_n}\ (x_j)\left(F(\ +\ _{(\mathbf{x}_n{}^j)}\ \ \ _{(\mathbf{x}_n)}\right)d\ (\mathbf{x}_n)$$

$$+N\ \left[\left(\left\{\frac{(f0g)}{N}\right\}\Phi_0\left(\frac{}{N}\right)(F(\ +\ _{(1;0)}\ \ \ _{(0)}))\right)+\sum_{n=1}^{C}\sum_{j=1}^{n+1}\int\ \ \int_{\mathsf{U}_n}\left\{\frac{1}{N(n+1)}\right\}\Phi_n\left(\frac{}{N}\right)\right.$$

$$F(\ +\ _{(x_n^j;0)}\quad _{(x_n)})\,d\ (x_n)+\int\quad\int_{U_C}\frac{1}{N}\Phi_C\left(\frac{}{N}\right)F(\ )\,d\ (x_C)\Bigg]: \quad (2.52)$$

For $\ 2\ C_b^1(U)$, it can be seen that the function $\ 2\ M_F(U)$ 7! $h\ ;\ i\ 2\ R$ belongs to the domain of $A^{(N)}(\ )$

Proposition 2.3. *For all* $\ 2\ C_b^1(U)$*, the process* $(M_t^{(N)}(\ );t\quad 0)$ *given by*

$$M_t^{(N)}(\ )=h\ _t^{(N)};\ \ i\quad h\ _0^{(N)};\ \ i\quad\int_{s=0}^{t}A^{(N)}h\ _s^{N};\ \ i\,ds \quad (2.53)$$

is an RCLL square integrable F_t^N *martingale. For* $\ 2\ C_b^1(U)$*, the quadratic variation of* $(M_t^{(N)}(\ );t\quad 0)$ *is given by*

$$\begin{aligned}<M^{(N)}(\ )>_t=&\int_{s=0}^{t}\left(\sum_{n=1}^{C}\sum_{j=1}^{n}\int\quad\int_{U_n}(x_j)\left(\ (x_n^{\,j})\quad (x_n)\right)^2\,d\ _s^{(N)}(x_n)\right.\\&+N\ \left[\left(\left\{\frac{_s^{(N)}(f0g)}{N}\right\}\Phi_0\left(\frac{_s^{(N)}}{N}\right)(\ (1;0)\quad (0)\right)^2\right)\\&\left.+\sum_{n=1}^{C}\sum_{j=1}^{1}\sum_{}^{n+1}\int\quad\int_{U_n}\left\{\frac{1}{N(n+1)}\right\}\Phi_n\left(\frac{_s^{(N)}}{N}\right)(\ (x_n^j;0)\quad (x_n))^2\,d\ _s^{(N)}(x_n)\right]\Bigg)\,ds: \end{aligned}\quad (2.54)$$

Proof. From the Dynkin's formula [56], the process $(M_t^{(N)}(\ );t\quad 0)$ defined by

$$M_t^{(N)}(\ )=h\ _t^{(N)};\ \ i\quad h\ _0^{(N)};\ \ i\quad\int_{s=0}^{t}A^{(N)}h\ _s^{N};\ \ i\,ds \quad (2.55)$$

is an RCLL F_t^N local martingale. Therefore, by simplification we get

$$\begin{aligned}M_t^{(N)}(\ )=&h\ _t^{(N)};\ \ i\quad h\ _0^{(N)};\ \ i\quad\int_{s=0}^{t}h\ _s^{(N)};r_1\ \ i\,ds\quad\int_{s=0}^{t}\left(\sum_{n=1}^{C}\sum_{j=1}^{n}\int\quad\int_{U_n}(x_j)\right.\\&\left.\left(\ (x_n^{\,j})\quad (x_n)\right)d\ _s^{(N)}(x_n)\right.\\&+N\ \left[\left(\frac{_s^{(N)}(f0g)}{N}\Phi_0\left(\frac{_s^{(N)}}{N}\right)(\ (1;0)\quad (0)\right)\right.\\&\left.\left.+\sum_{n=1}^{C}\sum_{j=1}^{1}\sum_{}^{n+1}\int\quad\int_{U_n}\left\{\frac{1}{N(n+1)}\right\}\Phi_n\left(\frac{_s^{(N)}}{N}\right)(\ (x_n^j;0)\quad (x_n))\,d\ _s^{(N)}(x_n)\right]\right)ds: \end{aligned}\quad (2.56)$$

By choosing $F\ (\ _t^{(N)})=h\ _t^{(N)};\ \ i$, from [79, Theorem 7.15], we have

$$<M^{(N)}(\ )>_t=\int_{s=0}^{t}\left(A^{(N)}F^2(\ _s^{(N)})\quad 2F\ (\ _s^{(N)})A^{(N)}F\ (\ _s^{(N)})\right)ds: \quad (2.57)$$

After simplifications, we get (2.54). Finally, since $\ 2\ C_b^1(U)$ and $\ 2\ C_b(R_+)$, we have $E\left[<M^{(N)}(\ )>_t\right]<\ 1$ and hence, $(M_t^{(N)}(\ );t\quad 0)$ is a square integrable martingale. □

2.8.2 Proof of Lemma 2.2

We first show that any process $(\ _t; t \quad 0)$ that satisfies (2.10) also satisfies (2.11). By using the fundamental theorem of calculus [69, p.144], for $\ 2\ C_b^1(U)$, a real valued process $(h\ _t;\ i; t \quad 0)$ satisfying equation (2.10) is a solution to the following differential equation (2.58) if the integrand in equation (2.10) is a continuous function of s,

$$\frac{dh\ _t;\ i}{dt} = h\ _t; r\ _1\ \ i + \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \ (x_j) \left(\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n) \right) d\ _t(\mathbf{x}_n) \right.$$
$$+ \left[\ _t(f0g)\ \ \Phi_0(\ _t)\ (\ (1; 0) \quad (0)) + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{j=1}^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \right.$$
$$\left. \left. \Phi_n(\ _t)(\ (\mathbf{x}_n^j; 0) \quad (\mathbf{x}_n))\, d\ _t(\mathbf{x}_n) \right] \right): \quad (2.58)$$

Therefore, we need to show that the two terms on the right side of (2.58) are continuous functions of t. Since $\ 2\ C_b^1(U)$ and the mapping $t\ 7!\ \ _t$ is continuous, the first term $h\ _t; r\ _1\ \ i$ is a continuous function of t. In the second term, the expression related to the case of departures can be written as

$$\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \ (x_j) \left(\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n) \right) d\ _t(\mathbf{x}_n) = h\ _t;\ \ _1 i;$$

where the function $\ _1$ is defined as

$$_1(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = 0; \\ \sum_{j=1}^{n} \ (x_j)((\ (\mathbf{x}_n{}^j) \quad (\mathbf{x}_n)) & \text{otherwise:} \end{cases}$$

Since $\ 2\ C_b^1(U)$ and $\ 2\ C_b^1(R_+)$, we have that $\ _1\ 2\ C_b(U)$. Therefore the mapping $t\ 7!\ h\ _t;\ \ _1 i$ is continuous. The expression that corresponds to the case of arrivals can be written as

$$h\ _t;\ \ _{(\ t)} i = (\ _t(f0g)\ \ \Phi_0(\ _t)\ (\ (1; 0) \quad (0)))$$
$$+ \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{j=1}^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \ \Phi_n(\ _t)(\ (\mathbf{x}_n^j; 0) \quad (\mathbf{x}_n))\, d\ _t(\mathbf{x}_n);$$

where $\ _{(\ t)}$ is defined as

$$_{(\ t)}(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = C; \\ \left\{ \frac{n(\ t)}{n+1} \right\} \sum_{j=1}^{n+1} (\ (\mathbf{x}_n^j; 0) \quad (\mathbf{x}_n)) & \text{otherwise:} \end{cases} \quad (2.59)$$

For given $_t$, since $\ 2\ C_b(U)$, we have that $\ _{(\ t)}\ 2\ C_b(U)$. Hence, for any constant $a\quad 0$, the mapping $t\ 7!\ h\ _t;\ _{(\ a)}i$ is continuous.

We next prove that the mapping $t\ 7!\ h\ _t;\ _{(\ t)}i$ is continuous, i.e., we need to prove that $h\ _{t+b};\ _{(\ t+b)}i\ !\ h\ _t;\ _{(\ t)}i$ as $b\ !\ 0$. We have

$$\left|h\ _{t+b};\ _{(\ t+b)}i\quad h\ _t;\ _{(\ t)}i\right|\quad \left|h\ _{t+b};\ _{(\ t+b)}i\quad h\ _{t+b};\ _{(\ t)}i\right|+\left|h\ _{t+b};\ _{(\ t)}i\quad h\ _t;\ _{(\ t)}i\right|: \tag{2.60}$$

Since $\ _{(\ t)}\ 2\ C_b(U)$, we have that $\lim_{b\ !\ 0}\left|h\ _{t+b};\ _{(\ t)}i\quad h\ _t;\ _{(\ t)}i\right|=0$: We next prove that $\lim_{b\ !\ 0}\left|h\ _{t+b};\ _{(\ t+b)}\quad\ _{(\ t)}i\right|=0$:

For $L\ >\ 0$, let

$$U^{(L)}=fx_n\ 2\ U_n:n\quad 1;x_i\ >\ L\ \text{for all}\ 1\quad i\quad ng:$$

For given $\ >\ 0$, since $\ _t$ is tight, we can find some $L\ >\ 0$ such that $h\ _t;I_{fU^{(L)}g}i\ <\ :$ Furthermore, from the continuity of the mapping $t\ 7!\ \ _t$, we can find some $h_1\ >\ 0$ such that for all $b\ 2\ [\ \ \min(t;h_1);h_1]$,

$$h\ _{t+b};I_{fU^{(L)}g}i\ <\ : \tag{2.61}$$

By using the fact that the mapping $t\ 7!\ \overline{R}_n(\ _t)\ =\ h\ _t;I_{f\bigcap_{j=n}^{c}U_jg}i$ is continuous, we next show that the mapping $t\ 7!\ \ _{(\ t)}$ is continuous. For this, we need to show that $k\ _{(\ t+b)}\quad\ _{(\ t)}k\ !\ 0$ as $b\ !\ 0$. From (2.59), we have

$$\begin{aligned}k\ _{(\ t+b)}\quad\ _{(\ t)}k\quad & 2\ k\ k\max_n(j\Phi_n(\ _{t+b})\quad\Phi_{n+1}(\ _t)j)\\ & 4d\ k\ k\max_n(\left|\overline{R}_n(\ _{t+b})\quad\overline{R}_n(\ _t)\right|):\end{aligned} \tag{2.62}$$

Since $\left|\overline{R}_n(\ _{t+b})\quad\overline{R}_n(\ _t)\right|\ !\ 0$ as $b\ !\ 0$ for all n, $k\ _{(\ t+b)}\quad\ _{(\ t)}k\ !\ 0$. This proves that the mapping $t\ 7!\ \ _{(\ t)}$ is continuous. As a consequence, we have that $\ _{(\ t+b)}$ is uniformly continuous on the interval $b\ 2\ [\ \ \min(t;h_1);h_1]$ and $\mathbf{u}\ 2\ \overline{U}^{(L)}$ (the complement of $U^{(L)}$). As a result, there exists some $h_2\ 2\ (0;h_1)$ such that for $b\ 2\ [\ \ \min(t;h_2);h_2]$, $\mathbf{u}\ 2\ \overline{U}^{(L)}$, we have

$$\left|\ _{(\ t+b)}(\mathbf{u})\quad\ _{(\ t)}(\mathbf{u})\right|\ <\ : \tag{2.63}$$

Using (2.61)-(2.63), for $b\ 2\ [\ \ \min(t;h_2);h_2]$, we have

$$\left|h\ _{t+b};\ _{(\ t+b)}\quad\ _{(\ t)}i\right|\quad h\ _{t+b};I^{n}{}_{\overline{U}^{(L)}}{}^{o}i+4d\ k\ k\quad\ +4d\ k\ k: \tag{2.64}$$

By letting $b\ !\ 0$ and then $\ !\ 0$ in (2.60), we have the continuity of the mapping $t\ 7!\ h\ _t;\ _{(\ t)}i$.

We next show that a solution to (2.58) is also a solution to an another differential equation obtained by applying a method of change of variables. For $r \quad t$, we have

$$\begin{aligned}\frac{dh\ _{r};\ _{t\ r}\ i}{dr} &= \lim_{h\ !\ 0}\frac{fh\ _{r+h};\ _{t\ r\ h}\ i \quad h\ _{r};\ _{t\ r}\ ig}{h} \\ &= \lim_{h\ !\ 0}\frac{fh\ _{r+h};\ _{t\ r\ h}\ i \quad h\ _{r+h};\ _{t\ r}\ ig}{h} + \lim_{h\ !\ 0}\frac{fh\ _{r+h};\ _{t\ r}\ i \quad h\ _{r};\ _{t\ r}\ ig}{h}\end{aligned} \tag{2.65}$$

We now obtain the expression for the first term on the right side of (2.65). We can write

$$h\ _{r+h};\ _{t\ r\ h}\ i \quad h\ _{r+h};\ _{t\ r}\ i = h\ _{r+h};\hat{w}i;$$

where $\hat{w}$ is defined such that $\hat{w}(\mathbf{y}_n) = \ _{t\ r\ h}\ (\mathbf{y}_n) \quad \ _{t\ r}\ (\mathbf{y}_n)$: We further simplify the function $\hat{w}$ by using the following definition, let

$$\frac{@}{@s_i}(\mathbf{y}_n) = \lim_{h\ !\ 0}\frac{(\mathbf{y}_n{}^{j};y_i + h) \quad (\mathbf{y}_n)}{h}:$$

We can write

$$\hat{w}(\mathbf{y}_n) = \ (\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n)) \quad ((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};y_1{+}t \quad r){+}\ ((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};y_1{+}t \quad r) \quad (\ _{t}^{+}{}_{r}(\mathbf{y}_n)):$$

Further, we have

$$(\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n)) \quad ((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};y_1 + t \quad r) = \int_{y_1+t\ \ r\ \ h}^{y_1+t\ \ r}\frac{@}{@s_1}((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};s_1)\,ds_1:$$

By replacing s_1 with $y_1 + t \quad r \quad hv$, we get

$$(\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n)) \quad ((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};y_1{+}t \quad r) = \ h\int_{v=0}^{1}\frac{@}{@s_1}((\ _{t}^{+}{}_{r\ \ h}(\mathbf{y}_n))\ ^{1};y_1{+}t \quad r \quad hv)\,dv:$$

Similarly, we can write

$$\begin{aligned}&(n;y_1 + t \quad r;\quad ;y_{i\ 1} + t \quad r;y_i + t \quad r \quad h;y_{i+1} + t \quad r \quad h;\quad ;y_n + t \quad r \quad h) \\ &\qquad (n;y_1 + t \quad r;\quad ;y_i + t \quad r;y_{i+1} + t \quad r \quad h;\quad ;y_n + t \quad r \quad h) \\ = \ h\int_{v=0}^{1}&\frac{@}{@s_i}(n;y_1{+}t \quad r;\quad ;y_{i\ 1}{+}t \quad r;y_i{+}t \quad r \quad hv;y_{i+1}{+}t \quad r \quad h;\quad ;y_n{+}t \quad r \quad h)\,dv:\end{aligned}$$

For $1 \quad i \quad n$, let

$$w_{(i;t;r;h;v)}(\mathbf{y}_n) = \frac{@}{@s_i}(n;y_1{+}t \quad r;\quad ;y_{i\ 1}{+}t \quad r;y_i{+}t \quad r \quad hv;y_{i+1}{+}t \quad r \quad h;\quad ;y_n{+}t \quad r \quad h):$$

As a consequence, after simplifications we have

$$\hat{w}(\mathbf{y}_n) = \quad h \int_{v=0}^{1} \sum_{i=1}^{n} \left(w_{(i;t;r;h;v)}(\mathbf{y}_n) \right) dv:$$

Let the function $w_{(t;r;h;v)} \; 2 \; C_b(U)$ be defined as

$$w_{(t;r;h;v)}(\mathbf{y}_n) = \begin{cases} 0 & \text{if } n = 0; \\ \sum_{i=1}^{n} \left(w_{(i;t;r;h;v)}(\mathbf{y}_n) \right) & \text{otherwise:} \end{cases}$$

Now we can see that

$$\lim_{h \; ! \; 0} \frac{f h_{\; r+h;\; t \; r \; h} \; i \quad h_{\; r+h;\; t \; r} \; ig}{h} = \quad \lim_{h \; ! \; 0} \int_{v=0}^{1} h_{\; r+h;} w_{(t;r;h;v)} i \, dv:$$

Since $h \; 7! \; h_{\; r+h;} w_{(t;r;h;v)} i$ is continuous, by the dominated convergence theorem, we have

$$\lim_{h \; ! \; 0} \frac{f h_{\; r+h;\; t \; r \; h} \; i \quad h_{\; r+h;\; t \; r} \; ig}{h} = \quad h_{\; r;} \; r_{\; 1} \tilde{} \; i: \tag{2.66}$$

We now focus on the second term on the right side of (2.65). We have

$$h_{\; r+h;\; t \; r} \; i \quad h_{\; r;\; t \; r} \; i = \int_{u=r}^{r+h} \frac{@}{@u} h_{\; u;\; t \; r} \; i \, du:$$

By using (2.58), we have

$$\begin{aligned} h_{\; r+h;\; t \; r} \; i \quad h_{\; r;\; t \; r} \; i = \int_{u=r}^{r+h} \Bigg(& h_{\; u;} \; r_{\; 1 \; t \; r} \; i \\ & + \sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \; (x_j) \left({}_{t \; r} \; (\mathbf{x}_n{}^{j}) \quad {}_{t \; r} \; (\mathbf{x}_n) \right) d \; {}_{u}(\mathbf{x}_n) \\ & + \Bigg[\; {}_{u}(f0g) \;\; \Phi_0(\; {}_{u}) \left({}_{t \; r} \; (1;0) \quad {}_{t \; r} \; (0) \right) + \sum_{n=1}^{C} \sum_{j=1}^{1} \sum_{} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \\ & \qquad \Phi_n(\; {}_{u})(\; {}_{t \; r} \; (\mathbf{x}_n^j; 0) \quad {}_{t \; r} \; (\mathbf{x}_n)) \, d \; {}_{u}(\mathbf{x}_n) \Bigg] \Bigg) \, du: \end{aligned}$$

Again, by applying the method of change of variables, we have

$$\begin{aligned} h_{\; r+h;\; t \; r} \; i \quad h_{\; r;\; t \; r} \; i = h \int_{v=0}^{1} & h_{\; r+hv;} \; r_{\; 1 \; t \; r} \; i \\ & + \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} \; (x_j) \left({}_{t \; r} \; (\mathbf{x}_n{}^{j}) \quad {}_{t \; r} \; (\mathbf{x}_n) \right) d \; {}_{r+hv}(\mathbf{x}_n) \right. \end{aligned}$$

$$+\left[{}_{r+hv}(f0g)\ \Phi_0({}_{r+hv})({}_t{}_r(1;0) \quad {}_t{}_r(0)) + \sum_{n=1}^{C}\sum_{j=1}^{1}\sum^{n+1}\int \quad \int_{U_n}\left\{\frac{1}{n+1}\right\}\right.$$
$$\left.\Phi_n({}_{r+hv})({}_t{}_r(x_n^j;0) \quad {}_t{}_r(x_n))\, d\ {}_{r+hv}(x_n)\right]\Bigg)\, dv:$$

As a result, by using the dominated convergence theorem we have

$$\lim_{h\,!\,0}\frac{fh\ {}_{r+h};\ {}_t{}_r\ i \quad h\ {}_r;\ {}_t{}_r\ ig}{h} = h\ {}_r;\ r\ {}_1\ {}_t{}_r\ i$$
$$+\left(\sum_{n=1}^{C}\sum_{j=1}^{n}\int \quad \int_{U_n}(x_j)\left({}_t{}_r(x_n{}^j) \quad {}_t{}_r(x_n)\right)\, d\ {}_r(x_n)\right.$$
$$+\left[{}_r(f0g)\ \Phi_0({}_r)({}_t{}_r(1;0) \quad {}_t{}_r(0)) + \sum_{n=1}^{C}\sum_{j=1}^{1}\sum^{n+1}\int \quad \int_{U_n}\left\{\frac{1}{n+1}\right\}\right.$$
$$\left.\left.\Phi_n({}_r)({}_t{}_r(x_n^j;0) \quad {}_t{}_r(x_n))\, d\ {}_r(x_n)\right]\right): \quad (2.67)$$

Finally, by using (2.66) and (2.67), we have

$$\frac{dh\ {}_r;\ {}_t{}_r\ i}{dr} = \sum_{n=1}^{C}\sum_{j=1}^{n}\int \quad \int_{U_n}(x_j)\left({}_t{}_r(x_n{}^j) \quad {}_t{}_r(x_n)\right)\, d\ {}_r(x_n)$$
$$+\left[{}_r(f0g)\ \Phi_0({}_r)({}_t{}_r(1;0) \quad {}_t{}_r(0)) + \sum_{n=1}^{C}\sum_{j=1}^{1}\sum^{n+1}\int \quad \int_{U_n}\left\{\frac{1}{n+1}\right\}\right.$$
$$\left.\Phi_n({}_r)({}_t{}_r(x_n^j;0) \quad {}_t{}_r(x_n))\, d\ {}_r(x_n)\right]:$$

By integrating $\frac{dh\ {}_r;\ {}_t{}_r\ i}{dr}$ with respect to r from 0 to t, we get (2.11) for $\ 2\ C_b^1(U)$. Then the result can be extended to the simple functions by using the monotone convergence theorem and then to the class of functions $C_b(U)$ from the standard arguments by using the Dynkin - theorem [56, page 497] or from the fact that $C_b^1(U)$ is dense in $C_b(U)$.

We next prove that for $\ 2\ C_b^1(U)$, the solution $(h^-{}_t;\ i;t \quad 0)$ to (2.11) is also a solution to (2.10). For this, it is enough to prove the differentiability of $h^-{}_t;\ i$ with respect to t. Since $\ 2\ C_b^1(U)$, the existence of $\frac{dh^-{}_0;\ {}_t\ i}{dt}$ follows from the dominated convergence theorem. By using the Leibniz integral rule, we now verify the existence of the derivative of the second term on the right side of (2.11) with respect to t. According to this rule, the first condition is that the integrand needs to be continuous with respect to both the variables r and t. This follows from the same arguments used in the proof of the continuity of the

integrand in equation (2.10). The second condition is that the derivative of the integrand with respect to t must exist and the differential should be continuous with respect to both r and t. The derivative of the integrand with respect to t exists from the dominated convergence theorem since $\quad 2\ C_b^1(U)$ and also, it is continuous with respect to r and t from the same arguments that we have used to prove the continuity of the integrand in (2.10). Therefore, any process $(\ _t; t \quad 0)\ 2\ C_{M_1(U)}([0;\ 1\ ))$ is a solution to (2.10) if and only if it is a solution to (2.11). Further, note that $\quad$ need not be a differentiable function in (2.11).

2.8.3 Proof of Theorem 2.1

From equation (2.11), we first make it clear that for all $\quad 2\ C_b(U)$, the operator $\quad 7!\ h\ _t;\ i$ is a linear operator with $\ _t(U) = 1$. Hence from the Riesz-Markov-Kakutani theorem (Theorem A.4) [69, Theorem 2.14], for $\ _t\ 2\ M_1(U)$, the existence of the unique operator $\quad 7!\ h\ _t;\ i$ implies the existence of the unique probability measure $\ _t$. The uniqueness of $\ _t$ also follows from the fact that $C_b(U)$ is a separating class of $M_1(U)$ [56, p.111]. If $\ _1;\ _2\ 2\ M_1(U)$ satisfies $h\ _t;\ i = h\ _1;\ i$ and $h\ _t;\ i = h\ _2;\ i$ for all $\quad 2\ C_b(U)$, then we have $\ _1 = \ _2$.

Given an initial measure $\ _0$, we next prove that there exists at most one mean-field solution by showing that there exists at most one real valued process $h\ _t;\ i$ corresponding to the mean-field. Suppose $(\ ^1_t; t \quad 0); (\ ^2_t; t \quad 0)$ are two solutions to the MFEs with initial points $\ ^1_0;\ ^2_0$, respectively. For $\quad 2\ C_b(U)$, we then have

$$
\begin{aligned}
h\ ^1_t \quad ^2_t;\ i = h\ ^1_0 \quad ^2_0;\ _t\ i + \int_{s=0}^{t} \Bigg(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{U_n} (x_j) \left(\ _t\ _s\ (\mathbf{x}_n{}^j) \quad _t\ _s\ (\mathbf{x}_n)\right) \\
d(\ ^1_s \quad ^2_s)(\mathbf{x}_n) \Bigg) ds \\
+ \int_{s=0}^{t} \Bigg(\Bigg[\ ^1_s(f0g)\ \ \Phi_0(\ ^1_s) \left(\ _t\ _s\ (1;0) \quad _t\ _s\ (0)\right) \\
+ \sum_{n=1}^{C} \sum_{j=1}^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \ \Phi_n(\ ^1_s)(\ _t\ _s\ (\mathbf{x}_n^j; 0) \quad _t\ _s\ (\mathbf{x}_n))\, d\ ^1_s(\mathbf{x}_n) \Bigg] \\
\left[\ ^2_s(f0g)\ \ \Phi_0(\ ^2_s) \left(\ _t\ _s\ (1;0) \quad _t\ _s\ (0)\right) \right. \\
\left. \sum_{n=1}^{C} \sum_{j=1}^{n+1} \int \quad \int_{U_n} \left\{ \frac{1}{n+1} \right\} \ \Phi_n(\ ^2_s)(\ _t\ _s\ (\mathbf{x}_n^j; 0) \quad _t\ _s\ (\mathbf{x}_n))\, d\ ^2_s(\mathbf{x}_n) \Bigg] \right) ds: \quad (2.68)
\end{aligned}
$$

The first term on the right side of (2.68) can be bounded as $jh\ \ _0^1\quad\ _0^2;\ _t\ ij\quad k\ \ _0^1\quad\ _0^2kk\ k$: To simplify the second term corresponding to departures, we define a function $h_{t;s}$ as follows:

$$h_{t;s}(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = 0; \\ \sum_{k=1}^{n} \ (x_k)(\ _t\ _s\ (\mathbf{x}_n{}^j) \quad\ _t\ _s\ (\mathbf{x}_n)) & \text{otherwise}: \end{cases}$$

Then since $\ 2\ \mathsf{C}_b(\mathsf{U})$ and $\ 2\ \mathsf{C}_b(\mathsf{R}_+)$, we have $h_{t;s}\ 2\ \mathsf{C}_b(\mathsf{U})$. Further, we have $kh_{t;s}k\quad 2Ck\ kk\ k$: Using the definition of $h_{t;s}$, we can write

$$\int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \quad \int_{\mathsf{U}_n} \ (x_j) \Big(\ _t\ _s\ (\mathbf{x}_n{}^j) \quad\ _t\ _s\ (\mathbf{x}_n) \Big)\, d(\ _s^1 \quad\ _s^2)(\mathbf{x}_n)\, ds \right.$$
$$\left. = \int_{s=0}^{t} h\ \ _s^1 \quad\ _s^2; h_{t;s} i\, ds: \right.$$

To simplify the third term that corresponds to arrivals, we define a function $f_{t;s;\ }$ as follows: for $0\quad n\quad C\quad 1$,

$$f_{t;s;\ }(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = C; \\ \sum_{j=1}^{n+1} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ )(\ _t\ _s\ (\mathbf{x}_n^j; 0) \quad\ _t\ _s\ (\mathbf{x}_n)) & \text{otherwise}: \end{cases}$$

Then the third term is equal to $\int_{s=0}^{t} \ \left(h\ \ _s^1; f_{t;s;\ _s^1} i \quad h\ \ _s^2; f_{t;s;\ _s^2} i\right) ds$. Further, we can write

$$\left|h\ \ _s^1; f_{t;s;\ _s^1} i \quad h\ \ _s^2; f_{t;s;\ _s^2} i\right| \quad \left|h\ \ _s^1 \quad\ _s^2; f_{t;s;\ _s^1} i\right| + \left|h\ \ _s^2; f_{t;s;\ _s^1} \quad f_{t;s;\ _s^2} i\right|$$
$$k\ \ _s^1 \quad\ _s^2 kk f_{t;s;\ _s^1} k + k\ \ _s^2 kk f_{t;s;\ _s^1} \quad f_{t;s;\ _s^1} k:$$

Since $\ _s^2$ is a probability measure, $k\ \ _s^2 k = 1$. Furthermore, $k f_{t;s;\ _s^1} k \quad 2dk\ \ k$ and

$$\left| f_{t;s;\ _s^1}(\mathbf{x}_n) \quad f_{t;s;\ _s^2}(\mathbf{x}_n) \right| \quad 2d^2 k\ \ k \left(\left| \overline{R}_n(\ _s^1) \quad \overline{R}_n(\ _s^2) \right| + \left| \overline{R}_{n+1}(\ _s^1) \quad \overline{R}_{n+1}(\ _s^2) \right| \right):$$

We can write $\overline{R}_n(\ _s^1) = h\ \ _s^1; f\ \ i$ where $f\ \ $ is a function defined as

$$f\ \ (\mathbf{x}_m) = \begin{cases} 1 & \text{if } m \quad n; \\ 0 & \text{otherwise}: \end{cases}$$

We then have $\left|\overline{R}_n(\ _s^1) \quad \overline{R}_n(\ _s^2)\right| \quad k\ \ _s^1 \quad\ _s^2 kk f\ \ k = k\ \ _s^1 \quad\ _s^2 k$:

Finally, by using bounds for all the terms, we get

$$\left|h\ \ _t^1 \quad\ _t^2;\ \ i\right| \quad \left(k\ \ _0^1 \quad\ _0^2 k + \int_{s=0}^{t} 2k\ \ kCk\ \ _s^1 \quad\ _s^2 k\, ds \ + \int_{s=0}^{t} 8d^2\ \ k\ \ _s^1 \quad\ _s^2 k\, ds \right) k\ \ k:$$

Therefore we have

$$k \ {}^{1}_{t} \quad {}^{2}_{t}k \quad k \ {}^{1}_{0} \quad {}^{2}_{0}k + (2Ck \ k + 8d^2 \ ) \int_{s=0}^{t} k \ {}^{1}_{s} \quad {}^{2}_{s}k \, ds: \tag{2.69}$$

From the Gronwall's inequality (Theorem A.7) [56, Theorem 5.1, p.498], for some $b; c > 0$, $t \ 2 \ [0; T]$, if $k \ {}^{1}_{t} \quad {}^{2}_{t}k \quad b + c \int_{s=0}^{t} k \ {}^{1}_{s} \quad {}^{2}_{s}k ds;$ then it follows that $k \ {}^{1}_{t} \quad {}^{2}_{t}k \quad b e^{ct}:$ Therefore, from (2.69), we have $k \ {}^{1}_{t} \quad {}^{2}_{t}k \quad k \ {}^{1}_{0} \quad {}^{2}_{0}k e^{(2Ck \ k+8d^2 \ )t}:$ Hence, starting from an initial measure ${}_{0}$, there exists at most one solution for the MFEs.

We now prove that there exists a process $(\ {}_{t}; t \quad 0) \ 2 \ \mathsf{C}_{\mathsf{M}_1(\mathsf{U})}([0; 1 \ ))$ satisfying the mean-field model equations. This follows from the relative compactness of the sequence $f{}^{-(N)}_{t}; t \quad 0g$ in $\mathsf{D}_{\mathsf{M}_1(\mathsf{U})}([0; 1 \ ))$ from the proof of Theorem 2.3. In particular, we have that every limit point of the sequence $f{}^{-(N)}_{t}; t \quad 0g$ satisfies (2.11). Further, each limiting point is almost surely continuous. This concludes that there exists a solution to the MFEs.

2.8.4 Proof of Theorem 2.4

The first part of Theorem 2.4 is a special case of the second part. Hence, it is sufficient to prove the second part.

For $t \quad 0$, ${}^{-(N)}_{t}$ is the empirical probability measure on U such that ${}^{-(N)}_{t}(f\mathbf{u}g)$ for $\mathbf{u} \ 2 \ \mathsf{U}$ denotes the fraction of servers lying in state $\mathbf{u} \ 2 \ \mathsf{U}$ at time t. From the dynamics of the system under the SQ(d) scheme and the exchangeability of $f\mathbf{S}_k(N)(0); 1 \quad k \quad Ng$, the collection $f\mathbf{S}^{(N)}_{k}(t); 1 \quad k \quad Ng$ is also exchangeable for all $t \ 2 \ [0; 1 \ )$. Further, from Theorem 2.2, we have ${}^{-(N)}_{t} \ ) \ {}^{-}_{t}$ for $t \ 2 \ [0; 1 \ )$ as $N \ ! \ 1$.

To prove the result, it is sufficient to show that the following holds:

$$\mathsf{E}\left[\prod_{k=1}^{l} \ {}_{k}(\mathbf{S}^{(N)}_{k}(t))\right] \ ! \ \prod_{k=1}^{l} h{}^{-}_{t}; \ {}_{k}i; \tag{2.70}$$

for all continuous bounded mappings ${}_{k} : \mathsf{U} \ ! \ \mathsf{R}$ as $N \ ! \ 1$.

We can write

$$\left|\mathsf{E}\left[\prod_{k=1}^{l} \ {}_{k}\left(\mathbf{S}^{(N)}_{k}(t)\right)\right] \quad \prod_{k=1}^{l} h{}^{-}_{t}; \ {}_{k}i\right| \left|\mathsf{E}\left[\prod_{k=1}^{l} \ {}_{k}\left(\mathbf{S}^{(N)}_{k}(t)\right)\right] \quad \mathsf{E}\left[\prod_{k=1}^{l} h{}^{-(N)}_{t}; \ {}_{k}i\right]\right| + \left|\mathsf{E}\left[\prod_{k=1}^{l} h{}^{-(N)}_{t}; \ {}_{k}i\right] \quad \prod_{k=1}^{l} h{}^{-}_{t}; \ {}_{k}i\right|: \tag{2.71}$$

From Theorem 2.2, the second term on the right hand side of the above inequality vanishes as $N \to \infty$. Let $Q(r;n)$ be the set of all permutations of the numbers $\{1;2;\ldots;n\}$ taken r at a time. Now, due to exchangeability, the permutation of states between servers does not affect the joint distribution. Then $|Q(r;n)| = (n)_r$ where $(N)_k = N(N-1)\ldots(N-k+1)$. Hence, we have

$$\mathsf{E}\left[\prod_{k=1}^{l} \phi_k\left(\mathrm{S}_k^{(N)}(t)\right)\right] = \frac{1}{(N)_l}\mathsf{E}\left[\sum_{\alpha \in Q(l;N)} \prod_{k=1}^{l} \phi_k\left(\mathrm{S}_{(k)}^{(N)}(t)\right)\right].$$

Also, by the definition of $\eta_t^{(N)}$, we have

$$\mathsf{E}\left[\prod_{k=1}^{l} \langle \eta_t^{(N)}, \phi_k \rangle\right] = \mathsf{E}\left[\left(\prod_{k=1}^{l} \frac{1}{N}\sum_{j=1}^{N} \phi_k\left(\mathrm{S}_j^{(N)}(t)\right)\right)\right]. \tag{2.72}$$

Also, let $D(r;n)$ be the set of all r-tuples that are formed by elements chosen from $\{1;2;\ldots;n\}$. Then $|D(r;n)| = n^r$. Let $\overline{Q(r;n)}$ be the set of elements that are present in $D(r;n)$ but not in $Q(r;n)$. Then from (2.72), we can write

$$\begin{aligned}\mathsf{E}\left[\prod_{k=1}^{l} \langle \eta_t^{(N)}, \phi_k \rangle\right] &= \frac{1}{N^l}\mathsf{E}\left[\sum_{\alpha \in D(l;N)} \prod_{k=1}^{l} \phi_k\left(\mathrm{S}_{(k)}^{(N)}(t)\right)\right] \\ &= \frac{1}{N^l}\mathsf{E}\left[\sum_{\alpha \in Q(l;N)} \prod_{k=1}^{l} \phi_k\left(\mathrm{S}_{(k)}^{(N)}(t)\right)\right] + \frac{1}{N^l}\mathsf{E}\left[\sum_{\alpha \in \overline{Q(l;N)}} \prod_{k=1}^{l} \phi_k\left(\mathrm{S}_{(k)}^{(N)}(t)\right)\right].\end{aligned}$$

Let $\max_k \|\phi_k\| = B$, then we have the bound

$$\begin{aligned}\left|\mathsf{E}\left[\prod_{k=1}^{l} \phi_k\left(\mathrm{S}_k^{(N)}(t)\right)\right] - \mathsf{E}\left[\prod_{k=1}^{l} \langle \eta_t^{(N)}, \phi_k \rangle\right]\right| &\le \left(\frac{1}{(N)_l} - \frac{1}{N^l}\right)|Q(l;N)|\,B^l + \frac{1}{N^l}\left|\overline{Q(l;n)}\right| B^l \\ &\le 2B^l\left(1 - \frac{(N)_l}{(N)^l}\right).\end{aligned}$$

The result follows since $\left(1 - \frac{(N)_l}{(N)^l}\right) \to 0$ as $N \to \infty$. This completes the proof.

2.8.5 Proof of Lemma 2.4

The mean-field $(\eta_t; t \ge 0)$ is a solution to (2.11). Let us consider the function $\hat{\phi} = I_{\{\mathbf{u}_n \in \mathsf{U}_n : 0 \le u_i \le y_i; \forall i\}}$. For $s \ge 0$, let η_s be an absolutely continuous measure which has no atoms, we have $\langle \eta_s, \hat{\phi}_i \rangle = \langle \eta_s, \phi_i \rangle$ where $\phi = I_{\{\mathbf{u}_n \in \mathsf{U}_n : 0 < u_i \le y_i; \forall i\}}$. Since there exists a sequence of functions $\{f_n, g_n\}_{n \ge 1} \subset \mathsf{C}_b(\mathsf{U})$ that increase point wise to $I_{\{f \in B\}}$ where B is an open set in

U_n, n 1, by using the monotone convergence theorem [69, Theorem 1.26] and (2.11), we have that (2.11) is true even for the function . Furthermore, since the measure $^{-}_{s}$ is absolutely continuous for all s 0, we have that equation (2.11) is true even for the function $\hat{}$. Therefore, using (2.11), we can obtain the evolution equations of the process $(P_t; t$ 0) where $P_t = (P_t(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ and $P_t(\mathbf{y}_n) = h^{-}_{t}; \hat{}\, i$. We can further simplify the expression of the process $(P_t(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U}; t$ 0) obtained from (2.11) using the fact that

$$\begin{aligned} h^{-}_{s};\ {}_{b}I_{f\mathbf{x}_n 2 \mathsf{U}_n : 0\ x_i\ y_i; 8ig} i &= h^{-}_{s}; I_{f\mathbf{x}_n 2 \mathsf{U}_n : 0\ x_i + b\ y_i; 8ig} i \\ &= h^{-}_{s}; I_{f\mathbf{x}_n 2 \mathsf{U}_n : 0\ x_i\ y_i\ b; 8ig} i: \end{aligned}$$

By differentiating $P_t(\mathbf{y}_n)$ with respect to t and after simplifications, it is verified that the process $(P_t; t$ 0) satisfies equations (2.18)-(2.20).

2.8.6 Proof of Lemma 2.3

From Remark 2.5, we recall that the MFEs are the dynamics of the probability distribution of a single server loss system with capacity C in which jobs arrive according to a Poisson process with rate $\Phi_n(^{-}_{t})$ (n 0) when there are n progressing jobs. We have that the initial distribution # has a density function and our objective is to show that for given $t = r$, $^{-}_{r}$ has a density function w.r.t. the Lebesgue measure. For n 1, $\mathbf{u}_n = (n; u_1;\ ; u_n)$, we now prove that $^{-}_{r}$ has density at $\mathbf{u}_n$. For $_i > 0$ and 1 i n, let

$$B\ ,\ f(n; y_1;\ ; y_n) : u_i\ y_i < u_i + \ _i; 1\ i\ n g: \tag{2.73}$$

At time $t = r$, the probability that there are n progressing jobs and the i^{th} job has age y_i such that $y_i\ 2\ [u_i; u_i + \ _i)$, i 1, is equal to $^{-}_{r}(B)$. Out of the n progressing jobs that are present at time $t = r$, let J_1 be the set of indices of all the progressing jobs that entered the system at a time $t > 0$ and J_2 be the set of indices of all the progressing jobs which are present in the system from time $t = 0$. Precisely,

$$J_1\ ,\ f i : r > u_i; 1\ i\ n g;$$

and

$$J_2\ ,\ f i : r\ u_i; 1\ i\ n g:$$

Essentially, if $i\ 2\ J_1$, it implies that the age of the i^{th} job is less than r. As the ages of progressing jobs increase linearly with time at a unit rate, the i^{th} job must have entered

the system at a time $t > 0$. Precisely, if the i^{th} job's age y_i satisfies $y_i \in [u_i; u_i + \ _i)$ and $i \in J_1$ at time $t = r$, then it implies that the i^{th} job must have entered the system in the time interval $(r \quad u_i \quad \ _i; r \quad u_i]$ and its service is not finished by the time $t = r$. On the other hand, if $j \in J_2$, then it means that the j^{th} job is present in the system from time $t = 0$. At time $t = r$, if the j^{th} job's age y_j satisfies $y_j \in [u_j; u_j + \ _j)$ and $j \in J_2$, then its age should lie in the interval $[u_j \quad r; u_j + \ _j \quad r)$ at time $t = 0$.

Using the sets J_1 and J_2, we now obtain an upper bound on $^{-}_r(B)$ from which we conclude that there exists a density function. Let $D_1 = jJ_1j$ and $D_2 = jJ_2j$.

