Topics in Modeling and Analysis of Low-Latency Systems

by

Samira Ghanbarian

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, 2023

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: Douglas Down
Professor, Dept. of Computing and Software
McMaster University

Supervisor(s): Ravi Mazumdar
Professor, Dept. of Electrical and Computer Engineering
University of Waterloo

Internal Member: Patrick Mitran
Professor, Dept. of Electrical and Computer Engineering
Andrew Heunis
Adjunct Professor, Dept. of Electrical and Computer Engineering
University of Waterloo

Internal-External Member: Hossein Abouee Mehrizi
Professor, Dept. of Management Sciences, University of Waterloo

Author's Declaration

Abstract

Cloud-based architectures have become integral elements of modern networking infrastructure and are characterized by a large number of servers operating in parallel. Optimizing performance in these systems, with a particular focus on specific metrics such as system response time and the probability of loss, is critical to ensure user satisfaction. To address this challenge, this thesis analyzes load balancing policies that are designed to efficiently assign incoming user requests to the servers such that the system performance is optimized. In particular, the thesis focuses on a specialized category known as ”randomized dynamic load balancing policies”. These policies optimize system performance by dynamically adapting assignment decisions based on the current state of the system while interacting with a randomly selected subset of servers. Given the complex interdependencies among servers and the large size of these systems, an exact analysis of these systems is intractable. Consequently, the thesis studies these systems in the system size limit. It employs relevant limit theorems, including mean-field techniques and Stein's approach, as crucial mathematical tools. Furthermore, the thesis evaluates the accuracy of these limits when applied to systems of finite size, providing valuable insights into the practical applicability of the proposed load balancing policies.

Motivated by different types of user requests or jobs, the thesis focuses on two main job categories: single-server jobs which can only run on a single server to represent non-parallelizable requests, and multiserver jobs, which can run on multiple servers simultaneously modeling parallelizable requests.

The first part of the thesis studies single-server jobs in a system comprising a large number of processor sharing servers operating in parallel, where servers have different processing speeds and unlimited queueing buffers. The objective is to design randomized load balancing policies that minimize the average response time of jobs. A novel policy is introduced that allocates incoming jobs to servers based on predefined thresholds, state information from a randomly sampled subset of servers, and their processing speeds. The policy subsumes a broad class of other load balancing policies by adjusting the threshold levels, offering a unified framework for concurrent analysis of multiple load balancing policies. It is shown that under this policy, the system achieves the maximal stability region. Moreover, it is shown that as the system size approaches infinity, the transient and stationary stochastic occupancy measure of the system converges to a deterministic mean-field limit and the unique fixed point of this mean-field limit, respectively. As a result, the study of the asymptotic average response time of jobs becomes feasible through the fixed point of the mean-field limit. The analysis continues by studying error estimation related to asymptotic values in finite-sized systems. It is shown that when the mean delay of the finite-size system is approximated by its asymptotic value, the error is proportional to the inverse square root of the system size.

Subsequently, the thesis analyzes adaptive multiserver jobs in loss systems, where they can be parallelized across a variable number of servers, up to a maximum degree of parallelization. In loss systems, each server can process only a finite number of jobs simultaneously and blocks any additional jobs beyond this capacity. Therefore, the goal is to devise randomized job assignment schemes that optimize the average response time of accepted jobs and the blocking probability while interacting with a sampled subset of servers. A load balancing policy is proposed, where the number of allocated servers for processing each job depends on the state information of a randomly sampled subset of servers and the maximum degree of parallelization. Employing Stein's method, it is shown that, provided that the sampling size grows at an appropriate rate, the difference between the steady-state system and a suitable deterministic system that exhibits optimality, decreases to zero as the system size increases. Thus, as the system size approaches infinity, the steady-state system achieves a zero blocking probability and optimal average response time for accepted jobs. Additionally, the thesis analyzes error estimation for these asymptotic values in finite-sized systems and establishes the error bounds as a function of the number of servers in the system.

Acknowledgements

First and foremost, I would like to extend my heartfelt appreciation to my supervisor, Professor Ravi Mazumdar. His unwavering commitment to excellence and his guidance throughout my studies have been nothing short of inspiring. Professor Mazumdar's insights and mentorship have been invaluable, shaping my research and academic growth.

I am also deeply grateful to Professor Arpan Mukhopadhyay and Dr. Fabrice Guilleman for their outstanding collaboration in the development of this thesis, particularly in chapters 4 and 5. The depth of our discussions, their continuous support, and the meaningful questions they raised were not only intellectually stimulating but also integral to the development of my research.

I would like to acknowledge Professor Thirupathaiah Vasantam for his early guidance, which set the foundation for my research. His contributions and guidance were crucial to the success of this thesis. Additionally, my fellow graduate student, Dr. Ryan Kinnear, has been a reliable source of intellectual exchange, and our discussions have greatly enriched my understanding of the subject matter.

My heartfelt appreciation extends to my committee members: Professor Patrick Mitran, Professor Andrew Heunis, and Professor Hossein Abouee Mehrizi. Their meticulous review of my thesis and their constructive feedback have been instrumental in refining the quality of this work.

I am profoundly grateful to the close-knit community of fellow students and friends at the University of Waterloo who have made this academic journey not only productive but also enjoyable. Their camaraderie, shared experiences, and ongoing support have been an integral part of my personal and academic growth.

To my parents, your unwavering support, encouragement, and belief in my abilities have been a driving force behind my accomplishments. Your sacrifices have not gone unnoticed, and I am eternally grateful for your love and guidance.

Lastly, I want to express my deepest gratitude to my husband Miad, for his understanding, unwavering support, and for keeping me grounded and sane during the challenging years of my doctoral studies. Your presence was my pillar of strength.

This thesis would not have been possible without the collective support of these exceptional individuals. I am profoundly thankful to each and every one of you for your contributions to my academic journey.

Dedication

This thesis is dedicated to my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cloud-based architectures have become an essential component of modern networking, offering a versatile framework for a wide array of applications. These applications span from non-parallelizable tasks like sequential machine learning in areas such as text processing, speech analysis, and financial analysis [1, 2], to parallelizable tasks exemplified by Google's Borg system [3], the MapReduce framework [4], TensorFlow [5], and erasure codes [6]. With the increasing scale and complexity of these applications (jobs), computing clusters often consist of hundreds of thousands of servers. Service providers who excel in delivering robust system performance, particularly in terms of minimizing latency and reducing the probability of loss, achieve a remarkable competitive advantage by ensuring user satisfaction [7, 8]. A central question to be addressed is how to effectively schedule the incoming load and network traffic in these large and complex systems to optimize specific performance metrics.

A promising class of solutions includes load balancing policies that try to distribute the incoming workload across servers. These policies determine which servers an incoming job should join based on a predefined assignment scheme. Load balancing policies can be categorized as static, dynamic, or combining these two, depending on the level of system information they use. Static load balancing policies assign jobs to servers following rules that are oblivious to the system's state; however, dynamic load balancing policies adopt their job assignment scheme based on the system's state at job arrival instants. Generally, there exists a trade-off between the amount of information a policy uses and the system's performance. In other words, dynamic load balancing policies exploit the extra data from the system to make more informed decisions for routing incoming tasks, leading to improved performance.

Ideally, a dynamic load balancing policy that has knowledge of the entire system's state should achieve optimal system performance. This arises from the ability to compare the states of all servers in the system and select the ones with the quickest processing time as destination servers. However, in real systems, it is not always possible for a job to access the state information of all servers, due to the high communication overhead involved. For instance, consider the well-known load balancing policy, the Join-the-Shortest-Queue (JSQ) policy. Under the JSQ policy, each incoming task requires access to the state information of all servers to join the server with the fewest jobs waiting in its queue [9, 10]. However, in a system with n servers where jobs arrive at the system at a rate proportional to the number of servers, $O(n)$, this policy requires a message rate of $O(n)$ messages per unit of time. This messaging overhead can quickly become prohibitive and make the policy expensive, or even impossible to implement in large-scale systems.

To account for the communication overheads and implementation complexities of load balancing policies, it is imperative to concurrently evaluate both the policy's performance and its scalability. Considering this fact, *randomized* dynamic load balancing policies offer a practical solution. In these policies, upon arrival of a job, only a random subset of servers is accessible to incoming jobs, and the assignment scheme is based on the state information of the sampled servers. Incorporating random subset access significantly reduces the communication overhead, while still leveraging dynamic system information to achieve almost as good performance. In this dissertation, we focus on the design and analysis of randomized dynamic load balancing policies with the objective of optimizing system performance in terms of specific metrics of interest in large-scale networks.

The design of randomized load balancing policies relies on the specific nature of jobs in the system, which can be broadly categorized into two main types, single-server jobs and multiserver jobs. Single-server jobs are employed to model non-parallelizable applications. These jobs are limited to running on a single server, determined by the job assignment scheme at their arrival instant. Once assigned to a server, the job remains there for the entire duration of its processing. Figure 1.1 illustrates a schematic of such systems, featuring a single dispatcher responsible for directing jobs to servers following a specific load balancing policy and highlighting three parallel servers. In this particular scenario, the JSQ policy is employed and the incoming job is assigned to server 2 which currently has the fewest number of jobs waiting in its queue.

On the other hand, multiserver jobs are employed to represent parallelizable applications. In systems with multiserver jobs, each job can run on multiple servers simultaneously. The selection of destination servers, which can be more than one, is determined by the job assignment scheme at the time of job arrival. Jobs are divided into smaller parts referred to as sub-jobs or tasks, with each sub-job being allocated to one of the destination servers.

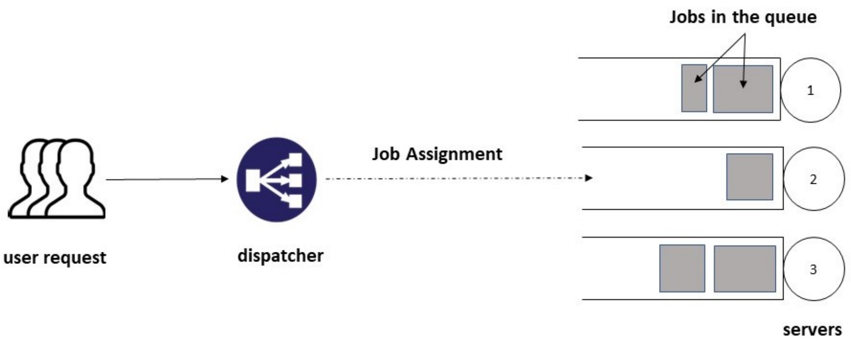


Figure 1.1: Schematic of a system with a single dispatcher, 3 parallel servers, and single-server jobs.

Once sub-jobs join the servers, they remain on their respective server until their processing is complete. Figure 1.2 illustrates a schematic of such systems, featuring a single dispatcher and three servers. In this schematic, the incoming job is assigned to servers 2 and 3 based on a predefined rule.

This dissertation focuses on the development and performance evaluation of randomized dynamic load balancing policies for single-server and multiserver jobs in large-scale systems. A drawback inherent in dynamic randomized schemes is the new interactions they introduce among system components. In particular, using system state information to assign incoming jobs leads to interdependencies among servers. This complexity, combined with the system's large size, makes the precise analysis of such networks challenging. For Markovian load balancing policies, which make assignments based on only the *current* state of the system, the time evolution of the system can often be described through a Markov process. Consequently, a comprehensive study of the system requires an analysis of the underlying Markov process. In cases where the system size is large, it becomes feasible to study the system's asymptotic behavior through the application of limit theorems suitable for Markov processes. Two particularly relevant techniques for this purpose are the mean-field technique and Stein's method. These methods show the convergence of the underlying Markov process to an appropriate process in transient and stationary regimes, respectively.

By applying the mean-field technique and Stein's method, this thesis shows the effectiveness of randomized load balancing policies in achieving excellent system performance with minimal communication overhead as the system size approaches infinity. These results provide strong approximations for large systems but are estimates for finite-sized systems.

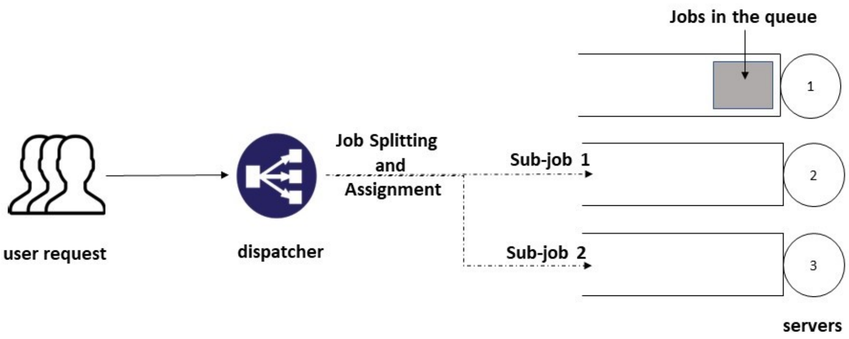


Figure 1.2: Schematic of a system with a single dispatcher, 3 parallel servers, and multi-server jobs, where the job is split into two sub-jobs and joins servers 2 and 3.

While prior research typically ends here, an important question remains: how accurate are these approximations when assessing performance for large but finite systems? In this dissertation, we not only develop new policies and evaluate their asymptotic performance but also study the accuracy of these approximations in relation to the system size, offering practical insights for real-world, large-scale systems.

1.1 Two Key Models

In this dissertation, we study two key models suitable for addressing single-server jobs and multiserver jobs. Our primary objective is to study job assignment schemes for large-scale systems and evaluate their performance in the context of finite-sized systems. We achieve this by characterizing each model by an appropriate Markov process. It is important to highlight that addressing the distinct challenges presented by each model necessitates the use of different methodologies.

1.1.1 Processor Sharing Systems with Single-Server Jobs

Many jobs inherently resist parallelization and must be executed on single servers. A practical design approach to reduce the processing time of these jobs (i.e., the time it takes for a server to process a job after it joins the server's queue) involves increasing the capacity of individual servers and efficiently distributing this capacity among all concurrently processed jobs by that server. A conceptually straightforward yet highly effective resource-sharing approach is the round-robin scheduling technique. In this approach, each job is allocated a specific time slot to use the available server capacity. If a job is not completed in its allocated time slot, it gets temporarily interrupted but retains its position in the queue to continue work in the next allocated time slot. As a result, smaller tasks do not become delayed by larger ones before they start processing, as they are given their turn for processing. As the time slot duration in the round-robin fashion approaches zero, it can be closely approximated by the Processor Sharing (PS) service discipline [11, 12, 13]. Consequently, processor sharing servers are appropriate mathematical models for practical systems employing the round-robin approach.