We next obtain bounds on the probability that there exists D_2 jobs at time $t = r$ such that their ages lie in the set J_2. Let B_1 be the event that there exists D_2 jobs at time $t = r$ such that their ages lie in the set J_2. Note that the total number of jobs say q that are present at time $t = 0$ can be more than D_2 but only D_2 of them should not expire by the time $t = r$. Let l_i be the i^{th} smallest element of the set J_2. Also, let i_j be the index of the job out of q jobs which will not expire by the time $t = r$ and its age will lie in the interval $[u_{l_j}; u_{l_j} + \ _{l_j})$ at time $t = r$. For the event B_1 to occur, the state of the server at time $t = 0$ should belong to the set V where

$$V \ , \ f(q; x_1; \quad ; x_q) : x_m \in \mathsf{R}_+ \text{ for } m \in f i_1; \quad ; i_{D_2} g$$
$$\text{and } x_m \in [u_{l_a} \quad r; u_{l_a} \quad r + \ _{l_a}) \text{ for } m = i_a; 1 \quad a \quad D_2; 1 \quad m \quad q; q \quad D_2 g:$$

A job with age x at time $t = 0$ will stay in the system at time $t = r$ with probability $\frac{\overline{G}(x+r)}{\overline{G}(x)}$. Then by using all the above arguments, we get the following bound where $f_\# = (f_\#(\mathbf{u}); \mathbf{u} \in \mathsf{U})$ is the pdf of #,

$$\mathsf{P}(B_1) \quad \sum_{q:q=D_2}^{C} \left(\sum_{(i_1; \quad ; i_{D_2}) \in f1;2; \quad ; qg} \int \quad \int_V f_\#(q; x_1; \quad ; x_q) \left(\left\{ \prod_{m=1}^{D_2} \frac{\overline{G}(x_{i_m} + r)}{\overline{G}(x_{i_m})} \right\} \right) dx_1 \quad dx_q \right) : \tag{2.74}$$

We now focus on the jobs that belong to the set J_1. Let B_2 be the event that for each $j \in J_1$, there is an arrival in the time interval $(r \quad u_j \quad \ _j; r \quad u_j]$ and furthermore, this job is not expired by the time $t = r$. Since the arrival process is a Poisson process with rate $\Phi_n(^{-}_t)$ when there are n jobs and $\Phi_n(^{-}_t) \quad d$ for all $n \quad 0$, for any time interval $[t_1; t_2]$, we have

$$\mathsf{P}(\mathbf{Z_1}) \quad \mathsf{P}(\mathbf{Z_2});$$

where $\mathbf{Z_1}$ denotes the number of arrivals to the server in the interval $[t_1; t_2]$ and $\mathbf{Z_2}$ denotes the number of arrivals in the interval $[t_1; t_2]$ when the arrival process is a Poisson process

with rate d. Let k_i be the i^{th} smallest element of the set J_1. Then since the arrival instants have uniform distribution conditioned on the number of arrivals over a time interval [77, page 325], we get

$$\mathsf{P}(B_2) \quad \frac{(\ d)^{D_1}}{D_1!}\left(\prod_{j=1}^{D_1}\overline{G}(u_{k_j})\ \ _{k_j}\right): \tag{2.75}$$

Finally, from (2.74) and (2.75), we have

$$\begin{aligned}{}^{-}{}_t(B) \quad \Bigg(\sum_{q:q=D_2}^{C}\Bigg(\sum_{(i_1;\ \ ;i_{D_2})2f1;2;\ \ ;qg}\int \quad \int_V f_{\#}(n;x_1;\quad ;x_n)\\ \left(\prod_{m=1}^{D_2}\left\{\frac{\overline{G}(x_{i_m}+r)}{\overline{G}(x_{i_m})}\right\}\right)dx_1 \quad dx_q\Bigg)\Bigg)\left(\frac{(\ d)^{D_1}}{D_1!}\left(\prod_{j=1}^{D_1}\overline{G}(u_{k_j})\ \ _{k_j}\right)\right): \end{aligned} \tag{2.76}$$

Clearly, ${}^{-}{}_t$ has density at $\mathbf{u}$ since ${}^{-}{}_t(B)\ !\ 0$ as $_j\ !\ 0$ for $1\quad j\quad n$.

2.9 Conclusions

. In this chapter, we have provided a measure-valued process approach to establish the mean-field behavior of loss systems with SQ(d) load balancing and general service time requirements. The extension of these results to multi-class systems where servers are classified into different classes based on their capacities and jobs are classified into different classes based on their service requirements follows in a similar manner *mutatis mutandis* from the approach used here. Establishing the global asymptotic stability of the unique fixed-point remains an open problem.

Chapter 3

On Occupancy Based Randomized Load Balancing for Large Processor Sharing Systems

In this chapter, we show that the mean-field analysis of the SQ(d) randomized load balancing schemes can be extended to randomized schemes that use the occupancy information of a finite number of randomly sampled servers to dispatch an incoming arrival in a large-scale system of parallel servers. Such policies besides the SQ(d) policy also contain threshold based policies, and d-adaptive policies and generalizations of them. In this chapter, we present a mean-field analysis of occupancy based routing schemes for a system with a large number of processor sharing (PS) servers as an archetype of shared resource systems. The PS case is interesting not only because of its use in server-farm applications but also because PS scheduling is also known to be ordinarily insensitive. The analysis of PS schemes is harder than for loss models. As in the loss model case, the MFEs are PDEs. We show that the probability measure of occupancy defined on the set of non-negative integers $\mathbb{Z}_+$ obtained from a fixed-point of the mean-field also satisfies the stationary mean-field equations when the job lengths are exponentially distributed with the same average length. If the mean-field in the exponential case has a unique fixed-point, then the fixed point is *insensitive* to the job length distribution. The approach is also via a measure-valued Markov process approach. We also provide simulation results to justify our analysis.

Organization of the Chapter The rest of the chapter is organized as follows. In Section 3.1, we first introduce the system model, and then we present a generic framework

for occupancy based randomized routing policies. We give additional notation used in this chapter in Section 3.2. After that, we provide a mathematical formulation to the time evolution of the system in Section 3.3. We then present the main results of this chapter in Section 3.4, where we also illustrate how to use our analysis to study four occupancy based randomized load balancing policies. We study the SQ(d) policy and three threshold based policies that are adaptive to variations in the system load. We give proofs of the main results in Section 3.7. In Section 3.5, we provide numerical results to study the impact of the routing policies on the average response time, implementation complexity, and the insensitivity of the stationary distribution as the number of servers converges to 1. In Section 3.6, we provide numerical results which support that the mean-field is globally asymptotically stable (GAS) for the SQ(d) policy and also, we give some insights into the behavior of the system in the stationary regime. Finally, we conclude in Section 3.8.

3.1 System Model and Routing Policy

In this section, we introduce the system model along with a general framework for a class of occupancy based randomized routing policies which use the information about the occupancy of a finite number of randomly selected servers upon an arrival to dispatch the arrival to a server.

3.1.1 System Model

Consider a large-scale system with N servers where each server has an infinite buffer and jobs are served according to the PS service discipline. Jobs arrive according to a Poisson process with rate N to a central job dispatcher. Upon an arrival, the job dispatcher collects information about the occupancy of a finite number of randomly chosen servers to decide the destination server based on a predefined routing policy. We assume that the dispatcher has no buffer and an arrival is assumed to join a server's buffer immediately although the routing decision incurs some delay overhead in practice.

We assume that every server has capacity to process jobs at a unit rate and hence, from the assumption that the service discipline is PS, if a server has n jobs in its buffer, then every job is served at a rate of $\frac{1}{n}$. Furthermore, a job that is routed to a server with occupancy n joins at a position chosen uniformly at random from $n+1$ positions. We

assume that the job lengths are distributed according to a general distribution function $G(\ )$ with density function $g(\ )$ having the mean $\frac{1}{}$. To ensure the stability of the system, we assume that $\quad <\quad$. The hazard rate function of the job length distribution is denoted by $\quad(\ )$ defined in (2.1). Furthermore, we assume that $\quad$ satisfies Assumption 2.1.1.

3.1.2 Routing Policy

Before defining the routing policy precisely, we first introduce a general framework for occupancy dependent randomized routing policies, according to which the number of randomly selected servers also referred to as potential destination servers is not a fixed value as in the SQ(d) policy, but it is a dynamic value that takes into account the system's congestion level. In this framework, we assume that there are at most M stages to decide the destination server for an arrival. In each stage, the dispatcher first samples a finite number of servers and then, it decides whether to route the arrival to one of the sampled servers or not based on the occupancies of the sampled servers. In stage i of the routing policy, the dispatcher selects d_i servers uniformly at random from all the N servers, and then, it decides whether to select or not to select a server from the list of the potential destination servers as the destination server, based on the occupancy information of the potential destination servers that are sampled up to stage i. In stage i, if the occupancies of the potential destination servers satisfy certain criteria as described in the routing policy, then the dispatcher selects a potential destination server as the destination server. Otherwise, the routing policy enters stage $i+1$ in which the dispatcher samples d_{i+1} servers and the same procedure as in stage i is repeated until stage M. The value of d_{i+1} may depend on the occupancies of the potential destination servers that are sampled in stages 1 to i to take into account the congestion level of the system. We assume that d_1 is a fixed-value but d_i for $i \quad 2$ may depend on the occupancies of the servers that are sampled up to stage $i \quad 1$. The routing policy is stopped at stage M in which a potential destination server is selected as the destination server.

Let $\mathbf{n}^{[i]} = (n_1^{[i]}; \quad ; n_{d_i}^{[i]})$ be the vector of occupancies of servers that are randomly selected in stage i, where $n_j^{[i]}$ is the occupancy of the j^{th} potential destination server selected in stage i. Let $S^{[i]}$ be the set of occupancies of all the potential destination servers selected up to the i^{th} stage of the routing policy given by

$$S^{[i]} = f(\mathbf{n}^{[1]}; \quad ; \mathbf{n}^{[i]}) : \text{ occupancies of sampled servers in stage } j \text{ is } \mathbf{n}^{[j]}; \ 1 \quad j \quad ig: \tag{3.1}$$

We assume that the dispatcher can sample at most $d^{(max)}$ servers to decide the destination server for an arrival. Therefore, we have $\sum_{i=1}^{M} d_i \quad d^{(max)}$. Without loss of generality, from now onwards, for a vector of real numbers, we refer to the number of elements in the given vector as its size. Let $S^{[i;j]}$ be the set of all occupancy vectors in $S^{[i]}$ with size j. Then for $1 \quad i \quad M$,

$$S^{[i]} = \lceil_{j=1}^{d^{(max)}} S^{[i;j]}; \tag{3.2}$$

In the i^{th} stage, based on the occupancies of the potential destination servers, the dispatcher may decide to route the arrival to one of the potential destination servers if the occupancies meet certain criteria as stated in the predefined routing policy. We indicate whether the dispatcher selects the destination server in stage i or not by a vector of non-negative real numbers. If the number of sampled potential destination servers up to stage i is k, then the outcome of stage i is represented by a vector of size $k+1$. If $(n_1; \quad ; n_k)$ is the occupancy vector of potential destination servers where k is the total number of potential destination servers selected up to stage i, then the outcome at the end of stage i according to the routing policy is given by $(m; p_1; \quad ; p_k)$ where m is a non-negative integer value and p_i is a non-negative real number for $1 \quad i \quad k$. If the dispatcher decides to route the arrival to one of the sampled potential destination servers in stage i as per the routing policy, then $m = 0$ and $(p_1; \quad ; p_k)$ indicates the routing probability vector satisfying $\sum_{j=1}^{k} p_j = 1$ where p_i is the probability with which the i^{th} potential destination server is chosen as the destination server for the arrival. In this case, we say that the outcome of stage i is a routing probability vector. On the other hand, if the dispatcher decides to enter stage $i+1$ according to the routing policy, then $m = d_{i+1}$ $(m \quad 1)$ which indicates the number of servers to be sampled in stage $i+1$ and $(p_1; \quad ; p_k)$ is the null-vector satisfying $\sum_{j=1}^{k} p_j = 0$. In this case, we say that the outcome of stage i is the null-vector.

Let $Y^{[i]}$ be the set of all occupancy vectors in the set $S^{[i]}$ for which the outcome of i^{th} stage is a routing probability vector. Further, let $Y^{[i;j]}$ be the set of elements of size j in $Y^{[i]}$. Then for $1 \quad i \quad M$,

$$Y^{[i]} = \lceil_{j=1}^{d^{(max)}} Y^{[i;j]}; \tag{3.3}$$

Let $\overline{Y^{[i]}}$ be the complement of the set $Y^{[i]}$ in $S^{[i]}$. In the i^{th} stage, the routing policy is a mapping $\Lambda^{(i)}$ that maps $Y^{[i;j]}$ as follows:

$$\Lambda^{(i)} : Y^{[i;j]} \; 7! \; (0; M_1(f1; 2; \quad ; jg)); \tag{3.4}$$

where $M_1(f1; 2; \quad ; jg)$ denotes the set of probability measures on the set $f1; 2; \quad ; jg$.

For $1 \quad k \quad d^{(max)}$, if $(r_1; \quad ; r_k) \; 2 \; \overline{Y^{[i]}}$, then

$$\Lambda^{(i)}((r_1; \quad ; r_k)) = (d_{i+1}; 0; \quad ; 0); \tag{3.5}$$

where d_{i+1} is the number of servers to be sampled in stage $i+1$ and the null vector $(0; \quad ; 0)$ has size k.

Note that in the M^{th} stage, the outcome is always a routing probability vector since the arrival is always routed to one of the potential destination servers. We now state precisely the class of routing policies that we study in this chapter under a common framework.

Definition 3.1. *Occupancy Based Randomized Routing Policy:*

Upon an arrival, in stage i, the routing policy samples d_i servers uniformly at random with replacement from N servers. The randomly selected or sampled servers are referred to as the potential destination servers for the arrival. [1] *Then the routing policy in stage i is a mapping function for $1 \quad i \quad M$ and $1 \quad j \quad d^{(max)}$,*

$$\Lambda^{(i)} : Y^{[i;j]} \; ! \; (0; p_1; \quad ; p_j)) \; 2 \; (0; M_1(f1; 2; \quad ; jg)); \tag{3.6}$$

where p_i denotes the probability with which the i^{th} potential destination server is chosen as the destination server. Also,

$$\Lambda^{(i)} : \overline{Y^{[i]}} \setminus S^{[i;j]} \; ! \; (d_{i+1}; 0; \quad ; 0)) \; 2 \; R_+^{j+1}; \tag{3.7}$$

where d_{i+1} denotes the number of servers to be sampled in stage $i + 1$. In stage 1, the dispatcher always samples a fixed number of d_1 servers. If the routing policy enters the last stage M, then one of the potential destination servers is chosen as the destination server based on the mapping $\Lambda^{(M)}$.

3.2 Additional Notation and Terminology

In this section, we provide the required additional notation and terminology that is specific to this chapter.

For any $f \; 2 \; C_b^1(R_+^n)$, the function $f^{\natural}$ which will appear in the mean-field equation is defined as

$$f^{\natural}(x_1; \quad ; x_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{@f(x_1; \quad ; x_n)}{@x_i}$$

[1] As $N \; ! \; 1$, sampling potential servers with or without replacement would yield same mean-field limit and hence, we use sampling with replacement to simplify the notation.

Since the rate at which a job is served depends on the number of progressing jobs of the server that processes it, we keep track of the status of a progressing job by using its age defined as the amount of service received since its arrival. We then use the notation $(n; a_1; \quad ; a_n)$ to keep track of the status of a server where n is the number of progressing jobs and a_i is the age of the i^{th} progressing job. If a job arrives at time t, then its age a at time $t + h$ is given by

$$a = \int_{s=t}^{t+h} \frac{1}{r(s)}\, ds;$$

where $r(s)$ denotes the instantaneous number of progressing jobs at the server at time s. Let U_n be the set of all possible states of a server when there are n progressing jobs, i.e.,

$$\mathsf{U}_n = f(n; a_1; \quad ; a_n) : a_i \; 2 \; \mathsf{R}_+; 1 \quad i \quad ng:$$

If a server is idle, then its state lies in the set

$$\mathsf{U}_0 = f0g:$$

Let U be the set defined as

$$\mathsf{U} = [_{n \;\; 0} \mathsf{U}_n;$$

and hence, U is the set of all possible server states. Without loss of generality, we write an element of the form $(n; u_1; \quad ; u_n)$ as $\mathbf{u}_n$ to indicate that it belongs to the set U_n for $n \quad 0$ and also, we write $(n; u_1; \quad ; u_n)$ as $\mathbf{u}$ to say that it is an element in U. We associate the space U with the metric d_{U} where for $\mathbf{u} = (n; u_1; \quad ; u_n)$ and $\mathbf{v} = (m; v_1; \quad ; v_m)$, $d_{\mathsf{U}}(\mathbf{u}; \mathbf{v})$ is defined as

$$d_{\mathsf{U}}(\mathbf{u}; \mathbf{v}) = \begin{cases} \sum_{i=1}^{n} ju_i \quad v_ij & \text{if } n = m \\ jm \quad nj & \text{otherwise:} \end{cases}$$

For $n \quad 1$, U_n is a complete and separable, and Polish space. Furthermore, U is a Polish space as it is the union of a countable set of disjoint Polish spaces. Also, U is separable and complete.

The restriction of a measure $\quad 2 \; \mathsf{M}_F(\mathsf{U})$ to the space U_0 is a Dirac measure with mass at (0). Also, if $\quad (f\mathbf{x}_n g) = 0$ for $\mathbf{x}_n \; 2 \; \mathsf{U}_n$, $n \quad 1$, then we say that $\quad$ is absolutely continuous at $\mathbf{x}_n$. If $\quad (f\mathbf{y}_n g) = 0$ for all $\mathbf{y}_n \; 2 \; \mathsf{U}_n$, $n \quad 1$, then we say that $\quad$ is absolutely continuous with respect to Lebesgue measure. For every $i \quad 1$ and $(x_1; \quad ; x_i) \; 2 \; \mathsf{R}_+^i$, if $\frac{@f(i;x_1; \quad ;x_i)}{@x_j}$ is defined for all $1 \quad j \quad i$, then f is said to be differentiable. From this definition, our analysis uses the fact that $f = I_{f\mathsf{U}_n g}$ is differentiable for $n \quad 1$.

Let $\Xi : \mathsf{U} \not\to \mathsf{R}$ and $\quad : \mathsf{U} \not\to \mathsf{R}$ be defined as for $(n; x_1; \quad ; x_n) \; 2 \; \mathsf{U}_n$,

$$\Xi(n; x_1; \quad ; x_n) = n;$$

and

$$(n; x_1; \quad ; x_n) = \begin{cases} 0 & \text{for } n = 0; \\ \sum_{i=1}^{n} x_i & \text{otherwise :} \end{cases}$$

Note that if a server state has distribution $\quad$, then $h \; ; \Xi i$ and $h \; ; \; i$ indicate the average number of progressing jobs and the average sum of ages of the progressing jobs, respectively. The following transition operators on functions and measures play a crucial role in the mathematical modeling of the time evolution of the system. For $b > 0$, let $\tilde{}_b^{+} : \mathsf{U} \not\to \mathsf{U}$ be the transition operator defined as

$$\tilde{}_b^{+}(n; u_1; : : : ; u_n) = \begin{cases} 0 & \text{if } n = 0; \\ (n; v_1; \quad ; v_n) & \text{otherwise;} \end{cases}$$

where $v_i = u_i + \frac{b}{n}$, for $1 \quad i \quad n$. Also, for any $b > 0; f \; 2 \; \mathsf{K}_b(\mathsf{U})$, let $\tilde{}_b : \mathsf{K}_b(\mathsf{U}) \; ! \; \mathsf{K}_b(\mathsf{U})$ be the transition operator defined as

$$\tilde{}_b f(\mathbf{u}) = f(\tilde{}_b^{+} \mathbf{u});$$

for $\mathbf{u} \; 2 \; \mathsf{U}$. Similarly, let $\tilde{}_b \quad 2 \; \mathsf{M}_F(\mathsf{U})$ be the shifted measure of $\quad$ satisfying that for $B \; 2 \; \mathsf{B}(\mathsf{U})$

$$\tilde{}_b \; (B) = \quad (\tilde{}_b^{+}(B)):$$

For $\quad 2 \; \mathsf{M}_F(\mathsf{U})$, the measure $\tilde{}_b \quad 2 \; \mathsf{M}_F(\mathsf{U})$ satisfies

$$h \tilde{}_b \; ; f i = h \; ; \tilde{}_b f i \tag{3.8}$$

for all $f \; 2 \; \mathsf{K}_b(\mathsf{U})$. The uniqueness of the measure $\tilde{}_b \quad$ in (3.8) follows from the Riesz-Markov-Kakutani theorem given in Appendix A.4 [80]. We use (3.8) in computing the generator of the measure-valued Markov process $(\overline{}_t^{(N)}; t \quad 0)$ to be defined in Section 3.4 which describes the time evolution of the system.

3.3 System Dynamics

In this section, we first define the state of a server using which we then define the system state. The system state descriptor is defined such that the time evolution of the system

can be described by the time evolution of a measure-valued Markov process. As stated earlier, we keep track of the status of a progressing a job by using its age. We say that a server lies in the state $(n; a_1; \quad ; a_n)$ if it has n progressing jobs and the age of the i^{th} progressing job is equal to a_i, $1 \quad i \quad n$. If a server is idle, then its state is equal to (0). Since the routing policy is symmetric to servers as server identities have no role to play, it is sufficient to keep track of the total number of servers lying in each possible server state instead of each individual server's state to model the time evolution of the system by a measure-valued Markov process. Hence, we use the following system state descriptor.

Definition 3.2. *System State Descriptor:*
Let $\mathbf{S}_i^{(N)}(t)$ be the state of the i^{th} server at time t, then the state of the system at time t is given by

$$ {}_t^{(N)} = \sum_{i=1}^{N} \mathbf{s}_i^{(N)}(t). \tag{3.9} $$

Hence, the number of servers lying in a given state $(m; z_1; \quad ; z_m)$ at time t is equal to ${}_t^{(N)}(f(m; z_1; \quad ; z_m)g)$.

For given ${}_t^{(N)}$ and $h > 0$, the value of ${}_{t+h}^{(N)}$ depends on the events that occur in the interval $(t; t+h]$. In the time interval $(t; t+h]$, there can be no event (arrival or departure) or some events (arrivals and departures) can occur in the system. By considering that h is sufficiently small, the probability that multiple events occur in the interval $(t; t+h]$ can be neglected. The departure events can be modeled by using the hazard rate function $(x) = \frac{g(x)}{\overline{G}(x)}$ which indicates the instantaneous rate of departure of a job conditioned on its age is x. Precisely, if a job has age x at time t, then the probability that this job departs the system when its age is less than $x+y$ is equal to $\frac{G(x+y) \quad G(x)}{\overline{G}(x)} = \quad (x)y + o(y)$. If the number of progressing jobs at a server with state say $(n; a_1; \quad ; a_n)$ at time t remains constant in the interval $(t; t+h]$, then its state becomes ${}_h^{\sim +}(n; a_1; \quad ; a_n)$ at time $t+h$.

3.4 Main Results

In this section, we give the main results of this chapter.

Let us consider an occupancy based randomized routing policy which maps occupancies of the potential destination servers up to stage k to an outcome using a mapping function $\Lambda^{(k)}$. In stage k, for $\mathbf{n} = (n_1; \quad ; n_i)$ 2 $\mathsf{Y}^{[k;i]}$, let $q^{[k;i]}(\mathbf{n}; j)$ be the probability with which

the j^{th} potential destination server is chosen as the destination server out of i potential destination servers. We begin with the result on the probability with which the destination server of an arrival lies in the given state $(m; x_1; \quad ; x_m)$.

Lemma 3.1. *For* $^{(N)}_t = \quad$, *let* $Q_n(\frac{}{N}) = \frac{(f \cup_n g)}{N}$ *be the fraction of servers with n jobs at time t. Then, an arrival at time t is routed to a server with state* $(m; x_1; \quad ; x_m)$ *with probability equal to*

$$p_r(\ ; m; x_1; \quad ; x_m) = \frac{f(m; x_1; \quad ; x_m)g}{N} \Phi_m\left(\frac{}{N}\right); \tag{3.10}$$

where

$$\Phi_m\left(\frac{}{N}\right) = \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y^{[k;i]}} \sum_{j=1}^{i} I_{f n_j = m g} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \neq j}^{i} Q_{n_r}\left(\frac{}{N}\right): \tag{3.11}$$

Proof. See Section 3.7.1. □

We consider the filtration

$$F_t^{(N)} = \backslash_{\ >0}(\ (\ ^{(N)}_s : s \quad t + \ )): \tag{3.12}$$

Let $A^{(N)}(\ )$ be the generator of the Markov process $(\ ^{(N)}_t; t \quad 0)$. Using the Dynkin's formula [56], we have the following result.

Theorem 3.1. *The process* $(\ ^{(N)}_t; t \quad 0)$ *is a Feller-Dynkin process in* $\mathsf{D}_{\mathsf{M}_F(\mathsf{U})}([0; 1\ ))$. *Let* $2\ \mathsf{C}^1_b(\mathsf{U})$, *then the process* $(\mathbf{M}^{(N)}_t(\ ); t \quad 0)$ *de ned as*

$$\mathbf{M}^{(N)}_t(\ ) = h\ ^{(N)}_t;\ \ i \quad h\ ^{(N)}_0;\ \ i \quad \int_{s=0}^{t} A^{(N)} h\ ^{(N)}_s;\ \ i\, ds \tag{3.13}$$

is a square integrable $F^{(N)}_t$*-martingale and it is an RCLL process.*

Proof. See Section 3.7.3. □

From the process $(\ ^{(N)}_t; t \quad 0)$, let us define the normalized process $(-^{(N)}_t; t \quad 0)$ where

$$-^{(N)}_t = \frac{^{(N)}_t}{N}: \tag{3.14}$$

The main objective of this chapter is to obtain the limiting process of the normalized sequence of processes $f(-^{(N)}_t; t \quad 0)g_{N\ \ 1}$ when $N \ ! \ 1$. For our analysis, we make the following assumption.

Assumption 3.4.1. *The sequence* $f\mu_0^{(N)}g_{N\ 1}$ *satisfies*

$$(\mu_0^{(N)};h\mu_0^{(N)};\Xi i;h\mu_0^{(N)};\ i)\ )\ (\mu_0;h\mu_0;\Xi i;h\mu_0;\ i); \tag{3.15}$$

where $\mu_0\ 2\ \mathsf{M}_1(\mathsf{U})$ *is an absolutely continuous measure with* $h\mu_0;\Xi i\ <\ 1$ *and* $h\mu_0;\ i\ <\ 1$.

We can interpret the conditions $h\mu_0;\Xi i\ <\ 1$ and $h\mu_0;\ i\ <\ 1$ as follows. If μ_0 is the probability measure of the state of a server, then both the average number of jobs $h\mu_0;\Xi i$ and the average sum of ages of jobs $h\mu_0;\ i$, are finite.

On comparing Assumptions 3.4.1 and 2.4.1, we make an extra assumption to analyze the PS model that the average occupancy $h\mu_0;\Xi i$ w.r.t. the limiting measure μ_0 is finite. This condition is used to show that the sequence $f(\mu_t^{(N)};t\quad 0)g_{N\ 1}$ is tight.

We now state the main result of the chapter on the convergence of sequence $f(\mu_t^{(N)};t\ 0)g_{N\ 1}$ as $N\ !\ 1$ to the mean-field limit.

Theorem 3.2. *Under Assumptions 2.1.1 and 3.4.1, we show that* $(\mu_t^{(N)};t\quad 0)\ )\ (\mu_t;t\ 0)$ *where the process* $(\mu_t;t\quad 0)$ *is a unique solution to the following equation with the initial point* $\mu_0\ 2\ \mathsf{M}_1(\mathsf{U})$, *for all* $\quad 2\ \mathsf{C}_b(\mathsf{U})$,

$$\begin{aligned}
h\mu_t;\ i = h\mu_0;\tilde{}_t\ i + \int_{s=0}^{t}\Bigg(\sum_{n=1}^{1}\sum_{i=1}^{n}\int_{x_1}\quad\int_{x_n}\left\{\frac{(x_i)}{n}\right\}(\tilde{}_t\ _s\ (n\quad 1;x_1;\quad ;x_{i\ 1};x_{i+1};\quad ;x_n)\\
\tilde{}_t\ _s\ (n;x_1;\quad ;x_n))\,d\mu_s(n;x_1;\quad ;x_n) + \Bigg[(\mu_s(f0g)\Phi_0(\mu_s)(\tilde{}_t\ _s\ (1;0)\quad \tilde{}_t\ _s\ (0)))\\
+\sum_{n=1}^{1}\sum_{i=1}^{n+1}\int_{x_1}\quad\int_{x_n}\left\{\frac{1}{n+1}\right\}\Phi_n(\mu_s)(\tilde{}_t\ _s\ (n+1;x_1;\quad ;x_{i\ 1};0;x_i;\quad ;x_n)\\
\tilde{}_t\ _s\ (n;x_1;\quad ;x_n))\,d\mu_s(n;x_1;\quad ;x_n)\Bigg]\Bigg)\,ds:
\end{aligned} \tag{3.16}$$

We refer to equation (3.16) as the mean- eld equation and its unique solution is referred to as the mean- eld limit.

Proof. See Section 3.7.4. □

For any $T\ >\ 0$, under the assumption of exchangeability of random variables that represent servers' states at time $t = 0$, any finite set of servers at time $t = T$ are asymptotically independent. This proof is similar to the proof of Theorem 2.4 and hence, we omit the proof. Furthermore, μ_t represents the distribution of a server's state as $N\ !\ 1$ at time t.

We now show that the mean-field limit is the unique solution of a set of partial differential equations. We next show that the measure $^{-}_{t}$ at time t is absolute continuous w.r.t. the Lebesgue measure and hence, it has a density function.

Lemma 3.2. *The measure $^{-}_{t}$ at time t has a density function w.r.t. the Lebesgue measure for almost all $\mathbf{u}_n$ 2 U_n, n 1.*

Proof. See Section 3.7.6. □

Let $p_t(n; x_1; \quad ; x_n)$ be the density of $^{-}_{t}$ at $(n; x_1; \quad ; x_n)$ for n 1 and let $P_t(0)$ be equal to $^{-}_{t}(f0g)$. Then we construct a probability distribution function $P_t = (P_t(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ defined as

$$P_t(n; y_1; \quad ; y_n) = \int_{x_1=0}^{y_1} \quad \int_{x_n=0}^{y_n} p_t(n; x_1; \quad ; x_n)\, dx_1 \quad dx_n; \tag{3.17}$$

where $P_t(n; y_1; \quad ; y_n)$ denotes the probability that at time t, a server has n jobs and the i^{th} job $(1 \quad i \quad n)$ has age at most y_i when $N \ ! \ 1$. From the mean-field equation, since there exists a sequence of bounded continuous functions that converge monotonically to an indicator function of an open set, by using the monotone convergence theorem [69, Theorem 1.26], we obtain the following differential equations. The proof follows mutatis mutandis from the arguments of proof of Corollary 2.4 except that we replace with ˜.

Corollary 3.1. *The probability distribution function $P_t = (P_t(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ satis es the PDEs*

$$\frac{dP_t(0)}{dt} = \int_{y=0}^{1} (y) \left(\frac{@P_t(1; y)}{@y} \right) dy \quad \Phi_0(P_t)P_t(0) \tag{3.18}$$

and for n 1,

$$\begin{aligned} \frac{dP_t(n; y_1; \quad ; y_n)}{dt} = & \sum_{i=1}^{n} \left\{ \frac{1}{n} \right\} \frac{@P_t(n; y_1; \quad ; y_n)}{@y_i} \\ & + \sum_{i=1}^{n+1} \int_{x_i=0}^{1} \left\{ \frac{(x_i)}{n+1} \right\} \left(\frac{@P_t(n+1; y_1; \quad ; y_{i\ 1}; x_i; y_i; \quad ; y_n)}{@x_i} \right) dx_i \\ & \sum_{i=1}^{n} \int_{x_i=0}^{y_i} \left\{ \frac{(x_i)}{n} \right\} \left(\frac{@P_t(n; y_1; \quad ; y_{i\ 1}; x_i; y_{i+1}; \quad ; y_n)}{@x_i} \right) dx_i \\ & + \sum_{i=1}^{n} \Phi_{n\ 1}(P_t)P_t(n \quad 1; y_1; \quad ; y_{i\ 1}; y_{i+1}; \quad ; y_n) \quad \Phi_n(P_t)P_t(n; y_1; \quad ; y_n); \end{aligned} \tag{3.19}$$

where $\Phi_n(P_t)$ is the same as in equation (3.11) except that we replace $Q_n(\overline{N})$ with $Q_n(P_t) = \lim_{b \to 1} P_t(n; b; \quad ; b)$.

The above mean-field PDEs (3.18)-(3.19) correspond to the dynamics of a single server PS system in which the job arrival process is a Poisson process with rate $\Phi_n(P_t)$ when the server has n jobs and the current distribution is P_t at time t. Hence, the MFEs represent the dynamics of a non-linear Markov process whose generator is a function of the current distribution of the Markov process.

The stationary behavior of the mean-field is discussed below. Let be a fixed-point of the mean-field $(P_t; t \quad 0)$ and let $\overline{R}_n(\ ) = \sum_{j=n}^{1} \lim_{b \to 1} (j; b; \quad ; b)$. Then we consider the following class of fixed-points Y of the mean-field defined as

$$Y = \left\{ \ : \sum_{n \ 1} \overline{R}_n(\ ) < 1 \text{ and } \sum_{n=1}^{1} \prod_{i=1}^{n} \frac{\Phi_{i \ 1}(\ )}{} < 1 \right\}: \tag{3.20}$$

If $2 Y$, since $\sum_{n \ 1} \overline{R}_n(\ ) < 1$, the average occupancy obtained from is finite. Furthermore, as $\sum_{n=1}^{1} \prod_{i=1}^{n} \frac{_{i \ 1}(\ )}{} < 1$, we can obtain a distribution from as in (3.124) and (3.125) with arrival rate $\Phi_n(\ )$ when there are n jobs. We exploit this logic to obtain the following result.

Theorem 3.3. *Any xed-point $= (\ (\mathbf{u}); \mathbf{u} \ 2 \ \mathsf{U}) \ 2 \ Y$ of the mean- eld $(P_t; t \quad 0)$ satis es*

$$(n; y_1; \quad ; y_n) = \Gamma_n \ ^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i) \, dx_i: \tag{3.21}$$

where $\Gamma = (\Gamma_n; n \quad 0)$ is a xed-point of the mean- eld when job lengths are exponentially distributed with the same mean $\frac{1}{}$. Under the assumption of exponential JLDs, if the mean- eld of the occupancy process has a unique xed-point, then the xed-point is unique when job lengths are generally distributed. Furthermore, the xed-point is insensitive since $\int_{x=0}^{1} \overline{G}(x) \, dx = \frac{1}{}$ which implies

$$\lim_{b \to 1} (n; b; \quad ; b) = \Gamma_n: \tag{3.22}$$

Proof. See Section 3.7.2. □

The benefit of our mean-field analysis of occupancy based randomized routing schemes is that, to obtain the mean-field limit for any routing policy, we just need to obtain the rate function $\Phi_m(\ )$ for $m \quad 0$ given in (3.11) and the final mean-field equation is then given by (3.19).

We now conclude our analysis by applying our results to four policies. For a given vector $\mathbf{n} = (n_1; \quad ; n_k)$, let $min(\mathbf{n})$ be the minimum value of the set $fn_1; \quad ; n_k g$ defined as

$$min(\mathbf{n}) = \min_{1 \quad j \quad k} n_j$$

and let

$$B_{min}(\mathbf{n}) = fj : n_j = min(\mathbf{n}); 1 \quad j \quad kg:$$

Suppose $jB_{min}(\mathbf{n})j$ be the size of the set $B_{min}(\mathbf{n})$.

Definition 3.3. *Policy 1: This policy is the SQ(d) routing policy de ned in De nition 2.1.1. For this policy, the number of stages is equal to one. For* $\mathbf{n} = (n_1; \quad ; n_d)$, *we have*

$$q^{[1;d])}(\mathbf{n}; j) = \begin{cases} \frac{1}{jB_{min}(\mathbf{n})j} & \text{if } n_j = min(\mathbf{n}); \\ 0 & \text{otherwise:} \end{cases} \tag{3.23}$$

As a consequence, upon simpli cation, we get that

$$\Phi_m(P_t) = \frac{f(\overline{R}_m(P_t))^d \quad (\overline{R}_{m+1}(P_t))^d g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g}; \tag{3.24}$$

where $\overline{R}_m(P_t) = \sum_{j=m}^{1} P_t(j; 1; \quad ; 1)$.

We presented the analysis of the SQ(d) policy for the case of general JLDs in [41].

Definition 3.4. *Policy 2: In this policy, upon an arrival, d servers are chosen uniformly at random. Out of these potential destination servers, a server is chosen as the destination server uniformly at random from the list of the potential destination servers whose occupancy is less than or equal to if the list is not empty. Otherwise, the potential destination server that has the least occupancy is chosen as the destination server with ties broken uniformly at random. In this policy, there is only one stage and the number of sampled servers is equal to d. Then for* $\mathbf{n} = (n_1; \quad ; n_d)$ *and let* $a = \sum_{i=1}^{d} I_{fn_i \quad g}$, *we have*

$$q^{[1;d]}(\mathbf{n}; j) = \begin{cases} \frac{1}{a} & \text{if } n_j \quad ; \\ \frac{1}{jB_{min}(\mathbf{n})j} & \text{else if } n_j > \quad \text{and } n_j = min(\mathbf{n}); \\ 0 & \text{otherwise:} \end{cases} \tag{3.25}$$

As a result, upon simpli cation,

$$\Phi_m(P_t) = \begin{cases} \frac{f1 \quad (\overline{R}_{\quad +1}(P_t))^d g}{f1 \quad \overline{R}_{\quad +1}(P_t) g} & \text{if } m \qquad ; \\ \\ \frac{f(\overline{R}_m(P_t))^d \quad (\overline{R}_{m+1}(P_t))^d g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g} & \text{otherwise:} \end{cases} \tag{3.26}$$

Definition 3.5. *Policy 3: In this policy, upon an arrival, the dispatcher selects a server uniformly at random. If the chosen server has occupancy i, then dispatcher selects further* $\min(i; d \quad 1)$ *servers uniformly at random, and the potential destination server with the minimum occupancy among all the potential destination servers is chosen as the destination server with ties broken uniformly at random. In this policy, there are two stages. In the rst stage, the dispatcher rst selects a server uniformly at random, and this server is chosen as the destination if it has no jobs. Otherwise, the policy enters the second stage in which* $\min(i; d \quad 1)$ *servers are selected uniformly at random. The destination server in the second stage is the server with the least occupancy among all the* $\min(i+1; d)$ *potential destination servers. Therefore, we have*

$$q^{[1;1]}(n_1; 1) = \begin{cases} 1 & \text{if } n_1 = 0; \\ 0 & \text{otherwise;} \end{cases} \tag{3.27}$$

and for $\mathbf{n} = (n_1; \quad ; n_l)$,

$$q^{[2;l]}(\mathbf{n}; j) = \begin{cases} \frac{1}{jB_{min}(\mathbf{n})j} & \text{if } n_j = min(\mathbf{n}); \\ 0 & \text{otherwise:} \end{cases} \tag{3.28}$$

Then $\Phi_m(\ )$ *is given by*

$$\Phi_0(P_t) = 1 + \sum_{i=1}^{d \quad 2} Q_i(P_t) \frac{f(\overline{R}_0(P_t))^i \quad (\overline{R}_1(P_t))^i g}{f\overline{R}_0(P_t) \quad \overline{R}_1(P_t) g} + \overline{R}_{d \quad 1}(P_t) \frac{f(\overline{R}_0(P_t))^{d \quad 1} \quad (\overline{R}_1(P_t))^{d \quad 1} g}{f\overline{R}_0(P_t) \quad \overline{R}_1(P_t) g}; \tag{3.29}$$

for $m \quad d \quad 1$,

$$\Phi_m(P_t) = (\overline{R}_m(P_t))^m + \sum_{i=m+1}^{d \quad 1} Q_i(P_t) \frac{f(\overline{R}_m(P_t))^i \quad (\overline{R}_{m+1}(P_t))^i g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g} + \overline{R}_d(P_t) \frac{f(\overline{R}_m(P_t))^{d \quad 1} \quad (\overline{R}_{m+1}(P_t))^{d \quad 1} g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g}; \tag{3.30}$$

and for $m \quad d$,

$$\Phi_m(P_t) = \frac{f(\overline{R}_m(P_t))^d \quad (\overline{R}_{m+1}(P_t))^d g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g}; \tag{3.31}$$

where $Q_i(P_t) = \overline{R}_i(P_t) \quad \overline{R}_{i+1}(P_t)$.

Definition 3.6. *Policy 4: In this policy, the dispatcher selects sequentially at most d servers uniformly at random, and the routing decision is made after every selection. For* $1 \quad i \quad d \quad 1$, *after the* i^{th} *selection, the arrival is routed to the* i^{th} *potential destination server if its occupancy is less than or equal to . Otherwise, the dispatcher selects a server uniformly at random for the* $(i+1)^{th}$ *time, and the same procedure is repeated as in the case of the* i^{th} *potential destination server. If the* $(d \quad 1)^{th}$ *potential destination has occupancy greater than , then another server is selected uniformly at random, and the potential destination server with the minimum occupancy among all the d potential destination servers is chosen as the destination server with ties broken uniformly at random. In this case, the policy has d stages. In the* i^{th} $(1 \quad i \quad d \quad 1)$ *stage, the dispatcher selects a server uniformly at random, and this server is selected as the destination if its occupancy is less than or equal to . In the* d^{th} *stage, the server with the minimum occupancy is chosen as the destination. For* $1 \quad i \quad d \quad 1$ *and* $\mathbf{n} = (n_1; \quad ; n_i)$, *we have*

$$q^{[i;i]}(\mathbf{n}; j) = \begin{cases} 0 & \text{if } i \neq j; \\ 1 & \text{else if } i = j; n_i \quad ; \\ 0 & \text{else if } i = j; n_i > \quad ; \end{cases} \tag{3.32}$$

and if $i = d$, *then*

$$q^{[d;d]}(\mathbf{n}; j) = \begin{cases} \frac{1}{jB_{min}(\mathbf{n})j} & \text{if } n_j = min(\mathbf{n}) \\ 0 & \text{otherwise:} \end{cases} \tag{3.33}$$

As a result, upon simpli cation, if $m \quad$, *then*

$$\Phi_m(P_t) = \frac{1 \quad \overline{R}^d{}_{+1}(P_t)}{1 \quad \overline{R}_{+1}(P_t)}. \tag{3.34}$$

On the other hand, if $m >$, *then*

$$\Phi_m(P_t) = \frac{f(\overline{R}_m(P_t))^d \quad (\overline{R}_{m+1}(P_t))^d g}{f\overline{R}_m(P_t) \quad \overline{R}_{m+1}(P_t) g}. \tag{3.35}$$

Remark 3.1. *Since the arrival rate function $\Phi_m(\,)$ is the same for both Policy 2 and Policy 4, they result in the same mean- eld equation. Hence, the xed-point of the mean- eld coincides for these two policies. However, the number of servers that are randomly selected to implement Policy 4 is adaptive to the system's tra c level whereas Policy 2 always selects d servers to decide the destination server for an arrival. Hence, Policy 4 is preferable over Policy 2 due to its lower complexity.*

The main challenge in the mean-field analysis of a routing policy is to show that the fixed-point of the mean-field is unique. Furthermore, we need to show that the stationary distribution of a server state as $N \to 1$ coincides with the unique fixed-point of the mean-field. Under the assumption of exponential JLDs, for Policy 1, it was shown in [9] (as in the FCFS case in [7]) that the stationary distribution of a server's state coincides with the unique fixed-point of the mean-field. In this chapter, we show that this result is true for the routing policies 2 and 4 based on the result that the mean-field is quasi-monotonic in these cases. For Policy 3, we could not show the uniqueness of a fixed-point of the mean-field. Furthermore, the mean-field equation is a complex equation to prove its monotonicity. For this policy, proving the uniqueness of a fixed-point of the mean-field and its global stability are still open problems.