In the processor sharing service discipline, once a job joins a server's queue, it immediately starts processing, and the server's processing capacity is distributed proportionally among all jobs in its queue. For example, if a server with a unit speed has two jobs in its queue, each job is allocated a processing speed of 1/2. Furthermore, we assume each server has an infinite queue, ensuring that every incoming job is accepted into the system without rejection. Since job loss is not a concern in this system, and jobs receive variable processing speeds based on the total number of jobs in the server's queue (referred to as its occupancy), the primary objective is to develop effective load balancing policies that minimize the occupancy of each server. This optimization ensures that each job receives the maximum processing speed, thus minimizing the average response time in the system (the time elapsed between a job's arrival and its departure from the system).

To achieve this objective, we propose a novel randomized load balancing policy determined by predefined thresholds. This policy randomly samples a finite number of servers and adjusts its behavior depending on whether the sampled servers' occupancies are below or above the specified thresholds. The ability to control these threshold levels in the policy allows for the design of a wide range of load balancing policies, both existing and new. Consequently, this approach enables the comprehensive study of different load balancing policies within a unified and systematic framework.

When the system size is large and the threshold-based load balancing policy is implemented, the interaction between each individual server and the rest of the system becomes weak. This characteristic allows for the application of the mean-field technique to analyze the underlying Markov process describing the system. The technique is based on the Law of Large Numbers (LLN) and is briefly explained in Section 1.5.1. As the system size approaches infinity, the Markov process converges to a deterministic process known as its mean-field limit. This limit serves as a good approximation for the transient behavior of the large-scale system. Importantly, when the equilibrium points of this mean-field limit are globally asymptotically stable, they effectively describe the system's stationary distribution. This property facilitates a detailed analysis of the system's steady-state performance.

Considering that the aforementioned results are based on limits and asymptotic conditions, it is imperative to evaluate the accuracy of mean-field approximations for systems of finite size. To measure this gap, we study the fluctuations in the empirical distribution of the system around its established mean-field limit, in both transient and stationary regimes. By applying Functional Central Limit Theorems (FCLT), we quantify the accuracy of these estimations as a function of the system's size.

1.1.2 Loss Systems with Multiserver Jobs

As discussed in the preceding subsection, system designers have effectively utilized the power of high-speed computation by continuously improving hardware capabilities and expanding server capacities. While this approach has undoubtedly reduced response times for job requests, it has also highlighted a fundamental limitation: the processing time of an individual job is constrained by the capacity of a single server and surpassing this limitation is unattainable. This constraint becomes particularly apparent when we consider multiserver jobs that can be parallelized across multiple servers. This paradigm shift underscores the importance of exploiting the inherent parallelism of jobs whenever possible to achieve better system performance.

Multiserver jobs, amenable to parallelization, consist of smaller constituent components known as sub-jobs or tasks. Each of these sub-jobs is executed on separate servers, running in parallel with other sub-jobs. These jobs are distinguished into two primary categories, rigid and adaptive, based on how their sub-job quantity and server requirements depend on the dynamic state of the system. Rigid multiserver jobs maintain a fixed and predetermined number of sub-jobs, irrespective of the current system state [14]. In contrast, adaptive multiserver jobs exhibit flexibility, adjusting the number of sub-jobs based on the system state and available resources. This adaptability can significantly reduce unnecessary delays, unlike rigid jobs, which may experience stalls when resource requirements are unmet.

Within the category of adaptive multiserver jobs, a further distinction is made between moldable and malleable jobs. In moldable jobs, once job processing commences, the resource allocation remains constant throughout the job's execution [15, 16]. Conversely, malleable jobs allow for resource allocation modifications while the job is actively running [17, 18]. As a result, designing load balancing policies for moldable jobs is relatively simpler, requiring resource allocation decisions only at job arrival times, while still delivering high-quality performance through adaptability rooted in the system's state for determining the number of sub-jobs.

We study the behavior of moldable multiserver jobs in an Erlang loss system as our mathematical framework, aiming to comprehend the impact of parallelism on system performance. In the loss model, servers can concurrently process a finite number of jobs at a constant rate and there is no queueing buffer in the system. Consequently, arriving jobs that cannot find an available server immediately are blocked. In loss systems, jobs confront two distinct outcomes, they are either blocked and discarded from the system permanently, or admitted into the system and start their service immediately. Traditionally, in loss systems with single-server or rigid multiserver jobs, the primary focus was on minimizing the blocking probability, given that the response time of an accepted job remained constant and equal to the service rate of a single-server. However, with the advent of adaptive multiserver jobs, characterized by their ability to adaptively parallelize and accommodate varying job response times, the objective is twofold: designing load balancing policies that simultaneously reduce both the blocking probability and average response time.

To achieve this objective, we propose a randomized load balancing policy. This policy involves randomly selecting a subset of servers and constraining each job's subdivision to a specific threshold determined by the maximum degree of parallelism. This limitation arises from the realization that attempting to parallelize a job beyond this degree can substantially impair the job's response time due to the excessive communication overhead involved. Furthermore, not all jobs possess the inherent characteristics to be subdivided into an arbitrary number of sub-jobs. By comparing the number of available resources in the sampled set against this threshold, the policy determines the degree of parallelism for each job, indicating the number of parallel servers allocated for its execution.

The introduction of job splitting into the load balancing policy gives rise to strong coupling among servers, requiring the utilization of analytical techniques other than the mean-field technique. One relevant technique in this context is Stein's method, which is briefly outlined in Section 1.5.2. Stein's method quantifies the distance between the steady-state distribution of the underlying Markov process and an ideal target distribution that we aim for the system to converge toward. Showing that this distance diminishes as the system size approaches infinity enables the establishment of bounds on steady-state system performance as a function of the system size. Consequently, this method not only provides performance approximations for large steady-state systems but also allows for the derivation of error bounds for these approximations.

1.2 Related Literature

We have structured our literature review into three primary categories: load balancing policies for single-server jobs, load balancing policies for multiserver jobs, and the evaluation of these policies' performance in finite-sized systems.

Load Balancing for Single-Server Jobs The first works of load balancing problems go back to [19] and [20] in which the JSQ policy was considered for a system of two identical First-Come First-Serve (FCFS) servers. The stationary distribution of server occupancy was established for each server. Subsequently, the optimality of the JSQ policy was studied for a two-server system in [21], and for an n-server system in [9]. It was shown that this policy minimizes the expected response time of jobs. The JSQ policy was also studied under the processor sharing service discipline in [10] for general job length distributions. The queue length distribution at each server was approximated and through simulations, it was shown that the stationary distribution is near insensitive to the distribution of job size as long as the mean remains the same.

To address the problem of high communication overhead associated with the JSQ policy, randomized JSQ(d) policy was proposed in seminal works [22] and [23], where every arrival randomly sampled a finite number of servers, $d \leq n$, in a system of n servers and joined the shortest queue among the sampled servers. A system of n FCFS servers with Poisson job arrivals and exponential job sizes was studied in the limit (as $n \to \infty$) using mean-field techniques. It was shown that in the case $d = 2$, the stationary queue sizes decay doubly exponentially which is an exponential improvement over the case $d = 1$, whereas $d = 3$ only has a constant factor improvement over the case $d = 2$. This phenomenon is known as the "*the power of two choices*".

Generalization of the JSQ(d) problem to queueing networks with general service time distributions and different service disciplines, was established in [24]. Using an ansatz of asymptotic independence of any finite subset of servers, it was shown that under the processor sharing service discipline, the queue length distribution at each server is insensitive to the service distribution of jobs. However, this ansatz was initially proved only for the FCFS case with the service time of jobs having a decreasing hazard rate [1]. In [25], this ansatz was proved for the processor sharing service discipline and general service time distributions, and the insensitivity of the equilibrium point of the mean-field limit to the job length distributions was established.