Theorem 3.4. *Under the assumption of exponential JLDs, for Policy 2 and Policy 4, the stationary distribution of a server state as $N \to 1$ coincides with the unique global asymptotically stable xed-point of the mean- eld.*

Proof. See Section 3.7.8. □

Remark 3.2. *For Policies 1, 2, and 4, from Theorem 3.3 and Theorem 3.4, since the exponential case has a unique xed- xed point, there exists a unique xed-point for the case of general JLDs. Furthermore, the xed-point is insensitive to the JLDs. We conjecture that this also holds for Policy 3.*

3.5 Numerical Results

In this section, we present numerical results to investigate the impact of the four routing policies on the system performance in terms of the average response time for an arrival, complexity in terms of the average number of sampled servers per arrival, and the insensitivity of the stationary distribution of the stochastic system when $N \to 1$. Our numerical

results suggest that as $N \to \infty$, the stationary distribution of occupancy of a server is insensitive for all the four policies which we study in this chapter. We choose the following parameters for our results: $d = 3$, $= 1$, $N = 300$, and $= 1$.

System Performance Versus Complexity

We now study the trade-off between the system performance and complexity of a routing policy. For this, we assume that job lengths are exponentially distributed with mean $\frac{1}{\mu}$. The unique fixed-point of the mean-field can be computed numerically by computing the unique probability vector $\Gamma = (\Gamma_n; n \geq 0)$ that satisfies for $n \geq 0$,

$$\Phi_n(\Gamma) = \Gamma_{n+1} \ ; \tag{3.36}$$

where $\Phi_n(\Gamma)$ is the same as in equation (3.11) except that we replace $Q_n(\frac{}{N})$ with Γ_n. Then the average response time $E(T_s)$ is given by

$$E(T_s) = \frac{1}{} \left\{ \sum_{i \geq 1} \left(\sum_{j \geq i} \Gamma_j \right) \right\} : \tag{3.37}$$

In Figure 3.1, we plot the resulting average response time versus the system load for all the four routing policies and the optimal JSQ policy.

We now present results on the average number of probed servers in deciding the destination server for an arrival in Table 3.1. From Figure 3.1 and Table 3.1, it is clear that a

Table 3.1: The average number of probed servers per arrival

	Policy 1	Policy 2	Policy 3	Policy 4	JSQ
0.6	3	3	1.7875	1.3511	300
0.7	3	3	2.0003	1.5238	300
0.8	3	3	2.2535	1.7823	300
0.9	3	3	2.5769	2.2020	300
0.95	3	3	2.7690	2.5238	300

policy achieves better system performance if the dispatcher probes higher number of servers and selects the server with the maximum number of available resources as the destination server. Hence, the JSQ policy which chooses the server with the least occupancy out of N servers as the destination for an arrival achieves the best system performance [5, 6]. We observe that Policy 1 that selects the server with the least occupancy from a set of d

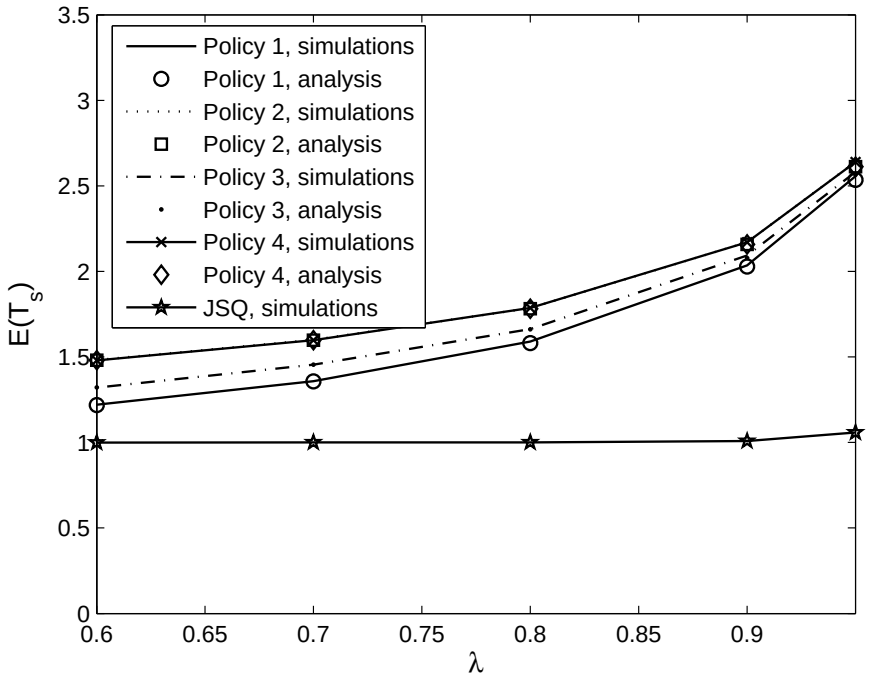

Figure 3.1: The average response time versus

randomly selected servers achieves good system performance but with a significant drop in complexity over the JSQ policy. From Figure 3.1, it can be seen that Policies 2 and 4 achieve the same system performance as they result in the same arrival rate function $\Phi_n(\ )$ for $n \quad 0$, but Policy 4 probes a significantly smaller number of servers to dispatch an arrival when compared to Policy 2. On the other hand, Policy 3 probes higher number of servers when compared to Policy 4 and hence, it achieves better system performance. Both Policies 3 and 4 are robust to the variations in the system load as they probe lower number of servers when the system load is low, and they probe higher number of servers when the system load is high. Hence, all the Policies 1, 3, and 4 result in similar system performance when the system load is high. It is clear that Policy 3 is desirable in practice when there is a variation in the system load as it achieves the better trade-off between the performance and the complexity.

Insensitivity of the System When $N \ ! \ 1$

We now study numerically that for different classes of JLDs, when $N \ ! \ 1$, the stationary distribution of occupancy of a server coincides with the unique fixed-point of the mean-field

of the case when job lengths are exponential with the same average job length.

Let $\Gamma^{(N)} = (\Gamma_i^{(N)}; i \quad 0)$ be the stationary distribution of occupancy of a server in the system with parameter N. Here, $\Gamma_i^{(N)}$ denotes the stationary probability that there are i jobs in progress at a server in the system with N servers in which arriving jobs are routed according to a predefined routing policy. In simulation results, we assume that servers are selected randomly without replacement, whereas in our analysis, we assume that servers are selected with replacement. We use PASTA property to compute $\Gamma^{(N)}$ by simulating the system up to $5 \quad 10^6$ job arrivals. Let $\Gamma^{(exp)} = (\Gamma_i^{(exp)}; i \quad 0)$ be the fixed point of the mean-field limit of the occupancy process and we compute $\Gamma^{(exp)}$ numerically by the stationary mean-field equation. We now compute the total variation distance between $\Gamma^{(N)}$ and $\Gamma^{(exp)}$ defined as

$$\#_{tv}(\Gamma^{(N)}; \Gamma^{(exp)}) = \sum_i \left| \Gamma_i^{(N)} \quad \Gamma_i^{(exp)} \right| : \tag{3.38}$$

We consider the following JLDs in our study: Exponential (Exp), Constant (Const), Power-law (PL), and mixed-Erlang (ME) distributions. The Power-law distribution has CDF

$$G(y) = \begin{cases} 1 \quad \frac{1}{3y^{\frac{3}{2}}} & \text{if } y \quad \frac{1}{3} \\ 0 & \text{otherwise:} \end{cases} \tag{3.39}$$

In the mixed-Erlang case, with probability p_i the distribution has i $(1 \quad i \quad 2)$ exponential phases and each exponential phase has intensity $\quad_p$. We choose $p_1 = :4; p_2 = 0:6$, and the value of $\quad_p$ is computed by using the formula of the average job length given by

$$\sum_{i=1}^{2} \frac{i p_i}{p} = \frac{1}{-}:$$

From Tables 3.2 and 3.3, it is evident that the fixed-point of the mean-field in the exponential case approximates $\Gamma^{(N)}$ for different classes of JLDs having the same average job length.

3.6 On the Stationary Regime under the SQ(d) Policy

In this section, for the SQ(d) policy, we provide some numerical results to show that the fixed-point of the mean-field is GAS. The results of this section were presented in [41]. We consider the case of mixed-Erlang distributions that allow for efficient numerical evaluation

Table 3.2: $\#_{tv}(\Gamma^{(N)}; \Gamma^{(exp)})$ for different JLDs for $= 0.7$ and $N = 300$.

Policy	Exp	$Const$	PL	ME
Policy 1	0.0020	0.0017	0.0018	0.0023
Policy 2	0.0017	0.0017	0.0026	0.0015
Policy 3	0.0029	0.0015	0.0016	0.0035
Policy 4	0.0016	0.0021	0.0020	0.0016

Table 3.3: $\#_{tv}(\Gamma^{(N)}; \Gamma^{(exp)})$ for different JLDs for $= 0.8$ and $N = 300$.

Policy	Exp	$Const$	PL	ME
Policy 1	0.0071	0.0044	0.0050	0.0061
Policy 2	0.0038	0.0036	0.0038	0.0031
Policy 3	0.0040	0.0041	0.0070	0.0066
Policy 4	0.0049	0.0051	0.0040	0.0039

of the MFEs. We also discuss the propagation of chaos in the stationary regime and some existing related works. If one can prove the GAS of the mean-field, then we can exploit the Prohorov's theorem [56] to conclude that the stationary distribution of a server's occupancy as $N \to 1$ coincides with the fixed-point of the mean-field [9]. Proving the GAS of the mean-field is extremely difficult since the mean-field does not possess any monotonicity properties when job lengths are generally distributed unlike the exponential case studied in [9, 37]. We also provide simulation results to support the convergence of the stationary distribution of a server's occupancy in a finite N system to the fixed-point of the mean-field when $N \to 1$. Taken together these results provide evidence of the insensitivity of the stationary distribution of the limiting system as $N \to 1$.

The numerical evaluation of the MFEs when JLDs are mixed-Erlang using the Euler's method was performed with a step size of $2 \quad 10^{-3}$. From the case of exponential distributions, we know that the stationary probability that there are at least k jobs at a server under the SQ(d) policy in the limiting system is given by $(-)^{\frac{d^k - 1}{d - 1}}$ [9] for given values of $-$ (< 1). We assume that the servers have a finite buffer size of C chosen such that $(-)^{\frac{d^C - 1}{d - 1}}$ is negligible. We consider the system parameters as follows. The job lengths have Mixed-Erlang distributions under which a job length is sampled with probability p_i ($i \in \{1, 2, \ldots, D\}$) from an Erlang distribution having i exponential phases with rate $_p$.

Let us define the state of a server with n progressing jobs having l_j phases remaining

for the j^{th} progressing job by $\mathbf{l} = (n; l_1; : : : ; l_n)$ with $1 \quad l_j \quad D$, $1 \quad j \quad n$. For $n \quad 1$, let $\mathsf{S}_n = f(n; l_1; : : : ; l_n) : 1 \quad l_i \quad M; 1 \quad i \quad ng$ be the set of all possible states of a server when there are n progressing jobs and $\mathsf{S}_0 = f(0)g$ denotes the state of a server when there are no progressing jobs. We then define S to be the set of all possible states of servers given by

$$\mathsf{S} = \lceil_{n=0}^{C} \mathsf{S}_n: \tag{3.40}$$

We can model the system evolution by using a Markov process $\mathbf{X}^N(t) = (\mathbf{X}_{\mathbf{l}}^N(t); \mathbf{l} \; 2 \; \mathsf{S})$ where $\mathbf{X}_{\mathbf{l}}^N(t)$ denotes the fraction of servers lying in state $\mathbf{l}$ at time t. Since the underlying space S is countable and finite-dimensional, the mean-field limit can be established by the same procedure as that of the exponential case [9].

It can be shown that the mean-field equations are given by the following system of ordinary differential equations,

$$\mathbf{x}(0; \mathbf{u}) = \mathbf{u}; \tag{3.41}$$

$$\dot{x}_{\mathbf{l}}(t; \mathbf{u}) = h_{\mathbf{l}}(\mathbf{x}(t; \mathbf{u})); \tag{3.42}$$

and $\mathbf{h} = (h_{\mathbf{l}}; \mathbf{l} \; 2 \; \mathsf{S})$ with the mapping $h_{\mathbf{l}}$ given by

$$\begin{aligned} h_{\mathbf{l}}(\mathbf{x}(t; \mathbf{u})) = & \sum_{b=1}^{n} \left\{ \frac{p_{l_b}}{n} \right\} x_{(n \;\; 1; l_1; l_2; \;\; ; l_b \;\; 1; l_{b+1}; \;\; ; l_n)}(t; \mathbf{u}) \;\; {}_{n \;\; 1}^{(ME)}(\mathbf{x}(t; \mathbf{u})) \;\; x_{\mathbf{l}}(t) \;\; {}_{n}^{(ME)}(\mathbf{x}(t; \mathbf{u})) I_{fn<Cg} \\ & + \sum_{b=1}^{n+1} \left\{ \frac{p}{n+1} \right\} I_{fn<Cg} x_{(n+1; l_1; \;\; ; l_b \;\; 1; 1; l_b; \;\; ; l_n)}(t; \mathbf{u}) \\ & + \sum_{b=1}^{n} \left\{ \frac{p}{n} \right\} x_{(n; l_1; \;\; ; l_b \;\; 1; l_b+1; l_{b+1}; \;\; ; l_n)}(t; \mathbf{u}) \qquad p x_{(n; l_1; \;\; ; l_n)}(t; \mathbf{u}); \end{aligned} \tag{3.43}$$

where

$${}_{n}^{(ME)}(\mathbf{r}) = \left\{ \frac{}{\sum_{\mathbf{l} 2 \mathsf{S}_n} r_{\mathbf{l}}} \right\} \left[\left(\sum_{i=n}^{C} \sum_{\mathbf{l} 2 \mathsf{S}_i} r_{\mathbf{l}} \right)^{d} \quad \left(\sum_{j=n+1}^{C} \sum_{\mathbf{l} 2 \mathsf{S}_j} r_{\mathbf{l}} \right)^{d} \right]: \tag{3.44}$$

We numerically solved the MFEs by choosing the following parameters: $d = 2, \quad = 1$, $C = 7$, $D = 2$, $p_1 = 0{:}4$ and $p_2 = 0{:}6$.

The unique fixed-point $\quad = (\;_{\mathbf{l}}; \mathbf{l} \; 2 \; \mathsf{S})$ is given by

$$_{(n; l_1; \;\; ; l_n)} = \;\; {}^{(exp)}(n) \prod_{i=1}^{n} \left\{ \frac{\sum_{j=l_i}^{D} p_j}{\sum_{r=1}^{D} r p_r} \right\}; \tag{3.45}$$

where $^{(exp)} = (\ ^{(exp)}(n); 0 \quad n \quad C)$ is the unique fixed-point of the mean-field in the exponential case.

Furthermore, we define a process $\mathbf{y}(t; \mathbf{v}) = (y_n(t; \mathbf{v}); 0 \quad n \quad C)$ referred to as the tail mean-field that satisfies $\mathbf{y}(0; \mathbf{v}) = \mathbf{v}$ and $y_j(t; \mathbf{v}) = \sum_{i=j}^{C} \sum_{1 2 S_i} x_1(t; \mathbf{u})$, and its fixed-point is denoted by . We now define following metrics,

$$d_{tv}(\mathbf{a}; \mathbf{b}) = \sum_{1 2 S} j a_1 \quad b_1 j; \tag{3.46}$$

$$\#_{tv}(\mathbf{w}; \mathbf{z}) = \sum_{0 \quad n \quad C} j w_n \quad z_n j; \tag{3.47}$$

$$\#_{ds}(\mathbf{w}; \mathbf{z}) = \sum_{0 \quad n \quad C} (w_n \quad z_n): \tag{3.48}$$

We observed that the mean-field is GAS. In particular, in Figure 3.2, we plot $d_{tv}(\mathbf{x}(t; \mathbf{u}); \ )$ as a function of t for three different initial points $\mathbf{u}_1; \mathbf{u}_2$ and $\mathbf{u}_3$. Similarly, in Figure 3.3, we plot $\#_{tv}(\mathbf{y}(t; \mathbf{v}); \ )$ as a function of t for three different initial points $\mathbf{v}_1; \mathbf{v}_2$ and $\mathbf{v}_3$. The mean-field $\mathbf{x}(t; \mathbf{u})$ and its tail mean-field $\mathbf{y}(t; \mathbf{v})$ converge to their fixed-points for three different initial points. This provides evidence for the result that $\mathbf{x}(t; \mathbf{u})$ and $\mathbf{y}(t; \mathbf{v})$ are globally stable. However, $\#_{tv}(\mathbf{y}(t; \mathbf{v}); \ )$ is not monotonically decreasing with t. Hence, the total variation distance cannot be used for constructing a Lyapunov function to show the global asymptotic stability of the mean-field $\mathbf{x}(t; \mathbf{u})$ and its tail mean-field $\mathbf{y}(t; \mathbf{v})$.

We also observed that the tail mean-field is not quasi-monotonic, i.e., if $\mathbf{v}_1 \quad \mathbf{v}_2$ (by component wise), then we do not always have $\mathbf{y}(t; \mathbf{v}_1) \quad \mathbf{y}(t; \mathbf{v}_2)$. We recall that when the job lengths are exponential, the tail mean-field is quasi-monotonic. Figure 3.4 plots $\#_{tv}(\mathbf{y}(t; \mathbf{v}); \ )$ and Figure 3.5 plots $\#_{ds}(\mathbf{y}(t; \mathbf{v}); \ )$ as a function of t for different initial points. In particular, we select v_4 and v_7 to be less than whereas $v_5; v_6; v_8$, and v_9 are chosen to be greater than . The violation of the quasi-monotonicity property is evident for the case of $= 0.9$ with initial point v_7. This is because since $v_7 <$, if we have quasi-monotonicity, $\#_{ds}(\mathbf{y}(t; \mathbf{v}); \ )$ must be always a negative value. This property does not hold as we can see in Figure 3.5 and hence, the mean-field is not quasi-monotonic.

The SQ(d) policy even for small values of d improves the system performance significantly. We plot the average response time ($E(T_s)$) of a job under the SQ(d) policy and the random routing scheme ($d = 1$) in Figure 3.6. The expression for $E(T_s)$ under the SQ(d) policy is given by

$$E(T_s) = \frac{1}{-} \sum_{i \quad 1} \left(- \right)^{\frac{d^i \quad 1}{d \quad 1}} \tag{3.49}$$

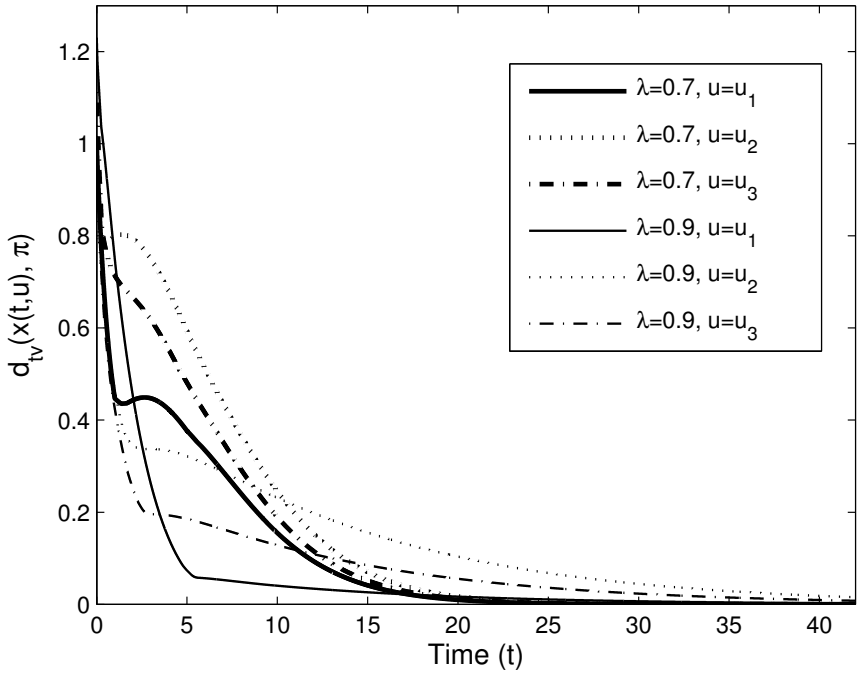

Figure 3.2: $d_{tv}(x(t;u);\ \ )$ versus t

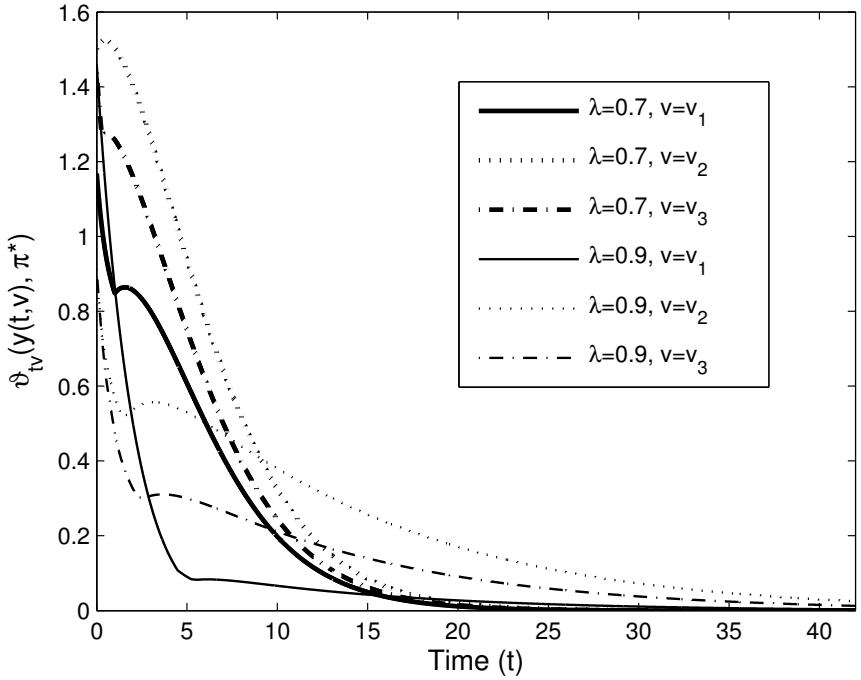

Figure 3.3: $\#_{tv}(y(t;v);\ \ \ )$ versus t

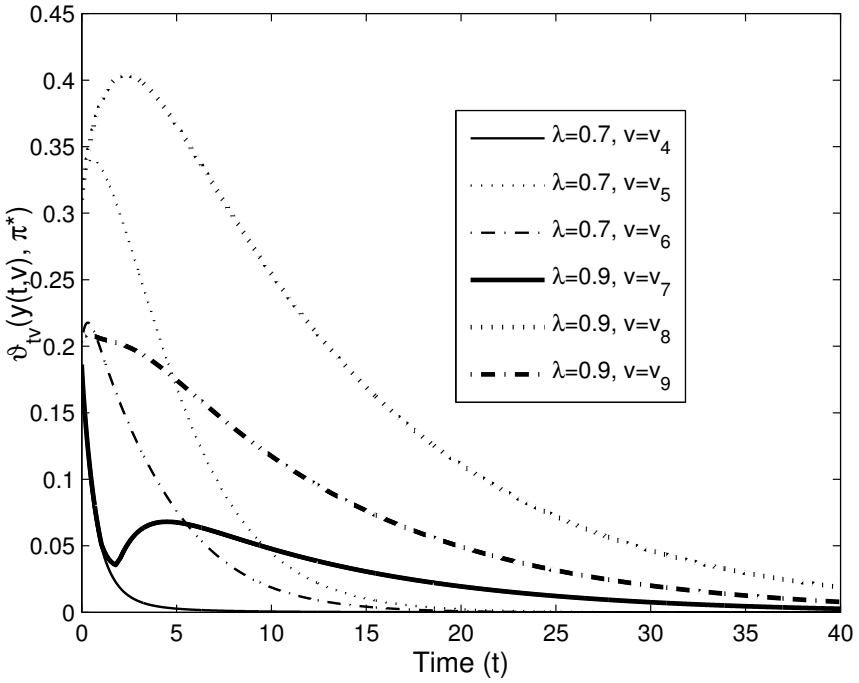

Figure 3.4: $\#_{tv}(y(t;v);\quad)$ versus t

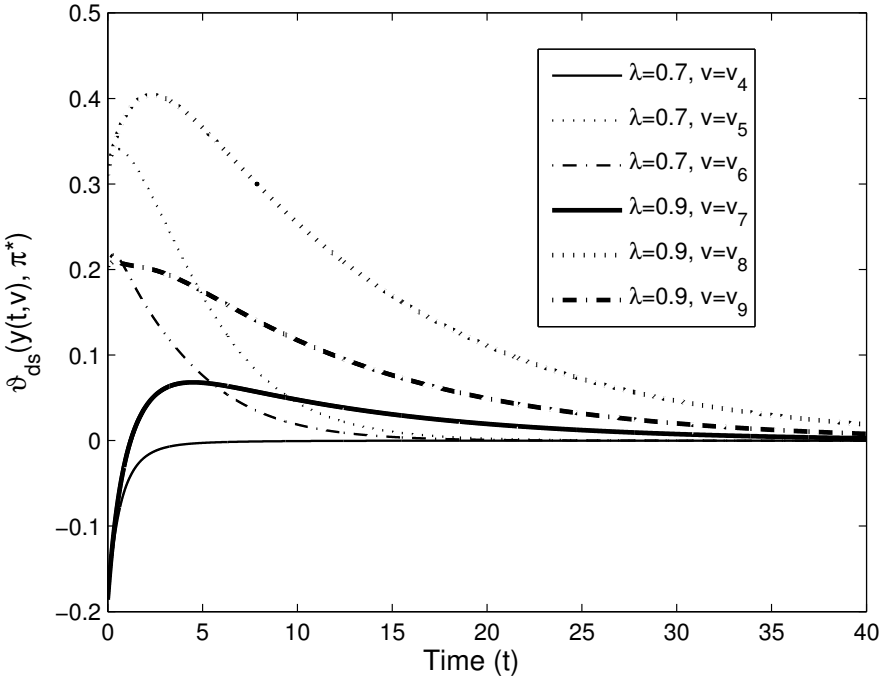

Figure 3.5: $\#_{ds}(y(t;v);\quad)$ versus t

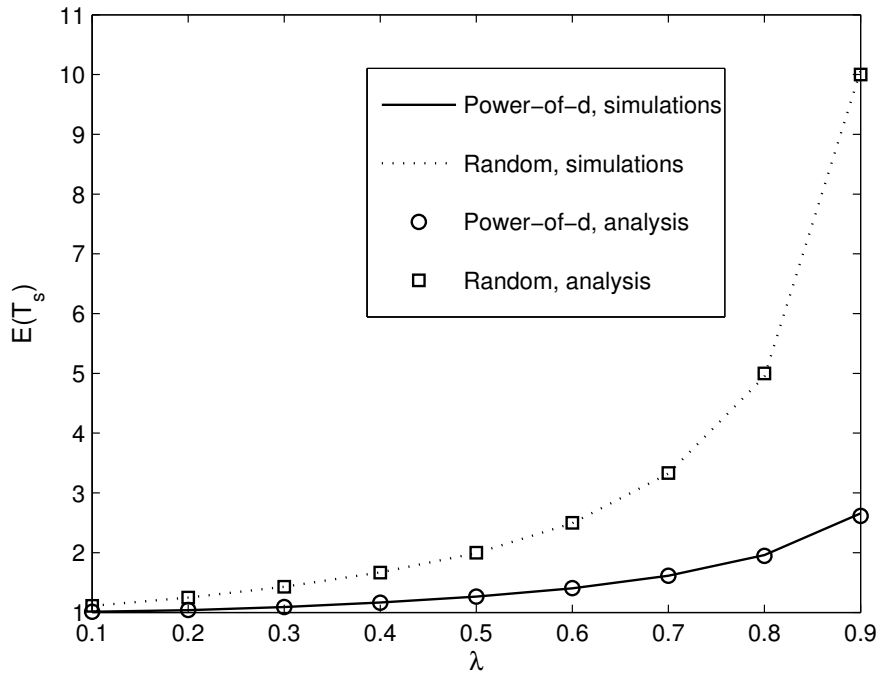

Figure 3.6: The average response time ($E(T_s)$) versus

and for the random routing, we have

$$E(T_s) = \frac{1}{-}\sum_{i\ 1}\left(-\right)^{i}: \tag{3.50}$$

We also plot the simulation results in Figure 3.6 by considering a system with $N = 100$ and exponential JLDs. It is clear that the SQ(d) policy reduces the average response time significantly over the random routing policy ($d = 1$).

3.6.1 On the Propagation of Chaos in the Stationary Regime

We now discuss the relationship between the propagation of chaos in the stationary regime, the tightness of the stationary distributions $f\ ^{(N)}g_{N\ 1}$ of the empirical process, and the global asymptotic stability of the fixed-point of the mean-field. For simplicity, we assume that the JLDs are mixed-Erlang and each server has finite buffer C and hence, there exists a unique invariant distribution $^{(N)}$ of the Markov process $(\mathbf{X}^{(N)}(t); t \quad 0)$ where $\mathbf{X}^{(N)}(t) = (\mathbf{X}_1^{(N)}(t); 1\ 2\ \mathsf{S})$. In this case, the mean-field equations are given by equations (3.41)-(3.43). However, the system does not exhibit monotonicity properties unlike the simple exponential case, and thus establishing the propagation of chaos is a challenging problem [40]. When

the mean-field is GAS, by invoking the Prohorov's theorem (Theorem A.2), we can establish ^{N}) where is the fixed-point of the mean-field [9]. This implies the validity of the interchange of limits

$$\lim_{N\ !\ 1}\lim_{t\ !\ 1}\mathbf{X}^{(N)}(t)=\lim_{t\ !\ 1}\lim_{N\ !\ 1}\mathbf{X}^{(N)}(t):$$

Furthermore, this would then imply that the propagation of chaos holds in the stationary regime [9, 61]. Thus it appears that the global stability of the mean-field, propagation of chaos, and the coincidence of the stationary distribution as $N\ !\ 1$ with the fixed point of the mean-field are all inter-related. We now discuss what happens when we cannot show that the mean-field is GAS.

Since the space of probability measures on S denoted by $\mathsf{M}_1(\mathsf{S})$ with metric induced by total variation distance is compact, from the Prohorov's theorem [57], the sequence $f\ ^{(N)}g_{N\ 1}$ is tight in $\mathsf{M}_1(\mathsf{M}_1(\mathsf{S}))$ under the topology induced by the weak convergence. Let $f\ ^{(N_k)}g_{k\ 1}$ be a converging subsequence with limiting point $Z\ 2\ \mathsf{M}_1(\mathsf{M}_1(\mathsf{S}))$. Then we say that for the sequence of systems with index $fN_kg_{k\ 1}$, the limiting system is said to have the stationary distribution Z for the empirical random variable. Then from Theorem 1 of [61] and Section 1.3, Z is an invariant distribution of the mean-field $\mathbf{x}(t;\ )$ that means

$$\int_{\mathsf{M}_1(\mathsf{S})} f\left(\mathbf{x}(t;\mathbf{u})\right)dZ(\mathbf{u})=\int_{\mathsf{M}_1(\mathsf{S})} f\left(\mathbf{u}\right)dZ(\mathbf{u}): \tag{3.51}$$

Furthermore, from Theorem 3 of [62], the support of Z is a compact set included inside the Birkhoff center of the mean-field that includes the existing limit cycles, fixed-points of the mean-field.

Let $\mathbf{S}_i^{(N_k)}(\ 1\ )$ be the random variable that denotes the state of the i^{th} server in the stationary regime in a finite N_k system. Let $\mathbf{V}^{(N_k)}(\ 1\ )$, $\mathbf{V}(\ 1\ )$ be random variables with distributions $^{(N_k)}$ and Z, respectively. Note that since the system behavior is symmetric to servers as servers' labels do not play any role, the set $(\mathbf{S}_i^{(N_k)}(\ 1\ );1\quad i\quad N_k)$ is exchangeable irrespective of the initial conditions on $(\mathbf{S}_i^{(N_k)}(t);1\quad i\quad N_k)$ in the transient regime. Let us consider continuous bounded mappings $_i:\mathsf{S}\ !\ \mathsf{R}_+,\ 1\quad i\quad I$.

Theorem 3.5. *If* $^{(N_k)}$) Z *as* $k\ !\ 1$, *then*

$$\mathsf{E}\left[\prod_{i=1}^{I}\ _i(\mathbf{S}_i^{(N_k)}(\ 1\ ))\right]\ !\ \mathsf{E}\left[\prod_{i=1}^{I}h\mathbf{V}(\ 1\ );\ _i i\right] \tag{3.52}$$

as $k\ !\ 1$. *Any nite set of servers* $(n_i)_{1\ i\ I}$ *in the limiting system of the sequence* $f\ ^{(N_k)}g_{k\ 1}$ *are mutually independent if and only if* Z *is a Dirac measure. Furthermore, if*

$Z =$ *for some* $2\ \mathsf{M}_1(\mathsf{S})$*, then each server's state is a random variable with distribution* $.$

Proof. See Section 3.5. □

Since Z is an invariant distribution of the mean-field $\mathbf{x}(t;\ )$, from equation (3.51)

$$\mathsf{E}\left[\prod_{i=1}^{l}\ {}_i(\mathbf{S}_i^{(N_k)}(\ 1\ ))\right]\ \ !\ \int_{\mathbf{u}2\ M_1(S)}\left(\prod_{i=1}^{l}h\mathbf{x}(t;\mathbf{u});\ \ {}_i i\right)\ dZ(\mathbf{u}) \tag{3.53}$$

as $k\ \ !\ \ 1$. From (3.53), at any time t in the stationary regime , servers are coupled through the position of the mean-field $\mathbf{x}(t;\ )$ which is a random element since its initial point is random with distribution Z. Furthermore, in the limiting system, at any instant t in the stationary regime, conditioned on the position of the mean-field, each servers' state is a random variable with distribution coinciding with the position of the mean-field, and any finite set of servers are independent. However, if the position of the initial point of the mean-field is random, then at any time t, the position of the mean-field is random thereby any finite set of servers are coupled through the position of the mean-field. For example, if the support of Z contains existing limit cycles or multiple fixed-points of the mean-field, then at any instant in the stationary regime, the position of the mean-field is random, as a consequence, any finite set of servers are coupled through the position of the mean-field. Since the mean-field has a unique fixed-point under the assumption of mixed-Erlang JLDs, we must have $Z =$ to have the propagation of chaos in the stationary regime. If we show that the fixed-point is GAS, then we have $Z =$.

Corollary 3.2. *If* $^{(N_k)}$ *)* Z*, from* (3.53)*, by taking* $l = 1$*, we get that* $\mathsf{E}[\mathbf{V}(\ 1\ )]$ *is the stationary distribution of a server in the limiting system of the sequence of systems* $fN_k g_{k\ 1}$*. That is, the average position of the mean- eld with an invariant distribution* Z *indicates the stationary distribution of a single particle when* Z *is not a Dirac measure.*

Remark 3.3. *In [61], an example with the Birkho center containing a limit cycle and an unstable unique xed-point was found. In this case, it was argued that the distributions of any two particles' states coincide and oscillate around the limit cycle. Hence, the two particles are correlated. But a proof was not provided. More importantly, they did not give any details about the distribution of a single particle when* Z *is not a Dirac measure.*

3.6.2 Discussion on a Related Work

In this section, we provide some insights by relating our work with an existing work for the sake of clarity. In the literature, insensitivity of the stationary distribution of the limiting system is claimed in [24] based on an *ansatz* that has not been shown to hold for PS models to the best of our knowledge. Using Theorem 3.5, we demonstrate that the *ansatz* in [24] implies the necessary and sufficient conditions required for the insensitivity of the limiting system. Therefore, we need to show that the *ansatz* holds to conclude insensitivity.

Let $\mathbf{Y}^{(N)}(t) = (\mathbf{r}^{1;N}(t); \mathbf{r}^{2;N}(t); \quad ; \mathbf{r}^{N;N}(t))$ be the joint queue-size process at time t where $\mathbf{r}^{i;N}(t)$ (notation $q^{i;n}(t)$ is used in [24]) denotes the number of jobs at server i at time t. Let $\Upsilon^{(N)}$ (Υ is replaced with in [24]) be the stationary distribution of $\mathbf{Y}^{(N)}(t)$. We now recall exactly the *ansatz* stated in [24].

The Ansatz in [24]: *Demonstrate* $\Upsilon^{(N)}$) Υ *as* N ! 1 *, where* Υ *is a stationary and ergodic measure on* Z_+^1*. Show that the limit* Υ *is unique, depending only on the service distribution, service discipline and load balancing rule. Let* $\Upsilon_{(k)}$ *be the restriction of* Υ *to its rst k coordinates, with* $= \Upsilon_{(1)}$ *being the one-dimensional marginal of* Υ*. Show that, for every k,*

$$\Upsilon_{(k)} = \quad ^k: \tag{3.54}$$

Let $\overline{\mathbf{W}}^{(N)}(1) = (\overline{\mathbf{W}}_i^{(N)}(1); 0 \quad i \quad C)$ where $\overline{\mathbf{W}}_i^{(N)}(1)$ denotes the random variable in the stationary regime indicating the faction of servers with i jobs. Then from Theorem 3.5 (also Proposition 2.1 of [32]), $\overline{\mathbf{W}}^{(N)}(1)$) where is a deterministic measure in $\mathsf{M}_1(\mathsf{S})$. Since the SQ(d) policy uses the queue-size information of a finite set of d randomly sampled servers that are independent with identical distributions coinciding with , the arrival process is a Poisson process to any particular server which is a necessary condition to have insensitivity in PS systems [76]. Furthermore, the arrival process to each server is a state-dependent Poisson process with rate $_k = \frac{f(\sum_{j=k}^{C} \ _j)^d \ (\sum_{j=k+1}^{C} \ _j)^d g}{k}$ when there are k jobs at the server. Therefore, the set of arrival rates $\Lambda = (\ _k; 0 \quad k \quad C)$ can be written as a function of as

$$\Lambda = \tilde{F}_1(\ ): \tag{3.55}$$

Further, for a given set of arrival rates Λ, the stationary distribution for a server occupancy in the limiting system can be written through a mapping $\tilde{F}_2$ as

$$= \tilde{F}_2(\Lambda): \tag{3.56}$$

Therefore, must be a unique fixed-point of the mapping $\tilde{F}_2(\tilde{F}_1)$ for the case of general JLDs which was not shown in [24] except for the case of FCFS queues with service time distributions having decreasing hazard rate functions. To have insensitivity, must be the same for all the general JLDs having the same average job length. In [24], from uniqueness of in *ansatz*, insensitivity is concluded from reversibility since the arrival process to each server is a state-dependent Poisson process. Note that since the mappings $\tilde{F}_2$, $\tilde{F}_1$ are the same for exponential and general distributions, the uniqueness of the fixed-point of $\tilde{F}_2(\tilde{F}_1)$ follows from the result that the mean-field is GAS in the exponential case. Therefore, the fixed-point of the mapping $\tilde{F}_2(\tilde{F}_1)$ is the same for both the exponential and general job length distributions when they have the same average job lengths. Since the *ansatz* implies the Poisson arrival process to servers and uniqueness of the stationary distribution in the limiting system, insensitivity follows immediately. However, the proof of the *ansatz* remains an open problem and has been shown only for the case of FCFS queueing models with service time distributions having decreasing hazard rate functions, and for more general systems with small arrival rates in [40]. It remains an open problem for loss and PS models. For these systems, we do not have ordering relationships on servers' states in U. For FCFS models, an element in U is two-dimensional of the form $(n; x)$ where n indicates the number of jobs at the queue and x is the age or residual service time of the progressing job. Hence, we can define ordering relationships for elements in U for FCFS models in that for two elements $(n_1; x_1)$ and $(n_2; x_2)$, we say $(n_1; x_1)$ $(n_2; x_2)$ if n_1 n_2 and x_1 x_2. We can not define such relationships for the loss and PS models. Even the analysis for JLDs with decreasing hazard rate functions is also an open problem for loss and PS models.

Unlike the approach of [24], which makes several assumptions on the stationary regime, we first focus on establishing the mean-field limit by studying the transient regime. We then show that the probability distribution of occupancy obtained from every fixed-point of the mean-field is a fixed-point of the mapping $\tilde{F}_2(\tilde{F}_1)$ which is the same mapping for the exponential case. Since we know that the exponential case has a unique fixed-point, the fixed-point is insensitive. If we can show that the mean-field is GAS, then the fixed-point is also the stationary distribution of a server when $N \; ! \; 1$.

3.7 Proofs of Main Results

In this section, we provide proofs of the results sated in Section 3.4.

3.7.1 Proof of Lemma 3.1

Proof. An arrival can be routed in any one of the stages to a potential destination server. Let k be the index of the stage in which the arrival is routed to a server. Furthermore, let us assume that the dispatcher has sampled i potential destination servers by the end of stage k and let the vector of occupancies of the potential destination servers be equal to $\mathbf{n} = (n_1; \quad ; n_i) \; 2 \; Y^{[k;i]}$. Also, let the j^{th} potential destination server be in the state $(m; x_1; \quad ; x_m)$. Then since each server can be chosen as a potential destination server with probability $\frac{1}{N}$, the probability that the potential destination servers' occupancies form the vector $\mathbf{n} = (n_1; \quad ; n_i)$ and the j^{th} server lies in the state $(m; x_1; \quad ; x_m)$ is equal to

$$\left\{ \frac{(f(m; x_1; \quad ; x_m)g)}{N} \right\} \prod_{r=1;\, r \neq j}^{i} \left\{ \frac{(U_{n_r})}{N} \right\} :$$

Then the j^{th} potential destination server can be chosen as the destination server with probability $q^{[k;i]}(\mathbf{n}; j)$ and hence, the probability that the occupancy vector of potential destination servers is $\mathbf{n}$ and the destination server lies in the j^{th} position with state $(m; x_1; \quad ; x_m)$ is equal to

$$q^{[k;i]}(\mathbf{n}; j) \left\{ \frac{(f(m; x_1; \quad ; x_m)g)}{N} \right\} \prod_{r=1;\, r \neq j}^{i} \left\{ \frac{(U_{n_r})}{N} \right\} :$$

Finally, by summing over all possible values of $k; \mathbf{n}; i$, and j, we get (3.10). □

3.7.2 Proof of Theorem 3.3

Let $= (\ (\mathbf{u}); \mathbf{u} \; 2 \; U)$ be a fixed-point of the process $(P_t; t \quad 0)$ in the set Y. Let $v_n = \prod_{i=1}^{n} \frac{\quad}{i \;\; 1(\ )}$. Since $\quad 2 \; Y$, we have $\sum_{m=1}^{1} v_m < 1$. We first show that

$$(n; y_1; : : : ; y_n) = \frac{v_n}{1 + \sum_{m=1}^{1} v_m} \; {}^{n} \prod_{j=1}^{n} \int_{x_j=0}^{y_j} \overline{G}(x_j)\, dx_j; \tag{3.57}$$

and

$$(0) = \frac{1}{1 + \sum_{m=1}^{1} v_m} : \tag{3.58}$$

We prove (3.57)-(3.58) as follows. Consider a system that has one PS server where jobs arrive according to a Poisson process with rate $\Phi_n(\ )$ when there are n jobs and the job length distribution is G. Then from [22], there exists a unique stationary distribution

$^{(single)}$. Then $^{(single)}(n; y_1; \quad ; y_n)$ is equal to right side of (3.57). From the stationary mean-field dynamics, we have that is also another stationary distribution of the system. Due to the stability of the single server system, we must have that $=$ $^{(single)}$. This implies that (3.57)-(3.58) are true.

Let $\Gamma = (\Gamma_n; n \quad 0)$ where $\Gamma_n = \lim_{b \,!\, 1} (n; b; \quad ; b))$ and $\Gamma_0 = (0)$. Then from (3.57)-(3.58), we have

$$\Gamma_n = \frac{v_n}{1 + \sum_{j=1}^{1} v_j}$$

and

$$\Gamma_0 = \frac{1}{1 + \sum_{j=1}^{1} v_j}:$$

Further, we can write v_n in terms of Γ as $v_n = \prod_{i=1}^{n} \frac{}{i \;\; 1(\;)}$ where

$$\Psi_l(\Gamma) = \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y^{[k;i]}} \sum_{r=1}^{i} I_{fn_r = lg} q^{[k;i]}(\mathbf{n}; r) \prod_{j=1; j \in r}^{i} \Gamma_{n_j}:$$

It can be seen that for $n \quad 0$, we have

$$\Psi_n(\Gamma)\Gamma_n = \quad \Gamma_{n+1}: \tag{3.59}$$

Then from (3.18)-(3.19), for exponential JLDs with the same mean $\frac{1}{}$, a fixed-point $\mathbf{w} = (w_n; n \quad 0)$ of the mean-field satisfies that for $n \quad 0$,

$$\Psi_n(\mathbf{w})w_n = \quad w_{n+1}: \tag{3.60}$$

Therefore, from (3.59) and (3.60), we have that Γ is a fixed-point of the mean-field under the assumption of exponential JLDs with average job length $\frac{1}{}$. Furthermore, from (3.57)-(3.58), we obtain (3.21).

3.7.3 Proof of Theorem 3.1

The proof is divided into three steps. In the first step, we obtain the semigroup operator of the Markov process $(\ _t^{(N)}; t \quad 0)$. The Feller property of the Markov process $(\ _t^{(N)}; t \quad 0)$ is established in the second step. We then compute the generator $A^{(N)}(\ )$ and prove the martingale property of the process defined in equation (3.69) by using the Dynkin's lemma [56, Proposition 1.7, p.162] in the third step.

We now compute the semigroup operator $T_h^{(N)}(\ )$ which is defined as

$$T_h^{(N)} f(\ ) = \mathsf{E}\left[f(\ {}_h^{(N)}) j\ {}_0^{(N)} = \ \right];$$

where the mapping $f : \mathsf{M}_F(\mathsf{U})$?! R is a bounded continuous mapping. For ${}_0^{(N)} = \ $, in the interval $[0; h]$, let $\mathbf{A}_h$ and $\mathbf{D}_h$ be the number of arrivals and departures, respectively. Let us assume that measure contains mass at m distinct points denoted by $\mathbf{u}^{(l)} = (n_l; u_1^{(l)}; \ ; u_{n_l}^{(l)})$, $1 \quad l \quad m$. Then the number of servers with state $\mathbf{u}^{(l)}$ is equal to $(f\mathbf{u}^{(l)}g)$. If a server lies in the state $\mathbf{b} = (n; b_1; \ ; b_n)$ at time t, conditioned on the event that there are no arrivals in the interval $[0; h]$ at the server, let $p_{ND}(\mathbf{b}; h)$ be the probability that there are no departures at the server in the interval $[0; h]$. Then

$$p_{ND}(\mathbf{b}; h) = \prod_{i=1}^{n} \left\{ \frac{\overline{G}(b_i + \frac{h}{n})}{\overline{G}(b_i)} \right\}:$$

From the definition of the hazard rate function, we have

$$p_{ND}(\mathbf{b}; h) = \prod_{i=1}^{n} \left\{ 1 \quad \frac{(b_i)h}{n} \right\} + o(h): \tag{3.61}$$

In the following result, $B(\mathbf{u}^{(j)}; i; y; h)$ denotes the state of a server at $t = h$ given that its state is $\mathbf{u}^{(j)}$ at time $t = 0$ and the job at the i^{th} position departs when its age is y. Similarly, $C(\mathbf{u}^{(j)}; i; x; h)$ denotes the state of a server at time $t = h$ given that its state is $\mathbf{u}^{(j)}$ at time $t = 0$ and a job arrives to this server at time $t = x$ which joins the server at the i^{th} position.

Lemma 3.3. *For f 2 $\mathsf{C}_b(\mathsf{M}_F(\mathsf{U}))$, the semigroup operator $T_h^{(N)}(\ )$ is given by*

$$\begin{aligned}
T_h^{(N)} f(\ ) = (1 \quad N \ \ h) & \prod_{j=1;\, n_j>0}^{m} (p_{ND}(\mathbf{u}^{(j)}; h))^{\ (f\mathbf{u}^{(j)}g)} f(\tilde{}_h \ ) \\
& + (1 \quad N \ \ h) \sum_{j=1;\, n_j>0}^{m} (f\mathbf{u}^{(j)}g) \sum_{i=1}^{n_j} \int_{y=u_i^{(j)}}^{u_i^{(j)}+\frac{h}{n_j}} \left\{ \frac{g(y)}{\overline{G}(u_i^{(j)})} \right\} \\
& \prod_{k=1;\, k\in i}^{n_j} \left\{ \frac{\overline{G}\left(u_k^{(j)} + (y \quad u_i^{(j)}) + \frac{h \quad (y \quad u_i^{(j)})n_j}{n_j \quad 1} \right)}{\overline{G}(u_k^{(j)})} \right\} \\
& \left((p_{ND}(\mathbf{u}^{(j)}; h))^{(\ (f\mathbf{u}^{(j)}g) \ \ 1)} \right) \left(\prod_{r=1;\, r\in j;\, n_r>0}^{m} (p_{ND}(\mathbf{u}^{(r)}; h))^{\ (f\mathbf{u}^{(r)}g)} \right) \\
& f\left(\tilde{}_h \ + \ \left(B(\mathbf{u}^{(j)}; i; y; h) \right) \quad (n_j; u_1^{(j)} + \tfrac{h}{n_j}; \ ; u_{n_j}^{(j)} + \tfrac{h}{n_j}) \right) dy
\end{aligned}$$

$$+N\ h\int_{x=0}^{h}\frac{1}{h}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j+1}\left\{\frac{1}{n_j+1}\right\}p_r(\tilde{}_x\ :\ \tilde{}_x^{+}\mathbf{u}^{(j)})\left[f\left(\tilde{}_h\ +\ fC(\mathbf{u}^{(j)};i;x;h)g\qquad(\tilde{}_h^{+}\mathbf{u}^{(j)})\right)\right.\right.$$
$$\left.\left.\left(\prod_{q=1}^{n_j}\left\{\frac{\overline{G}(u_q^{(j)}+\frac{x}{n_j}+\frac{h\quad x}{n_j+1})}{\overline{G}(u_q^{(j)})}\right\}\right)\overline{G}\left(\frac{h\quad x}{n_j+1}\right)(p_{ND}(\mathbf{u}^{(j)};h))^{\ (f\mathbf{u}^{(j)}g)\ \ 1}\right.$$
$$\left.\prod_{k=1;k\in j;n_k>0}^{m}(p_{ND}(\mathbf{u}^{(k)};h))^{\ (f\mathbf{u}^{(k)}g)}\right]\Bigg)dx+\ (\ ;h);\quad(3.62)$$

where

$$B(\mathbf{u}^{(j)};i;y;h)=\Bigg(n_j\quad 1;u_1^{(j)}+(y\quad u_i^{(j)})+\frac{fh\quad n_j(y\quad u_i^{(j)})g}{n_j\quad 1};\quad ;$$
$$u_i^{(j)}{}_1+(y\quad u_i^{(j)})+\frac{fh\quad n_j(y\quad u_i^{(j)})g}{n_j\quad 1};u_{i+1}^{(j)}+(y\quad u_i^{(j)})+\frac{fh\quad n_j(y\quad u_i^{(j)})g}{n_j\quad 1};\quad ;$$
$$u_{n_j}^{(j)}+(y\quad u_i^{(j)})+\frac{fh\quad n_j(y\quad u_i^{(j)})g}{n_j\quad 1}\Bigg);\quad(3.63)$$

$$C(\mathbf{u}^{(j)};i;x;h)=\Bigg(n_j+1;u_1^{(j)}+\frac{x}{n_j}+\frac{h\quad x}{n_j+1};\quad ;u_i^{(j)}{}_1+\frac{x}{n_j}+\frac{h\quad x}{n_j+1};$$
$$\frac{h\quad x}{n_j+1};u_i^{(j)}+\frac{x}{n_j}+\frac{h\quad x}{n_j+1};\quad ;u_{n_j}^{(j)}+\frac{x}{n_j}+\frac{h\quad x}{n_j+1}\Bigg)\quad(3.64)$$

and $(\ ;h)$ *is* $o(h)$.

Proof. We write $T_h^{(N)}f(\ )$ as

$$T_h^{(N)}f(\ )=\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=0g}j\ {}_0^{(N)}=\ \right]+\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=1g}j\ {}_0^{(N)}=\ \right]$$
$$+\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=1g}I_{f\mathbf{D}_h=0g}j\ {}_0^{(N)}=\ \right]+\sum_{i\ 1}\sum_{j\ 1}\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=ig}I_{f\mathbf{D}_h=jg}j\ {}_0^{(N)}=\ \right]:\quad(3.65)$$

We first obtain the expression for $\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=0g}j\ {}_0^{(N)}=\ \right]$ which is associated with the case when there are no arrivals and no departures in the interval $[0;h]$. In this situation, since the number of progressing jobs is constant, ${}_h^{(N)}=\tilde{}_h$. Hence,

$$\mathsf{E}\left[f(\ {}_h^{(N)})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=0g}j\ {}_0^{(N)}=\ \right]=\mathsf{P}(f\mathbf{A}_h=0;\mathbf{D}_h=0gj\ {}_0^{(N)}=\ )f(\tilde{}_h\ ):$$

As the arrival process is a Poisson process, we can write

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=0g}j\ {}^{(N)}_{0}=\ \right]$$
$$=\mathsf{P}(f\mathbf{A}_h=0g)\mathsf{P}(f\mathbf{D}_h=0gj\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ )f(\tilde{}_h\ ):$$

Let us define

$$_1(\ ;h)=(\mathsf{P}(f\mathbf{A}_h=0g)\quad(1\quad N\ \ h))\mathsf{P}(f\mathbf{D}_h=0gj\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ )f(\tilde{}_h\ ):$$

Then

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=0g}j\ {}^{(N)}_{0}=\ \right]$$
$$=(1\quad N\ \ h)\mathsf{P}(f\mathbf{D}_h=0gj\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ )f(\tilde{}_h\ )+\ _1(\ ;h):$$

In the time interval $[0;h]$, conditioned on the event that there are no arrivals, the probability that there are no departures is equal to

$$\mathsf{P}(f\mathbf{D}_h=0gj\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ )=\prod_{i=1;\,n_i>0}^{m}(p_{N\,D}(\mathbf{u}^{(i)};h))^{\ (f\mathbf{u}^{(i)}g)}:$$

Since $f\ 2\ \mathsf{C}_b(\mathsf{M}_F(\mathsf{U}))$ and $(\mathsf{P}(f\mathbf{A}_h=0g)\quad(1\quad N\ \ h))=o(h)$ as the arrival process is Poisson with rate $N\ $, we conclude $_1(\ ;h)$ is $o(h)$.

Next, consider the second case where there is no arrival and one job departs in the interval $[0;h]$. Let us define

$$_2(\ ;h)=(\mathsf{P}(f\mathbf{A}_h=0g)\quad(1\quad N\ \ h))\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=1g}j\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ \right]:$$

Then $_2(\ ;h)$ is $o(h)$. We can write

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=1g}j\ {}^{(N)}_{0}=\ \right]=(1\quad N\ \ h)\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=1g}j\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ \right]$$
$$+\ _2(\ ;h):$$

We next simplify the expression of $\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=1g}j\mathbf{A}_h=0;\ {}^{(N)}_{0}=\ \right]$. Suppose that a job in the i^{th} position with age x departs at a server whose state was $(n;b_1;\quad ;b_n)$ at time $t=0$. This implies that the departure has occurred at time $(x\quad b_i)n$. In this case, at time h, the server where a job departs will lie in the state $\big(n\quad 1;b_1+(x\quad b_i)+\frac{fh\quad n(x\quad b_i)g}{n\quad 1};\quad ;b_i\ _1+(x\quad b_i)+\frac{fh\quad n(x\quad b_i)g}{n\quad 1};b_{i+1}+(x\quad b_i)+\frac{fh\quad n(x\quad b_i)g}{n\quad 1};\quad ;b_n+(x\quad b_i)+\frac{fh\quad n(x\quad b_i)g}{n\quad 1}\big)$. Using this, we can write the following equation by assuming that a job at the i^{th} position with age y departs from a server that had state $\mathbf{u}^{(j)}=(n_j;u_1^{(j)};\quad ;u_{n_j}^{(j)})$ at time $t=0$,

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h=0g}I_{f\mathbf{D}_h=1g}j\ {}^{(N)}_{0}=\ \right]=(1\quad N\ \ h)\sum_{j=1;\,n_j>0}^{m}(f\mathbf{u}^{(j)}g)\sum_{i=1}^{n_j}\int_{y=u_i^{(j)}}^{u_i^{(j)}+\frac{h}{n_j}}\left\{\frac{g(y)}{\overline{G}(u_i^{(j)})}\right\}$$

$$\prod_{k=1;\,k\in i}^{n_j}\left\{\frac{\overline{G}\left(u_k^{(j)}+(y\quad u_i^{(j)})+\frac{h\quad(y\quad u_i^{(j)})n_j}{n_j\quad 1}\right)}{\overline{G}(u_k^{(j)})}\right\}\left((p_{N\,D}(\mathbf{u}^{(j)};h))^{(\ (f\mathbf{u}^{(j)}g)\ \ 1)}\right)$$

$$\left(\prod_{l=1;\,l\in j;\,n_l>0}^{m}(p_{N\,D}(\mathbf{u}^{(l)};h))^{\ (f\mathbf{u}^{(l)}g)}\right)f\left({}^{\sim}_{h}\ +\ {}_{\left(B(\mathbf{u}^{(j)};i;y;h)\right)\qquad n_j;u_1^{(j)}+\frac{h}{n_j};\quad ;u_{n_j}^{(j)}+\frac{h}{n_j}}\right)dy$$

$$+\ {}_2(\ ;h):$$

Let us now consider the third case that corresponds to the event that there is no departure but an arrival occurs in the interval $[0;h]$. Let ${}_3(\ ;h)$ be defined as

$${}_3(\ ;h)=(\mathsf{P}(f\mathbf{A}_h=1g)\quad N\ \ h)\,\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=0g}j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ \right]:$$

Since $\mathsf{P}(f\mathbf{A}_h=1g)\quad N\ \ h=o(h)$ and $f\ 2\ \mathsf{C}_b(\mathsf{M}_F(\mathsf{U}))$, ${}_3(\ ;h)$ is $o(h)$. We can write

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h=1g}I_{f\mathbf{D}_h=0g}j\ {}^{(N)}_{0}=\ \right]=N\ \ h\,\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{fD_h=0g}jA_h=1;\ {}^{(N)}_{0}=\ \right]$$

$$+\ {}_3(\ ;h);\quad(3.66)$$

We next obtain the expression for $\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=0g}j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ \right]$. We recall that since the arrival process is Poisson, conditioned on the event that there is one arrival in $[0;h]$, the time at which the job arrives is uniformly distributed in $[0;h]$ [77]. Let x be the time at which the job arrives and let i be its position at the destination server that had state $(n;b_1;\quad ;b_n)$ at time $t=0$. Then the state of the destination server of the arrival at time h will be equal to $(n+1;b_1+\frac{x}{n}+\frac{h\ \ x}{n+1};\quad ;b_{i\ \ 1}+\frac{x}{n}+\frac{h\ \ x}{n+1};\frac{h\ \ x}{n+1};b_i+\frac{x}{n}+\frac{h\ \ x}{n+1};\quad ;b_n+\frac{x}{n}+\frac{h\ \ x}{n+1})$. Note that the state of the system at time x is equal to ${}^{\sim}_{x}$ and hence, the dispatcher uses ${}^{\sim}_{x}$ as the system state while implementing the routing policy. As a consequence, we get the following equation where we use the notation that a job arrives at time x and it is routed to a server that was lying in state $\mathbf{u}^{(j)}$ at time $t=0$, and the job joins at the i^{th} position. Then

$$\mathsf{E}\left[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=0g}j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ \right]=\int_{x=0}^{h}\frac{1}{h}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j+1}\left\{\frac{1}{n_j+1}\right\}p_r({}^{\sim}_{x}\ :\ {}^{\sim+}_{x}\mathbf{u}^{(j)})\right.$$

$$\left[f\left({}^{\sim}_{h}\ +\ {}_{fC(\mathbf{u}^{(j)};i;x;h)g\qquad({}^{\sim+}_{h}\mathbf{u}^{(j)})}\right)\left(\prod_{q=1}^{n_j}\left\{\frac{\overline{G}(u_q^{(j)}+\frac{x}{n_j}+\frac{h\ \ x}{n_j+1})}{\overline{G}(u_q^{(j)})}\right\}\right)\overline{G}\left(\frac{h\quad x}{n_j+1}\right)\right.$$

$$(p_{ND}(\mathbf{u}^{(j)};h))^{(f\mathbf{u}^{(j)}g)\ \ 1}\prod_{k=1;k\in\mathcal{E}j;n_k>0}^{m}(p_{ND}(\mathbf{u}^{(k)};h))^{(f\mathbf{u}^{(k)}g)}\Bigg]\Bigg)\,dx: \quad (3.67)$$

We now consider the case when there are multiple events in the interval $[0;h]$. We show that $\mathsf{E}\Big[f(\ {}^{(N)}_{h})I_{f\mathbf{A}_h\ \ 1;\mathbf{D}_h\ \ 1g}j\ {}^{(N)}_{0}=\ \Big]$ is $o(h)$ denoted by ${}_4(\ ;h)$. Since f is bounded, it is sufficient to show that $\mathsf{P}(\mathbf{A}_h\ \ 1;\mathbf{D}_h\ \ 1j\ {}^{(N)}_{0}=\ )$ is $o(h)$. From the fact that $\mathsf{P}(\mathbf{A}_h\ \ 2)$ is $o(h)$ due to the Poisson arrival process assumption, we get that $\mathsf{P}(\mathbf{A}_h\ \ 2;\mathbf{D}_h\ \ 1j\ {}^{(N)}_{0}=\ )$ is $o(h)$. Also, we can write

$$\mathsf{P}(\mathbf{A}_h=1;\mathbf{D}_h\ \ 1j\ {}^{(N)}_{0}=\ )=\mathsf{P}(\mathbf{A}_h=1)(1\ \ \mathsf{P}(\mathbf{D}_h=0j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ ))$$

We show $\lim_{h\,!\,0}\mathsf{P}(\mathbf{D}_h=0j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ ))=1$. From (3.67), we replace x with hz, then

$$\mathsf{E}\Big[f(\ {}^{(N)}_{h})I_{f\mathbf{D}_h=0g}j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ \Big]=\int_{z=0}^{1}\Bigg(\sum_{j=1}^{m}\sum_{i=1}^{n_j+1}\Big\{\frac{1}{n_j+1}\Big\}p_r(\tilde{}_{hz}\ :\tilde{}^{+}_{hz}\mathbf{u}^{(j)})$$
$$\Bigg[f\Big(\tilde{}_h\ +\ {}_{fC(\mathbf{u}^{(j)};i;hz;h)g}\ \ (\tilde{}^{+}_{h}\mathbf{u}^{(j)})\Big)\Bigg(\prod_{q=1}^{n_j}\Big\{\frac{\overline{G}(u_q^{(j)}+\frac{hz}{n_j}+\frac{h\ \ hz}{n_j+1})}{\overline{G}(u_q^{(j)})}\Big\}\Bigg)\overline{G}\Big(\frac{h\ \ hz}{n_j+1}\Big)$$
$$(p_{ND}(\mathbf{u}^{(j)};h))^{(f\mathbf{u}^{(j)}g)\ \ 1}\prod_{k=1;k\in\mathcal{E}j;n_k>0}^{m}(p_{ND}(\mathbf{u}^{(k)};h))^{(f\mathbf{u}^{(k)}g)}\Bigg]\Bigg)\,dz: \quad (3.68)$$

From (3.68), it can be seen that $\lim_{h\,!\,0}\mathsf{P}(\mathbf{D}_h=0j\mathbf{A}_h=1;\ {}^{(N)}_{0}=\ )=1$, and hence, we get $\mathsf{P}(\mathbf{A}_h=1;\mathbf{D}_h\ \ 1j\ {}^{(N)}_{0}=\ )=o(h)$.

Finally, by defining

$$(\ ;h)=\ {}_1(\ ;h)+\ {}_2(\ ;h)+\ {}_3(\ ;h)+\ {}_4(\ ;h);$$

we get (3.62). □

By the same arguments as in the proof of Theorem 2.8.1, we can show that the process $(\ {}^{(N)}_{t};t\ \ 0)$ is a Feller-Dynkin process.

We now focus on establishing the third result. We first recall the definition of $A^{(N)}(\ )$. For any $F\ 2\ \mathsf{C}(\mathsf{M}_F(\mathsf{U}))$, the generator $A^{(N)}(\ )$ is defined as

$$A^{(N)}F(\ )=\lim_{h\,!\,0}\frac{\mathsf{E}\Big[F(\ {}^{(N)}_{h})j\ {}^{(N)}_{0}=\ \Big]\ \ F(\ )}{h};$$

where $F\ 2\ \mathsf{C}(\mathsf{M}_F(\mathsf{U}))$ is chosen such that the limit is well defined.

Lemma 3.4. *For* $\in C_b^1((U))$, *the process* $(M_t^{(N)}(\ ); t \quad 0)$ *de ned as*

$$M_t^{(N)}(\ ) = h\ {}_t^{(N)};\ i \quad h\ {}_0^{(N)};\ i \quad \int_{s=0}^{t} A^{(N)} h\ {}_s^{(N)};\ i ds \tag{3.69}$$

is a square integrable $F_t^{(N)}$*-martingale and it is an RCLL process. Furthermore, for* $\in C_b^1(U)$, *the quadratic variation of* $(M_t^{(N)}(\ ); t \quad 0)$ *is given by*

$$\begin{aligned} < M^{(N)}(\ ) >_t = \int_{s=0}^{t} \Bigg(& \sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \\ & (\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))^2 \, d\ {}_s^{(N)}(n; x_1; \quad ; x_n) \\ & + N \left[\left\{ \frac{{}_s^{(N)}(f0g)}{N} \right\} \Phi_0 \left(\frac{{}_s^{(N)}}{N} \right) (\ (1; 0) \quad (0))^2 \right. \\ & + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{N(n+1)} \right\} \Phi_n \left(\frac{{}_s^{(N)}}{N} \right) \\ & (\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))^2 \\ & \left. d\ {}_s^{(N)}(n; x_1; \quad ; x_n) \right] \Bigg) ds: \end{aligned} \tag{3.70}$$

.

Proof. We first obtain the expression for the generator $A^{(N)}(\ )$. From (3.62) and since the set of linear combinations of Q_f for $f \in C_s^1(U)$ where $Q_f(\ ) = e^{\ h\ ; f\ i}$ is dense in the set $C(M_F(U))$ [78, proposition 7:10], from the expression of $A^{(N)} Q_f(\ )$, we get

$$\begin{aligned} A^{(N)} F(\ ) = & \lim_{h\ !\ 0} \left(\frac{F(\tilde{}_h\ ) \quad F(\ )}{h} \right) \quad N \quad F(\ ) \\ & \sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} F(\ ) d\ (n; x_1; \quad ; x_n) + \sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \\ & \left(F(\ + {}_{(n\ 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n)} \quad {}_{(n; x_1; \quad ; x_n)}) \right) d\ (n; x_1; \quad ; x_n) \\ + N & \left[\left\{ \frac{(f0g)}{N} \right\} \Phi_0 \left(\frac{}{N} \right) F(\ + {}_{(1;0)} \quad {}_{(0)}) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{N(n+1)} \right\} \Phi_n \left(\frac{}{N} \right) \right. \\ & \left. F(\ + {}_{(n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n)} \quad {}_{(n; x_1; \quad ; x_n)}) d\ (n; x_1; \quad ; x_n) \right] : \end{aligned}$$

For $\in C_b^1(U)$, $\in M_F(U)$, then $A^{(N)} h\ ;\ i$ is well defined. From the Dynkin's formula [56, Proposition 1.7, p.162], $(M_t^{(N)}(\ ); t \quad 0)$ defined as

$$M_t^{(N)}(\ ) = h\ {}_t^{(N)};\ i \quad h\ {}_0^{(N)};\ i \quad \int_{s=0}^{t} A^{(N)} h\ {}_s^{(N)};\ i ds \tag{3.71}$$

is an RCLL $\mathcal{F}_t^{(N)}$-local martingale. It can be checked that after simplifications, we get

$$\begin{aligned}
\mathbf{M}_t^{(N)}(\ ) = h\ {}_t^{(N)};\ \ i \quad h\ {}_0^{(N)};\ \ i \quad \int_{s=0}^{t} h\ {}_s^{(N)};\ \ {}^{\natural} i\, ds \quad \int_{s=0}^{t} \Bigg(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \\
(\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; : : : ; x_n) \quad (n; x_1; \quad ; x_n))\, d\ {}_s^{(N)}(n; x_1; \quad ; x_n) \\
+ N \left[\left(\left\{ \frac{{}_s^{(N)}(f0g)}{N} \right\} \Phi_0 \left(\frac{{}_s^{(N)}}{N} \right) (\ (1; 0) \quad (0)) \right) \right. \\
+ \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{N(n+1)} \right\} \Phi_n \left(\frac{{}_s^{(N)}}{N} \right) \\
\left. (\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\, d\ {}_s^{(N)}(n; x_1; \quad ; x_n) \right] \Bigg)\, ds;
\end{aligned} \tag{3.72}$$

where

$$h\ {}_s^{(N)};\ \ {}^{\natural} i = \sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \frac{1}{n} \frac{@\ (n; x_1; : : : ; x_n)}{@ x_i}\, d\ {}_s^{(N)}(n; x_1; \quad ; x_n): \tag{3.73}$$

By choosing $F\ (\ {}_t^{(N)}) = h\ {}_t^{(N)};\ \ i$, from [79, Theorem 7.15], we have

$$< \mathbf{M}^{(N)}(\ ) >_t = \int_{s=0}^{t} \left(A^{(N)} F^2(\ {}_s^{(N)}) \quad 2F\ (\ {}_s^{(N)}) A^{(N)} F\ (\ {}_s^{(N)}) \right) ds: \tag{3.74}$$

We get (3.70) after simplifications. In the next section, we show in (3.78) that $\sup_m \Phi_m(\ ) < 1$. Using this result, and from the fact that $2\ \mathcal{C}_b^1(\mathcal{U})$ and $2\ \mathcal{C}_b(\mathbb{R}_+)$, we have $\mathbb{E}\left[< \mathbf{M}^{(N)}(\ ) >_t \right] < 1$. Hence, $(\mathbf{M}_t^{(N)}(\ ); t \quad 0)$ is a square integrable martingale. □

3.7.4 Proof of Theorem 3.2

We first show that there exists a unique mean-field solution. This step is required to prove the convergence of $(\frac{{}_t^{(N)}}{N}; t \quad 0)$ as $N \ ! \ 1$. We then prove the convergence of $(\frac{{}_t^{(N)}}{N}; t \quad 0)$ as $N \ ! \ 1$ to the unique mean-field solution.

Existence and Uniqueness of a Mean-field Solution

From (3.16), for $2\ \mathcal{C}_b(\mathcal{U})$, the operator $7! \ h^{-}{}_t;\ \ i$ is a linear operator and ${}^{-}{}_t(\mathcal{U}) = 1$. As a consequence, from the Riesz-Markov-Kakutani theorem [69,80] given in Theorem A.4, for

$^{-}_{t}\ 2\ \mathsf{M}_1(\mathsf{U})$, the uniqueness of the probability measure $^{-}_{t}$ is equivalent to the uniqueness of the operator $\ 7!\ h^{-}_{t};\ \ i$. Also, since $\mathsf{C}_b(\mathsf{U})$ is a separating class of $\mathsf{M}_1(\mathsf{U})$ [56, p. 111], if two measures $_1;\ _2\ 2\ \mathsf{M}_1(\mathsf{U})$ satisfies $h\ _1;\ \ i = h\ _2;\ \ i$ for all $\ 2\ \mathsf{C}_b(\mathsf{U})$, then $_1 = \ _2$. Therefore it is enough to show the uniqueness of the operator $\ 7!\ h^{-}_{t};\ \ i$.

For given $^{-}_{0}$, we now show that there exists at most one real valued process $h(^{-}_{t};\ \ i;t$ $0)$ that satisfies the MFEs. Let $(^{-1}_{t};t\ \ 0);(^{-2}_{t};t\ \ 0)$ be the two mean-field solutions with initial points $^{-1}_{0};^{-2}_{0}$, respectively. Then for $\ 2\ \mathsf{C}_b(\mathsf{U})$, we get

$$
\begin{aligned}
h^{-1}_{t}\ \ ^{-2}_{t};\ \ i = h^{-1}_{0}\ \ ^{-2}_{0};\tilde{}_{t}\ \ i + \int_{s=0}^{t}\Bigg(\sum_{n=1}^{1}\sum_{i=1}^{n}\int_{x_1}\ \ \int_{x_n}\left\{\frac{(x_i)}{n}\right\}\\
(\tilde{}_{t\ s}\ (n\ \ 1;x_1;\ \ \ ;x_{i\ \ 1};x_{i+1};\ \ \ ;x_n)\ \ \tilde{}_{t\ s}\ (n;x_1;\ \ \ ;x_n))\\
d(^{-1}_{s}\ \ ^{-2}_{s})(n;x_1;\ \ \ ;x_n)\Bigg)\,ds\\
+\int_{s=0}^{t}\Bigg(\ \Big[(^{-1}_{s}(f0g)\Phi_0(^{-1}_{s})\,(\tilde{}_{t\ s}\ (1;0)\ \ \tilde{}_{t\ s}\ (0))\Big)\\
+\sum_{n=1}^{1}\sum_{i=1}^{n+1}\int_{x_1}\ \ \int_{x_n}\frac{1}{(n+1)}\Phi_n(^{-1}_{s})(\tilde{}_{t\ s}\ (n+1;x_1;\ \ \ ;x_{i\ \ 1};0;x_i;\ \ \ ;x_n)\\
\tilde{}_{t\ s}\ (n;x_1;\ \ \ ;x_n))\,d^{-1}_{s}(n;x_1;\ \ \ ;x_n)\Big]\\
\Big[(^{-2}_{s}(f0g)\Phi_0(^{-2}_{s})\,(\tilde{}_{t\ s}\ (1;0)\ \ \tilde{}_{t\ s}\ (0))\Big)\\
\sum_{n=1}^{1}\sum_{i=1}^{n+1}\int_{x_1}\ \ \int_{x_n}\left\{\frac{1}{(n+1)}\right\}\Phi_n(^{-2}_{s})(\tilde{}_{t\ s}\ (n+1;x_1;\ \ \ ;x_{i\ \ 1};0;x_i;\ \ \ ;x_n)\\
\tilde{}_{t\ s}\ (n;x_1;\ \ \ ;x_n))\,d^{-2}_{s}(n;x_1;\ \ \ ;x_n)\Big]\Bigg)\,ds: \quad (3.75)
\end{aligned}
$$

Our objective is to obtain a result of the form

$$k^{-1}_{t}\ \ ^{-2}_{t}k\ \ b + c\int_{s=0}^{t}k^{-1}_{s}\ \ ^{-2}_{s}k\,ds \quad (3.76)$$

for some $b;c > 0$, $t\ 2\ [0;T]$. Then from Gronwall's inequality stated in Theorem A.7 [56], we would get

$$k^{-1}_{t}\ \ ^{-2}_{t}k\ \ be^{ct} \quad (3.77)$$

for $t\ 2\ [0;T]$. In this direction, by using the first term on the right side of equation (3.75), we can write

$$\left|h^{-1}_{0}\ \ ^{-2}_{0};\tilde{}_{t}\ \ i\right|\ \ k^{-1}_{0}\ \ ^{-2}_{0}kk\ \ k:$$

We now simplify the second term on the right side of equation (3.75). For this, we define a function $w_{t;s}$ as follows:
for $n > 0$,

$$w_{t;s}(n; x_1; \quad ; x_n) = \sum_{i=1}^{n} \left\{ \frac{(x_i)}{n} \right\} (\tilde{}_{t\ s}\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n) \quad \tilde{}_{t\ s}\ (n; x_1; \quad ; x_n)$$

and for $n = 0$,

$$w_{t;s}(0) = 0:$$

Then $w_{t;s}\ 2\ \mathsf{C}_b(\mathsf{U})$ since $\quad 2\ \mathsf{C}_b(\mathsf{U})$ and $\quad 2\ \mathsf{C}_b(\mathsf{R}_+)$. Furthermore, we have

$$kw_{t;s}k \quad 2k\ \ kk\ \ k:$$

It can be seen that the second term can be written as $\int_{s=0}^{t} h^{-1}_{s} \quad {}^{-2}_{s}; w_{t;s} i\, ds$. This implies that we can write

$$\left| \int_{s=0}^{t} h^{-1}_{s} \quad {}^{-2}_{s}; w_{t;s} i\, ds \right| \quad \int_{s=0}^{t} 2k\ \ kk\ \ kk^{-1}_{s} \quad {}^{-2}_{s} k\, ds:$$

Let us now consider the third term on the right side of (3.75) that corresponds to the case of arrivals. Let $h_{t;s;}$ be the function defined as

$$h_{t;s;}\ (n; x_1; \quad ; x_n) = \sum_{i=1}^{n+1} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ ) \ (\tilde{}_{t\ s}\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad \tilde{}_{t\ s}\ (n; x_1; \quad ; x_n))$$

for all $(n; x_1; \quad ; x_n)$. Then the third term is equal to $\int_{s=0}^{t} \quad \left(h^{-1}_{s}; h_{t;s;^{-1}_{s}} i \quad h^{-2}_{s}; h_{t;s;^{-2}_{s}} i\right) ds$. We can bound this term using the fact that

$$\left|h^{-1}_{s}; h_{t;s;^{-1}_{s}} i \quad h^{-2}_{s}; h_{t;s;^{-2}_{s}} i\right| \quad \left|h^{-1}_{s} \quad {}^{-2}_{s}; h_{t;s;^{-1}_{s}} i\right| + \left|h^{-2}_{s}; h_{t;s;^{-1}_{s}} \quad h_{t;s;^{-2}_{s}} i\right|:$$

Also, we can write

$$\left|h^{-1}_{s}; h_{t;s;^{-1}_{s}} i \quad h^{-2}_{s}; h_{t;s;^{-2}_{s}} i\right| \quad k^{-1}_{s} \quad {}^{-2}_{s} kkh_{t;s;^{-1}_{s}} k + k^{-2}_{s} kkh_{t;s;^{-1}_{s}} \quad h_{t;s;^{-2}_{s}} k:$$

We have that $k^{-2}_{s} k = 1$ as ${}^{-2}_{s}$ is a probability measure.

We now obtain an upper bound on $kh_{t;s;^{-1}_{s}} k$. It can be seen that we have

$$kh_{t;s;^{-1}_{s}} k \quad 2k\ \ k \sup_{n\ 0} j\Phi_n(\ )j:$$

We now show that $\sup_{n \ 0} j\Phi_n(\ )j$ is bounded by a finite value. For $\ 2\ \mathsf{M}_1(\mathsf{U})$,

$$\Phi_m(\ ) = \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 \mathsf{Y}^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \in j}^{i} Q_{n_r}(\ ):$$

Let $\mathsf{Y}_m^{[k;i]}$ be defined as

$$\mathsf{Y}_m^{[k;i]} = f\mathbf{n} \ 2 \ \mathsf{Y}^{[k;i]} : 9j \text{ such that } n_j = mg:$$

Then we can write

$$\Phi_m(\ ) = \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 \mathsf{Y}_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \in j}^{i} Q_{n_r}(\ ):$$

We can bound the above term as

$$\Phi_m(\ ) \quad \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 \mathsf{Y}_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} \prod_{r=1; r \in j}^{i} Q_{n_r}(\ ):$$

Out of i sampled servers upto stage k, let b be the number of servers with m jobs. Then we can write

$$\begin{aligned}
\Phi_m(\ ) \quad & \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{b=1}^{i} b\binom{i}{b} (Q_m(\ ))^{b \ 1} (1 \quad Q_m(\ ))^{i \ b}; \\
& \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{b=1}^{i} b\binom{i}{b}; \\
& M(d^{(max)})^2 \sum_{b=1}^{d^{(max)}} \binom{d^{(max)}}{b}: \\
& M(d^{(max)})^3 (d^{(max)}!): \qquad (3.78)
\end{aligned}$$

Therefore, we have

$$kh_{t;s;\frac{-1}{s}}k \quad 2Mk \ k(d^{(max)})^3(d^{(max)}!):$$

Similarly, we can write

$$kh_{t;s;\frac{-1}{s}} \quad h_{t;s;\frac{-2}{s}}k \quad 2k \ k \sup_{m \ 0} \left|\Phi_m(\tfrac{-1}{s}) \quad \Phi_m(\tfrac{-2}{s})\right|:$$

We have

$$\left|\Phi_m(\tfrac{-1}{s}) \quad \Phi_m(\tfrac{-2}{s})\right|$$

$$= \left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \left(\prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad \prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \right) \right| :$$

We can write

$$\begin{aligned} &\left|\Phi_m\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad \Phi_m\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right)\right| \\ &\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \left(\prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_{n_1}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \prod_{r=2;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \right) \right| \\ &+ \left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \left(Q_{n_1}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \prod_{r=2;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad \prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \right) \right| : \end{aligned} \tag{3.79}$$

We next bound the first term on the right side of (3.79). For each j, we have

$$\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y_m^{[k;i]}} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \left(\prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_{n_1}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \prod_{r=2;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \right) \right|$$
$$M d^{(max)} \sum_{m \quad 0} \left|Q_m\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_m\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right)\right| : \tag{3.80}$$

Let us define a function $r_{(\begin{smallmatrix} -1 \\ s \end{smallmatrix}, \begin{smallmatrix} -2 \\ s \end{smallmatrix})} \; 2 \; \mathsf{C}_b(\mathsf{U})$ such that for $\mathbf{u} \; 2 \; \mathsf{U}_n$,

$$r_{(\begin{smallmatrix} -1 \\ s \end{smallmatrix}, \begin{smallmatrix} -2 \\ s \end{smallmatrix})}(\mathbf{u}) = \begin{cases} 1 & \text{if } Q_n\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_n\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \\ 1 & \text{otherwise:} \end{cases}$$

Then $\sum_{m \quad 0} j Q_m\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_m\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) j = h^{-1}_{s} \quad {}^{-2}_{s}; r_{(\begin{smallmatrix} -1 \\ s \end{smallmatrix}, \begin{smallmatrix} -2 \\ s \end{smallmatrix})} i$. Hence,

$$\sum_{m \quad 0} \left|Q_m\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_m\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right)\right| \quad k^{-1}_{s} \quad {}^{-2}_{s} k:$$

Therefore

$$\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y_m^{[k;i]}} \sum_{j=1}^{i} I_{fn_j = mg} q^{[k;i]}(\mathbf{n}; j) \left(\prod_{r=1;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad Q_{n_1}\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right) \prod_{r=2;\, r \in j}^{i} Q_{n_r}\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \right) \right|$$
$$M (d^{(max)})^2 k^{-1}_{s} \quad {}^{-2}_{s} k:$$

By similar arguments, the second term on the right side of (3.79) is bounded by $M (d^{(max)})^2 (d^{(max)} 1) k^{-1}_{s} \quad {}^{-2}_{s} k$. Hence,

$$\left|\Phi_m\left(\begin{smallmatrix} -1 \\ s \end{smallmatrix}\right) \quad \Phi_m\left(\begin{smallmatrix} -2 \\ s \end{smallmatrix}\right)\right| \quad M (d^{(max)})^3 k^{-1}_{s} \quad {}^{-2}_{s} k:$$

Therefore we obtain that

$$kh_{t;s;\bar{}_s^{-1}} \quad h_{t;s;\bar{}_s^{-2}}k \quad 2M\,k \;\; k(d^{(max)})^3 k_s^{-1} \quad \bar{}_s^{-2}k:$$

Finally, by using the above bounds, we get

$$\left|h_t^{-1} \quad \bar{}_t^{-2};\;\; i\right| \quad \left(k_0^{-1} \quad \bar{}_0^{-2}k + \int_{s=0}^{t} 2k \;\; kk_s^{-1} \quad \bar{}_s^{-2}k\,ds \right.$$
$$\left. +4M\,(d^{(max)})^3(d^{(max)}!) \int_{s=0}^{t} \quad k_s^{-1} \quad \bar{}_s^{-2}k\,ds\right) k \;\; k:$$

Hence

$$k_t^{-1} \quad \bar{}_t^{-2}k \quad k_0^{-1} \quad \bar{}_0^{-2}k + (2k \;\; k + 4M\,(d^{(max)})^3(d^{(max)}!) \;\;) \int_{s=0}^{t} k_s^{-1} \quad \bar{}_s^{-2}k\,ds:$$

From equation (3.76), we get

$$k_t^{-1} \quad \bar{}_t^{-2}k \quad k_0^{-1} \quad \bar{}_0^{-2}k\,e^{(2k \;\; k+4M\,(d^{(max)})^3(d^{(max)}!) \;\;)t}:$$

This implies that for given initial measure $\bar{}_0$, there exists at most one mean-field solution.

The existence of a mean-field solution follows from the proof of the relative compactness of the sequence $f\bar{}_t^{(N)};t \quad 0g$ in $\mathsf{D}_{\mathsf{M}_1(\mathsf{U})}([0;\,1\,))$. We show that every limit point of the sequence $f\bar{}_t^{(N)};t \quad 0g$ has sample paths that are $a:s:$ solutions of equation (3.16). This implies that there exists a solution to the MFEs.