The homogeneous system was generalized in [26] and [27] to heterogeneous processor

[1] A service distribution with Cumulative Distribution Function (CDF) F, and Probability Density Function (PDF) f, has a hazard rate h which is given by $h(t) = f(t)/(1 - F(t))$.

sharing systems, where it was shown that a naive random selection of servers with different capacities can cause instability. However, in later work in [28], it was shown that by guaranteeing the existence of every different server type in the selection set, named as type-based JSQ(d) policy, the maximal stability region (also called the maximal throughput region in contemporary literature) can be recovered. The maximum stability region corresponds to the stability region by pooling the resources of all the servers.

Despite the favorable job response time properties of the JSQ(d) policy, it was shown in [29] that achieving *zero* asymptotic mean waiting delay under this policy in a system with n servers is impossible when the number of selected servers, $d < n$, remains constant and there is no memory at the dispatcher. This observation classifies research into two distinct directions, whether involving a variable amount of d or incorporating memory at the dispatcher. The JSQ($d(n)$) policy was studied in [30] where the number of selected servers $d(n)$, grows with the number of servers n and there is no memory at the dispatcher. It was shown that the fluid limit of the system under the JSQ($d(n)$) policy converges to that of the JSQ policy as long as $d(n)$ grows to n, but the exact growth rate of $d(n)$ is not important. However, for the diffusion limit of the system to converge to that of the JSQ system, a minimum growth rate must be satisfied by $d(n)$. A sufficient condition for this convergence is that $d(n)/\sqrt{n}\log(n) \to \infty$.

In contrast, in [31] a new pull-based load balancing policy was introduced by incorporating memory at the dispatcher. This policy is called Join-the-Idle-Queue (JIQ) policy in which the memory keeps track of all idle servers and the incoming job joins the queue of one of the idle servers. If there is no idle server, it joins any one of the servers uniformly at random. It was shown that the JIQ policy effectively reduces the communication overhead compared to the JSQ(2) policy. In [32], the JIQ policy was studied using the mean-field fluid limit methods, and a set of differential equations to describe the system behavior was proposed as $n \to \infty$. Through mean-field analysis, the asymptotic optimality of this policy was studied in [33, 34], where it was shown that the steady-state probability of an arriving customer waiting for service vanishes as $n \to \infty$.

When the system is operating under heavy traffic conditions, i.e., the arrival rate of the system is approaching its total capacity, the JIQ policy performs poorly compared to other state-dependent policies like JSQ(d), since the probability of finding no idle server available in the system is high. Motivated by this observation, a more general threshold-based scheme was studied in [35, 36] where an incoming job is routed to a server whose occupancy is below some threshold. If all servers are loaded above the threshold, the job is assigned uniformly to one of the servers. This policy is known as the Join-Below-Threshold (JBT) policy and can be considered as an extension of the JIQ policy. It was shown that the JBT policy achieves heavy-traffic delay optimality in homogeneous systems.

By heavy-traffic delay optimality, it means that the delay corresponds to the delay in a resource-pooled system. This policy was also studied in [37] where a slightly different version of the policy was studied. The incoming job joins the server whose occupancy is below a common threshold. However, if there is no such server, the job is routed to the shortest queue among d randomly selected servers. Such a modification thus subsumes the JSQ(d) policy. The JBT policy was also studied in [38] for a heterogeneous resource-sharing system and the authors showed excellent delay performance for such systems.

Load Balancing for Multiserver Jobs The problem of load balancing for multiserver jobs has been of great interest to researchers. Different models and policies are introduced for these jobs. One classic theoretical model is Fork-Join (FJ) system. In an FJ system with n servers, an incoming job is split into n independent tasks at the fork station, and each task joins one server, processed in an FCFS order. When all tasks are complete, the delay of the job is determined by the maximum delay of its tasks at the join station. The stationary joint workload distribution of FJ systems with two servers, Poisson job arrivals and exponential service times is studied in [39]. However, the analysis of FJ systems in the general case of more than two servers is extremely challenging due to the complex correlations between the fork and join stations. The existing literature only provides approximations and bounds on the performance metrics of the system such as upper bounds on the delay performance of the system, as can be seen in studies like [40] and [41]. A comprehensive survey of the existing results on FJ systems is presented in [42].

A generalization of FJ systems is partial or limited FJ systems, where in a system with n servers, the incoming job is split into $k < n$ tasks. When the number of tasks is fixed, upper bounds on the tail distribution of the response time of jobs are obtained in [41]. These bounds are derived for a policy that assigns tasks randomly to servers and for four scenarios of system parameters by combining renewal and non-renewal arrivals with non-blocking and blocking servers. The study is extended in [43] to heterogeneous systems with slow and fast servers. A probabilistic routing policy for task assignment is proposed and bounds on the average completion time of jobs are calculated. Additionally, when the number of tasks is not fixed and can vary with the number of servers n, it is shown in [44] that when $k(n) = o(n^{1/4})$, any $k(n)$ subset of servers becomes asymptotically independent as $n \to \infty$. This asymptotic independence leads to an upper bound on the mean response time of jobs in the non-asymptotic regime. Another policy named batch-filling policy which is a variant of the JSQ(d) policy for multiserver jobs, is proposed in [45]. In this policy, tasks are assigned sequentially to the shortest queue in a sampled subset of servers. This policy achieves zero queueing delay, meaning that every task of the job starts its process immediately.

In limited FJ systems, each server has its own queue, and tasks wait in the queue of the server to start their service. However, an alternative queueing model is when there is a central queue from which tasks are dispatched to servers. If the number of available servers is less than the number of job's tasks, the job will wait in the central queue until there are enough resources available. The stationary distribution of the number of jobs in systems with two servers is studied in [14] and [46]. In these systems, the job at the head of the queue may not fit into the system immediately, therefore blocking later arrivals from service and leaving servers idle. This incurs an extra waiting time in the system. However, it is shown in [47] and [48] as the system load, the number of tasks, and the number of servers scale, jobs can achieve zero asymptotic waiting time. Analogous results are studied in [49], where the mean response time is minimized under heavy traffic limit, while the number of servers remains unscaled. An alternative approach to address the extra waiting time is by dropping the job at the head of the queue if it cannot fit into the system. Multiserver-job systems with a central queue and blocking have been of considerable interest and are studied well such as in [50] and [51], where the goal is to minimize the blocking probability of each job.

Another class of queueing models for multiserver jobs involves variable server allocation. Prior research in this domain has predominantly focused on the use of speedup functions; these functions measure the ratio of a job's original service time to its accelerated service time when parallelized. In other words, speedup functions denote the amount of acceleration jobs can get by joining multiple servers. One well-known example of a speedup function is Amdahl's law [52], which accounts for scenarios where only a part of the job is parallelizable, while the remaining part receives no parallelization. In [53], the optimality of concave and sublinear speedup functions for exponential job sizes is studied. It is shown that when the jobs can be parallelized into any number of sub-jobs, the policy that shares servers equally among the jobs in the system minimizes the mean response time. In [54], it is assumed that some jobs can split into any number of parts, while others follow a threshold parallelization approach where the level of parallelization is limited to a specific threshold. Optimal policies are derived for this speedup function when job sizes are exponential. Additionally, in [55], a work-conserving finite skip framework is introduced, which includes threshold parallelism as a special case. The mean response time of the system is characterized by heavy traffic conditions. Nevertheless, these studies make specific assumptions each job has access to the full set of servers, consider linear speedup functions and assume no system blocking.

Performance of Load Balancing Policies in Finite Systems Many of the studies mentioned earlier have primarily employed limit theorems to evaluate the system performance in the asymptotic regime, i.e., as the system size n approaches infinity. However, to obtain a more comprehensive understanding of how these load balancing policies function in finite-sized systems, it is essential to investigate the accuracy of these asymptotic results. Researchers typically employ two approaches to investigate these accuracies: first, studying accuracy at the process level for any finite time duration; second, analyzing accuracy at a given instant, such as in the stationary regime.

When the focus is to derive approximation accuracies at a process level, explicit bounds on the deviation of Discrete-Time Markov Chains (DTMCs) from their mean-field limit were studied in [56], both in transient time and in stationary regime. This study focuses on numerical methods to compute the error bounds and does not provide a general analytic result. In [57], the accuracy of the mean-field limit in a system of n servers was studied under the JSQ(d) policy and FCFS server discipline. The fluctuation process was constructed as the gap between the empirical measure of the system and its mean-field limit. Suitable FCLTs were obtained to show that under a light traffic regime (i.e., when the arrival rate of the system is strictly below the total capacity of the system), the scaled fluctuation process converges to an Ornstein–Uhlenbeck (OU) process as the system size grows to infinity. However, the author did not consider the performance issues in steady-state. Such FCLTs were employed in [58] to study the accuracy of blocking probability approximations in Erlang-Loss models under the JSQ(d) policy and light traffic regime. It was shown that the gap between the exact average blocking probability of a job in the system with n servers and the limiting average blocking probability is $O(\frac{1}{\sqrt{n}})$. The results were extended in [59] to the Halfin-Whitt regime (a heavy-traffic regime) [60] and new bounds on the mean-squared difference between the blocking probability of the system and its asymptotic value were obtained.

The alternate approach is to focus on the accuracy only in the stationary regime. The problem of mean-field accuracy was studied in [61] for finite-dimensional Continuous-Time Markov Chains (CTMCs). Using Stein's method, it was shown that under some mild conditions, the mean-square difference between the stationary distribution of a system of size n and the fixed point of its mean-field limit is $O(\frac{1}{n})$. The results were extended in [62] to study this convergence rate under heavy traffic conditions. Stein's method was also used in [63] where the convergence rate for expectations of performance functionals was obtained in both finite and infinite-dimensional systems. It was shown that if the dynamical system is asymptotically exponentially stable, the convergence rate for a system of size n is $O(\frac{1}{n})$.

Unlike the order-wise scaling results, researchers were able to obtain calculable error bound on the accuracy of mean-field limits. This problem was addressed in [64] and [65], where the mean-field approximations were refined by finding the constant related to the terms $\frac{1}{n}$ and $\frac{1}{n^2}$, respectively. A combination of Stein's method and the State Space Concentration was used in [66] to obtain calculable error bounds on the heavy-traffic mean-field approximations for a class of FCFS systems with the JSQ(2) policy. The mean-field accuracy and its refinement for heterogeneous systems were studied in [67], using Stein's method. It was shown that in a system of size n, the mean-field accuracy is $O(\frac{1}{n})$, and the accuracy of refined mean-field limit is $O(\frac{1}{n^2})$.

1.3 Notation

In this section, we establish the notation and terminology that will be used throughout the dissertation. We denote $\mathbb{R}$ as the set of real numbers, $\mathbb{Z}$ as the set of integers, and $\mathbb{N}$ as the set of natural numbers. Additionally, we use $\mathbb{R}_+$ and $\mathbb{Z}_+$ to represent the sets of non-negative real numbers and non-negative integers, respectively. Furthermore, for an integer m, we use $[m]$ to represent the set $\{1, 2, \ldots, m\}$.

For the analysis in Chapter 2, we define the space $\mathbb{U}$ as the space of tail probabilities for non-negative integer-valued random variables with finite mean. Additionally, we introduce the space $\mathbb{U}^{(n)}$ to denote the space of tail occupancy probabilities for each server in a system of n servers.

In the context of two vector spaces $\mathcal{X}$ and $\mathcal{Y}$ equipped with some norm $\|.\|$, a function $f : \mathcal{X} \to \mathcal{Y}$ is said to be Lipschitz with respect to that norm if for any $a, b \in \mathcal{X}$, there exists some finite constant $K > 0$ such that $\|f(b) - f(a)\| \leq K \|b - a\|$. Also, for any metric space $\mathcal{X}$, $C(\mathcal{X})$ denotes the space of all continuous functions $f : \mathcal{X} \to \mathbb{R}$. For any bounded function $f \in C(\mathcal{X})$, we define its supremum norm as

$$\|f\|_\infty = \underset{x \in X}{sup} |f(x)|.$$

Consider two stochastic processes $x^{(n)}$ and x defined on probability spaces $\left(\Omega^{(n)}, \mathcal{F}^{(n)}, \mathbb{P}^{(n)}\right)$ and $(\Omega, \mathcal{F}, \mathbb{P})$, respectively. The sequence $x^{(n)}$ is said to converge in distribution to x and is written $x^{(n)} \Rightarrow x$, if for all bounded, real-valued and continuous functions f, we have $\lim_{n\to\infty} \int_{\Omega^{(n)}} f\left(x^{(n)}\right) d\mathbb{P}^{(n)} = \int_\Omega f(x)\, d\mathbb{P}$. For a martingale $M(t)$, its quadratic variation will be denoted by $< M >_t$.

For the asymptotic notation, let $f(n)$ and $g(n)$ be positive real-valued increasing functions. We define $f(n)$ to be $o\left(g(n)\right)$ if $\limsup_{n\to\infty} \frac{f(n)}{g(n)} = 0$, $f(n)$ to be $O\left(g(n)\right)$ if $\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty$, $f(n)$ to be $\omega\left(g(n)\right)$ if $\liminf_{n\to\infty} \frac{f(n)}{g(n)} = \infty$, and $f(n)$ to be $\Omega\left(g(n)\right)$ if $\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$. Moreover, we define $f(n)$ to be $\Theta\left(g(n)\right)$ if $f(n)$ is $O\left(g(n)\right)$ and $\Omega\left(g(n)\right)$.

1.4 Contributions and Outline

In this section, we provide a brief overview of the contributions of this thesis. We start by studying single-server job systems in Chapter 2. Our focus is on systems consisting of a single central dispatcher and n processor sharing servers working in parallel. The servers in this system exhibit heterogeneity in their processing speeds, where they are categorized into distinct groups, with servers in each group sharing the same processing speed. We introduce a novel randomized type-based join below threshold policy, that addresses the heterogeneity in the system by selecting a finite number of servers from each group. Based on the sampled servers' occupancies normalized to their speeds and predefined thresholds, the destination server is selected. Our primary evaluation metric is the average response time of jobs in the system. Our contributions can be summarized as follows.

- **Introduction of the Randomized Type-Based JBT Policy:** We propose a randomized type-based JBT policy, where assignment decisions depend on predefined threshold levels and the instantaneous occupancy of sampled servers normalized to their processing speeds. This policy accommodates changes in threshold levels, enabling the derivation of various load balancing policies suitable for heterogeneous systems, such as JSQ(d), JIQ, and more. Consequently, it provides a unified and systematic framework for studying a wide range of load balancing policies in heterogeneous environments.

- **System Stability Analysis:** We show that for every arrival rate below the total system capacity, the processor sharing system employing the type-based JBT policy is stable, irrespective of the threshold values in the system. In this context, stability means that the average response time of jobs remains bounded. This result holds true for every arrival rate of $n\lambda(n)$ where $\lambda(n) < 1$.

- **Mean-Field Limit Analysis:** We show that the stochastic empirical measure of servers' occupancies converges to a deterministic mean-field limit in large system asymptotic. This limit is described by a set of Ordinary Differential Equations (ODEs) that capture the transient behavior of the system at any finite time $t \geq 0$ as the system size grows ($n \to \infty$).