Convergence of $(\bar{}_t^{(N)};t \quad 0)$

We now study the convergence of $(\bar{}_t^{(N)};t \quad 0)$ in $\mathsf{D}_{\mathsf{M}_1(\mathsf{U})}([0;\,1\,)$ as $N \;\; ! \;\; 1$. Let $(\overline{F}_t^{(N)};t$ $0)$ be the right continuous filtration associated with the process $(\bar{}_t^{(N)};t \quad 0)$. The main logic behind the proof is that we first show the relative compactness of the sequence $f(\bar{}_t^{(N)};t \quad 0)g_{N \;\; 1}$ and then we show every limit point $(\;\; _t;t \quad 0)$ has sample paths evolving almost surely according to the MFEs with the initial point $\bar{}_0$. From the uniqueness of the mean-field solution for a given initial measure, all the limiting points have almost surely identical sample paths coinciding with the unique mean-field solution with the initial point $\bar{}_0$. This implies that the sequence $f(\bar{}_t^{(N)};t \quad 0)g_{N \;\; 1}$ converges in distribution to the unique mean-field solution with the initial point $\bar{}_0$, and this mean-field solution is referred to as the mean-field limit denoted by $(\bar{}_t;t \quad 0)$.

Let $\overline{A}^{(N)}(\ )$ be the generator of the Markov process $(\overline{\ }_t^{(N)}; t \quad 0)$. Then the following process $(\overline{\mathbb{M}}_t^{(N)}(\ ); t \quad 0)$ for $\quad 2\ \mathsf{C}_b^1(\mathsf{U})$ is an RCLL square integrable $\overline{F}_t^{(N)}$-martingale,

$$\overline{\mathbb{M}}_t^{(N)}(\ ) = h\overline{\ }_t^{(N)};\ \ i \quad h\overline{\ }_0^{(N)};\ \ i \quad \int_{s=0}^{t} \overline{A}^{(N)} h\overline{\ }_s^{(N)};\ \ i\, ds: \tag{3.81}$$

By using the expression of $\overline{A}^{(N)}(\ )$, we get

$$\begin{aligned}
\overline{M}_t^{(N)}(\ ) = {} & h\overline{\ }_t^{(N)};\ \ i \quad h\overline{\ }_0^{(N)};\ \ i \quad \int_{s=0}^{t} h\overline{\ }_s^{(N)};\ \ {}^{\mathbb{P}}\, i\, ds \quad \int_{s=0}^{t} \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \right. \\
& (\ (n \quad 1; x_1; \quad ; x_{i\ \ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\ d\overline{\ }_s^{(N)}(n; x_1; : : : ; x_n) \\
& + \left[(\overline{\ }_s^{(N)}(f 0 g)\Phi_0(\overline{\ }_s^{(N)})(\ (1; 0) \quad (0))) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\overline{\ }_s^{(N)}) \right. \\
& (\ (n+1; x_1; \quad ; x_{i\ \ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n)) \\
& \left. \left. d\overline{\ }_s^{(N)}(n; x_1; \quad ; x_n) \right] \right) ds: \quad (3.82)
\end{aligned}$$

For $\quad 2\ \mathsf{C}_b^1(\mathsf{U})$, we get

$$\begin{aligned}
< \overline{\mathbb{M}}^{(N)}(\ ) >_t = {} & \frac{1}{N} \left[\int_{s=0}^{t} \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \right. \right. \\
& (\ (n \quad 1; x_1; \quad ; x_{i\ \ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))^2\ d\overline{\ }_s^{(N)}(n; x_1; \quad ; x_n) \\
& + \left[\left(\overline{\ }_s^{(N)}(f 0 g)\Phi_0(\overline{\ }_s^{(N)})(\ (1; 0) \quad (0))^2 + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\overline{\ }_s^{(N)}) \right. \right. \\
& \left. \left. \left. (\ (n+1; x_1; \quad ; x_{i\ \ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))^2\, d\overline{\ }_s^{(N)}(n; x_1; \quad ; x_n) \right] \right) ds \right]:
\end{aligned} \tag{3.83}$$

From the Prohorov's theorem [57], since the space $\mathsf{D}_{\mathsf{M}_1(\mathsf{U})}([0; 1))$ is complete and separable, proving the relative compactness of $f(\overline{\ }_t^{(N)}; t \quad 0)g_{N\ \ 1}$ is equivalent to showing the tightness of this sequence. To show the tightness, we now recall the Jakubowski's criteria (From Theorem 4:6 of [74]) in Appendix A.3, which give the necessary and sufficient conditions J1 and J2. We first show that the condition J2 is satisfied. For this, we prove that the two sufficient conditions C1 and C2 stated in Appendix A.3 are satisfied. For any $T > 0$, $t\ 2\ [0; T]$, we have

$$h\overline{\ }_t^{(N)};\ \ i \quad k\ \ k_1 h\overline{\ }_t^{(N)}; \mathbf{1} i;$$

and since $h^{-(N)}_t; \mathbf{1}i = 1$, with $b = k \;\; k_1$, the condition C1 is satisfied.

We now focus on the proof of the condition C2. From (3.83) and Doob's inequality given in Theorem A.8, for $\quad > 0$, we get

$$\mathsf{P}\left(\sup_{t \;\; T}\left|\overline{\mathbf{M}}^{(N)}_t(\;\;)\right| \qquad\right) \quad \frac{4}{2}\mathsf{E}\left[< \overline{\mathbf{M}}^{(N)}(\;\;) >_T\right]$$
$$\frac{4T}{2}k \;\; k^2\frac{1}{N}(k \;\; k + d^{(max)2}(d^{(max)}!) \;\;) \;\; ! \;\; 0$$

as $N \;\; ! \;\; 1$. Therefore the sequence $f(\overline{\mathbf{M}}^{(N)}_t(\;\;); t \quad 0)g_{N \;\; 1}$ converges in distribution to the null process from the convergence criterion in $\mathsf{D}_{\mathsf{R}}([0; T])$ [56, Theorem 1.4, p.339]. Furthermore, since the sequence $f(\overline{\mathbf{M}}^{(N)}_t(\;\;); t \quad 0)g_{N \;\; 1}$ is tight in $\mathsf{D}_{\mathsf{R}}([0; T])$, there exists $_1 > 0$ and $N_1 > 0$ such that for $N \quad N_1$, we have

$$\mathsf{P}\left(\sup_{u;v \;\; T;ju \;\; vj \;\; _1}\left|\overline{\mathbf{M}}^{(N)}_v(\;\;) \quad \overline{\mathbf{M}}^{(N)}_u(\;\;)\right| \quad \frac{}{2}\right) \quad \frac{}{2}: \tag{3.84}$$

Also, for any u and v with $u < v \quad T$, from (3.82), we get

$$\left|h^{-(N)}_v; \;\; i \quad h^{-(N)}_u; \;\; i\right| \quad \int_{s=u}^{v}\left|h^{-(N)}_s; \;\; {}^{\mathbb{0}}\mathsf{P}\, i\right| ds + 2k \;\; kk \;\; kju \quad vj$$
$$+ 2Mk \;\; k \;\; (d^{(max)})^3(d^{(max)}!)ju \quad vj + \left|\overline{\mathbf{M}}^{(N)}_v(\;\;) \quad \overline{\mathbf{M}}^{(N)}_u(\;\;)\right|:$$

The above equation can be simplified as

$$\left|h^{-(N)}_v; \;\; i \quad h^{-(N)}_u; \;\; i\right| \quad jv \quad uj k \;\; k_1(1 + 2Mk \;\; k + 2(d^{(max)})^3(d^{(max)}!) \;\;)$$
$$+ \left|\overline{\mathbf{M}}^{(N)}_v(\;\;) \quad \overline{\mathbf{M}}^{(N)}_u(\;\;)\right|: \tag{3.85}$$

By using (3.84) and (3.85), we can find $_2 > 0$ and $N_2 > 0$ such that for $N \quad N_2$, we have

$$\mathsf{P}\left(\sup_{u;v \;\; T;ju \;\; vj \;\; _2}\left|h^{-(N)}_v; \;\; i \quad h^{-(N)}_u; \;\; i\right| \qquad\right) \quad :$$

This completes the proof of the condition J2.

We now focus on the proof of the condition J1. Let $(\mathbf{n}_i(t); \mathbf{V}_{i1}(t); \quad ; \mathbf{V}_{i\mathbf{n}_i(t)}(t))$ be the state of the i^{th} server where $\mathbf{V}_{ij}(t)$ is the age of the j^{th} progressing job. Then

$$h^{-(N)}_t; \;\; i = \frac{1}{N}\sum_{i=1; \mathbf{n}_i(t)>0}^{N}(\mathbf{V}_{i1}(t) + \quad + \mathbf{V}_{i\mathbf{n}_i(t)}(t)):$$

Let us consider a random variable $\mathbf{H}_t$ to denote the age of a progressing job at time t. Further, let $\mathbf{X}$ be a random variable with the same distribution as the job length distribution G. Then for any b 0, we have

$$\mathsf{P}(\mathbf{H}_t \quad b) \quad \mathsf{P}(\mathbf{X} \quad b):$$

Also, we have

$$h^{-(N)}_t; \Xi i = \sum_{n=0}^{1} \int_{x_1} \quad \int_{x_n} n \, d^{-(N)}_t(n; x_1; \quad ; x_n):$$

At any time t, we can classify the progressing jobs into two classes. The first class of jobs are the ones which are in service starting from the time $t = 0$, and the number of such jobs is upper bounded by its initial value say $\mathbf{W}$. If a job belonging to this class has age a at time $t = 0$, then its age is at most $a + t'$ at time $t = t'$. The second class of jobs are the ones which entered the system after the time $t = 0$, and the number of such jobs is bounded by the total number of arrivals up to time t denoted by $\mathbf{E}^{(N)}(t)$. Hence, we can write

$$\mathsf{P}((h^{-(N)}_t; \Xi i + h^{-(N)}_t; \ i) \quad b)$$
$$\mathsf{P}\left(\left[h^{-(N)}_0; \Xi i + \frac{\mathbf{E}^{(N)}(t)}{N} + h^{-(N)}_0; \ i + t h^{-(N)}_0; \Xi i + \frac{\sum_{j=1}^{\mathbf{E}^{(N)}(t)} \mathbf{X}_j}{N}\right] \quad b\right)$$

where $(\mathbf{X}_j; 1 \quad j \quad \mathbf{E}^{(N)}(t))$ are i.i.d. random variables with the distribution G. We now consider the convergence of $\frac{\mathsf{P}_{j=1}^{\mathbf{E}^{(N)}(t)} \mathbf{X}_j}{N}$ in distribution sense. Since the job arrival process is a Poisson process with rate N , it is the same as the sum of N independent Poisson processes with rate . Hence,

$$\mathbf{E}^{(N)}(t) = \sum_{i=1}^{N} \mathbf{E}_{(i)}(t);$$

where $\mathbf{E}_{(i)}(t)$ $(1 \quad i \quad N)$ is the number of arrivals from the i^{th} Poisson process whose rate is . Therefore

$$\frac{\sum_{j=1}^{\mathbf{E}^{(N)}(t)} \mathbf{X}_j}{N} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{k=1}^{\mathbf{E}_{(i)}(t)} \mathbf{X}_{ik} \right);$$

where $f\mathbf{X}_{ik} g_{i \ 1; k \ 1}$ are i.i.d. random variables with distribution G. From the law of large numbers, we get

$$\frac{\sum_{j=1}^{\mathbf{E}^{(N)}(t)} \mathbf{X}_j}{N} \ ) \ \ \frac{t}{};$$

and

$$\frac{\mathbf{E}^{(N)}(t)}{N} \ ) \quad t:$$

From Assumption 3.4.1, we get

$$h^{-(N)}_0;\ \ i + (t+1)h^{-(N)}_0; \Xi i + \frac{\sum_{j=1}^{\mathbf{E}^{(N)}(t)} \mathbf{X}_j}{N} + \frac{\mathbf{E}^{(N)}(t)}{N}\ \ )\ \ h^{-}_0;\ \ i + (t+1)h^{-}_0; \Xi i + \ \ t + \frac{t}{-}:$$

Also, using Assumption 3.4.1, we can find some constant M_0 such that

$$\liminf_{N\ !\ 1} \mathsf{P}(\max(h^{-(N)}_0; \Xi i; h^{-(N)}_0;\ \ i) < M_0) > 1 \qquad :$$

Let $M_T = M_0(2+T) + 2\ \ T + \frac{2\ T}{}$, then

$$\liminf_{N\ !\ 1} \mathsf{P}(\sup_{t2[0;T]} (h^{-(N)}_t; \Xi i + h^{-(N)}_t;\ \ i) < M_T) > 1 \qquad : \tag{3.86}$$

For given $0 < \ \ < 1$, let

$$\mathsf{W}_{T;\ \ \ ,\ f\ \ 2\ \mathsf{M}_1(\mathsf{U}) : (h\ ; \Xi i + h\ ;\ \ i) < M_T g:$$

For an integer n, let

$$C^{(n)}_k = f(k; y_1; \qquad ; y_k) : y_i\ 2\ \mathsf{R}_+ \text{ and } y_i \quad n \text{ for all } 1 \quad i \quad k g;$$

and let $C^{(n)} = f(0) g\ [\ (\ [^{n}_{k=1} C^{(n)}_k)$ and let $\overline{C^{(n)}}$ be the complement of $C^{(n)}$, then

$$(\overline{C^{(n)}}) \quad \frac{2M_T}{n};$$

and hence,

$$\lim_{n\ !\ 1} \sup_{2\mathsf{W}_{T;}} (\overline{C^{(n)}}) = 0:$$

From Lemma A7.5 of [75], we have that $\mathsf{W}_{T;}$ is relatively compact in $\mathsf{M}_1(\mathsf{U})$. Further, from equation (3.86), we have

$$\liminf_{N\ !\ 1} \mathsf{P}(^{-(N)}_t\ 2\ \mathsf{W}_{T;}\ ; 8\, t\ 2\ [0; T]) > 1 \qquad :$$

Let $\mathsf{K}_{T;}$ be the closure of $\mathsf{W}_{T;}$, then $\mathsf{K}_{T;}$ is a compact set with

$$\liminf_{N\ !\ 1} \mathsf{P}(^{-(N)}_t\ 2\ \mathsf{K}_{T;}\ ; 8\, t\ 2\ [0; T]) \quad 1 \qquad :$$

This proves the condition J1. Hence, we have proved that the sequence of processes $f(^{-(N)}_t; t \quad 0) g_{N\ \ 1}$ is is relatively compact.

Let $(\ \ _t; t \quad 0)$ be the limiting point of a converging subsequence of $f(^{-(N)}_t; t \quad 0) g_{N\ \ 1}$. Then $\ \ _0$ almost surely coincides with $^{-}_0$ from Assumption 3.4.1. The continuous mapping theorem implies that

$$h\ _t;\ \ i = h\ _0;\ \ i + \int_{s=0}^{t} h\ _s;\ \ ^{\mathbb{P}}\, i\, ds \quad \int_{s=0}^{t} \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \right.$$
$$(\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\ d\ _s(n; x_1; \quad ; x_n)$$
$$+ \left[(\ _s(f0g)\Phi_0(\ _s)\,(\ (1; 0) \quad (0))) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ _s) \right.$$
$$\left. \left. (\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\, d\ _s(n; x_1; \quad ; x_n) \right] \right) ds: \quad (3.87)$$

Finally, we prove that the sample paths of $(\ _t; t \quad 0)$ coincide almost surely with the unique mean-field solution with the initial point $^{-}_0$. Since $\mathsf{C}^1_b(\mathsf{U})$ is a separating class of $\mathsf{M}_1(\mathsf{U})$, we have that $(\ _t; t \quad 0)\ 2\ \mathsf{C}_{\mathsf{M}_1(\mathsf{U})}([0; \ 1))$. We then prove that for $\ _0$, any process $(\ _t; t \quad 0)\ 2\ \mathsf{C}_{\mathsf{M}_1(\mathsf{U})}([0; \ 1))$ is a solution to equation (3.87) iff it is a solution to the mean-field equation (3.16). We give a proof of this in Section 3.7.5. Since there exists a unique solution to the mean-field equation for the given initial point $^{-}_0$, from Assumption 3.4.1, all the limiting points have almost surely identical sample paths coinciding with the unique mean-field solution. Therefore, $f(^{-(N)}_t; t \quad 0)g_{N\ 1}$ converges in distribution to the unique mean-field solution denoted by $(^{-}_t; t \quad 0)$.

3.7.5 Evolution of $(h\ _t; \ i; t \quad 0)$ for $\ 2\ \mathsf{C}_b(\mathsf{U})$

In this section, we prove that $(\ _t; t \quad 0)$ is a solution to equation (3.87) if and only if it is a solution to equation (3.16).

We first show that any process $(\ _t; t \quad 0)$ that satisfies (3.87) also satisfies (3.16). For $2\ \mathsf{C}^1_b(\mathsf{U})$, if the integrand in equation (3.87) is a continuous function of s, any real valued process $(h\ _t; \ i; t \quad 0)$ that satisfies (3.87) also satisfies the following differential equation

$$\frac{dh\ _t; \ i}{dt} = h\ _t; \ ^{\mathbb{P}}\, i + \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \right.$$
$$(\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\ d\ _t(n; x_1; \quad ; x_n)$$
$$+ \left[(\ _t(f0g)\Phi_0(\ _t)\,(\ (1; 0) \quad (0))) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ _t) \right.$$
$$\left. \left. (\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\, d\ _t(n; x_1; \quad ; x_n) \right] \right): \quad (3.88)$$

Hence, we need to prove that the two terms on the right side of (3.88) are continuous in t. Since $\quad 2\ C_b^1(U)$, the first term $h\ _t;\ ^{\Phi}\ i$ is a continuous function of t. Let $\tilde{}$ be the function defined as

$$\tilde{}(0) = 0$$

and for $n \quad 1$

$$\tilde{}(n; x_1; \quad ; x_n) = \sum_{i=1}^{n} \left\{ \frac{(x_i)}{n} \right\} ((\ (n \quad 1; x_1; : : : ; x_{i\ 1}; x_{i+1}; : : : ; x_n) \quad (n; x_1; : : : ; x_n)):$$

Then the second term on the right side of (3.88) that corresponds to departures can be written as

$$h\ _t;\ \tilde{}\ i = \sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\}$$
$$(\ (n \quad 1; x_1; \quad ; x_{i\ 1}; x_{i+1}; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\ d\ _t(n; x_1; \quad ; x_n):$$

It can be seen that $\tilde{}\ 2\ C_b(U)$ and hence, $h\ _t;\ \tilde{}\ i$ is a continuous function of t.

The term related to arrivals can be written as below. Let $\ _{(\ t)}$ be the function defined as

$$_{(\ t)}(n; x_1; \quad ; x_n)$$
$$= \left\{ \frac{}{n+1} \right\} \Phi_n(\ _t) \sum_{i=1}^{n+1} (\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n)):$$

Then

$$h\ _t;\ _{(\ t)}i = \left[(\ _t(f0g)\Phi_0(\ _t)\ (\ (1; 0) \quad (0))) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ _t) \right.$$
$$(\ (n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\ d\ _t(n; x_1; \quad ; x_n):$$

Since $\ _{(\ t)}\ 2\ C_b(U)$, $t\ 7!\ h\ _t;\ _{(\ b)}i$ is continuous when b is a fixed-value. The mapping $t\ 7!\ h\ _t;\ _{(\ t)}i$ is continuous if $h\ _{t+b};\ _{(\ t+b)}i\ !\ h\ _t;\ _{(\ t)}i$ as $b\ !\ 0$. We can write

$$\left|h\ _{t+b};\ _{(\ t+b)}i \quad h\ _t;\ _{(\ t)}i\right| \quad \left|h\ _{t+b};\ _{(\ t+b)}i \quad h\ _{t+b};\ _{(\ t)}i\right| + \left|h\ _{t+b};\ _{(\ t)}i \quad h\ _t;\ _{(\ t)}i\right|: \tag{3.89}$$

Since $\ _{(\ t)}\ 2\ C_b(U)$, we get

$$\lim_{b\ !\ 0} \left|h\ _{t+b};\ _{(\ t)}i \quad h\ _t;\ _{(\ t)}i\right| = 0: \tag{3.90}$$

To prove the continuity of the mapping $t \mapsto h_t; {}_{(t)}i$, it remains to be shown that

$$\lim_{b \to 0} \left| h_{t+b}; {}_{(t+b)} \quad {}_{(t)} i \right| = 0.$$

For an integer $L > 0$, let

$$V^{(L)} = \{(n; x_1; \ldots; x_n) \in \cup_n : n \quad 1; x_i \quad L \text{ for all } 1 \quad i \quad n \text{ where } 1 \quad n \quad L\}.$$

Let $\overline{V}^{(L)}$ be the complement of $V^{(L)}$. Then for given $\quad > 0$, there exists $L > 0$ such that

$$h_t; I^n_{\overline{V}^{(L)}} i < \quad . \tag{3.91}$$

Since the mapping $t \mapsto \quad_t$ is continuous, we can find some $r_1 > 0$ such that for all $b \in [\quad \min(t; r_1); r_1]$,

$$h_{t+b}; I^n_{\overline{V}^{(L)}} i < \quad . \tag{3.92}$$

We now show that there exists $r_2 \in (0; r_1)$ such that for $b \in [\quad \min(t; r_2); r_2]$ and $\mathbf{u} \in V^{(L)}$, the following result holds

$$\left| {}_{(t+b)}(\mathbf{u}) \quad {}_{(t)}(\mathbf{u}) \right| < 6 \quad M(d^{(max)})^3 k \quad k \quad . \tag{3.93}$$

For $\mathbf{u} \in V^{(L)}$,

$$\left| {}_{(t+b)}(\mathbf{u}) \quad {}_{(t)}(\mathbf{u}) \right| < 2 \quad k \quad k \max_{0 \quad m \quad L} |\Phi_m(\quad_{t+b}) \quad \Phi_m(\quad_t)| .$$

Let us now obtain bounds on $\max_{0 \quad m \quad L} |\Phi_m(\quad_{t+b}) \quad \Phi_m(\quad_t)|$. We have

$$|\Phi_m(\quad_{t+b}) \quad \Phi_m(\quad_t)| = \left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} \in Y^{[k;i]}} \sum_{j=1}^{i} I_{\{n_j = m\}} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \neq j}^{i} (Q_{n_r}(\quad_{t+b}) \quad Q_{n_r}(\quad_t)) \right| .$$

For each j, we can write

$$\begin{aligned}
&\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} \in Y^{[k;i]}} I_{\{n_j = m\}} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \neq j}^{i} (Q_{n_r}(\quad_{t+b}) \quad Q_{n_r}(\quad_t)) \right| \\
&\qquad \left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} \in Y^{[k;i]}} I_{\{n_j = m\}} q^{[k;i]}(\mathbf{n}; j) (Q_{n_1}(\quad_{t+b}) \quad Q_{n_1}(\quad_t)) \prod_{r=2; r \neq j}^{i} Q_{n_r}(\quad_{t+b}) \right| \\
&+ \left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} \in Y^{[k;i]}} I_{\{n_j = m\}} q^{[k;i]}(\mathbf{n}; j) Q_{n_1}(\quad_t) \left(\prod_{r=2; r \neq j}^{i} (Q_{n_r}(\quad_{t+b}) \quad Q_{n_r}(\quad_t)) \right) \right| .
\end{aligned} \tag{3.94}$$

We can bound the first term on the right side of (3.94) as

$$\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y^{[k;i]}} I_{f n_j = mg} q^{[k;i]}(\mathbf{n}; j)(Q_{n_1}(\ _{t+b}) \quad Q_{n_1}(\ _t)) \prod_{r=2; r \in j}^{i} Q_{n_r}(\ _{t+b}) \right|$$
$$M d^{(max)} \sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j :$$

By similar arguments, we obtain

$$\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y^{[k;i]}} I_{f n_j = mg} q^{[k;i]}(\mathbf{n}; j) Q_{n_1}(\ _t) \left(\prod_{r=2; r \in j}^{i} (Q_{n_r}(\ _{t+b}) \quad Q_{n_r}(\ _t)) \right) \right|$$
$$M d^{(max)} (d^{(max)} \quad 1) \sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j :$$

Hence, for each j, we have

$$\left| \sum_{k=1}^{M} \sum_{i=1}^{d^{(max)}} \sum_{\mathbf{n} 2 Y^{[k;i]}} I_{f n_j = mg} q^{[k;i]}(\mathbf{n}; j) \prod_{r=1; r \in j}^{i} (Q_{n_r}(\ _{t+b}) \quad Q_{n_r}(\ _t)) \right|$$
$$M (d^{(max)})^2 \sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j :$$

As a result, we have

$$j \Phi_m(\ _t) \quad \Phi_m(\ _{t+b}) j \quad M (d^{(max)})^3 \sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j :$$

We now show

$$\lim_{b\ !\ 0} \sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j = 0.$$

We can write

$$\sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j \quad \sum_{0\ i\ L\ 1} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j + \sum_{i\ L} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j$$
$$\sum_{0\ i\ L\ 1} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j + \sum_{i\ L} j Q_i(\ _{t+b}) j + \sum_{i\ L} j Q_i(\ _t) j :$$

From (3.91)-(3.92), for $b\ 2\ [\quad \min(t; r_1); r_1]$, we have

$$\sum_{i\ 0} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j \quad \sum_{0\ i\ L\ 1} j Q_i(\ _{t+b}) \quad Q_i(\ _t) j + \quad + \ :$$

Again, as $\lim_{b \to 0} |Q_i(\ _{t+b}) \quad Q_i(\ _t)| = 0$ for $0 \quad i \quad L \quad 1$, there exists $r_2 \ 2 \ (0; r_1)$ such that for $b \ 2 \ [\ \ \min(t; r_2); r_2]$,

$$0 \sum_{i \ L \ 1} |Q_i(\ _{t+b}) \quad Q_i(\ _t)| \quad :$$

Hence, we have

$$\sum_{i \ 0} |Q_i(\ _{t+b}) \quad Q_i(\ _t)| \quad 3$$

for $b \ 2 \ [\ \ \min(t; r_2); r_2]$. Therefore,

$$0 \max_{m \ L \ 1} |\Phi_m(\ _t) \quad \Phi_m(\ _{t+b})| \quad 3M(d^{(max)})^3 :$$

From (3.91)-(3.93) and for $b \ 2 \ [\ \ \min(t; r_2); r_2]$, since $\ _{(\ _{t+b})} \ 2 \ C_b(U)$, we can write

$$\begin{aligned} |\langle h\ _{t+b};\ _{(\ _{t+b})} \quad _{(\ _t)} i| \quad & 6 \ M(d^{(max)})^3 k \ k \ h\ _{t+b}; I_{f V^{(L)}} g i + 2(d^{(max)})^3 (d^{(max)}!) \ k \ k \\ & 8 \ M(d^{(max)})^3 (d^{(max)}!) k \ k : \end{aligned}$$

By taking $b \ ! \ 0$ and then $\ ! \ 0$ in equation (3.89), the mapping $t \ 7! \ \langle h\ _t;\ _{(\ _t)} i$ is continuous.

We now show that any solution to (3.88) also satisfies an alternative equation. For this, we apply the method of change of variables. For $\ 2 \ C_b^1(U)$ and $r \quad t$, let $\tilde{\ }$ be defined as

$$\tilde{\ }(n; x_1; : : : ; x_n) = \begin{cases} (0) & \text{if } n = 0; \\ \tilde{\ }_{t \ r} \ (n; x_1; \quad ; x_n) & \text{otherwise:} \end{cases} \tag{3.95}$$

Let us consider the change in $\langle h\ _r; \tilde{\ } i$ due to change in 'r'. We can write

$$\frac{d\langle h\ _r; \tilde{\ } i}{dr} = \lim_{h \ ! \ 0} \frac{f\langle h\ _{r+h}; \tilde{\ }_{t \ r} \ i \quad \langle h\ _r; \tilde{\ }_{t \ r} \ ig}{h} + \lim_{h \ ! \ 0} \frac{f\langle h\ _{r+h}; \tilde{\ }_{t \ r} \ h \ i \quad \langle h\ _{r+h}; \tilde{\ }_{t \ r} \ ig}{h} : \tag{3.96}$$

In the above equation, the first term on the right side can be computed by using (3.88) and the second term is equal to $\langle h\ _t; \tilde{\ } \mathbb{P} i$. Hence, we obtain

$$\begin{aligned} \frac{d\langle h\ _r; \tilde{\ } i}{dr} = & \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} \right. \\ & \left. \left(\tilde{\ }(n \quad 1; x_1; : : : ; x_{i \ 1}; x_{i+1}; : : : ; x_n) \quad \tilde{\ }(n; x_1; : : : ; x_n) \right) d\ _r(n; x_1; \quad ; x_n) \right) \\ + & \left[\left(\ _r(f0g) \Phi_0(\ _r) \left(\tilde{\ }(1; 0) \quad \tilde{\ }(0) \right) \right) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \Phi_n(\ _r) \right. \end{aligned}$$

$$\left(\tilde{}(n+1; x_1; \quad ; x_{i\ 1}; 0; x_i; \quad ; x_n) \quad \tilde{}(n; x_1; \quad ; x_n)) \, d_{\ r}(n; x_1; \quad ; x_n) \right] \Bigg): \quad (3.97)$$

By integrating $\frac{dh_{\ r;\tilde{}i}}{dr}$ with respect to r from 0 to t, we obtain (3.16).

For $2\ C_b^1(U)$, we now show that any solution $(h_{\ t};\ i; t \quad 0)$ to (3.16) also satisfies (3.87). For this, it is enough to show that the derivative of $h_{\ t};\ i$ with respect to t exists. Since $2\ C_b^1(U)$, the dominated convergence theorem [69, Theorem 1.34] implies that $\frac{dh_{\ 0;\tilde{}t\ i}}{dt}$ is well defined. We can apply the Leibniz integral rule to prove that the derivative of the second term on the right side of equation (3.16) with respect to t exists. According to this rule, the integrand should be continuous with respect to both the variables r and t. This proof is similar to the proof of the integrand in equation (3.87). We then need to show that the derivative of the integrand with respect to t exists, and further, the differential should be continuous with respect to both the variables r and t. Since $2\ C_b^1(U)$, the dominated convergence theorem implies that the integrand is differentiable and further, it is continuous with respect to r and t. This proof is similar to that of the continuity of the integrand in (3.87). Hence, any solution $(\ _t; t \quad 0)\ 2\ C_{M_1(U)}([0; 1))$ is a solution to (3.87) if and only if it is a solution to (3.16). In equation (3.16), need not be differentiable.

3.7.6 Proof of Lemma 3.2

We prove this result based on induction arguments. Let $S(k; n)$ be the set of all k ordered numbers chosen from $f1; 2; \quad ; ng$. For $y_i\ 2\ R_+$, $i \quad 1$, $\mathbf{b} = (b_1; \quad ; b_k)\ 2\ S(k; n)$ and $= (\ _1; \quad ; \ _k)$, let

$$B^{[k;\ ;n;\mathbf{b}]}(y_1; \quad ; y_k)$$
$$= f(n; x_1; \quad ; x_n) : x_i\ 2\ R_+ \text{ for } i\ 2\ fb_1; \quad ; b_k g \text{ and } y_j \quad x_{b_j} \quad y_j + \ _j; 1 \quad j \quad kg:$$

Further, we define

$$\bar{}_t(B^{[k;\ ;n]}) = \max_{\mathbf{b}2S(k;n)} \sup_{(y_1; \ ;y_k)2R_+^k} \bar{}_t(B^{[k;\ ;n;\mathbf{b}]}(y_1; \quad ; y_k)): \quad (3.98)$$

We now show that for each k, $1 \quad j \quad k$,

$$\lim_{j\ !\ 0} \sum_{n\ \ k} \bar{}_t(B^{[k;\ ;n]}) = 0: \quad (3.99)$$

From (3.99), we obtain that the measure $\bar{}_t$ is absolute continuous w.r.t. the Lebesgue measure and hence, it has a density function. Here, $\sum_n {}_k \bar{}_t(B^{[k;\ ;n]})$ is well-defined since $\bar{}_t(B^{[k;\ ;n]}) \quad \bar{}_t(\mathsf{U}_n)$ and $\sum_n {}_k \bar{}_t(\mathsf{U}_n) \quad 1$.

Since there existence sequence of bounded continuous functions that decrease pointwise to indicators of closed sets, by using monotone convergence theorem, we can write from (3.16) as

$$\begin{aligned}
\bar{}_t(B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)) &= h\bar{}_0; \tilde{}_t I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g} i \\
&\int_{s=0}^{t} \Bigg(\sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \frac{(x_i)}{n} \tilde{}_t \ {}_s I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_n)\, d\bar{}_s(n;x_1;\quad ;x_n) \\
&\qquad + \sum_{j=1}^{n+1} \int_{x_1} \quad \int_{x_{n+1}} \frac{(x_j)}{n+1} \\
&\tilde{}_t \ {}_s I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_{j\ 1};x_{j+1};\quad ;x_{n+1})\, d\bar{}_s(n+1;x_1;\quad ;x_{n+1}) \\
&\int_{x_1} \quad \int_{x_n} \Phi_n(\bar{}_s) \tilde{}_t \ {}_s I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_n)\, d\bar{}_s(n;x_1;\quad ;x_n) \\
&\qquad + \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_{i\ 1}} \int_{x_{i+1}} \quad \int_{x_n} \frac{\Phi_{n\ 1}(\bar{}_s)}{n} \\
&\tilde{}_t \ {}_s I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_{i\ 1};0;x_{i+1};\quad ;x_n) \\
&\qquad d\bar{}_s(n\quad 1;x_1;\quad ;x_{i\ 1};x_{i+1};\quad ;x_n) \Bigg)\, ds; \qquad (3.100)
\end{aligned}$$

We now obtain bounds on every term of the right side of (3.100). By using the fact that

$$\tilde{}_t I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_n) = I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}\left(n;x_1+\frac{t}{n};\quad ;x_n+\frac{t}{n}\right); \qquad (3.101)$$

we obtain the following bounds

$$h\bar{}_0; \tilde{}_t I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g} i \quad \bar{}_0(B^{k;\ ;n}); \qquad (3.102)$$

$$\left\{ \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \frac{(x_i)}{n} \tilde{}_t \ {}_s I_{f_{B^{[k;\ ;n;\mathbf{b}]}(y_1;\quad ;y_k)}g}(n;x_1;\quad ;x_n)\, d\bar{}_s(n;x_1;\quad ;x_n) \right\}$$
$$k \ k\bar{}_s(B^{k;\ ;n}); \qquad (3.103)$$

$$\left\{ \sum_{j=1}^{n+1} \int_{x_1} \quad \int_{x_{n+1}} \frac{(x_j)}{n+1} \right.$$

$$\tilde{}_{t\ \ s} I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g(n; x_1;\ \ \ ;x_{j\ \ 1}; x_{j+1};\ \ \ ;x_{n+1})\, d^{-}_{s}(n+1; x_1;\ \ \ ;x_{n+1}) \Big\}$$
$$k\ \ k^{-}_{s}(B^{k;\ \ ;n+1}); \quad (3.104)$$

and from (3.78) we obtain

$$\int_{x_1} \ \ \int_{x_n} \Phi_n(^{-}_{s})\, \tilde{}_{t\ \ s} I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g(n; x_1;\ \ \ ;x_n)\, d^{-}_{s}(n; x_1;\ \ \ ;x_n)$$
$$M\,(d^{(max)})^3 (d^{(max)}!)^{-}_{s}(B^{k;\ \ ;n}): \quad (3.105)$$

We now obtain bounds on the last term. We have

$$\tilde{}_{t\ \ s} I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g(n; x_1;\ \ \ ;x_{i\ \ 1}; 0; x_{i+1};\ \ \ ;x_n)$$
$$= I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g\left(n; x_1 + \frac{t\ \ s}{n};\ \ \ ;x_{i\ \ 1} + \frac{t\ \ s}{n}; \frac{t\ \ s}{n}; x_{i+1} + \frac{t\ \ s}{n};\ \ \ ;x_n + \frac{t\ \ s}{n}\right): \quad (3.106)$$

We now investigate the range of values of x_i, $i \quad 1$, for which we get

$$\tilde{}_{t\ \ s} I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g(n; x_1;\ \ \ ;x_{i\ \ 1}; 0; x_{i+1};\ \ \ ;x_n) = 1:$$

There are two possibilities: either $i \ 2\ f b_1;\ \ \ ;b_k g$ or $i \not\in f b_1;\ \ \ ;b_k g$. We begin with the first case. Let us assume that $i = b_r$ for some $1 \quad r \quad k$. Then we have

$$y_r \quad \frac{t\ \ s}{n} \quad y_r + \ _r$$

As a result, we conclude $t \quad (n y_r + n\ _r) \quad s \quad t \quad n y_r$. Also, for $j = b_m$ with $j \not= b_r$, we obtain that x_j satisfies $y_m \quad x_{b_m} + \frac{t\ \ s}{n} \quad y_m + \ _m$: For $j \not\in f b_1;\ \ \ ;b_k g$, we have $x_j + \frac{t\ \ s}{n}\ 2\ \mathsf{R}_+$.

In the second case when $i \not\in f b_1;\ \ \ ;b_k g$, we have $\frac{t\ \ s}{n}\ 2\ \mathsf{R}_+$ implying that $s \quad t$. Also, for $j = b_m$ with $1 \quad m \quad k$, we obtain that x_j satisfies $y_m \quad x_{b_m} + \frac{t\ \ s}{n} \quad y_m + \ _m$ and for $j \not\in f b_1;\ \ \ ;b_k g$, we have $x_j + \frac{t\ \ s}{n}\ 2\ \mathsf{R}_+$.

From the above discussion, we can show that

$$\Big\{ \int_{s=0}^{t} \sum_{i=1}^{n} \Big(\int_{x_1} \ \ \int_{x_{i\ \ 1}} \int_{x_{i+1}} \ \ \int_{x_n} \frac{\Phi_{n\ \ 1}(^{-}_{s})}{n}$$
$$\tilde{}_{t\ \ s} I_{f B^{[k;\ \ ;n;\flat]}(y_1;\ \ ;y_k)} g(n; x_1;\ \ \ ;x_{i\ \ 1}; 0; x_{i+1};\ \ \ ;x_n)$$
$$d^{-}_{s}(n \quad 1; x_1;\ \ \ ;x_{i\ \ 1}; x_{i+1};\ \ \ ;x_n) \Big)\, ds \Big\}$$

$$M\,(d^{(max)})^3(d^{(max)}!)\Bigg(\sum_{l=1}^{k}\int_{s=0}^{t}\frac{1}{n}I_{f[t\ \ n(y_l+\ _l);t\ \ ny_l]g}(s)^{-}{}_s\left(B^{[k\ \ 1;\quad {}^{l};n\ \ 1]}\right)\,ds$$
$$+\frac{n\quad k}{n}\int_{s=0}^{t}{}^{-}{}_s\left(B^{[k;\ ;n\ \ 1]}\right)\,ds\Bigg);\quad (3.107)$$

where $^{l}=(\ _1;\quad ;\ _l\ 1;\ _{l+1};\quad ;\ _k)$. We next show the following result by using (3.102)-(3.107) for the case $k=1$ and $\quad=(\ _1)$,

$$\lim_{j\ !\ 0}\sum_{n\ \ 1}{}^{-}{}_t(B^{[1;\ ;n]})=0:\quad (3.108)$$

From (3.102)-(3.107), we obtain

$${}^{-}{}_t(B^{[1;\ ;n;\mathrm{b}]}(y_1))\quad {}^{-}{}_0(B^{[1;\ ;n]})+k\ \ k\int_{s=0}^{t}{}^{-}{}_s(B^{[1;\ ;n}])\,ds+k\ \ k\int_{s=0}^{t}{}^{-}{}_s(B^{[1;\ ;n+1]})\,ds$$
$$+\quad M\,(d^{(max)})^3(d^{(max)}!)\int_{s=0}^{t}{}^{-}{}_s(B^{[1;\ ;n]})\,ds$$
$$+\quad M\,(d^{(max)})^3(d^{(max)}!)\Bigg(\sum_{l=1}^{1}\int_{s=0}^{t}\frac{1}{n}I_{f[t\ \ n(y_l+\ _l);t\ \ ny_l]g}(s)^{-}{}_s\,(\mathsf{U}_{n\ \ 1})\,ds$$
$$+\frac{n\quad 1}{n}\int_{s=0}^{t}{}^{-}{}_s\left(B^{[1;\ ;n\ \ 1]}\right)\,ds\Bigg):\quad (3.109)$$

We can also write

$$\int_{s=0}^{t}\frac{1}{n}I_{f[t\ \ n(y_l+\ _l);t\ \ ny_l]g}(s)^{-}{}_s\,(\mathsf{U}_{n\ \ 1})\,ds\quad \int_{s=0}^{t}\frac{1}{n}I_{f[t\ \ (y_l+\ _l);t\ \ y_l]g}(s)^{-}{}_s\,(\mathsf{U}_{n\ \ 1})\,ds:\quad (3.110)$$

From (3.109)-(3.110), we get

$${}^{-}{}_t(B^{[1;\ ;n]})\quad {}^{-}{}_0(B^{[1;\ ;n]})+k\ \ k\int_{s=0}^{t}({}^{-}{}_s(B^{[1;\ ;n}])+{}^{-}{}_s(B^{[1;\ ;n+1]}))\,ds$$
$$+\quad M\,(d^{(max)})^3(d^{(max)}!)\Bigg(\int_{s=0}^{t}{}^{-}{}_s(B^{[1;\ ;n]})\,ds$$
$$+\left(\sup_{y_1}\sum_{l=1}^{1}\int_{s=0}^{t}I_{f[t\ \ (y_l+\ _l);t\ \ y_l]g}(s)^{-}{}_s\,(\mathsf{U}_{n\ \ 1})\,ds+\int_{s=0}^{t}{}^{-}{}_s\left(B^{[1;\ ;n\ \ 1]}\right)\,ds\right)\Bigg):\quad (3.111)$$

From (3.111), by summing over $n\quad 1$, we obtain

$$\left(\sum_{n\ \ 1}{}^{-}{}_t(B^{[1;\ ;n]})\right)\quad \left(\sum_{n\ \ 1}{}^{-}{}_0(B^{[1;\ ;n]})\right)+2k\ \ k\int_{s=0}^{t}\left(\sum_{n\ \ 1}{}^{-}{}_s(B^{[1;\ ;n]})\right)\,ds$$
$$+2\ \ M\,(d^{(max)})^3(d^{(max)}!)\Bigg(\int_{s=0}^{t}\left(\sum_{n\ \ 1}{}^{-}{}_s(B^{[1;\ ;n]})\right)\,ds+\ _1\Bigg):\quad (3.112)$$

By using Gronwall's inequality, we have

$$\left(\sum_{n=1}^{} \bar{}_t(B^{[1;\ ;n]})\right) \quad \left(\left(\sum_{n=1}^{} \bar{}_0(B^{[1;\ ;n]})\right) + 2\ \ M\,(d^{(max)})^3(d^{(max)}!)\ \ {}_1\right) e^{(2k\ \ k+2\ \ M\,(d^{(max)})^3(d^{(max)}!))t}: \tag{3.113}$$

Since the measure $\bar{}_0$ has a density function, we have

$$\lim_{{}_1\ !\ 0} \left(\sum_{n=1}^{} \bar{}_t(B^{[1;\ ;n]})\right) = 0: \tag{3.114}$$

We now use induction arguments. Suppose, for $= [\ {}_1;\quad ;\ {}_m]$, we have

$$\lim_{r\ !\ 0} \left(\sum_{n\ \ m}^{} \bar{}_t(B^{[m;\ ;n]})\right) = 0 \text{ and } \left(\sum_{n\ \ m}^{} \bar{}_t(B^{[m;\ ;n]})\right) \quad \hat{z}_m e^{\hat{\nu}_m t}; \tag{3.115}$$

for all $1 \quad r \quad m$ and $\hat{z}_m; \hat{\nu}_m \quad 0$. Then we will show that for all $1 \quad r \quad m+1$, there exists $\hat{z}_{m+1}; \hat{\nu}_{m+1} \quad 0$ such that

$$\lim_{r\ !\ 0} \left(\sum_{n\ \ m+1}^{} \bar{}_t(B^{[m+1;\ ;n]})\right) = 0 \text{ and } \left(\sum_{n\ \ m+1}^{} \bar{}_t(B^{[m+1;\ ;n]})\right) \quad \hat{z}_{m+1} e^{\hat{\nu}_{m+1} t}: \tag{3.116}$$

From (3.102)-(3.107), we obtain

$$\begin{aligned} f \sum_{n\ \ m+1}^{} \bar{}_t(B^{[m+1;\ ;n]})g \quad & f \sum_{n\ \ m+1}^{} \bar{}_0(B^{[m+1;\ ;n]})g + 2k\ \ k \int_{s=0}^{t} f \sum_{n\ \ m+1}^{} \bar{}_s(B^{[m+1;\ ;n]})g\,ds \\ & + 2\ \ M\,(d^{(max)})^3(d^{(max)}!) \int_{s=0}^{t} f \sum_{n\ \ m+1}^{} \bar{}_s(B^{[m+1;\ ;n]})g\,ds \\ + \ M\,(d^{(max)})^3(d^{(max)}!) & \sup_{y_1;\ \ ;y_{m+1}} \left(\sum_{l=1}^{m+1} \int_{s=0}^{t} I_{f[t\ \ (y_l+\ {}_l);t\ \ y_l]g}(s) \left\{\sum_{n\ \ m+1}^{} \bar{}_s\left(B^{[m;\quad {}^l;n\ \ 1]}\right)\right\} ds\right): \end{aligned} \tag{3.117}$$

We can also write

$$\begin{aligned} f \sum_{n\ \ m+1}^{} \bar{}_t(B^{[m+1;\ ;n]})g \quad & f \sum_{n\ \ m+1}^{} \bar{}_0(B^{[m+1;\ ;n]})g + 2k\ \ k \int_{s=0}^{t} f \sum_{n\ \ m+1}^{} \bar{}_s(B^{[m+1;\ ;n]})g\,ds \\ & + 2\ \ M\,(d^{(max)})^3(d^{(max)}!) \int_{s=0}^{t} f \sum_{n\ \ m+1}^{} \bar{}_s(B^{[m+1;\ ;n]})g\,ds \\ & + \ M\,(d^{(max)})^3(d^{(max)}!) \left(\sum_{l=1}^{m+1} \left\{\sup_{0\ \ s\ \ t} \sum_{n\ \ m+1}^{} \bar{}_s\left(B^{[m;\quad {}^l;n\ \ 1]}\right)\right\}\ \ {}_l\right): \end{aligned} \tag{3.118}$$

Now by using Gronwall's inequality, we have

$$\left\{ \sum_{n \quad m+1} {}_{-t}(B^{[m+1;\ ;n]}) \right\}$$
$$\left\{ \left(\sum_{n \quad m+1} {}_{-0}(B^{[m+1;\ ;n]}) \right) + 2 \quad M\,(d^{(max)})^3 (d^{(max)}!) \right.$$
$$\left. \left(\sum_{l=1}^{m+1} \left\{ \sup_{0 \quad s \quad t} {}_{n} \sum_{m+1} {}_{-s} \left(B^{[m;\quad {}^{l};n \quad 1]} \right) \right\} \quad {}_{l} \right) \right\} e^{(2k \quad k+2 \quad M\,(d^{(max)})^3 (d^{(max)}!))t}. \quad (3.119)$$

From (3.115), we conclude that for $1 \quad r \quad m+1$,

$$\lim_{r \ ! \ 0} \left(\sum_{n \quad m+1} {}_{-t}(B^{[m+1;\ ;n]}) \right) = 0. \quad (3.120)$$

This concludes the proof.