- **Fixed Point Global Asymptotic Stability Analysis:** Exploiting the monotonicity of the system of ODEs, we prove that the mean-field limit has a unique fixed point that is globally asymptotically stable. This critical property enables us to establish the interchangeability of the order of limits in time (t) and system size (n). Consequently, we can analyze the stationary distribution of a system of n servers, showing that this distribution converges to the aforementioned fixed point as $n \to \infty$. Furthermore, the steady-state average response time of jobs in the system can be expressed in relation to this unique fixed point for large values of n.

In Chapter 3, we build upon the results from Chapter 2 by analyzing the gap between the stochastic empirical measure of the system and its corresponding mean-field limit. Specifically, we characterize the sensitivity of this gap to variations in the system's arrival rate $\lambda(n)$. This study is carried out for large yet finite system sizes n. To achieve our objective, we construct a fluctuation process that quantifies the scaled difference between the empirical occupancy measure of the system and its mean-field limit. Our contributions are outlined below.

- **Convergence of Fluctuation Process in the Transient Regime:** Employing FCLTs, we show that the fluctuation process in the transient regime converges to an OU process. The characteristics of this OU process, including its drift and diffusion coefficients, depend on the mean-field limit and system parameters. The limiting OU process effectively captures the asymptotic behavior of the gap between the empirical occupancy measure of the system and its mean-field limit in the transient regime. Importantly, this gap has been established to scale as $O(\frac{1}{\sqrt{n}})$ as the system size n approaches infinity.

- **Fixed Point Exponential Stability Analysis:** To study the fluctuations process in the stationary regime, it is imperative to establish the local exponential stability of the mean-field limit derived in Chapter 2 at its fixed point, due to technical constraints. We rigorously prove this result in the case of homogeneous systems (where all servers have the same speed). However, in the heterogeneous case, analytical confirmation of this property becomes infeasible, and we provide numerical evidence to support it. Such difficulties have been encountered in prior research by other scholars [61, 63, 64], where exponential stability of the fixed point was assumed.

- **Convergence of Fluctuation Process in the Stationary Regime:** Under the local exponential stability of the fixed point of the mean-field limit, we show that the fluctuation process in the stationary regime converges to a stationary OU process. The fixed point of the mean-field limit and system parameters entirely determine the mean and covariance of this OU process. This limiting OU process exhibits a discrepancy of $O(\frac{1}{\sqrt{n}})$ between the stationary measure of the finite system and the fixed point of the mean-field limit as n tends to infinity.

- **Error Estimation:** Finally, we use these results to derive error estimates for approximating the mean delay in finite systems when employing the mean-field distribution instead of the measure associated with finite systems. Results show that the error scales as $O(\frac{1}{\sqrt{n}})$, for which the constant can be precisely calculated.

In Chapter 4, we move to adaptive multiserver jobs that can be parallelized across a varying number of servers, up to a maximum degree of parallelism d. We introduce a speedup function to measure the amount of acceleration received in the processing time of a job through parallelization and assume that this speedup function is linear in the number of servers processing the job up to d. We consider systems consisting of a single central dispatcher and n parallel loss servers. Our objective is to devise job assignment schemes that optimize system performance in terms of average response time and blocking probability while interacting with a sampled subset of servers of size $k(n)$. Our key contributions in this chapter are as follows.

- **Introduction of the Greedy Job Assignment Scheme:** We introduce a job assignment scheme in which each incoming job uses as many servers as available in the sampled subset, up to the threshold d. If a job cannot find any available server upon arrival, it is blocked.

- **Full Server Access:** When jobs have access to all servers ($k(n) = n$), we show that all arriving jobs tend to find at least d available servers in the system. Specifically, our results indicate that the blocking probability approaches zero and the mean response time of jobs approaches its minimum possible value of $1/d$ with error bounds of $O(\frac{1}{\sqrt{n}})$ and $O(\frac{1}{n})$, respectively, as the system size increases.

- **Limited Server Access:** In the case each job has access to only a randomly sampled subset of size $k(n) \ll n$, we show that the greedy scheme still achieves the same performance asymptotically as long as the size $k(n)$ of the sampled subset of servers grows at an appropriate rate. In particular, we show that both the mean response time and the blocking probability exhibit convergence to their respective optimal values with error bounds of $O(n^{-(1-\alpha)/2})$ where α denotes the rate at which the arrival rate of jobs approaches to the critical load of the system.

- **Heterogeneous Workloads:** We extend our results to heterogeneous systems with multiple arrival streams, each associated with different job sizes and maximum degrees of parallelism. We show analogous asymptotic optimality results hold in this heterogeneous context as well.

In Chapter 5, we expand on the preceding chapter's results, considering scenarios where the speedup function is no longer linear. This can occur due to the communication overhead involved in breaking down a job into smaller sub-jobs and assigning them to multiple servers. Within the same system featuring a single dispatcher and n loss servers, we introduce a probabilistic assignment scheme and establish its asymptotic optimality in the system size limit. The key contributions in this chapter are as follows.

- **Criterion For Optimality:** In cases where jobs receive a nonlinear speedup, determining the optimal system behavior becomes less straightforward. We formulate an optimization problem and show that under specific conditions—namely when the speedup function is strictly increasing, concave, and sublinear— the optimal system can be at most two-dimensional. This result simplifies the study of the optimal behavior, enhancing our understanding of it.

- **No Parallelization in Heavy Traffic:** We show that when the system operates under heavy traffic conditions and the speedup function is strictly increasing, concave, and sublinear, no advantage is gained from parallelization. In such scenarios, the optimal strategy is to assign each job to individual servers.

- **Asymptotic Optimality:** We introduce a deterministic system that mimics the optimal behavior in the stationary regime and devise a probabilistic job assignment scheme to guide the original system toward this deterministic system. We prove that the system achieves optimal performance as the system size approaches infinity. Specifically, the blocking probability tends to zero, and the mean response time of jobs converges to the deterministic system's mean response time, which represents the optimal value for large values of n.

Chapter 6 includes a summary of our results and a discussion on potential future extensions. Additional material can be found in the Appendix.

1.5 Brief Overview of Techniques

1.5.1 Mean-Field Technique

The mean-field technique, initially introduced in the field of physics by [68], is a valuable tool for analyzing the behavior of stochastic systems characterized by a large number of interacting particles that exhibit random interactions. The fundamental concept underlying this technique involves replacing the collective impact of all interactions on an individual particle with a single averaged effect. As the number of particles approaches infinity, due to the LLN, this averaged effect becomes precise, exactly describing the original high-dimensional dynamical system. For the sake of simplicity and clarity, we provide an overview of this technique in the context of homogeneous systems, where all particles are identical, and each particle takes values in a finite set of states denoted as $E = \{1, 2, \dots, s\}$.

Let us consider a system comprising n particles, with $x_i^{(n)}(t)$ representing the fraction of these n particles in state $i \in E$ at time t. Consequently, $x^{(n)}(t) = \Big(x_1^{(n)}(t), \dots, x_s^{(n)}(t)\Big)$ characterizes the system's state at time t. Since the system's state is quantified by the scaled number of particles, it guarantees that each individual particle's contribution to the overall state changes of the system is $O(1/n)$, ensuring a weakly interacting system for large system sizes n.

We assume that the process $\big(x^{(n)}(t), t \geq 0\big)$, representing the empirical distribution of the system, behaves as a continuous-time Markov process and possesses a unique stationary distribution. Due to the Markovian nature of the system, it is certain that only one particle can change state and transition at any given time t, with probability one. Let $r_{u,v}(x)$ be independent of n and denote the rate at which a single particle switches from state $u \in E$ to state $v \in E$, given the current empirical measure of the system is x. Noting that there are $nx_u^{(n)}$ particles in state u undergoing this transition, the transition rate of the entire system is $O(n)$, while the size of each transition is $O(1/n)$.

We introduce the drift function f defined on the state space E as follows.

$$f(x) = \sum_{v \neq u} r_{u,v}(x) \cdot (v - u)\,.$$

This drift function is independent of the system size n. In [69, Theorem 2.1-p.456], it is shown that if the drift function f is Lipschitz continuous and the initial empirical distribution $x^{(n)}(0)$, converges in probability to a constant x_0, then the stochastic process $x^{(n)}(\cdot)$ converges in distribution to a deterministic process whose dynamics is governed by f as $n \to \infty$. In simpler terms, the drift function f captures the average changes in the original system $x^{(n)}(\cdot)$ that starts from the state x. This convergence implies that the Functional Law of Large Numbers (FLLN) applies to the entire path of the system, allowing us to analyze the time evolution of the original system at any finite time t.

The aforementioned convergence, however, does not provide any information about the stationary distribution of the system. Nonetheless, it is possible to study the system's stationary behavior through the equilibrium points or fixed points of the mean-field limit. Using Prohorov's theorem [70], it can be established that if the mean-field limit has a unique fixed point that is globally asymptotically stable (i.e., the mean-field limit converges to this fixed point as $t \to \infty$, regardless of the initial state of the system), then the order of limits with respect to time (t) and system size (n) can be interchanged for the empirical distribution. In other words, $\lim_{n\to\infty} \lim_{t\to\infty} x^{(n)}(t) = \lim_{t\to\infty} \lim_{n\to\infty} x^{(n)}(t)$. This implies that the system's stationary distribution converges to this unique fixed point as $n \to \infty$.

1.5.2 Stein's Method

As discussed in the previous subsection, the study of stochastic systems in the stationary regime using mean-field techniques involves a systematic approach comprising three key steps:

- **Transient Regime Convergence:** The initial step requires establishing the convergence of the empirical distribution of the system to its mean-field limit in the transient regime.
- **Global Asymptotic Stability:** Subsequently, it is imperative to demonstrate that the resulting mean-field limit possesses a unique fixed point that is globally asymptotically stable.
- **Interchange of Limits:** Finally, the possibility of interchanging the limits in system size (n) and time (t) must be verified.

When these three steps are combined effectively, it becomes possible to validate that

$$\lim_{n\to\infty} x^{(n)}(\infty) = P,$$

where $x^{(n)}(\infty)$ denotes the stationary distribution of the system, and P represents the fixed point of the mean-field limit.

However, establishing the global asymptotic stability of this fixed point can be challenging, particularly in cases where standard properties such as monotonicity are not satisfied. Additionally, for systems characterized by strong coupling among particles, it may be impossible to establish a mean-field limit. Systems discussed in Chapters 4 and 5, for instance, fall into this category.

A direct method to study the system in the stationary regime is Stein's method [71, 72]. Stein's method involves quantifying the distance between two distributions: one to be approximated (e.g., the stationary distribution of the interacting system) and another, a desired distribution to serve as an approximation target. The first application of Stein's method to study stationary distributions of Markov processes can be traced back to [73], which highlights its applicability whenever the distribution to be approximated is the stationary distribution of a Markov process. The power of Stein's method in approximating steady-state conditions has gained recognition in several recent publications, including [74, 75, 76, 61, 62, 63].

Considering that the empirical distribution of the system represented as $x^{(n)}(.)$, follows a Markov process, it can be deduced that the expected drift of any appropriate function V under the generator of this Markov process, denoted as G, becomes zero in the steady state. In other words,

$$\mathbb{E}\left[GV\left(x^{(n)}\right)\right] = 0,$$

where $x^{(n)}$ denotes the steady-state distribution of the system, with the explicit dependence on time (t) removed.

The method proceeds by selecting a simple deterministic dynamical system that mimics the behavior of the desired system for large values of n. Assuming this simple system is described by a set of ODEs $\dot{x}^{(n)} = f\left(x^{(n)}\right)$ for a suitable function f, the drift of the function V under this system can be expressed as $\frac{\partial V}{\partial x^{(n)}} f\left(x^{(n)}\right)$. By comparing the generator of the original system with that of the deterministic system, we can write:

$$\mathbb{E}\left[GV\left(x^{(n)}\right) - \frac{\partial V}{\partial x^{(n)}} f\left(x^{(n)}\right)\right] = -\mathbb{E}\left[\frac{\partial V}{\partial x^{(n)}} f\left(x^{(n)}\right)\right]. \tag{1.1}$$

Equation (1.1) is referred to as Stein's (or Poisson's) Equation. Let us represent our quantities of interest by a target function $h\left(x^{(n)}\right)$. Then solving the equation

$$-\frac{\partial V}{\partial x^{(n)}} f\left(x^{(n)}\right) = h\left(x^{(n)}\right)$$

allows us to select the function V in a way that the quantities of interest are expressed as the difference between the generator of the original system and the generator of the simple dynamical system, providing a means to establish bounds on these quantities. It should be noted that since Stein's method operates in the pre-limit regime, all results are established for systems of finite size. Consequently, it is also possible to determine error bounds by constructing suitable Lyapunov functions.

1.6 Conclusion

In this introductory chapter, we have introduced the framework for our study of processor sharing systems with single-server jobs and loss systems with multiserver jobs. We have outlined the core objectives along with the challenges associated with each, that will guide our research through the subsequent chapters. Employing analytical techniques such as the mean-field technique, FCLT, and Stein's method, our goal is to discover approaches for enhancing system performance and resource allocation in practical, real-world scenarios. The paper [77] presents the results for processor sharing systems with single-server jobs, while the study of loss systems with multiserver jobs is detailed in [78].

Chapter 2

Mean-Field Analysis of Threshold-Based Load Balancing Policies in Processor Sharing Systems

In this chapter, we study the problem of threshold-based randomized load balancing policies for processor sharing systems with a large number of servers working in parallel. Due to the heterogeneity of vendors and configurations, these servers are grouped into different types, each exhibiting a different processing speed. Under threshold-based policies, when a job arrives, it is permanently assigned to a single server based on predefined thresholds, instantaneous system state, and server speeds. The objective is to minimize the response time of jobs in the system while making them less dependent on the type of server that is used. To achieve this, we consider different thresholds for each server type and introduce a new load balancing policy named "type-based Join-Below-Threshold (type-based JBT)" policy, defined later in Section 2.1. This policy subsumes different classes of policies like JIQ, JSQ(d), etc., and provides a unified framework to analyze the system performance in the heterogeneous context. Previous research in processor sharing systems mainly concentrated on specific policies. Analyzing system performance requires understanding the system's mean-field limit and the stability of its fixed point. Establishing that the fixed point of the mean-field limit is globally asymptotically stable, is a challenging task and depends on the dynamics of the mean-field system. In this chapter, through the introduction of the type-based JBT policy, we establish a comprehensive framework to prove this stability property across a wide range of load balancing policies, both existing and new. The approach is based on the theory of Feller semigroup of operators for Markov processes for establishing the mean-field limit and exploiting the monotonicity of the mean-field limit for showing the stability of its fixed point.