3.7.7 Single Server System with State-dependent Arrival Rates

Let us consider a system that has a single PS server in which jobs arrive according to a Poisson process with rate $\ _n$ when there are n progressing jobs. The job length distribution is $G(\ )$ which is the same as in the system model of our multi-server system. Let $\ _t^{(single)}$ be the probability measure of the server state at time t, then we have the following Kolmogorov equations, for $\quad 2\ \mathsf{C}_b^1(\mathsf{U})$,

$$h\ {}_t^{(single)};\ \ i = h\ {}_0^{(single)};\ \ i + \int_{s=0}^{t} h\ {}_s^{(single)};\ \ {}^{\mathbb{0}\mathsf{P}}\, i\, ds$$
$$\int_{s=0}^{t} \left(\sum_{n=1}^{1} \sum_{i=1}^{n} \int_{x_1} \quad \int_{x_n} \left\{ \frac{(x_i)}{n} \right\} (\ (n \quad 1; x_1; \quad ; x_{i\ \ 1}; x_{i+1}; \quad ; x_n) \right.$$
$$(n; x_1; \quad ; x_n))\ d\ {}_s^{(single)}(n; x_1; \quad ; x_n)$$
$$\left. + \left[(\ \ {}_0\ {}_s^{(single)}(f0g)\,(\ \ (1;0) \quad (0))) + \sum_{n=1}^{1} \sum_{i=1}^{n+1} \int_{x_1} \quad \int_{x_n} \left\{ \frac{1}{n+1} \right\} \right.\right.$$
$$\left.\left. {}_n(\ \ (n+1; x_1; \quad ; x_{i\ \ 1}; 0; x_i; \quad ; x_n) \quad (n; x_1; \quad ; x_n))\, d\ {}_s^{(single)}(n; x_1; \quad ; x_n) \right] \right) ds. \quad (3.121)$$

Let $p_t^{(single)}(\mathbf{u})$ be the density of the measure $\ _t^{(single)}$ at $\mathbf{u}\ 2\ \mathsf{U}$. Then we obtain the differential equations satisfied by the density function $p_t^{(single)} = (p_t^{(single)}(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$

by using the same procedure as in [35, 67, 68, 81]. Let us define a process $P_t^{(single)} = (P_t^{(single)}(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ where $P_t^{(single)}(0)$ is equal to $_t^{(single)}(f0g)$ and

$$P_t^{(single)}(n; y_1; \quad ; y_n) = \int_{x_1=0}^{y_1} \quad \int_{x_n=0}^{y_n} p_t^{(single)}(n; x_1; \quad ; x_n)\, dx_1 \quad dx_n:$$

Then we get the following differential equations

$$\frac{dP_t^{(single)}(0)}{dt} = \int_{y=0}^{1} \quad (y) \left(\frac{@P_t^{(single)}(1; y)}{@y} \right) dy \quad {}_0P_t^{(single)}(0); \tag{3.122}$$

for $n \quad 1$,

$$\begin{aligned} \frac{dP_t^{(single)}(n; y_1; \quad ; y_n)}{dt} = & \quad \sum_{i=1}^{n} \frac{1}{n} \frac{@P_t^{(single)}(n; y_1; : : : ; y_n)}{@y_i} \\ & + \sum_{i=1}^{n+1} \int_{x_i=0}^{1} \frac{(x_i)}{n+1} \left(\frac{@P_t^{(single)}(n+1; y_1; \quad ; y_{i\ 1}; x_i; y_i; \quad ; y_n)}{@x_i} \right) dx_i \\ & \sum_{i=1}^{n} \int_{x_i=0}^{y_i} \frac{(x_i)}{n} \left(\frac{@P_t^{(single)}(n; y_1; \quad ; y_{i\ 1}; x_i; y_{i+1}; \quad ; y_n)}{@x_i} \right) dx_i \\ & + \sum_{i=1}^{n} \left(\frac{n\ 1}{n} \right) P_t^{(single)}(n \quad 1; y_1; \quad ; y_{i\ 1}; y_{i+1}; \quad ; y_n) \\ & {}_nP_t^{(single)}(n; y_1; \quad ; y_n): \end{aligned} \tag{3.123}$$

From [22], the single server PS system with state-dependent arrival rate $_i$ when there are i jobs in progress and job lengths are generally distributed with finite mean $\frac{1}{}$, has a unique stationary distribution $^{(single)} = (\ ^{(single)}(\mathbf{u}); \mathbf{u}\ 2\ \mathsf{U})$ given by,

$$^{(single)}(n; y_1; \quad ; y_n) = \frac{\left(\prod_{i=1}^{n} \frac{i\ 1}{} \right)}{1 + \sum_{m=1}^{1} \left(\prod_{i=1}^{m} \frac{i\ 1}{} \right)} \quad {}^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i; \tag{3.124}$$

and

$$^{(single)}(0) = \frac{1}{1 + \sum_{m=1}^{1} \left(\prod_{i=1}^{m} \frac{i\ 1}{} \right)}: \tag{3.125}$$

3.7.8 Proof of Theorem 3.4

To prove this result, we work with the mean-field of the tail occupancy process. The proof uses the result that the mean-field is quasi-monotonic for Policies 2 and 4. We show that the mean-field is GAS and we skip the rest of the details as they follow by similar

arguments of [11, Section 2.6]. We use the spaces $\mathsf{U}^{\emptyset}$ and U defined in (1.23) and (1.24), respectively. We equip the space $\mathsf{U}^{\emptyset}$ with the metric $!(\ ;\ )$ defined in (1.25). The space $\mathsf{U}^{\emptyset}$ equipped with the metric $!$ is compact, complete, and separable. For the stochastic system with parameter N, let $(\mathbf{X}_i^{(N)}(t); i \quad 0; t \quad 0)$ be the empirical process where $\mathbf{X}_i^{(N)}(t)$ denotes the fraction of servers with at least i progressing jobs. Then if $\mathbf{X}^{(N)}(0) \ ) \ \mathbf{u}$, then $(\mathbf{X}^{(N)}(t); t \quad 0) \ ) \ (\mathbf{x}(t;\mathbf{u}); t \quad 0)$ which is the unique solution to the following equations,

$$\mathbf{x}(0;\mathbf{u}) = \mathbf{u} \tag{3.126}$$

$$\frac{dx_i(t;\mathbf{u})}{dt} = F_i(\mathbf{x}(t;(u))) \tag{3.127}$$

$$= (x_{i\ 1}(t;\mathbf{u}) \quad x_i(t;\mathbf{u}))\ \Phi_{i\ 1}(\mathbf{x}(t;\mathbf{u})) \quad x_i(t;\mathbf{u}) + \ x_{i+1}(t;\mathbf{u}); \tag{3.128}$$

where $\Phi_i(\mathbf{x}(t;\mathbf{u}))$ is the arrival rate function given in (3.11) with $Q_j(\frac{}{N})$ replaced by $x_j(t;\mathbf{u}) \quad x_{j+1}(t;\mathbf{u})$. Let $\mathbf{F}(\mathbf{x}(t;\mathbf{u})) = (F_i(\mathbf{x}(t;\mathbf{u}); i \quad 1)$. Then any fixed-point $=$ $(\ _i; i \quad 0)$ of the mean-field satisfies $\mathbf{F}(\ ) = 0$.

We show that $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ for all $\mathbf{u}\ 2\ \mathsf{U}$. The following result is the key step to conclude that the asymptotic stationary distribution of a server coincides with the unique fixed-point of the mean-field.

Proposition 3.1. *For Policies 2 and 4, the mean- eld is quasi-monotonic. Furthermore,* $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ *for all* $\mathbf{u}\ 2\ \mathsf{U}$.

Proof. For Policy 2 and Policy 4, the mean-field equations are

$$\frac{dx_i(t;\mathbf{u})}{dt} = F_i(\mathbf{x}(t;\mathbf{u})); \tag{3.129}$$

where we have for $1 \quad i \quad +1$,

$$F_i(\mathbf{x}(t;\mathbf{u})) = \frac{(1 \quad x^d{}_{+1}(t;\mathbf{u}))}{(1 \quad x\ _{+1}(t;\mathbf{u}))}(x_{i\ 1}(t;\mathbf{u}) \quad x_i(t;\mathbf{u})) \quad x_i(t;\mathbf{u}) + \ x_{i+1}(t;\mathbf{u}); \tag{3.130}$$

and for $i \quad +1$,

$$F_i(\mathbf{x}(t;\mathbf{u})) = \ (x_{i\ 1}^d(t;\mathbf{u}) \quad x_i^d(t;\mathbf{u})) \quad x_i(t;\mathbf{u}) + \ x_{i+1}(t;\mathbf{u}): \tag{3.131}$$

Since $F_i(\ )$ is non-decreasing in $x_j(t;\mathbf{u})$ for $i \neq j$, the mean-field is quasi-monotone from [82, p. 70-74]. Hence if $\mathbf{u} \quad \mathbf{v}$ by element wise, then $\mathbf{x}(t;\mathbf{u}) \quad \mathbf{x}(t;\mathbf{v})$ by element wise. As a result, we have $\mathbf{x}(t;\min(\mathbf{u};\ )) \quad \mathbf{x}(t;\mathbf{u}) \quad \mathbf{x}(t;\max(\mathbf{u};\ ))$. Therefore, to show that $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ for all $\mathbf{u}\ 2\ \mathsf{U}$, it is sufficient to show $\lim_{t\ !\ 1} \mathbf{x}(t;\mathbf{u}) =$ for all $\mathbf{u}$ and for all $\mathbf{u}$.

Let us now define $z_n(t;\mathbf{u}) = \sum_{k \quad n} x_k(t;\mathbf{u})$ and $z_n(\mathbf{u}) = \sum_{k \quad n} u_k$. If $\mathbf{u} \; 2 \; \mathsf{U}$, then $\mathbf{x}(t;\mathbf{u}) \; 2 \; \mathsf{U}$ and further, we have for $1 \quad n \quad +1$,

$$\frac{dz_n(t;\mathbf{u})}{dt} = \frac{(1 \quad x^d{}_{+1}(t;\mathbf{u}))}{(1 \quad x_{\;+1}(t;\mathbf{u}))}(x_{n\;\;1}(t;\mathbf{u}) \quad x_{\;+1}(t;\mathbf{u})) + x^d{}_{+1}(t;\mathbf{u}) \quad x_n(t;\mathbf{u}); \quad (3.132)$$

and for $n \quad +2$,

$$\frac{dz_n(t;\mathbf{u})}{dt} = x^d_{n\;\;1}(t;\mathbf{u}) \quad x_n(t;\mathbf{u}): \quad (3.133)$$

The fixed-point satisfies the following equations, for $1 \quad n \quad +1$,

$$\frac{(1 \quad {}^d{}_{+1})}{(1 \quad {}_{+1})}(\;{}_{n\;\;1} \quad {}_{+1}) + {}^d{}_{+1} \quad {}_n = 0; \quad (3.134)$$

and for $n \quad +2$, we have

$${}^d_{n\;\;1} \quad {}_n = 0: \quad (3.135)$$

We first show that $z_n(t;\mathbf{u})$ is bounded uniformly in t for $n \quad 1$. Note that

$$\frac{dz_1(t;\mathbf{u})}{dt} = \quad x_1(t;\mathbf{u}):$$

If $\mathbf{u}$, then $\mathbf{x}(t;\mathbf{u}) \quad \mathbf{x}(t;\;)$ and hence, $z_1(t;\mathbf{u}) \quad z_1(\;)$. Since $z_n(t;\mathbf{u}) \quad z_{n+1}(t;\mathbf{u})$, we have that $z_n(t;\mathbf{u})$ is uniformly bounded in t. On the other hand, if $\mathbf{u}$, then $\mathbf{x}(t;\mathbf{u}) \quad \mathbf{x}(t;\;)$ and hence,

$$\frac{dz_1(t;\mathbf{u})}{dt} \quad {}_1 = 0:$$

This implies $z_1(t;\mathbf{u}) \quad z_1(\mathbf{u})$ and hence, $z_n(t;\mathbf{u})$ for $n \quad 1$ is uniformly bounded in t. When $\mathbf{u}$, since the derivative of $x_n(t;\mathbf{u})$ is bounded for all $n \quad 1$, then $\lim_{t \; ! \; 1} x(t;\mathbf{u}) =$ follows if we show for all $n \quad 1$,

$$\int_{t=0}^{1} (x_n(t;\mathbf{u}) \quad {}_n)\, dt < 1:$$

Similarly, for the case of $\mathbf{u}$, it is sufficient to show

$$\int_{t=0}^{1} (\;{}_n \quad x_n(t;\mathbf{u}))\, dt < 1$$

for all $n \quad 1$. We only give a proof for the case when $\mathbf{u}$ and the proof for the other case is similar.

We first begin with the proof of

$$\int_{t=0}^{1} (x_1(t;\mathbf{u}) \quad {}_1)\, dt < 1:$$

From (3.134) and (3.132), we have

$$\frac{dz_1(t;\mathbf{u})}{dt} = \quad (x_1(t;\mathbf{u}) \quad {}_1):$$

Then we have

$$\int_{t=0}^{T} (x_1(t;\mathbf{u}) \quad {}_1)\,dt = \int_{t=0}^{T} \frac{dz_1(t;\mathbf{u})}{dt}\,dt:$$

This implies

$$\int_{t=0}^{T} (x_1(t;\mathbf{u}) \quad {}_1)\,dt = z_1(\mathbf{u}) \quad z_1(T;\mathbf{u}):$$

Since $z_1(t;\mathbf{u})$ is uniformly bounded in t, we get

$$\int_{t=0}^{1} (x_1(t;\mathbf{u}) \quad {}_1)\,dt < 1$$

as $T \; ! \; 1$.

We now focus on the proof of

$$\int_{t=0}^{1} (x_n(t;\mathbf{u}) \quad {}_n)\,dt < 1$$

for $2 \quad n \quad +1$. We prove this by using the induction method. For $2 \quad K \quad +1$, let

$$\int_{t=0}^{1} (x_i(t;\mathbf{u}) \quad {}_i)\,dt < 1$$

is true for $1 \quad i \quad K \quad 1$, then we prove that

$$\int_{t=0}^{1} (x_K(t;\mathbf{u}) \quad {}_K)\,dt < 1:$$

From (3.132), we have

$$\frac{dz_1(t;\mathbf{u})}{dt} \quad \frac{dz_K(t;\mathbf{u})}{dt} = \frac{(1 \quad x^d{}_{+1}(t;\mathbf{u}))}{(1 \quad x \; {}_{+1}(t;\mathbf{u}))}(1 \quad x_K \; {}_1(t;\mathbf{u})) \quad x_1(t;\mathbf{u}) + \quad x_K(t;\mathbf{u}):$$

Then from (3.134), we have

$$\frac{dz_1(t;\mathbf{u})}{dt} \quad \frac{dz_K(t;\mathbf{u})}{dt} = \frac{(1 \quad x^d{}_{+1}(t;\mathbf{u}))}{(1 \quad x \; {}_{+1}(t;\mathbf{u}))}(1 \quad x_K \; {}_1(t;\mathbf{u})) \quad \frac{(1 \quad {}^d{}_{+1})}{(1 \quad {}_{+1})}(1 \quad {}_K \; {}_1)$$
$$x_1(t;\mathbf{u}) + \quad {}_1 + \quad x_K(t;\mathbf{u}) \quad {}_K:$$

Therefore, we obtain

$$\int_{t=0}^{T} (\; x_K(t;\mathbf{u}) \quad {}_K)\,dt = \int_{t=0}^{T} \Big(\frac{dz_1(t;\mathbf{u})}{dt} \quad \frac{dz_K(t;\mathbf{u})}{dt} \quad \frac{(1 \quad x^d{}_{+1}(t;\mathbf{u}))}{(1 \quad x \; {}_{+1}(t;\mathbf{u}))}(1 \quad x_K \; {}_1(t;\mathbf{u}))$$

$$+ \frac{(1 \quad {}^{d}{}_{+1})}{(1 \quad {}_{+1})}(1 \quad {}_{K\ 1}) + \quad x_1(t;\mathbf{u}) \quad {}_{1}\Big)\,dt:$$

Since $x_{\ +1}(t;\mathbf{u}) \quad {}_{+1}$, we have

$$\int_{t=0}^{T}(\ x_K(t;\mathbf{u}) \quad {}_{K})\,dt = \int_{t=0}^{T}\Big(\frac{dz_1(t;\mathbf{u})}{dt} \quad \frac{dz_K(t;\mathbf{u})}{dt} \quad \frac{(1 \quad x^{d}{}_{+1}(t;\mathbf{u}))}{(1 \quad x_{\ +1}(t;\mathbf{u}))}(1 \quad x_{K\ 1}(t;\mathbf{u}))$$
$$+ \frac{(1 \quad x^{d}{}_{+1}(t;\mathbf{u}))}{(1 \quad x_{\ +1}(t;\mathbf{u}))}(1 \quad {}_{K\ 1}) + \quad x_1(t;\mathbf{u}) \quad {}_{1}\Big)\,dt:$$

Further, as $\frac{(1 \quad x^{d}{}_{+1}(t;\mathbf{u}))}{(1 \quad x_{\ +1}(t;\mathbf{u}))} \quad d$, we get

$$\int_{t=0}^{T}(\ x_K(t;\mathbf{u}) \quad {}_{K})\,dt \quad \int_{t=0}^{T}\Big(\frac{dz_1(t;\mathbf{u})}{dt} \quad \frac{dz_K(t;\mathbf{u})}{dt}$$
$$+ d\ (x_{K\ 1}(t;\mathbf{u}) \quad {}_{K\ 1}) + \quad x_1(t;\mathbf{u}) \quad {}_{1}\Big)\,dt:$$

From the induction hypothesis and since $z_n(t;\mathbf{u})$ is bounded uniformly in t for $n \quad 1$, we get

$$\int_{t=0}^{1}(x_K(t;\mathbf{u}) \quad {}_{K})\,dt < 1$$

as $T \ ! \ 1$.

Finally, we show

$$\int_{t=0}^{1}(x_n(t;\mathbf{u}) \quad {}_{n})\,dt < 1$$

for $n \quad +2$. We prove this by using the induction method. For $K \quad +2$, let

$$\int_{t=0}^{1}(x_i(t;\mathbf{u}) \quad {}_{i})\,dt < 1$$

is true for $1 \quad i \quad K \quad 1$, then we prove that

$$\int_{t=0}^{1}(x_K(t;\mathbf{u}) \quad {}_{K})\,dt < 1:$$

From (3.133), we have

$$\int_{t=0}^{T} \quad (x_K(t;\mathbf{u}) \quad {}_{K})\,dt = \int_{t=0}^{T}\Big(\quad \frac{dz_K(t;\mathbf{u})}{dt} + \quad (x^{d}_{K\ 1}(t;\mathbf{u}) \quad {}^{d}_{K\ 1})\Big)\,dt:$$

From the induction hypothesis and since $z_K(t;\mathbf{u})$ is bounded uniformly in t, we get $\int_{t=0}^{1} \quad (x_K(t;\mathbf{u}) \quad {}_{K})\,dt < 1$ as $T \ ! \ 1$. This completes the proof.

3.7.9 Proof of Theorem 3.5

The proof mainly uses the logic that in the stationary regime, the servers' states are exchangeable random variables. We can write

$$\left|\mathsf{E}\left[\prod_{i=1}^{I} {}_i\left(\mathbf{S}_i^{(N_k)}(\ 1\ )\right)\right] \quad \mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}(\ 1\ );\ {}_i i\right]\right|$$
$$\left|\mathsf{E}\left[\prod_{i=1}^{I} {}_i\left(\mathbf{S}_i^{(N_k)}(\ 1\ )\right)\right] \quad \mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}^{(N_k)}(\ 1\ );\ {}_i i\right]\right|$$
$$+\left|\mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}^{(N_k)}(\ 1\ );\ {}_i i\right] \quad \mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}(\ 1\ );\ {}_i i\right]\right|: \quad (3.136)$$

Note that since $\mathbf{V}^{(N_k)}(\ 1\ )\ )\ \mathbf{V}(\ 1\ )$, the second term on the right hand side of the above inequality vanishes as $N_k\ !\ \ 1$. Now, due to exchangeability, the permutation of states between servers does not affect the joint distribution. Hence, we have

$$\mathsf{E}\left[\prod_{i=1}^{I} {}_i\left(\mathbf{S}_i^{(N_k)}(\ 1\ )\right)\right] = \frac{1}{(N_k)_I}\mathsf{E}\left[\sum_{2Q(I;N_k)}\prod_{i=1}^{I} {}_i\left(\mathbf{S}_{(i)}^{(N_k)}(\ 1\ )\right)\right];$$

where $(N)_j = N(N \quad 1):::(N \quad j+1)$, and $Q(r;n)$ denotes the set of all permutations of the numbers $f1;2;:::;ng$ taken r at a time. Also, by the definition of $\mathbf{V}^{(N_k)}(\ 1\ )$, we have

$$\mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}^{(N_k)}(\ 1\ );\ {}_i i\right] = \mathsf{E}\left[\left(\prod_{i=1}^{I}\frac{1}{N_k}\sum_{j=1}^{N_k} {}_i\left(\mathbf{S}_j^{(N_k)}(\ 1\ )\right)\right)\right]:$$

Hence, the first term on the right hand side of (3.136) can be bounded as follows

$$\left|\mathsf{E}\left[\prod_{i=1}^{I} {}_i\left(\mathbf{S}_i^{(N_k)}(\ 1\ )\right)\right] \quad \mathsf{E}\left[\prod_{i=1}^{I} h\mathbf{V}^{(N_k)}(\ 1\ );\ {}_i i\right]\right| \quad 2B^I\left(1 \quad \frac{(N_k)_I}{(N_k)^I}\right);$$

where $\max_i k\ {}_i k = B$. The result follows since $\left(1 \quad \frac{(N_k)_I}{(N_k)^I}\right)\ !\ 0$ as $N_k\ !\ \ 1$.

Finally, from (3.52), any finite set of servers are independent of each other iff Z is a Dirac measure. Otherwise they are coupled through the sample value of the random variable $\mathbf{V}(\ 1\ )$. If $Z = \quad$, then it implies that in the limiting system, the stationary empirical random variable $\mathbf{V}(\ 1\ )$ is a deterministic value coinciding with $\quad$. Then the following equation concludes that each server has distribution

$$\mathsf{E}\left[\prod_{i=1}^{I} {}_i(\mathbf{S}_i^{(N_k)}(\ 1\ ))\right]\ !\ \prod_{i=1}^{I} h\ ;\ {}_i i$$

as $N_k\ !\ \ 1$. This completes the proof.

3.8 Conclusions

In this chapter, we studied occupancy dependent policies that subsume threshold based policies and policies that are adaptive to variations in the system load using mean-field techniques. Our results conclude that the fixed-point of the mean-field is insensitive for any occupancy based policy as long as the mean-field when job lengths are exponentially distributed has a unique fixed-point. For general JLDs, when the fixed-point of the mean-field is unique, it is of great interest to show that the stationary distribution of a server as $N \to 1$ coincides with the unique fixed-point of the mean-field. This result seems to be true as it is observed in our simulation results for the policies studied in this chapter.

Chapter 4

A Functional Central Limit Theorem for Multi-Server Erlang Loss Systems Under SQ(d) Load Balancing Policy

In this chapter, we return to the multi-server Erlang loss system that has N servers, each server can serve only a finite number of C jobs simultaneously. We assume that the dispatcher employs the SQ(d) load balancing policy, according to which the destination server for an incoming job is the server with the least occupancy among d randomly sampled servers. The job arrival process is a Poisson process with rate $N\ ^{(N)}$, where $^{(N)}\ 2\ \mathsf{R}_+$.

The main contribution of this chapter is to derive a functional central limit theorem (FCLT) by considering the asymptotic regime, $N\ !\ 1$. The FCLT that we establish in this chapter is useful to obtain an approximation to the blocking probability for a job in the system with N servers when N is large. In earlier works [8,10], they obtained the blocking probability for a job when $N\ !\ 1$ with the help of mean-field techniques. Without loss of generality, we call the blocking probability obtained when $N\ !\ 1$ as the asymptotic blocking probability. Quantifying the error between the actual blocking probability for the system with N servers and the asymptotic blocking probability is of great interest. We use the FCLT which we establish in this chapter to quantify the resulting error between the actual blocking probabilities and the asymptotic blocking probabilities, as a function of N. We assume that $^{(N)} =\ \ \tilde{}=^{p}\overline{N}$ for $\tilde{}\ 2\ \mathsf{R}$ and $\ 2\ \mathsf{R}_+$. Such an assumption includes the Halfin-Whitt regime [45] as a special case which corresponds to the case when $\ = C$ and $\tilde{}\ 6= 0$. The Halfin-Whitt regime is so-called an efficiency driven heavy traffic regime that

allows us to study how the effects of increasing load affects performance.

Organization of the Chapter

The rest of the chapter is organized as follows: We introduce the system model in Section 4.1. We then present the required notation and some preliminary results in Section 4.2. In Section 4.3, we state the main results of the chapter. In this section, we begin first with some preliminary results on the transient regime and then we state the functional central limit theorem for the transient regime in Theorem 4.3. After that we present some preliminary results for the stationary regime and then we give the functional central limit theorem for the stationary regime in Theorem 4.6. We conclude Section 4.3 with the result on approximations to the average blocking probability of the system with N servers. We give proofs of the main results in Section 4.4. Finally, we conclude in Section 4.5

4.1 System Model

In this section, we give a detailed description of the system model. We study a large-scale multi-server Erlang loss system with a central job dispatcher which routes an incoming job request to one of the servers according to the SQ(d) policy. The number of servers in the system is equal to N, a large value in the order of 10,000. Each server can serve at most C jobs simultaneously. As a result, the system's capacity indicating the maximum number of jobs that can be processed simultaneously in the system is equal to NC. When a job arrives, the dispatcher routes the arrived job to one of the servers, and the destination server of the arrival accepts the job if its occupancy is less than C. Otherwise, the destination server of the arrival blocks the job from service. The system immediately discards a blocked job. If a server accepts a job, it will serve or process the job at a constant unit rate until the service of the job is completed. We assume that the service time distributions are exponential with unit mean. The incoming jobs arrive according to a Poisson process with rate $N\ ^{(N)}$, where $^{(N)}\ 2\ \mathsf{R}_+$ and it is defined precisely later in this section.

Remark 4.1. *For our model, it does not matter whether the dispatcher samples d servers without replacement or with replacement to dispatch an arrival as they lead to the same asymptotic results. The proof follows by the same arguments as in [47, pages 11-12]. Therefore for simplicity, we assume that the dispatcher samples d servers with replacement upon an arrival.*

We next introduce the parameter $^{(N)}$. For $2 \mathsf{R}_+$ and $\tilde{} \; 2 \mathsf{R}$, the parameter $^{(N)}$ is defined as

$$^{(N)} = \quad \tilde{p}\overline{N}: \tag{4.1}$$

We are interested in obtaining an FCLT for our model. The particular form of $^{(N)}$ in (4.1) allows us to study the following two regimes as special cases:

1. For $\tilde{} = 0$: The considered case corresponds to the case when the job arrival process is a Poisson process with rate N . The previous works [8, 10] focused on obtaining a functional law of large numbers result referred to as the mean-field limit. They derived the blocking probabilities for a job when $N \; ! \; 1$ as a function of the unique fixed-point of the mean-field. The FCLT that we establish in this chapter allows us to show that the error between the actual blocking probability for a job in the system with N servers and the asymptotic blocking probability obtained in terms of the fixed-point of the mean-field is $o(N^{\frac{1}{2}})$.

2. For $\tilde{} \; 6= 0$ and $= C$: In this case, the system is in a heavy-traffic regime when N is large. In fact, the resulting model corresponds to the Halfin-Whitt regime since the traffic intensity $^{(N)} = \frac{^{(N)}}{C} = 1 \quad \frac{\tilde{p}}{C\overline{N}}$ approaches one as $N \; ! \; 1$. Then the arrival rate of jobs $N \; ^{(N)}$ and the system capacity NC are related as $NC = N \; ^{(N)} + \tilde{} \; p\overline{N}$, and they converge to 1 as $N \; ! \; 1$. In the Halfin-Whitt regime, we obtain an approximation to the blocking probability for a job by exploiting the FCLT that we establish in this chapter.

4.2 Additional Notation and Preliminary Results

In this section, we introduce the required notation and provide some preliminary results.

4.2.1 Additional Notation

We now introduce the required additional notation.

Due to the fact that the SQ(d) policy uses only the occupancy information of servers but not their identities, we use the Markov process $(\mathbf{X}^{(N)}(t); t \quad 0)$ to describe the system

evolution, where $\mathbf{X}^{(N)}(t) = (\mathbf{X}_i^{(N)}(t); 0 \quad i \quad C)$ and $\mathbf{X}_i^{(N)}(t)$ denotes the fraction of the servers with at least i progressing jobs at time t. Let us define a space U as

$$\mathsf{U} \ , \ f(u_0; u_1; \quad ; u_C) : u_0 = 1 \quad u_1 \qquad u_C \quad 0g: \tag{4.2}$$

It is evident that $\mathbf{X}^{(N)}(t)$ lies in the space U. Without loss of generality, we write an element of the form $(u_0; \quad ; u_C)$ as $\mathbf{u}$. The space U is equipped with the metric generated by the euclidean norm $k \, k_2$ defined as

$$k\mathbf{u}k_2 = \sqrt{\sum_{i=0}^{C} ju_i j^2}; \tag{4.3}$$

where $\mathbf{u} = (u_0; \quad ; u_C)$. It can be verified that the space U is a Polish space.

4.2.2 Preliminary Results

We next provide a mathematical framework for the main problem and present some preliminary results.

We first present a mathematical formulation to the time evolution of the process $(\mathbf{X}^{(N)}(t); t \quad 0)$. Upon an arrival, if the system state is $\mathbf{b} = (b_0; \quad ; b_C)$ implying that the fraction of servers with at least i progressing jobs is equal to b_i for $0 \quad i \quad C$, then the destination server of the job will have occupancy n with probability $b_n^d \quad b_{n+1}^d$. Since the rate of the arrival process is $N \ ^{(N)}$, the total instantaneous rate of arrivals to servers having n jobs is $N \ ^{(N)}((\mathbf{X}_n^{(N)}{}_1(t))^d \quad (\mathbf{X}_n^{(N)}(t))^d)$. Furthermore, as job lengths have exponential distributions with unit rate, the total instantaneous rate of departures from servers having n jobs is $nN(\mathbf{X}_n^{(N)}(t) \quad \mathbf{X}_{n+1}^{(N)}(t))$. Then we can write the system dynamics using random time change of a set of mutually independent unit rate Poisson processes as described in [83, Section 2.1].

Let $f(N_i(t); t \quad 0)g_{i \ 1}$ be the set of mutually independent unit rate Poisson processes. We use $(N_i(t); t \quad 0)$ to model the arrival process to servers that have $i \quad 1$ progressing jobs. Similarly, let $f(D_i(t); t \quad 0)g_{i \ 1}$ be the collection of a set of mutually independent unit rate Poisson processes. Furthermore, the set of processes $f(D_i(t); t \quad 0)g_{i \ 1}$ is independent of the set of processes $f(N_i(t); t \quad 0)g_{i \ 1}$. Also, $f(N_i(t); t \quad 0)g_{i \ 1}$ and $f(D_i(t); t \quad 0)g_{i \ 1}$ are independent of $\mathbf{X}^{(N)}(0)$. We use $(D_i(t); t \quad 0)$ to model the departure process from servers that have i progressing jobs. Since the arrival process of jobs to the system is a

Poisson process with rate $N^{(N)}$ and the service time distributions are exponential with unit mean, as in [83], we can write

$$\mathbf{X}_0^{(N)}(t) = 1; \tag{4.4}$$

and for $n \geq 1$,

$$\mathbf{X}_n^{(N)}(t) = \mathbf{X}_n^{(N)}(0) + \frac{1}{N} N_n\Big(N^{(N)} \int_{s=0}^{t} ((\mathbf{X}_{n-1}^{(N)}(s))^d - (\mathbf{X}_n^{(N)}(s))^d)\, ds\Big) - \frac{1}{N} D_n\Big(N n \int_{s=0}^{t} ((\mathbf{X}_n^{(N)}(s)) - (\mathbf{X}_{n+1}^{(N)}(s)))\, ds\Big): \tag{4.5}$$

We choose the filtration $(F^{(N)}(t); t \geq 0)$ where

$$F^{(N)}(t) = \Big(\mathbf{X}^{(N)}(0); N_n\Big(N^{(N)} \int_{s=0}^{r} ((\mathbf{X}_{n-1}^{(N)}(s))^d - (\mathbf{X}_n^{(N)}(s))^d)\, ds\Big); D_n\Big(Nn \int_{s=0}^{r} ((\mathbf{X}_n^{(N)}(s)) - (\mathbf{X}_{n+1}^{(N)}(s)))\, ds\Big); n \geq 1; 0 \leq r \leq t\Big); \tag{4.6}$$

augmented by all null sets.

We now present results on the mean-field analysis of the model without proofs as they directly follow from the case $^{(N)} = b$ for $b \in \mathbb{R}_+$, studied in [10]. The mean-field equations (MFEs) in our case are the same as the $^{(N)} =$ case except that replaces in the MFEs given in 1.3.2. We recall the MFEs in this chapter as we use them frequently.

Theorem 4.1. *For $\mathbf{u} \in U$, if $\mathbf{X}^{(N)}(0) \Rightarrow \mathbf{u}$ as $N \to \infty$, then $(\mathbf{X}^{(N)}(t); t \geq 0) \Rightarrow (\mathbf{x}(t;\mathbf{u}); t \geq 0)$ as $N \to \infty$ where $(\mathbf{x}(t;\mathbf{u}); t \geq 0) = (x_n(t;\mathbf{u}); t \geq 0; 0 \leq n \leq C)$ is the unique solution to the following equations:*
for $\mathbf{h}(\mathbf{x}(t;\mathbf{u})) = (h_n(\mathbf{x}(t;\mathbf{u})); 0 \leq n \leq C)$,

$$\mathbf{x}(0;\mathbf{u}) = \mathbf{u}; \quad \frac{dx_n(t;\mathbf{u})}{dt} = h_n(\mathbf{x}(t;\mathbf{u})); \tag{4.7}$$

where

$$h_0(\mathbf{x}(t;\mathbf{u})) = 0; \tag{4.8}$$

and for $n \geq 1$,

$$h_n(\mathbf{x}(t;\mathbf{u})) = (x_{n-1}^d(t;\mathbf{u}) - x_n^d(t;\mathbf{u})) - n(x_n(t;\mathbf{u}) - x_{n+1}(t;\mathbf{u})); \tag{4.9}$$

with $x_0(t;\mathbf{u}) = 1$ and $x_{C+1}(t;\mathbf{u}) = 0$. The deterministic process $(\mathbf{x}(t;\mathbf{u}); t \geq 0)$ is referred to as the mean-field limit and equations (4.7)-(4.9) are referred to as the mean-field equations with initial point $\mathbf{u}$.

Without loss of generality, we say that a process $(\mathbf{y}(t); t \quad 0)$ is a solution to the differential equations (4.7)-(4.9), it means that it is the unique generic solution with initial point $\mathbf{y}(0)$.

The mean-field $(\mathbf{x}(t; \mathbf{u}); t \quad 0)$ has a unique global asymptotically stable fixed-point $= (\ _n; 0 \quad n \quad C)$ with $\ _0 = 1$. Also, the following exchange of limits holds

$$\lim_{N \ ! \ 1} \lim_{t \ ! \ 1} \mathbf{X}^{(N)}(t) = \lim_{t \ ! \ 1} \lim_{N \ ! \ 1} \mathbf{X}^{(N)}(t): \tag{4.10}$$

Using (4.10), under the assumption of the exchangeability of initial servers' states, we can show the independence of any finite set of servers as $N \ ! \ 1$. Also, for a server, $\mathbf{x}(t; \mathbf{u})$ denotes the distribution at time t and $\quad$ denotes the stationary distribution, as $N \ ! \ 1$. As a result, $\ _C$ denotes the stationary probability that a server is fully occupied when $N \ ! \ 1$. Since the dispatcher samples d servers when a job arrives, the stationary average blocking probability of a job as $N \ ! \ 1$ is equal to $\ ^d_C$, where we use the fact that the chosen d servers are independent of each other. Our objective is to find the gap between the actual blocking probability of the system with N servers and $\ ^d_C$ as a function of $\tilde{}$, $\quad$, $\quad$, N, and C.

We can find $\quad$ numerically as explained below. The fixed-point $\quad$ is the unique solution to the following equations

$$(\ ^d_n \ _1 \quad \ ^d_n) = n(\ _n \quad \ _{n+1}) \tag{4.11}$$

for $n \quad 1$ and $\ _{C+1} = 0$. Then from (4.11), we can also write

$$\frac{(\ ^d_n \ _1 \quad \ ^d_n)}{(\ _n \ _1 \quad \ _n)}(\ _n \ _1 \quad \ _n) = n(\ _n \quad \ _{n+1}) \tag{4.12}$$

for $n \quad 1$ and $\ _{C+1} = 0$. Let us define $\hat{}_n = \frac{(\ ^d_n \quad \ ^d_{n+1})}{(\ _n \quad \ _{n+1})}$. Then from (4.12), $\quad$ is the stationary distribution of the single server loss model with a Poisson arrival process of jobs having rate $\hat{}_n$ when there are n progressing jobs, and $\ _n$ is the probability that the server has at least n progressing jobs. Let $\mathsf{M}_1(f0; 1; \quad ; Cg)$ be the set of probability measures on $f0; 1; \quad ; Cg$. Then from [10], the fixed-point $\quad$ can be computed using the formula for the stationary distribution of a single server loss system with state-dependent arrival rates. We first define two mappings, $\Theta : \mathsf{M}_1(f0; 1; \quad ; Cg) \ 7! \ \mathsf{R}_+^{C+1}$ and $\widehat{\Xi} : \mathsf{R}_+^{C+1} \ 7! \ \mathsf{M}_1(f0; 1; \quad ; Cg)$ that are used in computing $\quad$. For every $(p_0; \quad ; p_C) \ 2 \ \mathsf{M}_1(f0; 1; \quad ; Cg)$, there exists $(r_0; \quad ; r_C) \ 2 \ \mathsf{R}_+^{C+1}$ such that

$$\Theta((p_0; \quad ; p_C)) = (r_0; \quad ; r_C); \tag{4.13}$$

where

$$r_n = \frac{((\sum_{j=n}^{C} p_j)^d \quad (\sum_{i=n+1}^{C} p_i)^d)}{((\sum_{j=n}^{C} p_j) \quad (\sum_{i=n+1}^{C} p_i))}: \tag{4.14}$$

Similarly, for every $(b_0; \quad ; b_C)$ 2 R_+^{C+1}, there exists $(a_0; \quad ; a_C)$ 2 $\mathsf{M}_1(f0; 1; \quad ; Cg)$ such that

$$\widehat{\Xi}((b_0; \quad ; b_C)) = (a_0; \quad ; a_C); \tag{4.15}$$

where

$$a_n = \left(\prod_{i=1}^{n} \left(\frac{b_i \ 1}{i} \right) \right) a_0 \tag{4.16}$$

for n 1 and $\sum_{i=0}^{C} a_i = 1$. Then can be computed from the fixed-point of the mapping $\widehat{\Xi}(\Theta)$ as in Lemma 1.1 except that we replace by .

We first study the fluctuation process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ as $N \ ! \ 1$ where

$$\mathbf{Z}^{(N)}(t) = \sqrt{N}(\mathbf{X}^{(N)}(t) \quad \mathbf{x}(t; \mathbf{u})): \tag{4.17}$$

It can be checked that $\mathbf{Z}^{(N)}(t)$ lies in the space V defined as

$$\mathsf{V} \ , \ f(r_0; \quad ; r_C) : r_0 = 0 \text{ and } r_i \ 2 \ \mathsf{R}; 1 \quad i \quad C g: \tag{4.18}$$

We equip the space V with the topology induced by the euclidean norm (4.3). For a linear operator $K : \mathsf{V} \ 7! \ \mathsf{V}$, let the operator norm kKk_2 be defined as

$$kKk_2 = \sup_{\mathbf{v} 2 \mathsf{V}} \frac{kK\mathbf{v}k_2}{k\mathbf{v}k_2}: \tag{4.19}$$

We next obtain the time evolution of the process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ by using (4.5), (4.7), and (4.17). For this, we first define the following three useful operators $W_1; W_2; W_3 : \mathsf{U} \ 7! \ \mathsf{V}$ as follows: for $\mathbf{b}$ 2 U,

$$(W_1(\mathbf{b}))_0 = 0; \ (W_2(\mathbf{b}))_0 = 0; \ (W_3(\mathbf{b}))_0 = 0; \tag{4.20}$$

and for n 1,

$$(W_1(\mathbf{b}))_n = \ (b_n^d \ _1 \quad b_n^d); \ (W_2(\mathbf{b}))_n = n(b_n \quad b_{n+1}); \ (W_3(\mathbf{b}))_n = \tilde{} (b_n^d \ _1 \quad b_n^d): \tag{4.21}$$

From (4.9) and (4.21), we have

$$h_n(\mathbf{b}) = (W_1(\mathbf{b}))_n \quad (W_2(\mathbf{b}))_n: \tag{4.22}$$

Furthermore, let $W : U \to V$ be the operator defined as

$$W = W_1 \quad W_2: \tag{4.23}$$

The operator W is Lipschitz continuous satisfying the following inequality for all $\mathbf{a}; \mathbf{b} \in U$,

$$kW(\mathbf{a}) \quad W(\mathbf{b})k_2 \quad B_W k\mathbf{a} \quad \mathbf{b}k_2; \tag{4.24}$$

where $B_W = 2d^{p}\overline{{}^2 + C^2}$.

We now define a set of independent square-integrable martingales $(\mathbf{M}^{(N)}(t); t \quad 0) = f(\mathbf{M}_i^{(N)}(t); t \quad 0)g_{(i2\{0;1; \quad ;C\})}$ adapted to the filtration $(F^{(N)}(t); t \quad 0)$ such that $(\mathbf{M}^{(N)}(t); t \quad 0)$ is independent of $\mathbf{Z}^{(N)}(0)$ and for $i \quad 1$,

$$< \mathbf{M}_i^{(N)} >_t = \int_{s=0}^{t} \left((W_1(\mathbf{X}^{(N)}(s)))_i + (W_2(\mathbf{X}^{(N)}(s)))_i \quad \frac{1}{N}(W_3(\mathbf{X}^{(N)}(s)))_i \right) ds: \tag{4.25}$$

Then from (4.5)-(4.9), and (4.17), we get

$$\mathbf{Z}^{(N)}(t) = \mathbf{Z}^{(N)}(0) + \int_{s=0}^{t} \sqrt{N}(W(\mathbf{X}^{(N)}(s)) \quad W(\mathbf{x}(s; \mathbf{u}))) \, ds \quad \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s)) \, ds + \mathbf{M}^{(N)}(t): \tag{4.26}$$

The following result concludes the stochastic boundedness of the process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ as $N \to 1$ and the stochastic boundedness is useful for proving the tightness of the sequence $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \quad 1}$.

Lemma 4.1. *For any $T > 0$, if $\limsup_{N \to 1} \mathsf{E}\left[kZ^{(N)}(0)k_2^2\right] < 1$, then*

$$\limsup_{N \to 1} \mathsf{E}\left[\sup_{0 \quad t \quad T} kZ^{(N)}(t)k_2^2\right] < 1: \tag{4.27}$$

Proof. See Section 4.4.2. □

4.3 Summary of Main Results

In this section, we give the main results and provide proofs in Section 4.4. We present the results related to the transient regime and the stationary regime in Sections 4.3.1 and 4.3.2, respectively.

4.3.1 Transient Regime

In this section, for the transient regime case, we show that the process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ converges to an Ornstein-Uhlenbeck (OU) process as $N \; ! \; 1$.

As we show later in this section, the limit of the sequence $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \; 1}$ depends on the linearization of (4.7)-(4.9) around a solution $(\mathbf{r}(t); t \quad 0)$ to (4.7)-(4.9) with an initial point $\mathbf{r}(0)$ given by

$$\frac{d\mathbf{s}(t)}{dt} = H(\mathbf{r}(t))\mathbf{s}(t); \tag{4.28}$$

where for $\mathbf{a} \; 2 \; \mathsf{U}$ and $\mathbf{b} \; 2 \; \mathsf{V}$, the linear operator $H(\mathbf{a}) : \mathsf{V} \; 7! \; \mathsf{V}$ is defined as

$$(H(\mathbf{a})\mathbf{b})_n = \quad da_n^{d \; 1} \; {}_1^1 b_{n \; 1} \quad (\quad da_n^{d \; 1} + n)b_n + nb_{n+1}; \tag{4.29}$$

$n \quad 1$. The limit of the sequence $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \; 1}$ depends on the process $(\mathbf{s}(t); t \quad 0)$ when $(\mathbf{r}(t); t \quad 0)$ in (4.28) is replaced with the mean-field $(\mathbf{x}(t; \mathbf{u}); t \quad 0)$. Note that any solution $(\mathbf{s}(t); t \quad 0)$ to (4.28) satisfies $\mathbf{s}(t) = \mathbf{w}(t) \quad \mathbf{r}(t)$, where $(\mathbf{w}(t); t \quad 0)$ is a solution to the equations (4.7)-(4.9). The operator $H(\mathbf{a})$ is a matrix in the canonical basis $(0; 1; 0; \quad ; 0)$, $(0; 0; 1; 0; \quad ; 0)$, $\ldots$, $(0; 0; \quad ; 0; 1)$, where the dimension of each vector is $C + 1$. We can write $H(\mathbf{a})$ as the following matrix of size $C \quad C$:

$$H(\mathbf{a}) = \begin{bmatrix} {}_1 & 1 & 0 & 0 & & 0 & 0 \\ 1 & {}_2 & 2 & 0 & & 0 & 0 \\ 0 & 2 & {}_3 & 3 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & 0 & {}_{C \; 2} & {}_{C \; 1} & C \quad 1 \\ 0 & 0 & & 0 & 0 & {}_{C \; 1} & {}_C \end{bmatrix};$$

where ${}_i = \quad da_i^{d \; 1}$ and ${}_i = \quad {}_i + i, \; 1 \quad i \quad C$.

Let $(\mathbf{M}(t); t \quad 0) = f(\mathbf{M}_i(t); t \quad 0)g_{i \, 2 \, f0;1; \quad ;Cg}$ be a collection of mutually independent real valued continuous and centered Gaussian martingales, determined in law by their deterministic quadratic variation process

$$< \mathbf{M}_n >_t = \int_{s=0}^{t} ((W_1(\mathbf{x}(s; \mathbf{u}))_n + (W_2(\mathbf{x}(s; \mathbf{u}))_n) \, ds; \tag{4.30}$$

$n \quad 0$. Note that both $\mathbf{M}(t)$ and $(< \mathbf{M}_i >_t; 0 \quad i \quad C)$ lie in V. From (4.30), the martingale $(\mathbf{M}(t); t \quad 0)$ is square integrable since $((W_1(\mathbf{b}))_n + (W_2(\mathbf{b}))_n)$ for $\mathbf{b} \; 2 \; \mathsf{U}$ is uniformly bounded in n and $\mathbf{b}$ as $0 \quad b_i \quad 1$.

We now introduce the stochastic differential equation (SDE) satisfied by the limit of the process $(\mathbf{Z}^{(N)}(t); t \quad 0)$ as $N \; ! \; 1$ in the transient regime.

Definition 4.1. *Stochastic Di erential Equation (SDE) for the Transient Regime: Let* $(\mathbf{Z}(t); t \quad 0)$ *be a solution to the following SDE*

$$\mathbf{Z}(t) = \mathbf{Z}(0) + \int_{s=0}^{t} H(\mathbf{x}(s;\mathbf{u}))\mathbf{Z}(s)\, ds \quad \int_{s=0}^{t} W_3(\mathbf{x}(s;\mathbf{u}))\, ds + \mathbf{M}(t): \tag{4.31}$$

A solution to (4.31) is an OU process. We next study the SDE (4.31) below.

Theorem 4.2. *We show*

1. *For* $\mathbf{a}$ *2* U*, the linear operator* $H(\mathbf{a})$ *satis es* $kH(\mathbf{a})k_2 < B_H$*, where* $B_H = \sqrt{32(\ ^2d^2 + C^2)}$.
2. *If* $\mathsf{E}[k\mathbf{Z}(0)k_2^2] < 1$*, then there exists a unique strong solution to* (4.31) *denoted by* $(\mathbf{Z}(t); t \quad 0)$ *that satis es* $\mathsf{E}\left[\sup_t \ _T k\mathbf{Z}(t)k_2^2\right] < 1$.

Proof. See Section 4.4.1. □

We now present the main result of this chapter for the transient regime.

Theorem 4.3. *If* $\mathbf{Z}^{(N)}(0) \ ) \ \mathbf{Z}(0)$*, then* $(\mathbf{Z}^{(N)}(t); t \quad 0) \ ) \ (\mathbf{Z}(t); t \quad 0)$ *where* $(\mathbf{Z}(t); t \quad 0)$ *is the unique solution to* (4.31) *with initial point* $\mathbf{Z}(0)$.

Proof. See Section 4.4.3. □

Remark 4.2. *For a constant* r *2* R_+*, if* $^{(N)}$ *is independent of* N *xed at* $^{(N)} = r$*, then* $= r$ *and* $\tilde{} = 0$*. As a result, for* $n \quad 1$*,*

$$(W_1(\mathbf{b}))_n = r(b_n^d \ _1 \quad b_n^d); \ (W_2(\mathbf{b}))_n = n(b_n \quad b_{n+1}); \ (W_3(\mathbf{b}))_n = 0:$$

We recover the SDE obtained in [55] for $^{(N)} = r$ *case from the SDE* (4.31) *given by*

$$\mathbf{Z}(t) = \mathbf{Z}(0) + \int_{s=0}^{t} H(\mathbf{x}(s;\mathbf{u}))\mathbf{Z}(s)\, ds + \mathbf{M}(t):$$

4.3.2 Stationary Regime

In this section, we present results pertaining to the stationary regime. In the stationary regime, the mean-field is located at , and hence we assume $\mathbf{u} =$. We recall that satisfies

$$W(\ ) = W_1(\ ) \quad W_2(\ ) = 0: \tag{4.32}$$

Our objective is to obtain the limit of the process $(\mathbf{Z}^{(N)}(t); t \ 0)$ defined in (4.17) as $N \ ! \ 1$ in the stationary regime. For this, by studying the process $(\mathbf{Q}^{(N)}(t); t \ 0)$ where

$$\mathbf{Q}^{(N)}(t) = {}^{p}\overline{N}(\mathbf{X}^{(N)}(t) \quad); \tag{4.33}$$

we conclude that the sequence $f\mathbf{Z}^{(N)}(t)g_{N \ 1}$ is tight in the stationary regime. We introduce an SDE (4.38) and show that there exists a unique solution to this SDE with a unique invariant law. We then use this result to prove that the limit of the process $(\mathbf{Z}^{(N)}(t); t \ 0)$ as $N \ ! \ 1$ in the stationary regime is a stationary OU process with the same invariant law as that of the solution to the proposed SDE (4.38).

We next state the exponential stability of the mean-field. We use the following result in the proof of the subsequent result stated in Lemma 4.3.

Lemma 4.2. *There exists* $_1 > 0$ *and* $D_3 < 1$ *such that for all* $\mathbf{u} \ 2 \ \mathsf{U}$*, the mean- eld* $(\mathbf{x}(t; \mathbf{u}); t \ 0)$ *satis es*

$$k\mathbf{x}(t; \mathbf{u}) \quad k_2 \quad e \ {}^{1t}D_3 k\mathbf{u} \quad k_2: \tag{4.34}$$

Proof. See Section 4.4.7. □

The following result shows the tightness of $f\mathbf{Z}^{(N)}(t)g_{N \ 1}$ in the stationary regime.

Lemma 4.3. *If* $\limsup_{N \ ! \ 1} \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right] < 1$*, then*

$$\limsup_{N \ ! \ 1} \sup_{t \ 0} \mathsf{E}\left[k\mathbf{Q}^{(N)}(t)k_2^2\right] < 1: \tag{4.35}$$

Consequently, in the stationary regime, we have

$$\limsup_{N \ ! \ 1} \mathsf{E}\left[k\mathbf{Z}^{(N)}(\ 1 \ )k_2^2\right] < 1: \tag{4.36}$$

Proof. See Section 4.4.4. □

We next state the SDE that is used to obtain the limit of the sequence $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \quad 1}$ in the stationary regime. For this, we first claim that by linearizing (4.9) around , we get

$$\frac{d\mathbf{s}(t)}{dt} = H(\quad)\mathbf{s}(t): \tag{4.37}$$

Let $\mathbf{B}(t) = (\mathbf{B}_i(t); 0 \quad i \quad C)$ with $\mathbf{B}_0(t) = 0$ where $f(\mathbf{B}_i(t); t \quad 0)g_{0 \quad i \quad C}$ are independent centered Brownian motions and $\mathsf{E}[\mathbf{B}_i^2(1)] = V_i = var(\mathbf{B}_i(1)) = 2n(\quad_i \quad _{i+1})$, $i \quad 1$. The infinitesimal covariance matrix of $(\mathbf{B}(t); t \quad 0)$ is diagonal diag$(\mathbf{V})$, where $\mathbf{V} = (V_n; 0 \quad n \quad C)$. From (4.30), the martingales $(\mathbf{M}(t); t \quad 0)$ with $\mathbf{x}(0; \quad) = \quad$ has the same law as $(\mathbf{B}(t); t \quad 0)$. We now define the following SDE which is used to study the process $(\mathbf{Z}(t); t \quad 0)$ in the stationary regime.

Definition 4.2. *An SDE for the Stationary Regime: The process* $(\mathbf{Q}(t); t \quad 0)$ *is a solution to the following SDE,*

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \int_{s=0}^{t} H(\quad)\mathbf{Q}(s)\,ds \quad \int_{s=0}^{t} W_3(\quad)\,ds + \mathbf{B}(t): \tag{4.38}$$

Any solution to (4.38) is an OU process.

For an arbitrary $\mathbf{Q}(0)$ in (4.38), we have the following result and the proof follows by the same arguments as in the proof of Theorem 4.2. Hence, we omit the proof.

Theorem 4.4. *We show*

1. $kH(\quad)k_2$ *is bounded.*
2. *If* $\mathsf{E}[k\mathbf{Q}(0)k_2^2] < 1$, *then* $(\mathbf{Q}(t); t \quad 0)$ *given by*

$$\mathbf{Q}(t) = e^{H(\quad)t}\mathbf{Q}(0) \quad \int_{s=0}^{t} e^{H(\quad)(t \quad s)}W_3(\quad)\,ds + \int_{s=0}^{t} e^{H(\quad)(t \quad s)}\,d\mathbf{B}(s); \tag{4.39}$$

is the unique strong solution to (4.38). *Furthermore,*

$$\mathsf{E}\left[\sup_{t \quad T} k\mathbf{Q}(t)k_2^2\right] < 1: \tag{4.40}$$

We point out that the transpose $H(\quad)$ of $H(\quad)$ is the generator of a finite state birth-death process and the birth, death, and killing rates in state i $(1 \quad i \quad C)$ are $_i$, $i \quad 1$, and 1, respectively. Let I be the identity matrix of dimension $C \quad C$. Then since $H(\quad) + I$ is the generator of a birth-death process with zero killing rates, all the eigenvalues of $H(\quad) + I$ are negative [84]. Hence, all the eigenvalues of $H(\quad)$ are less than 1. As a consequence, we have the following result due to the fact that all the eigenvalues are negative.

Lemma 4.4. *The unique solution to* (4.37) *is given by* $(\mathbf{s}(t); t \quad 0)$ *where* $\mathbf{s}(t) = e^{H(\ )t}\mathbf{s}(0)$. *Furthermore,* $(\mathbf{s}(t); t \quad 0)$ *satis es that for some* $_2 > 0$ *and* $D_4 < 1$ *,*

$$k\mathbf{s}(t)k_2 \quad e \quad {}^{2t}D_4 k\mathbf{s}(0)k_2: \tag{4.41}$$

From Lemma 4.4 and the unique solution given in Theorem 4.4, the following result follows immediately. Hence, we omit the proof.

Theorem 4.5. *The unique solution to* (4.38) *as* $t \ ! \ 1$ *has the invariant law coinciding with the law of a stationary Gaussian process with mean* $\int_0^1 e^{H(\ )s}W_3(\ )\,ds$ *and covariance matrix* $\int_0^1 e^{H(\ )s}diag(\mathbf{V})e^{H(\ )\ s}\,ds$.

We are now ready to state the main result on the FCLT for the stationary regime.

Theorem 4.6. *Under the assumption that the system with index* N *is in the stationary regime, from Theorem 4.5, we prove that the sequence* $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N\ 1}$ *as* $N \ ! \ 1$ *converges in law to the unique stationary OU process which solves* (4.38). *Furthermore, the limit of the sequence* $f(\mathbf{Z}^{(N)}(0))g_{N\ 1}$ *in the stationary regime has the same law as the invariant law of the solution to* (4.38).

Proof. See Section 4.4.5. □

We now use Theorem 4.6 to provide an approximation to the average blocking probability in the system with N servers.

Theorem 4.7. *Let* $P_{block}^{(N)}$ *be the average blocking probability in the system with* N *servers, then*

$$P_{block}^{(N)} = \ {}^{d}_{C} \quad \frac{1}{p\overline{N}}\left(\sum_{i=0}^{C} i(\ _i \quad \ _{i+1})\right) \quad \frac{\tilde{}}{p\overline{N}}(1 \quad {}^{d}_{C}) + o(N \ {}^{\frac{1}{2}}); \tag{4.42}$$

where the vector $= (\ _i; 0 \quad i \quad C) = \int_0^1 e^{H(\ )s}W_3(\ )\,ds$ *is the mean of the unique solution to* (4.38) *in the stationary regime.*

Proof. See Section 4.4.6. □

Remark 4.3. *From Theorem 4.7, we have*

1. *If* $= C$, *the result* (4.42) *corresponds to the Hal n-Whitt regime. In this case, we have*

$$\lim_{N \ ! \ 1} {}^{p}\overline{N}(P_{block}^{(N)} \quad {}^{d}_{C}) = \ \frac{1}{-}\left(\sum_{i=0}^{C} i(\ _i \quad \ _{i+1})\right) \quad \frac{\tilde{}}{-}(1 \quad {}^{d}_{C}):$$

2. *If* $\tilde{} = 0$, *then* $W_3(\ ) = 0$. *As a result,*

$$\lim_{N \to 1} \sqrt{N}(P_{block}^{(N)} \quad \overset{d}{C}) = 0:$$

We presented the results for the case $\tilde{} = 0$ *in [55]. For this case, we can also apply the results of [51] to conclude that* $P_{block}^{(N)} \quad \overset{d}{C} = O(\frac{1}{N})$. *In [51], an FCLT result was not studied..*

The significance of Theorem 4.7 is that although the exact blocking formula for $P_{block}^{(N)}$ is not known, and also it is difficult to characterize due to complex interactions between servers, but as N becomes large we can compute approximations to the blocking probability as a function of $\ , \tilde{}, N, \ , \text{and } C$.

4.4 Proofs of Main Results

4.4.1 Proof of Theorem 4.2

For $\mathbf{y} \in \mathsf{V}$, we have

$$kH(\mathbf{a})\mathbf{y}k_2^2 = \sum_{n=0}^{C} \left| \ da_n^d \ {}_1^1 y_n \ _1 \quad (\ da_n^d \ ^1 + n)y_n + ny_{n+1} \right|^2 :$$

Then it can be checked that $kH(\mathbf{a})\mathbf{y}k_2^2 \quad 32(\ ^2d^2 + C^2)kyk_2^2$ as $0 \quad a_i \quad 1$. This implies $kH(\mathbf{a})k_2^2 \quad B_H^2 = 32(\ ^2d^2 + C^2)$ and hence, $kH(\mathbf{a})k_2 \quad B_H$.

We next show the uniqueness of a solution to (4.31). For two solutions $(\mathbf{Z}^1(t); t \quad 0)$ and $(\mathbf{Z}^2(t); t \quad 0)$ with initial points $\mathbf{Z}^1(0)$ and $\mathbf{Z}^2(0)$, we obtain

$$\mathbf{Z}^1(t) \quad \mathbf{Z}^2(t) = \mathbf{Z}^1(0) \quad \mathbf{Z}^2(0) + \int_{s=0}^{t} H(\mathbf{x}(s; \mathbf{u}))(\mathbf{Z}^1(s) \quad \mathbf{Z}^2(s)) \, ds:$$

Then

$$k\mathbf{Z}^1(t) \quad \mathbf{Z}^2(t)k_2 \quad k\mathbf{Z}^1(0) \quad \mathbf{Z}^2(0)k_2 + B_H \int_{s=0}^{t} k\mathbf{Z}^1(s) \quad \mathbf{Z}^2(s)k_2 \, ds:$$

By using the Gronwall's Lemma (Theorem A.7) [56, Page 498], we get

$$k\mathbf{Z}^1(t) \quad \mathbf{Z}^2(t)k_2 \quad e^{B_H t}k\mathbf{Z}^1(0) \quad \mathbf{Z}^2(0)k_2:$$

Hence,

$$\mathsf{E}\left[k\mathbf{Z}^1(t) \quad \mathbf{Z}^2(t)k_2^2\right] \quad e^{2B_H t}\mathsf{E}\left[k\mathbf{Z}^1(0) \quad \mathbf{Z}^2(0)k_2^2\right] :$$

If $\mathbf{Z}^1(0) = \mathbf{Z}^2(0)$, then we obtain $\mathbf{Z}^1(t) = \mathbf{Z}^2(t)$ *a:s:* for all rational t. Finally, since $(\mathbf{Z}^1(t); t \quad 0)$ and $(\mathbf{Z}^2(t); t \quad 0)$ have continuous sample paths, we conclude that $\mathbf{Z}^1(t) = \mathbf{Z}^2(t)$ *a:s:*, for all $t \quad 0$.

We next use the Gronwall's lemma and the Doob's inequality (Theorem A.8) [56, p. 63] to show that $\mathsf{E}\left[\sup_{t \quad T} k\mathbf{Z}(t)k_2^2\right] < 1$. From (4.31), we have

$$k\mathbf{Z}(t)k_2 \quad k\mathbf{Z}(0)k_2 + B_H \int_{s=0}^{t} k\mathbf{Z}(s)k_2\, ds + \int_{s=0}^{t} kW_3(\mathbf{x}(s; \mathbf{u}))k_2 + k\mathbf{M}(\mathbf{t})k_2:$$

However, for $\mathbf{a} \; 2 \; \mathsf{U}$, we have $kW_3(\mathbf{a})k_2 \quad 2j \tilde{} j \overline{p\,C}$. Therefore, we get

$$k\mathbf{Z}(t)k_2^2 \quad 4\left(k\mathbf{Z}(0)k_2^2 + B_H^2\left(\int_{s=0}^{t} k\mathbf{Z}(s)k_2\, ds\right)^2 + 4t^{2\tilde{}2}C + k\mathbf{M}(t)k_2^2\right):$$

We can apply the Holder's inequality to $\left(\int_{s=0}^{t} k\mathbf{Z}(s)k_2\, ds\right)^2$ to obtain the following inequality

$$k\mathbf{Z}(t)k_2^2 \quad 4\left(k\mathbf{Z}(0)k_2^2 + B_H^2 t\int_{s=0}^{t} k\mathbf{Z}(s)k_2^2\, ds + 4t^{2\tilde{}2}C + k\mathbf{M}(t)k_2^2\right):$$

By using the Gronwall's Lemma, we obtain

$$k\mathbf{Z}(t)k_2^2 \quad 4\left(k\mathbf{Z}(0)k_2^2 + 4t^{2\tilde{}2}C + k\mathbf{M}(t)k_2^2\right)e^{4B_H^2 t^2}:$$

Therefore for any $T > 0$, we get

$$\sup_{0 \quad t \quad T} k\mathbf{Z}(t)k_2^2 \quad 4\left(k\mathbf{Z}(0)k_2^2 + 4T^{2\tilde{}2}C + \sup_{0 \quad t \quad T} k\mathbf{M}(t)k_2^2\right)e^{4B_H^2 T^2}:$$

Hence,

$$\mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{Z}(t)k_2^2\right] \quad 4e^{4B_H^2 T^2}\left(\mathsf{E}\left[k\mathbf{Z}(0)k_2^2\right] + 4T^{2\tilde{}2}C + \mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{M}(t)k_2^2\right]\right):$$

The Doob's inequality implies

$$\mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{Z}(t)k_2^2\right] \quad 4e^{4B_H^2 T^2}\left(\mathsf{E}\left[k\mathbf{Z}(0)k_2^2\right] + 4T^{2\tilde{}2}C + 4\mathsf{E}\left[k\mathbf{M}(T)k_2^2\right]\right):$$

Therefore

$$\mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{z}(t)k_2^2\right] \quad 4e^{4B_H^2 T^2}\left(\mathsf{E}\left[k\mathbf{z}(0)k_2^2\right] + 4T^{2\tilde{}2}C + 4\mathsf{E}\left[\sum_{i=1}^{C} < \mathbf{M}_i >_T\right]\right):$$

For every T and i, $< \mathbf{M}_i >_T$ is bounded. Hence, $\mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{Z}(t)k_2^2\right] < 1$.

The fact that there exists a solution to the SDE (4.31) follows from [85, page 354]. Let $(\tilde{F}(t); t \quad 0)$ be the unique solution of the following equation

$$\frac{dY(t)}{dt} = H(\mathbf{x}(t; \mathbf{u}))Y(t); \qquad Y(0) = I;$$

where I is the identity matrix. Then the unique solution to the SDE (4.31) is given by

$$\mathbf{Z}(t) = \tilde{F}(t)\left[\mathbf{Z}(0) \quad \int_{s=0}^{t} (\tilde{F}(s)) \quad {}^{1}W_3(\mathbf{x}(s; \mathbf{u}))\, ds + \int_{s=0}^{t} (\tilde{F}(s)) \quad {}^{1}\, d\mathbf{M}(s)\right]: \tag{4.43}$$

4.4.2 Proof of Lemma 4.1

We are given

$$\mathbf{Z}^{(N)}(t) = \mathbf{Z}^{(N)}(0) + \int_{s=0}^{t} \quad p\overline{N}(W(\mathbf{X}^{(N)}(s)) \quad W(\mathbf{x}(s; \mathbf{u})))\, ds \quad \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\, ds + \mathbf{M}^{(N)}(t); \tag{4.44}$$

and for $n \quad 1$,

$$< \mathbf{M}_n^{(N)} >_t = \int_{s=0}^{t} ((W_1(\mathbf{X}^{(N)}(s)))_n + (W_2(\mathbf{X}^{(N)}(s)))_n \quad p\frac{1}{N}(W_3(\mathbf{X}^{(N)}(s)))_n)\, ds:$$

From (4.44), we obtain

$$k\mathbf{Z}^{(N)}(t)k_2 \quad k\mathbf{Z}^{(N)}(0)k_2 + B_W \int_{s=0}^{t} k\mathbf{Z}^{(N)}(s)k_2\, ds + 2C\tilde{}t + k\mathbf{M}^{(N)}(t)k_2:$$

By using the Gronwall's Lemma,

$$k\mathbf{Z}^{(N)}(t)k_2 \quad (k\mathbf{Z}^{(N)}(0)k_2 + 2C\tilde{}t + k\mathbf{M}^{(N)}(t)k_2)e^{B_W t}:$$

As a result, we get

$$k\mathbf{Z}^{(N)}(t)k_2^2 \quad 3(k\mathbf{Z}^{(N)}(0)k_2^2 + 4C^{2}\tilde{}^{2}t^2 + k\mathbf{M}^{(N)}(t)k_2^2)e^{2B_W t}:$$

For $T > 0$, we have

$$\sup_{0 \quad t \quad T} k\mathbf{Z}^{(N)}(t)k_2^2 \quad 3(k\mathbf{Z}^{(N)}(0)k_2^2 + 4C^{2}\tilde{}^{2}T^2 + \sup_{0 \quad t \quad T} k\mathbf{M}^{(N)}(t)k_2^2)e^{2B_W T}:$$

Finally, the Doob's inequality implies the following inequality

$$\mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{Z}^{(N)}(t)k_2^2\right] \quad 3e^{2B_W T}\left(\mathsf{E}\left[k\mathbf{Z}^{(N)}(0)k_2^2\right] + 4C^{2}\tilde{}^{2}T^2 + 4\mathsf{E}\left[\sum_{i=1}^{C} < \mathbf{M}_i^{(N)} >_T\right]\right):$$

Since $\sup_{N \quad 1} \mathsf{E}\left[\sum_{i=1}^{C} < \mathbf{M}_i^{(N)} >_T\right] < 1$, we conclude that

$$\limsup_{N \; ! \; 1} \mathsf{E}\left[\sup_{0 \quad t \quad T} k\mathbf{Z}^{(N)}(t)k_2^2\right] < 1:$$

4.4.3 Proof of Theorem 4.3

We recall that since the space V is Polish, the space of càdlàg functions under the Skhorohod topology is a Polish space [56, Theorem 5.6, p.121]. Hence from the Prohorov's theorem (Theorem A.2) [56], tightness is equivalent to relative compactness. Therefore we first need to show the tightness and then we need to show that every limiting point has the same law as the unique OU process with the initial point $\mathbf{Z}(0)$.

To show the tightness of $(\mathbf{Z}^{(N)}(t); t \quad 0)$, we need to show that Theorem 4.1 of [56, page 354] holds. For this, we first establish several useful results.

Since $\mathbf{Z}^{(N)}(0) \ ) \ \mathbf{Z}(0)$, it implies that the sequence $\mathbf{Z}^{(N)}(0)$ is tight. Let $\tilde{B}(r)$ be the closed ball with radius r centered at $\mathbf{0}$. For every $\quad > 0$, there exists $r \ < \ 1$ such that $\mathsf{P}(\mathbf{Z}^{(N)}(0) \ 2 \ \tilde{B}(r \ )) > 1 \quad$ for all $N \quad 1$. We now define a random variable $\mathbf{X}^{(N; \ )}(0)$ such that it coincides with $\mathbf{X}^{(N)}(0)$ on $f\mathbf{Z}^{(N)}(0) \ 2 \ \tilde{B}(r \ )g$ and $\mathbf{Z}^{(N; \ )}(0)$ is uniformly bounded in N on $f\mathbf{Z}^{(N)}(0) \ 2 \ \tilde{B}(r \ )g$. Then by using coupling arguments, the processes $(\mathbf{Z}^{(N; \ )}(t); t \quad 0)$ and $(\mathbf{Z}^{(N)}(t); t \quad 0)$ coincide on $f\mathbf{Z}^{(N)}(0) \ 2 \ \tilde{B}(r \ )g$. Hence, without loss of generality, we assume that $\mathbf{Z}^{(N)}(0)$ is uniformly bounded in N. As a consequence, the result stated in Lemma 4.1 can be used in the rest of the proof.

We next recall the following useful result from [47, Lemma 3.3]. For a and h in R, let

$$\hat{B}(a; h) = (a + h)^d \quad a^d \quad da^{d \ 1}h:$$

Then if both a and $a + h$ lie in $[0; 1]$, we have

$$0 \quad \hat{B}(a; h) \quad h^d + (2^d \quad d \quad 2)ah^2: \tag{4.45}$$

We now define a mapping $\hat{G} : \mathsf{U} \quad \mathsf{V} \ ?! \ \mathsf{V}$ as follows: for $\mathbf{r} \ 2 \ \mathsf{U}$ and $\mathbf{y} \ 2 \ \mathsf{V}$,

$$(\hat{G}(\mathbf{r}; \mathbf{y}))_n = \quad \hat{B}(r_{n \ 1}; y_{n \ 1}) \quad \hat{B}(r_n; y_n): \tag{4.46}$$

Then if $\mathbf{r} + \mathbf{y} \ 2 \ \mathsf{U}$, we have

$$W(\mathbf{r} + \mathbf{y}) \quad W(\mathbf{r}) = H(\mathbf{r})\mathbf{y} + \hat{G}(\mathbf{r}; \mathbf{y}): \tag{4.47}$$

Note that since $\mathbf{Z}^{(N)}(t) = \overline{{}^{p}N}(\mathbf{X}^{(N)}(t) \quad \mathbf{x}(t; \mathbf{u}))$, we have

$$\mathbf{X}^{(N)}(t) = \mathbf{x}(t; \mathbf{u}) + \frac{\mathbf{Z}^{(N)}(t)}{\overline{p\overline{N}}}: \tag{4.48}$$

Here, $\mathbf{X}^{(N)}(t)$, $\mathbf{x}(t;\mathbf{u})$ 2 U and $\frac{\mathbf{Z}^{(N)}(t)}{p\overline{N}}$ 2 V. Hence, from (4.47), we have

$$W(\mathbf{X}^{(N)}(t) \quad W(\mathbf{x}(t;\mathbf{u})) = H(\mathbf{x}(t;\mathbf{u}))\frac{\mathbf{Z}^{(N)}(t)}{p\overline{N}} + \hat{G}\left(\mathbf{x}(t;\mathbf{u}); \frac{\mathbf{Z}^{(N)}(t)}{p\overline{N}}\right): \tag{4.49}$$

We now show that the conditions of [56, Theorem 4.1, page 354] are satisfied. Let us write $\mathbf{B}_n(t)$ and $\mathbf{A}_n^{i;j}(t)$ of [56, Theorem 4.1, page 354] as $\mathbf{D}_{(N)}(t)$ and $\mathbf{A}_{(N)}^{i;j}(t)$, respectively. Then from (4.26)

$$\mathbf{D}_{(N)}(t) = \int_{s=0}^{t} p\overline{N}(W(\mathbf{X}^{(N)}(s)) \quad W(\mathbf{x}(s;\mathbf{u}))\, ds \quad \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\, ds: \tag{4.50}$$

From (4.49), we can write

$$\begin{aligned}&\mathbf{D}_{(N)}(t)\\ &= \int_{s=0}^{t} H(\mathbf{x}(s;\mathbf{u}))\mathbf{Z}^{(N)}(s)\, ds + \int_{s=0}^{t} p\overline{N}\hat{G}\left(\mathbf{x}(s;\mathbf{u}); \frac{\mathbf{Z}^{(N)}(s)}{p\overline{N}}\right) ds \quad \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\, ds:\end{aligned} \tag{4.51}$$

Also, from (4.25)

$$\mathbf{A}_{(N)}^{i;i}(t) = \int_{s=0}^{t} \left((W_1(\mathbf{X}^{(N)}(s)))_i + (W_2(\mathbf{X}^{(N)}(s)))_i \quad \frac{1}{p\overline{N}}(W_3(\mathbf{X}^{(N)}(s)))_i\right) ds; \tag{4.52}$$

and $\mathbf{A}_{(N)}^{i;j}(t) = 0$ for i 6= j. Let us define

$$\mathbf{D}(t) = \int_{s=0}^{t} H(\mathbf{x}(s;\mathbf{u}))\mathbf{Z}(s)\, ds \quad \int_{s=0}^{t} W_3(\mathbf{x}(s;\mathbf{u}))\, ds; \tag{4.53}$$

and

$$A^{i;i}(t) = \int_{s=0}^{t} ((W_1(\mathbf{x}(s;\mathbf{u})))_i + (W_2(\mathbf{x}(s;\mathbf{u})))_i)\, ds; \tag{4.54}$$

with $A^{i;j}(t) = 0$ for i 6= j.

Since $(\mathbf{Z}_i^{(N)}(t); t \quad 0)$ has jumps of size $\frac{1}{p\overline{N}}$ and from the continuity of $\mathbf{D}_{(N)}(t)$ and $\mathbf{A}_{(N)}^{i;j}(t)$ in t, the conditions (4:3)-(4:5) of [56, Theorem 4.1, p. 354] are valid. From the condition (4:6) of [56, Theorem 4.1, p. 354], we need to show that for $T > 0$, $\mathbf{V}_{(N)}(t)$! 0 in probability where $\mathbf{V}_{(N)}(t) = \mathbf{D}_{(N)}(t) \quad \mathbf{D}(t)$ given by

$$\begin{aligned}&\mathbf{V}_{(N)}(t)\\ &= \left(\int_{s=0}^{t} H(\mathbf{x}(t;\mathbf{u}))\mathbf{Z}^{(N)}(t)\, ds + \int_{s=0}^{t} p\overline{N}\hat{G}\left(\mathbf{x}(t;\mathbf{u}); \frac{\mathbf{Z}^{(N)}(t)}{p\overline{N}}\right) ds \quad \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\, ds\right)\end{aligned}$$

$$\left(\int_{s=0}^{t} H(\mathbf{x}(s;\mathbf{u}))\mathbf{Z}^{(N)}(s)\,ds \quad \int_{s=0}^{t} W_3(\mathbf{x}(s;\mathbf{u}))\,ds\right): \quad (4.55)$$

From (4.27) and (4.45), we have $\int_{s=0}^{T} {}^{p}\overline{N}\hat{G}\left(\mathbf{x}(s;\mathbf{u}); \frac{\mathbf{Z}^{(N)}(s)}{p\overline{N}}\right) ds \; ! \; 0$ in probability. From the existence of the mean-field limit, $\int_{s=0}^{T} W_3(\mathbf{X}^{(N)}(s))\,ds \quad \int_{s=0}^{T} W_3(\mathbf{x}(s;\mathbf{u}))\,ds \; ! \; 0$ in probability. We conclude that the condition (4:6) of [56, Theorem 4.1, p. 354] is also true. The condition (4:7) of [56, Theorem 4.1, p. 354] follows as $\mathbf{A}^{i;i}_{(N)}(T) \quad A^{i;i}(T) \; ! \; 0$ in probability from the existence of the mean-field limit. From the fact that the SDE (4.31) has a unique solution, the sequence $(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \; 1}$ converges to the unique solution of the SDE (4.31) as $N \; ! \; 1$.

4.4.4 Proof of Lemma 4.3

The proof is based on Lemma 4.2. Although our main motivation is to establish the convergence of $f\mathbf{Z}^{(N)}(t)g_{N \; 1}$ as $N \; ! \; 1$ in the stationary regime, we prove the tightness of the stationary sequence $f\mathbf{Z}^{(N)}(\ 1\ )g_{N \; 1}$ by studying an alternative process $(\mathbf{Q}^{(N)}(t); t \quad 0)$ with the help of Lemma 4.2, where

$$\mathbf{Q}^{(N)}(t) = {}^{p}\overline{N}(\mathbf{X}^{(N)}(t) \qquad):$$

Let us write the solution to the mean-field equation (4.9) at time h with the initial point $\mathbf{v}$ as $\mathbf{y}(h;\mathbf{v})$. We have

$$\mathbf{y}(h;\mathbf{v}) = \mathbf{v} + \int_{s=0}^{h} W(\mathbf{y}(s;\mathbf{v}))\,ds: \quad (4.56)$$

Also, the process $(\mathbf{X}^{(N)}(t); t \quad 0)$ satisfies

$$\mathbf{X}^{(N)}(t) = \mathbf{X}^{(N)}(0) + \int_{s=0}^{t} W(\mathbf{X}^{(N)}(s))\,ds \quad \int_{s=0}^{t} \frac{1}{p\overline{N}} W_3(\mathbf{X}^{(N)}(s))\,ds + \frac{\mathbf{M}^{(N)}(t)}{p\overline{N}}: \quad (4.57)$$

For $t_0 \quad 0$, we obtain

$$\mathbf{Q}^{(N)}(t_0+h) = {}^{p}\overline{N}(\mathbf{X}^{(N)}(t_0+h) \quad \mathbf{y}(h;\mathbf{X}^{(N)}(t_0))) \; + \; {}^{p}\overline{N}(\mathbf{y}(h;\mathbf{X}^{(N)}(t_0)) \qquad):$$

By defining $\mathbf{Z}^{(N)}(t_0;h) = {}^{p}\overline{N}(\mathbf{X}^{(N)}(t_0+h) \quad \mathbf{y}(h;\mathbf{X}^{(N)}(t_0)))$, we obtain

$$\mathbf{Q}^{(N)}(t_0+h) = \mathbf{Z}^{(N)}(t_0;h) + {}^{p}\overline{N}(\mathbf{y}(h;\mathbf{X}^{(N)}(t_0)) \qquad): \quad (4.58)$$

Also, from (4.56),

$$\mathbf{Z}^{(N)}(t_0;h) = {}^{p}\overline{N}(\mathbf{X}^{(N)}(t_0+h) \quad \mathbf{X}^{(N)}(t_0)) \quad {}^{p}\overline{N}\int_{s=0}^{t} W(\mathbf{y}(s;\mathbf{X}^{(N)}(t_0)))\,ds: \quad (4.59)$$

From (4.57), we have

$$\mathbf{X}^{(N)}(t_0+h) \quad \mathbf{X}^{(N)}(t_0) = \int_{s=t_0}^{t_0+h} W(\mathbf{X}^{(N)}(s))\,ds + \frac{\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0)}{p\overline{N}}$$
$$\frac{1}{p\overline{N}} \int_{s=t_0}^{t_0+h} W_3(\mathbf{X}^{(N)}(s))\,ds: \quad (4.60)$$

Then by using (4.59) and (4.60), we get

$$\mathbf{Z}^{(N)}(t_0;h) = p\overline{N} \int_{s=0}^{h} W(\mathbf{X}^{(N)}(t_0+s))\,ds \quad p\overline{N} \int_{s=0}^{h} W(\mathbf{y}(s;\mathbf{x}^{(N)}(t_0)))\,ds$$
$$+ (\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0)) \quad \int_{s=t_0}^{t_0+h} W_3(\mathbf{X}^{(N)}(s))\,ds:$$

After simplifications, we get

$$k\mathbf{Z}^{(N)}(t_0;h)k_2 = p\overline{N} \int_{s=0}^{h} kW(\mathbf{X}^{(N)}(t_0+s)) \quad W(\mathbf{y}(s;\mathbf{X}^{(N)}(t_0)))k_2\,ds$$
$$+ k\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0))k_2 + j\tilde{}jd^{p}\overline{C}h:$$

Hence, we obtain

$$k\mathbf{Z}^{(N)}(t_0;h)k_2 \quad B_W \int_{s=0}^{h} k\mathbf{Z}^{(N)}(t_0;s)k_2\,ds + j\tilde{}jd^{p}\overline{C}h + k\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0))k_2:$$

For any $T \quad 0$, the Gronwall's inequality implies that there exists a constant S_T such that

$$\sup_{0 \quad h \quad T} k\mathbf{Z}^{(N)}(t_0;h)k_2 \quad S_T(S_T + \sup_{0 \quad h \quad T} k\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0))k_2): \quad (4.61)$$

By using Lemma 4.2, from (4.58), we get

$$k\mathbf{Q}^{(N)}(t_0+h)k_2 \quad k\mathbf{Z}^{(N)}(t_0;h)k_2 + e \quad {}^{1h}D_3k\mathbf{Q}^{(N)}(t_0)k_2: \quad (4.62)$$

From (4.61) and (4.62), we obtain

$$k\mathbf{Q}^{(N)}(t_0+h)k_2 \quad S_T(S_T + \sup_{0 \quad h \quad T} k\mathbf{M}^{(N)}(t_0+h) \quad \mathbf{M}^{(N)}(t_0))k_2) + e \quad {}^{1h}D_3k\mathbf{Q}^{(N)}(t_0)k_2:$$

As a result, we can find some constant L_T as a function of T such that for $0 \quad h \quad T$, we have

$$\mathsf{E}\left[k\mathbf{Q}^{(N)}(t_0+h)k_2^2\right] \quad L_T + 3e^{\ 2\ {}^{1h}D_3^2}\mathsf{E}\left[k\mathbf{Q}^{(N)}(t_0)k_2^2\right]: \quad (4.63)$$

We now select a large value of T such that $3e^{\ 2\ {}^{1T}D_3^2} \quad < 1$. Then for all $N \quad 1$ and an integer m, we get

$$\mathsf{E}\left[k\mathbf{Q}^{(N)}((m+1)T)k_2^2\right] \quad L_T + \ \mathsf{E}\left[k\mathbf{Q}^{(N)}(mT)k_2^2\right]:$$

By using the induction method, we obtain

$$\mathsf{E}\left[k\mathbf{Q}^{(N)}(mT)k_2^2\right] \quad L_T\Big(\sum_{j=1}^{m} {}^{j\ 1}\Big) + {}^{m}\mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right]$$
$$\frac{L_T}{1} + \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right]: \tag{4.64}$$

However, from (4.63),

$$\sup_{0\ h\ T} \mathsf{E}\left[k\mathbf{Q}^{(N)}(mT+h)k_2^2\right] \quad L_T + 3D_3^2\mathsf{E}\left[k\mathbf{Q}^{(N)}(mT)k_2^2\right]:$$

As a consequence, (4.64) implies

$$\sup_{0\ h\ T} \mathsf{E}\left[k\mathbf{Q}^{(N)}(mT+h)k_2^2\right] \quad L_T + 3D_3^2\left(\frac{L_T}{1} + \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right]\right):$$

Since m is arbitrary, we conclude

$$\sup_{t\ 0} \mathsf{E}\left[k\mathbf{Q}^{(N)}(t)k_2^2\right] \quad L_T + 3D_3^2\left(\frac{L_T}{1} + \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right]\right):$$

From ergodicity and the fatou's Lemma [56, p. 492], the stationary random variable $\mathbf{Z}^{(N)}(1)$ satisfies

$$\mathsf{E}\left[k\mathbf{Z}^{(N)}(1)k_2^2\right] \quad \liminf_{t\ 0} \mathsf{E}\left[k\mathbf{Q}^{(N)}(t)k_2^2\right]$$
$$\sup_{t\ 0} \mathsf{E}\left[k\mathbf{Q}^{(N)}(t)k_2^2\right]:$$

Finally, to show that $\limsup_{N\ !\ 1} \mathsf{E}\left[k\mathbf{Z}^{(N)}(1)k_2^2\right] < 1$, we need to find an $\mathbf{X}^{(N)}(0)$ such that $\limsup_{N\ !\ 1} \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right] < 1$. For $n \quad 1$, if we select $\mathbf{X}_n^{(N)}(0) = \frac{j}{N}$ so that $\frac{1}{2N} \quad n \quad \frac{j}{N} \quad \frac{1}{2N}$, then $\mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right] < \frac{C}{4N}$. Hence, $\limsup_{N\ !\ 1} \mathsf{E}\left[k\mathbf{Q}^{(N)}(0)k_2^2\right] = 0$. This completes the proof.

4.4.5 Proof of Theorem 4.6

From Lemma 4.3 and the Markov inequality, the sequence $f\mathbf{Z}^{(N)}(0)g_{N\ 1}$ is tight. As a result, from the Prohorov theorem [56, Page 104], the sequence $f\mathbf{Z}^{(N)}(0)g_{N\ 1}$ is relatively compact. Consider a converging subsequence and let $\mathbf{Z}^{(1)}(0)$ be its limiting point, which is square integrable. Then from Theorem 4.3, the considered converging subsequence converges in law to the unique OU process $(\mathbf{Z}^{(1)}(t); t \quad 0)$ satisfying the SDE (4.38) with initial point $\mathbf{Z}^{(1)}(0)$. But, we know from [56, Lemma 7.7 and Theorem 7.8, page

131] that the limit of a sequence of stationary processes is stationary. Hence the law of $(\mathbf{Z}^{(1)}(t); t \quad 0)$ should be the unique law of the stationary OU process solving the SDE (4.38). This argument applies for every converging subsequence. Hence, the sequence $f(\mathbf{Z}^{(N)}(t); t \quad 0)g_{N \ 1}$ in the stationary regime converges to the unique stationary OU process solving the SDE (4.38). This completes the proof.

4.4.6 Proof of Theorem 4.7

The proof is based on Little's law [76, Theorem 4.1], Theorem 4.5, and Theorem 4.6. Let us consider a random variable $\tilde{\mathbf{S}}^{(N)}(1)$ which denotes the number of progressing jobs in the system in stationary. From Little's law, we obtain

$$(N \ ^{(N)})(1 \quad P_{block}^{(N)}) = \mathsf{E}\left[\tilde{\mathbf{S}}^{(N)}(1)\right]:$$

We now consider a random variable $\mathbf{X}^{(N)}(1) = (\mathbf{X}_n^{(N)}(1); 0 \quad n \quad C)$, where $\mathbf{X}_n^{(N)}(1)$ is the fraction of servers with at least n progressing jobs in stationary. Then we can write

$$\tilde{\mathbf{S}}^{(N)}(1) = \sum_{n=0}^{C} Nn(\mathbf{X}_n^{(N)}(1) \quad \mathbf{X}_{n+1}^{(N)}(1)):$$

Hence, we have

$$(N \ ^{(N)})(1 \quad P_{block}^{(N)}) = \mathsf{E}\left[\sum_{n=0}^{C} Nn(\mathbf{X}_n^{(N)}(1) \quad \mathbf{X}_{n+1}^{(N)}(1))\right]:$$

Therefore, we obtain

$$^{(N)}(1 \quad P_{block}^{(N)}) = \sum_{n=0}^{C} n\mathsf{E}\left[\mathbf{X}_n^{(N)}(1) \quad \mathbf{X}_{n+1}^{(N)}(1)\right]: \tag{4.65}$$

Also, from Theorem 4.5, Lemma 4.3, and Theorem 4.6, since the diffusion limit in stationarity regime has the mean vector $= \int_0^1 e^{H(\ )s} W_3(\ ) ds$ and $\limsup_{N \ ! \ 1} \mathsf{E}\left[k\mathbf{Z}^{(N)}(1)k^2\right] < 1$, we have

$$\lim_{N \ ! \ 1} \sqrt{N}\left(\mathsf{E}\left[\mathbf{X}^{(N)}(1)\right] \quad \right) = \quad :$$

Therefore

$$\mathsf{E}\left[\mathbf{X}^{(N)}(1)\right] \quad = \frac{\sqrt{}}{N} + o(N^{-\frac{1}{2}}): \tag{4.66}$$

Then from (4.65) and (4.66),

$$^{(N)}(1 \quad P_{block}^{(N)}) = \sum_{n=0}^{C} n(\ _n \quad _{n+1}) + \frac{1}{\sqrt{N}} \sum_{n=0}^{C} n(\ _n \quad _{n+1}) + o(N^{-\frac{1}{2}}): \tag{4.67}$$

But, from the stationary mean-field equations, the fixed-point satisfies

$$_0 = 1 \text{ and } \sum_{n=0}^{C} n(\ _n \quad _{n+1}) = \ (1 \quad \tfrac{d}{C}): \tag{4.68}$$

Then from (4.67) and (4.68),

$$1 \quad P_{block}^{(N)} = \frac{1}{(N)}\left[\ (1 \quad \tfrac{d}{C}) + \tilde{p}\frac{1}{N}\sum_{n=0}^{C} n(\ _n \quad _{n+1})\right] + o(N^{\ \frac{1}{2}}):$$

However, by using $^{(N)} = \quad \tilde{p}\overline{N}$ and the fact that $\tilde{p}\overline{N} < 1$, we obtain

$$1 \quad P_{block}^{(N)} = \frac{1}{}\left(1 + \tilde{p}\overline{N} + o(N^{\ \frac{1}{2}})\right)\left[\ (1 \quad \tfrac{d}{C}) + \tilde{p}\frac{1}{N}\sum_{n=0}^{C} n(\ _n \quad _{n+1})\right] + o(N^{\ \frac{1}{2}}):$$

After simple calculations, we obtain

$$P_{block}^{(N)} = \ \tfrac{d}{C} \quad \tilde{p}\frac{1}{N}\left(\sum_{n=0}^{C} n(\ _n \quad _{n+1})\right) \quad \tilde{p}\overline{N}(1 \quad \tfrac{d}{C}) + o(N^{\ \frac{1}{2}}):$$

This completes the proof.

4.4.7 Proof of Lemma 4.2

The proof relies on the quasi-monotonicity of the mean-field. Let us write the unique solution to the MFEs with the initial point $\mathbf{v}$ as $(\mathbf{y}(t;\mathbf{v});t \quad 0)$. From the quasi-monotonicity of the mean-field, we have

$$\mathbf{y}(t;\min(\mathbf{v};\ )) \quad \mathbf{y}(t;\mathbf{v}) \quad \mathbf{y}(t;\max(\mathbf{v};\ )): \tag{4.69}$$

For $\mathbf{a};\mathbf{b}\ 2\ \mathsf{U}$, let

$$k\mathbf{a} \quad \mathbf{b}k_1 = \sum_{i=0}^{C} ja_i \quad b_ij:$$

From Lemma 4 of [8], since $\min(\mathbf{v};\ )$ and $\max(\mathbf{v};\ )$, we have

$$\begin{aligned} k\mathbf{y}(t;\min(\mathbf{v};\ )) \quad k_1 \quad e^{\ tk}\min(\mathbf{v};\ ) \quad k_1; \\ k\mathbf{y}(t;\max(\mathbf{v};\ )) \quad k_1 \quad e^{\ tk}\max(\mathbf{v};\ ) \quad k_1: \end{aligned} \tag{4.70}$$

For $t \quad 0$, let us define two sets $V_+(t)$ and $V\ (t)$ as

$$V_+(t) = fi : y_i(t;\mathbf{v}) \quad _ig;$$

$$V \ (t) = f i : y_i(t; \mathbf{v}) < \ _i g:$$

Then we can write

$$k\mathbf{y}(t; \mathbf{v}) \quad k_1 = \sum_{i 2 V_+(t)} (y_i(t; \mathbf{v}) \quad _i) + \sum_{j 2 V \ (t)} (\ _j \quad y_j(t; \mathbf{v})):$$

From (4.69) and (4.70), we write

$$\begin{aligned} k\mathbf{y}(t; \mathbf{v}) \quad k_1 & \sum_{i 2 V_+(t)} (y_i(t; \max(\mathbf{v}; \ )) \quad _i) + \sum_{j 2 V \ (t)} (\ _j \quad y_j(t; \min(\mathbf{v}; \ ))); \\ & k\mathbf{y}(t; \max(\mathbf{v}; \ )) \quad k_1 + k\mathbf{y}(t; \min(\mathbf{v}; \ )) \quad k_1; \\ & e \ ^t k \max(\mathbf{v}; \ ) \quad k_1 + e \ ^t k \min(\mathbf{v}; \ ) \quad k_1; \\ = e \ ^t & k\mathbf{v} \quad k_1: \end{aligned}$$

Finally, the result follows from the fact that the norms $k \ k_1$ and $k \ k_2$ are equivalent.

4.5 Conclusions

In this chapter, we studied a FCLT that characterizes the fluctuations of the stochastic empirical process for both the transient and stationary regimes. Our analysis covers the Halfin-Whitt regime as a special case. We then used this result to quantify the error between actual blocking probabilities and asymptotic blocking probabilities.

Chapter 5

Summary and Future Research

In this dissertation, we studied the robustness and accuracy of a mean-field approach in analyzing the performance of two important classes of models arising in applications.

We mainly focused on two themes. In the first theme, we addressed the case of general JLDs for the loss and PS models. For these systems, we showed that there exists a mean-field limit that is a unique solution of a set of PDEs. Furthermore, we characterized the fixed-points of the mean-field. We have shown that the robustness measured through insensitivity extends to any occupancy based randomized load balancing policy for which the mean-field possesses a globally stable fixed-point. In the second theme, we studied the accuracy of the mean-field approximations of the loss model with exponential JLDs under the assumption that the dispatcher uses the SQ(d) load balancing policy. We studied the fluctuation of the stochastic empirical process around the mean-field limit. We then used this analysis to quantify the error between the actual blocking probability and the asymptotic blocking probability obtained as a function of the fixed-point of the mean-field.

In Chapter 2, we investigated the impact of the SQ(d) policy for the loss model that can be used to model Infrastructure-as-a-Service (IaaS) clouds. Under certain assumptions on the initial empirical random measures, we showed that there exists a mean-field. We then exploited the existence of the mean-field limit to show the asymptotic independence of any finite set of servers as $N \to 1$ when the initial states of servers are exchangeable. After that, we proved the uniqueness of the fixed-point of the mean-field. The proof uses two existing results, the form of the stationary distribution of a single server system and the uniqueness of the fixed-point in the exponential case. From the stationary distribution of a single-server loss system with a Poisson arrival process having constant state-dependent

arrival rates, the distribution of occupancy obtained from any fixed-point of the mean-field satisfies the corresponding fixed-point equations in the exponential job length distributions (JLDs) case having the same average job length. Since the exponential case has a unique fixed-point, the occupancy distribution obtained from any fixed-point must coincide with the fixed-point in the exponential case implying insensitivity. Again from the form of the stationary distribution of state-dependent loss models, the fixed-point of the mean-field corresponds to a joint distribution of occupancy and ages that satisfies a product form. We also studied the mean-field equations (MFEs) numerically for mixed-Erlang JLDs, and we observed the GAS of the fixed-point. The numerical evidence for the GAS of the fixed-point supports the hypothesis that the fixed-point approximates the stationary distribution of the state of a server.

In Chapter 3, we assumed that the servers are PS servers. For this model, we studied occupancy based randomized load balancing policies under a common framework as various parameters that include complexity and variations in the traffic level influence the choice of the routing policy in practice. We showed that all occupancy based randomized load balancing policies could be studied in the transient regime under a common framework by using mean-field techniques. We proved the existence of a mean-field limit and the corresponding MFEs correspond to the dynamics of the distribution of a single server system with a Poisson job arrival process having a rate that depends both on the instantaneous occupancy as well as the instantaneous distribution. Every occupancy based policy affects only the job arrival rate function. Furthermore, we showed that every fixed-point corresponds to a distribution on Z_+, which is a fixed-point in the exponential case having the same average job length. If the exponential case has a unique fixed-point, then the fixed-point in the general case is unique and insensitive. Although one can study occupancy based policies under a common framework, the stationary analysis depends strongly on the chosen routing policy. Hence, for a routing policy that falls into our framework, one can directly obtain the mean-field equations from our analysis, and then the existence, uniqueness, and stability of the fixed-point should be studied separately. We applied our results to four policies and observed that all of them exhibit insensitivity.

Finally, we investigated the stationary behavior of the mean-field for the SQ(d) policy by using numerical methods for the case of mixed-Erlang JLDs. We observed that the mean-field is not quasi-monotonic, but still, the fixed-point is GAS. We also showed a generic result that if a sequence of the stationary distributions of the empirical measures converges to a distribution, then any finite set of servers in the corresponding limit system

will be independent of each other if and only if the limiting measure is Dirac measure. Otherwise, servers are coupled through the position of the mean-field whose position is random. Conditioned on the position of the mean-field, any finite set of servers are independent of each other, and their distributions coincide with the position of the mean-field.

In Chapter 4, we established an FCLT to characterize the fluctuation of the empirical process around the mean-field. We showed that in the transient regime, the limiting diffusion process is an OU process whose drift and diffusion coefficients depend on the mean-field. In the stationary regime, the limiting diffusion is a stationary OU process whose drift and diffusion coefficients depend on the fixed-point of the mean-field. For the analysis of the stationary regime, we used the exponential stability of the fixed-point of the mean-field. We then used the FCLT and Little's law to show that $\lim_{N \to \infty} \sqrt{N}(P^{(N)}_{block} - \pi^d_C) \neq 0$ in the Halfin-Whitt regime and $\lim_{N \to \infty} \sqrt{N}(P^{(N)}_{block} - \pi^d_C) = 0$ if $\tilde{\alpha} = 0$. This implies $\lim_{N \to \infty} \sqrt{N} P^{(N)}_{block} = 1$ whereas for the complete resource pooling case and for the JSQ policy, we have $\lim_{N \to \infty} \sqrt{N} P^{(N)}_{block} < 1$.

5.1 Future Research

A significant open problem in Chapters 2 and 3 is, showing the GAS of the fixed-point of the mean-field. It is easier to investigate this problem for the simple case of mixed-Erlang distributions. Showing this result will help us to better understand the general case due to the fact that mixed-Erlang distributions are dense in the class of general JLDs.

Our numerical and simulation results provide evidence that the fixed-point of the mean-field is GAS and the limiting stationary distribution coincides with the fixed-point of the mean-field. Since the fixed-point corresponds to the stationary distribution of a reversible Markov process, our investigations support the hypothesis that the non-reversible Markov process which models the dynamics with a finite number of servers is converging to a reversible Markov process as $N \to \infty$. Showing this result will lend credence to the results in [63] where it was shown that for any reversible Markov process that corresponds to a mean-field model, any limit point of the stationary distribution is concentrated at fixed-points of the mean-field.

In Chapter 4, we studied the gap between $P^{(N)}_{block}$ and π^d_C as a function of N. Future work should also address the Halfin-Whitt regime for the PS model. For this case, in the limiting system as $N \to \infty$, each server is critically loaded. Hence, the average response

time of a job approaches 1 as $N \to \infty$. But for each N, the average response time is finite. Hence, it is of interest to find the rate of convergence of the average response times to 1 as a function of N when $N \to \infty$. It is of interest to see whether the randomized schemes provide a slower increase over pure random routing or not.

References

[1] "Amazon EC2," http://aws.amazon.com/ec2/.

[2] "Microsoft Azure," http://www.microsoft.com/windowsazure/.

[3] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," in *The 29th International Symposium on Computer Performance, Modeling, Measurements and Evaluation*, October 2011.

[4] V. Gupta, M. Harchol Balter, K. Sigman, and W. Whitt, "Analysis of Join-the-shortest-queue routing for web server farms," *Perform. Eval.*, vol. 64, no. 9-12, pp. 1062–1081, Oct. 2007.

[5] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14(1), pp. 181–189, 1977.

[6] R. R. Weber, "On the optimal assignment of customers to parallel servers," *J. Appl. Probability*, vol. 15, no. 2, pp. 406–413, 1978.

[7] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: an asymptotic approach," *Problems of Information Transmission*, vol. 32, no. 1, pp. 20–34, 1996.

[8] Q. Xie, X. Dong, Y. Lu, and R. Srikant, "Power of d choices for large-scale bin packing: A loss model," in *Proceedings of the 2015 ACM SIGMETRICS*, 2015, pp. 321–334.

[9] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized join-the-shortest-queue (JSQ) schemes in large heterogeneous processor sharing systems," *IEEE Transactions on Control of Network Systems*, vol. 3(2), pp. 116–126, 2016.

[10] A. Mukhopadhyay, R. R. Mazumdar, and F. Guillemin, "The power of randomized routing in heterogeneous loss systems," in *Teletra c Congress (ITC 27), 2015 27th International*, 2015, pp. 125–133.

[11] Mukhopadhyay, Arpan, "Mean field interactions in heterogeneous networks," Ph.D. dissertation, 2016. [Online]. Available: http://hdl.handle.net/10012/10253

[12] M. Mitzenmacher, "The power of two choices in randomized load balancing," *PhD Thesis, Berkeley*, 1996.

[13] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.

[14] P. Kolesar, "Stalking the endangered cat: A queueing analysis of congestion at automatic teller machines," *Interfaces*, vol. 14, no. 6, pp. 16–26, 1984.

[15] S. R. E. Turner, "Resource pooling in stochastic networks," *Ph.D. dissertation, University of Cambridge*, 1996.

[16] D. A. Dawson, *Measure-valued Markov processes*, ser. École d'Été de Probabilités de Saint-Flour XXI—1991. Berlin: Springer, 1993, vol. 1541.

[17] P. Whittle, "Partial balance and insensitivity," *J. Appl. Probab.*, vol. 22, no. 1, pp. 168–176, 1985.

[18] T. Bonald and A. Proutiere, "Insensitivity in processor sharing networks," *Performance Evaluation*, vol. 49, no. 1, pp. 193–209, 2002.

[19] A. Mukhopadhyayay, A. Karthik, R. R. Mazumdar, and F. M. Guillemin, "Mean field and propagation of chaos in multi-class heterogeneous loss models," *Performance Evaluation*, vol. 91, pp. 117–131, September 2015.

[20] C. Graham and S. Méléard, "Propagation of chaos for a fully connected loss network with alternate routing," *Stochastic Processes and their Applications*, vol. 44, no. 1, pp. 159–180, 1993.

[21] ——, "Stochastic particle approximations for generalized boltzmann models and convergence estimates," *The Annals of Probability*, vol. 28, no. 1, pp. 115–132, 1997.

[22] S. L. Brumelle, "A generalization of Erlang's loss system to state dependent arrival and service rates," *Mathematics of Operations Research*, vol. 3, no. 1, pp. 10–16, 1978.

[23] T. Bonald, M. Jonckheere, and A. Proutiére, "Insensitive load balancing," in *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '04/Performance '04. New York, NY, USA: ACM, 2004, pp. 367–377.

[24] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," in *Proceedings of ACM SIGMETRICS*, 2010, pp. 275–286.

[25] F. I. Karpelevich and A. N. Rybko, "Thermodynamic limit for the mean field model of simple symmetrical closed queueing network," *Markov Processes and Related Fields*, vol. 6, pp. 89–105, 2000.

[26] R. Aghajani and K. Ramanan, "The hydrodynamic limit of a randomized load balancing network," *Ann. Appl. Probab.*, vol. 29, no. 4, pp. 2114–2174, 08 2019.

[27] H. C. Gromoll, A. L. Puha, and R. J. Williams, "The fluid limit of a heavily loaded processor sharing queue," *Ann. Appl. Probab.*, vol. 12, no. 3, pp. 797–859, 08 2002.

[28] H. C. Gromoll, P. Robert, and B. Zwart, "Fluid limits for processor-sharing queues with impatience," *Math. Oper. Res.*, vol. 33, no. 2, pp. 375–402, May 2008.

[29] L. Decreusefond and P. Moyal, "A functional central limit theorem for the M/GI/ *1* queue," *Ann. Appl. Probab.*, vol. 18, no. 6, pp. 2156–2178, 12 2008.

[30] J. Zhang, "Fluid models of many-server queues with abandonment," *Queueing Systems*, vol. 73, no. 2, pp. 147–193, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11134-012-9307-9

[31] H. Kaspi and K. Ramanan, "Law of large numbers limits for many-server queues," *Ann. Appl. Probab.*, vol. 21, no. 1, pp. 33–114, 02 2011.

[32] C. Graham, "Chaoticity on path space for a queueing network with selection of the shortest queue among several," *Journal of Applied Probability*, vol. 37, no. 1, pp. 198–211, 2000.

[33] T. Vasantam, A. Mukhopadhyay, and R. R. Mazumdar, "Mean-field analysis of loss models with mixed-Erlang distributions under Power-of-d routing," in *2017 29th International Teletra c Congress (ITC 29)*, vol. 1, Sept 2017, pp. 250–258.

[34] ——, "Insensitivity of the mean-field limit of loss systems under SQ(d) routeing," *Advances in Applied Probability*, vol. 51, no. 4, pp. 1–40, December 2019.

[35] S. F. Yashkov and A. S. Yashkova, "Processor sharing: A survey of the mathematical theory," *Automation and Remote Control*, vol. 68, no. 9, pp. 1662–1731, 2007.

[36] L. Kleinrock, "Time-shared systems: A theoretical treatment," *J. ACM*, vol. 14, no. 2, pp. 242–261, Apr. 1967. [Online]. Available: http://doi.acm.org/10.1145/321386.321388

[37] A. Mukhopadhyay and R. R. Mazumdar, "Rate-based randomized routing in large heterogeneous processor sharing systems," in *Proceedings of 26th International Teletra c Congress (ITC 26)*, 2014.

[38] M. Bramson, "Stability of join the shortest queue networks," *Ann. Appl. Probab.*, vol. 21, no. 4, pp. 1568–1625, 08 2011.

[39] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," in *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '10. New York, NY, USA: ACM, 2010, pp. 275–286.

[40] ——, "Asymptotic independence of queues under randomized load balancing," *Queueing Systems*, vol. 71, no. 3, pp. 247–292, 2012.

[41] T. Vasantam, A. Mukhopadhyay, and R. R. Mazumdar, "The mean-field behavior of processor sharing systems with general job lengths under the SQ(d) policy," *Performance Evaluation*, vol. 127-128, pp. 120 – 153, 2018.

[42] T. Vasantam and R. Mazumdar, "On occupancy based randomized routing schemes in large systems of shared servers," *2018 30th International Teletra c Congress (ITC 30)*, vol. 01, pp. 28–36, 2018.

[43] P. Gazdzicki, I. Lambadaris, and R. Mazumdar, "Blocking probabilities for large multi-rate Erlang loss systems," *Adv.Appl.Prob.*, vol. 25, pp. 997–1009, 1993.

[44] J. van Leeuwaarden, B. Mathijsen, and B. Zwart, "Economies-of-scale in many-server queueing systems: Tutorial and partial review of the qed halfin–whitt heavy-traffic regime," *SIAM Review*, vol. 61, no. 3, pp. 403–440, 2019. [Online]. Available: https://doi.org/10.1137/17M1133944

[45] W. Whitt, "Heavy-traffic approximations for service systems with blocking," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 5, pp. 689–708, May 1984.

[46] D. Mukherjee, S. Borst, J. S. H. van Leeuwaarden, and P. Whiting, "Asymptotic optimality of Power-of-*d* load balancing in large-scale systems," *CoRR*, 12/2016. [Online]. Available: http://arxiv.org/abs/1612.00722

[47] C. Graham, "Functional central limit theorems for a large network in which customers join the shortest of several queues," *Probability Theory and Related Fields*, vol. 131, no. 1, pp. 97–120, Jan 2005.

[48] P. J. Hunt, "Loss networks under diverse routing: The symmetric star network," *Advances in Applied Probability*, vol. 27, no. 1, pp. 255–272, 1995.

[49] L. Ying, "On the approximation error of mean-field models," in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, ser. SIGMETRICS '16. New York, NY, USA: ACM, 2016, pp. 285–297.

[50] N. Gast and B. Van Houdt, "A refined mean field approximation," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, pp. 33:1–33:28, Dec. 2017.

[51] N. Gast, "Expected values estimated via mean-field approximation are 1/n-accurate," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 1, pp. 17:1–17:26, Jun. 2017. [Online]. Available: http://doi.acm.org/10.1145/3084454

[52] P. Eschenfeldt and D. Gamarnik, "Join the shortest queue with many servers. The heavy-traffic asymptotics," *Math. Oper. Res.*

[53] D. Mukherjee, S. Borst, J. van Leeuwaarden, and P. Whiting, "Universality of Power-of-d load balancing in many-server systems," *Stochastic Systems*, vol. 8, no. 4, pp. 265–292, 2018.

[54] P. Eschenfeldt and D. Gamarnik, "Supermarket queueing system in the heavy traffic regime. short queue dynamics," 2016.

[55] T. Vasantam and R. R. Mazumdar, "Fluctuations around the mean-field for a large scale Erlang loss system under the SQ(d) load balancing, *accepted to Teletra c Congress (ITC 31), 2019 31st International*," 2019.

[56] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. John Wiley and Sons Ltd, 1985.

[57] P. Billingsley, *Convergence of probability measures*, 2nd ed., ser. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1999, a Wiley-Interscience Publication.

[58] T. G. Kurtz, "Solutions of ordinary differential equations as limits of pure jump markov processes," *Journal of Applied Probability*, vol. 7, no. 1, p. 4958, 1970.

[59] A.-S. Sznitman, "Topics in propagation of chaos," in *Ecole d'Ete de Probabilites de Saint-Flour XIX | 1989*, P.-L. Hennequin, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 165–251.

[60] J.-w. Cho, J.-Y. Le Boudec, and Y. Jiang, "On the Asymptotic Validity of the Decoupling Assumption for Analyzing 802.11 MAC Protocol," *arXiv e-prints*, p. arXiv:1106.6328, Jun 2011.

[61] M. Benaim and J.-Y. Le Boudec, "On mean field convergence and stationary regime," 11 2011.

[62] M. Benam and J.-Y. L. Boudec, "A class of mean field interaction models for computer and communication systems," *Performance Evaluation*, vol. 65, no. 11, pp. 823 – 838, 2008, performance Evaluation Methodologies and Tools: Selected Papers from ValueTools 2007.

[63] J.-Y. L. Boudec, "The stationary behaviour of fluid limits of reversible processes is concentrated on stationary points," p. 529, 2013. [Online]. Available: http://aimsciences.org//article/id/ad4efa6f-9312-4213-a0c5-216623b07682

[64] N. Gast, L. Bortolussi, and M. Tribastone, "Size expansions of mean field approximation: Transient and steady-state analysis," *Performance Evaluation*, vol. 129, pp. 60 – 80, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166531618302633

[65] L. Bortolussi and N. Gast, "Mean field approximation of uncertain stochastic models," in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, June 2016, pp. 287–298.

[66] A. Mukhopadhyay, R. R. Mazumdar, and R. Roy, "Binary opinion dynamics with biased agents and agents with different degrees of stubbornness," in *2016 28th International Teletra c Congress (ITC 28)*, vol. 01, Sep. 2016, pp. 261–269.

[67] Q.-L. Li and C. Lin, "The M/G/1 processor-sharing queue with disasters," *Computers & Mathematics with Applications*, vol. 51, no. 6, pp. 987 – 998, 2006.

[68] B. A. Sevastyanov, "An ergodic theorem for markov processes and its application to telephone systems with refusals," *Theory of Probability & Its Applications*, vol. 2, no. 1, pp. 104–112, 1957.

[69] W. Rudin, *Real and complex analysis*, 3rd ed. McGraw-Hill Book Co., New York, 1987.

[70] J. Gärtner, "On the mckean-vlasov limit for interacting diffusions," *Mathematische Nachrichten*, vol. 137, no. 1, pp. 197–248, 1988.

[71] K. Oelschlager, "A martingale approach to the law of large numbers for weakly interacting stochastic processes," *Ann. Probab.*, vol. 12, no. 2, pp. 458–479, 05 1984.

[72] P. M. Kotelenez and T. G. Kurtz, "Macroscopic limits for stochastic partial differential equations of mckean–vlasov type," *Probability Theory and Related Fields*, vol. 146, no. 1, p. 189, Dec 2008.

[73] V. N. Kolokoltsov, *Nonlinear Markov Processes and Kinetic Equations*, ser. Cambridge Tracts in Mathematics. Cambridge University Press, 2010.

[74] A. Jakubowski, "On the skorokhod topology," *Annales de l'I.H.P. Probabilites et Statistiques*, vol. 22, no. 3, pp. 263–285, 1986.

[75] O. Kallenberg, *Random measures*. Akademie-Verlag, 1983.

[76] S. Asmussen, *Applied Probability and Queues*, ser. Stochastic Modelling and Applied Probability. Springer, New York, 2003, vol. 51.

[77] S. M. Ross, *Introduction to Probability Models*. Academic Press; 10th edition, 2009.

[78] P. Robert, *Stochastic Networks and Queues*, ser. Stochastic Modelling and Applied Probability Series. Springer-Verlag, 2003.

[79] L. Decreusefond and P. M., *Stochastic Modeling and Analysis of Telecom Networks*, ser. ISTE, London, 2012.

[80] V. Varadarajan, "On a theorem of F. Riesz concerning the form of linear functionals," *Fund. Math.*, vol. 46, pp. 209–220, 1959).

[81] S. F. Yashkov, "A note on application of the method of supplementary variables to the analysis of a processor sharing system," *Automation and Remote Control*, vol. 69, no. 9, p. 1622, 2008.

[82] K. Deimling, *Ordinary di erential equations in Banach spaces*, ser. Lecture notes in mathematics. Springer-Verlag Berlin Heidelberg, 1977, vol. 596.

[83] G. Pang, R. Talreja, and W. Whitt, "Martingale proofs of many-server heavy-traffic limits for markovian queues," *Probab. Surveys*, vol. 4, pp. 193–267, 2007.

[84] W. Ledermann and G. E. H. Reuter, "Spectral theory for the differential equations of simple birth and death processes," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 246, no. 914, pp. 321–369, 1954.

[85] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, 2nd ed., ser. Graduate texts in mathematics; 113. Springer-Verlag, New York, 1991, vol. 113.

Appendix A

Some Key Background Results

In this Appendix, we collect some background material related to results that are frequently used in the dissertation. Consider a metric space $(\mathsf{U}; d_{\mathsf{U}})$. Let $\mathsf{B}(\mathsf{U})$ be the Borel -algebra of U and $\mathsf{M}_1(\mathsf{U})$ is the space of Borel probability measures on U.

A.1 Weak Convergence and Prohorov's Theorem

We first define the concept of weak convergence of probability measures.

Definition A.1. *([56], p.107): For a metric space $(\mathsf{U}; d_{\mathsf{U}})$, a sequence of probability measure in $\{P_n\}_{n \ 1}$ in $\mathsf{M}_1(\mathsf{U})$ is said to be weakly convergent to a probability measure $P \ 2 \ \mathsf{M}_1(\mathsf{U})$ if*

$$\lim_{n \ ! \ 1} \int_{u2\mathsf{U}} f(u)\, dP_n(u) = \int_{u2\mathsf{U}} f(u)\, dP(u); \tag{A.1}$$

for all $f \ 2 \ \mathsf{C}_b(\mathsf{U})$.

If $(\mathsf{U}; d_{\mathsf{U}})$ is separable, then the weak convergence of a sequence of probability measures is equivalent to the convergence w.r.t. a metric known as Prohorov metric. We next define the Prohorov metric.

Definition A.2. *([56], p.96): For P, $Q \ 2 \ \mathsf{M}_1(\mathsf{U})$, the Prohorov metric $_p(\ ;\ )$ is de ned as*

$$_p(P; Q) = \inf\{ \ > 0 : P(A) \quad Q(A \ ) + \ ; 8 A \ 2 \ \mathsf{C}\}; \tag{A.2}$$

where C is the collection of all closed subsets of U and

$$A \ = \{u \ 2 \ \mathsf{U} : \inf_{y2A} d_{\mathsf{U}}(y; u) < \ \}. \tag{A.3}$$

The next result shows that if $(\mathsf{U}; d_{\mathsf{U}})$ is separable and complete, then $(\mathsf{M}_1(\mathsf{U}); \ _p)$ is also separable and complete.

Theorem A.1. *([56], Theorem 1.7, p.101): If $(\mathsf{U}; d_{\mathsf{U}})$ is separable, then $(\mathsf{M}_1(\mathsf{U}); \ _p)$ is separable. In addition, if $(\mathsf{U}; d_{\mathsf{U}})$ is complete, then $(\mathsf{M}_1(\mathsf{U}); \ _p)$ is complete.*

We now define the concept of tightness of a set of probability measures that relates to the compactness of the set of measures.

Definition A.3. *([56], p.103): A probability measure $P \ 2 \ \mathsf{M}_1(\mathsf{U})$ is said to be tight if for every $\ > 0$, there exists a compact set $A \quad \mathsf{U}$ such that $P(A) \quad 1 \quad$. A family of probability measures $M \quad \mathsf{M}_1(\mathsf{U})$ is tight, if for every $\ > 0$, there exists a compact set A in U such that $\inf_{Q2M} Q(A) \quad 1 \quad$.*

We next see conditions on $(\mathbb{U}; d_{\mathbb{U}})$ that guarantee the tightness of a set of probability measures.

Lemma A.1. *([56], Lemma 2.1, p.104) If $(\mathbb{U}; d_{\mathbb{U}})$ is complete and separable, then every P 2 $M_1(\mathbb{U})$ is tight.*

We now state an important theorem known as Prohorov's theorem from [56, Theorem 2.2, p.104]. For a given topological space X, a set B X is said to be relatively compact if the closure of B is compact.

Theorem A.2. *Prohorov's Theorem: Let $(\mathbb{U}; d_{\mathbb{U}})$ be complete and separable. If M $M_1(\mathbb{U})$, then M is tight if and only if M is relatively compact.*

Finally, we state the result on the equivalence of the weak convergence of a sequence of probability measures and their convergence w.r.t. Prohorov metric.

Theorem A.3. *([56], Theorem 3.1, p.108): If $(\mathbb{U}; d_{\mathbb{U}})$ is separable, then for a sequence of probability measures $\{P_n\}_{n \ 1}$ in $M_1(\mathbb{U})$ and P 2 $M_1(\mathbb{U})$, we have P_n) P if and only if* $\lim_{n \ ! \ 1} \ _p(P_n; P) = 0$.

A.2 Riesz Markov Kakutani Theorem

Theorem A.4. *([69], Theorem 2.14): Let X be a locally compact Hausdro space and let be a positive linear functional on $C_s(X)$, then there exists a unique regular measure on X such that*

$$\Lambda f = \int_{x2X} f(x)\, d\ (x): \tag{A.4}$$

A.3 Jakubowski's Criteria

Theorem A.5. *([74], Theorem 4.6): Let $(\mathbb{U}; d_{\mathbb{U}})$ be a completely regular topological space. Then a sequence of stochastic processes $\{\mathbf{X}^N\}_{N \ 1}$ of $D_{M_1(\mathbb{U})}([0; 1))$-valued random elements de ned on $(\Omega; \mathbb{F}; \mathbb{P})$ is tight if and only if the following two conditions are satis ed:*

J1: For each $T > 0$ and > 0, there exists a compact set $K_{T;}$ $M_1(\mathbb{U})$ such that

$$\liminf_{N \ ! \ 1} \mathbb{P}(\mathbf{X}^N_t \ 2 \ K_{T;} \ ; 8\, t \ 2 \ [0; T]) > 1 \quad : \tag{A.5}$$

This condition is called the compact-containment condition.

J2: There exists a family $\mathcal{Q}$ *of real valued continuous functions* F *de ned on* $\mathcal{M}_1(\mathcal{U})$ *that separates points in* $\mathcal{M}_1(\mathcal{U})$ *and is closed under addition such that for every* $F \in \mathcal{Q}$, *the sequence* $\{F(\mathbf{X}_t^N); t \geq 0\}_{N \geq 1}$ *is tight in* $\mathcal{D}_{\mathbb{R}}([0, 1))$.

To prove the condition J2, we define a class of functions $\mathcal{Q}$ as follows:

$$\mathcal{Q} := \{F : \exists f \in C_b^1(\mathcal{U}) \text{ such that } F(\mu) = \langle f, \mu \rangle, \forall \mu \in \mathcal{M}_1(\mathcal{U})\}. \tag{A.6}$$

It can be seen that every function $F \in \mathcal{Q}$ is continuous w.r.t. the weak topology on $\mathcal{M}_1(\mathcal{U})$ and furthermore, the class of functions $\mathcal{Q}$ separates points in $\mathcal{M}_1(\mathcal{U})$ and is also closed under addition. We next recall the following result (From Theorem C.9, [78]) that is sufficient to prove the condition J2.

Theorem A.6. *Tightness in* $\mathcal{D}_{\mathbb{R}}([0, T])$: *Let* $\{P_n\}_{n \geq 1}$ *be a sequence of probability distributions on* $\mathcal{D}_{\mathbb{R}}([0, T])$, *then* $\{P_n\}_{n \geq 1}$ *is tight if for any* $\eta > 0$,

C1: There exists $b > 0$ *such that*

$$P_n(|X(0)| > b) \tag{A.7}$$

for all $n \in \mathbb{Z}_+$.

C2: For any $\eta > 0$, *there exists* $\delta > 0$ *such that*

$$P_n(w_X(\delta) > \eta) \tag{A.8}$$

for n *su ciently large, where*

$$w_X(\delta) = \sup\{|X(t) - X(s)| : s, t \leq T, |s - t| < \delta\} \tag{A.9}$$

and any limiting point P *satis es* $P(\mathcal{C}_{\mathbb{R}}([0, T])) = 1$.

A.4 Gronwall's Inequality

Theorem A.7. *([56], Theorem 5.1, p.498): Let* f *be a function on* $[0, 1)$ *that is bounded on bounded intervals. For* $\beta \geq 0$ *and* $K > 0$, *if*

$$0 \leq f(t) \leq \beta + K \int_{s=0}^{t} f(s)\, ds, \tag{A.10}$$

then

$$f(t) \leq \beta e^{Kt}. \tag{A.11}$$

A.5 Doob's Inequality

Theorem A.8. *([56], Corollary 2.17, p.64): Let* $(\mathbf{X}(t); t \quad 0)$ *be a right continuous martingale. Then for* > 1 *and* $T > 0$,

$$\mathsf{E}\left[\sup_{t \quad T} j\mathbf{X}(t)j \quad\right] \quad \left(\frac{\quad}{1}\right) \mathsf{E}\left[j\mathbf{X}(T)j \quad\right]: \tag{A.12}$$