The rest of the chapter is organized as follows. In Section 2.1, we describe the system model and introduce additional notation. Section 2.2 offers a brief overview of previous research results. In Section 2.3, we show that the stability region of the system corresponds to the maximal stability region achievable. In Section 2.4 we study the transient behavior of the system and show that for every finite time, the empirical measure of servers' occupancy converges weakly to its mean-field limit. In Section 2.5, we present results on the stability properties of the fixed point of the mean-field limit and the stationary behavior of the system. Finally, Section 2.6 includes concluding remarks.

2.1 System Model

We consider a single dispatcher that routes arriving jobs to n processor sharing servers working in parallel. Servers have different capacities and are clustered into M groups based on their capacities. We denote the set of all groups by the set $[M] = \{1, 2, ..., M\}$ where each server in the group $m \in [M]$ has capacity C_m. Thus the set $\mathcal{C} = \{C_1, C_2, ..., C_M\}$ indicates all different server capacities in the system. Without loss of generality, we assume that the set $\mathcal{C}$ is sorted in ascending order, i.e., $C_1 \le C_2 \le ... \le C_M$. Additionally, the fraction of servers in each group $m \in [M]$ is assumed to be fixed and is denoted by $\gamma_m \in [0, 1]$. Obviously, $\sum_{m\in[M]} \gamma_m = 1$. Moreover, we assume without loss of generality that the normalized system capacity $\sum_{m\in[M]} \gamma_m C_m = 1$.

Jobs arrive at the system following a Poisson process with the rate $n\lambda(n) = n(\lambda - \frac{\beta}{\sqrt{n}}) \ge 0$ where $\lambda \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$ are constants. Each job brings with it an amount of workload that is exponentially distributed with unit mean. Inter-arrival times and service requirements of jobs are independent of each other. Recent work [25] has established that the characterization of the mean-field limit does not depend on the exponential job length assumption but indeed for any general service time distribution with the same mean and finite second moments. Upon each arrival, the dispatcher assigns the job to one of the n servers based on the type-based JBT load balancing policy defined in Definition 2.1. Once the job joins the queue, it will start its service immediately. The job will be processed with a service rate reciprocal to the number of jobs in the queue, i.e., if there are N jobs in a type m server, the job will be processed at the rate $\frac{C_m}{N}$.

Definition 2.1. In the type-based JBT policy, at each arrival instant, $d_m \ge 1$ servers of type m are selected uniformly at random for all $m \in [M]$. The job is then routed to the server whose occupancy is less than or equal to some predefined server-dependant threshold α_m. If there are multiple servers below their thresholds, servers with the highest capacity are nominated. If there is more than one nominated server, one of them is selected uniformly at random as the destination server. In the case all $\sum_{m\in[M]} d_m$ selected servers are occupied above their thresholds, the job is routed to the server with the minimum expected delay, or in other words, the highest processing rate per unfinished job. That is, among d_m selected servers of the same type m, the server with the fewest number of jobs, $q_{m,min}$, is nominated. Then, among nominated servers of different types, the server with the largest value of $\frac{C_m}{q_{m,min}}$ is chosen as the destination server.

At any step of the selection process, ties between servers of different capacities are broken by selecting the server with the highest capacity, and between servers with the same capacity are broken by choosing any server uniformly at random.

Remark 2.1. We consider different thresholds α_m for different server types m. For example, we could choose thresholds such that $\frac{C_1}{\alpha_1} = \frac{C_2}{\alpha_2} = ... = \frac{C_M}{\alpha_M}$. In such a case, delays between different server types will be similar. In another setting, we could choose all α_m to be negative. In that case, the arrival will always be dispatched to the server with the minimum expected delay. If we choose all thresholds α_m to be zero, then Join-the-Idle-Queue (JIQ) policy is achieved, where in the absence of idle servers, the minimum expected delay policy is applied. Moreover, by setting only some thresholds α_m to zero, it is guaranteed that the job will be dispatched to that server of type m only when it is idle. These arguments show that by choosing thresholds α_m in a certain way, different load balancing policies can be achieved. Thus the type-based JBT policy in Definition 2.1 subsumes a class of different load balancing policies. The following analysis assumes that the thresholds α_m for type m servers are general.

2.1.1 Additional Notation

We define the following real sequence spaces.

$$\mathbb{U} = \{(u_k, k \in \mathbb{Z}_+), u_0 = 1, u_k \geq u_{k+1} \geq 0 \quad \forall k \in \mathbb{Z}_+, \quad \sum_k |u_k| < \infty\},$$

$$\mathbb{U}_m^{(n)} = \{(u_k, k \in \mathbb{Z}_+) \in \mathbb{U}, \; n\gamma_m u_k \in \mathbb{N} \quad \forall k \in \mathbb{Z}_+\}, \qquad m \in [M].$$

The space $\mathbb{U}$ is the space of tail distributions for non-negative integer-valued random variables with finite mean, and the space $\mathbb{U}^{(n)} = \prod_{m\in[M]} \mathbb{U}_m^{(n)}$ which is the M-fold Cartesian product of spaces $\mathbb{U}_m^{(n)}$, is the space of tail distributions for heterogeneous systems with n servers and M different types. Additionally, let $\mathbb{U}^M$ be the M-fold Cartesian product of the space $\mathbb{U}$. We equip spaces $\mathbb{U}^{(n)}$ and $\mathbb{U}^{M}$ with the metric induced by the ℓ_2-norm where for an element $\mathbf{u} = (u_{k,m}, k \in \mathbb{Z}_+, m \in [M])$, we have

$$\|\mathbf{u}\|_2^2 = \sum_{m\in[M]} \sum_{k\in\mathbb{Z}_+} |u_{k,m}|^2.$$

The space $\mathbb{U}^M$ is compact and hence complete and separable under ℓ_2-norm (proof is given in Appendix A.1). Also for $\mathbf{u}, \mathbf{v} \in \mathbb{U}^M$, the inequality $\mathbf{u} \leq \mathbf{v}$ implies element-wise comparison, i.e., $u_{k,m} \leq v_{k,m}$ for all $k \in \mathbb{Z}_+$ and $m \in [M]$.

2.2 A Brief Review of Previous Results

In this section, we provide an overview of prior mean-field analysis relevant to systems with infinite queueing buffers. Given that the type-based JBT policy in this specific format has not been previously studied, we shift our focus to a closely related policy that can offer valuable insights. As highlighted in Remark 2.1, the type-based JBT policy includes the minimum expected delay policy as a special case (when all thresholds are negative). This particular policy is equivalent to the JSQ(d) policy in homogeneous systems where all servers have identical processing speeds. Consequently, we revisit the analysis of homogenous systems under the JSQ(d) policy, as presented in [22, 23]. We restate their results as a fundamental basis for our study.

The system studied in [22, 23] includes n servers that follow the FCFS service discipline. Jobs arrive as a Poisson process with a rate of $n\lambda$, where $\lambda < 1$, and job sizes are distributed exponentially with a mean of 1. Upon arrival, each job independently and uniformly selects a constant number $d \geq 2$ of servers and joins the queue of the server with the minimum occupancy at that moment.

To align with our notation, let $x_i^{(n)}(t)$ denote the fraction of servers containing at least i jobs in a system of n servers at time t. Then the state of the system at any time t can be described by an infinite-dimensional vector $x^{(n)}(t) = \left(x_0^{(n)}(t), x_1^{(n)}(t), x_2^{(n)}(t), \ldots\right)$ and only requires information about the occupancy of queues.

First, the system was compared to another system in which each job randomly joins a server, i.e., when $d = 1$. We refer to the latter system as the static random system. By employing an argument based on majorization, it was demonstrated in [22] that the size of the longest queue in the original system with $d \geq 2$ is stochastically dominated by the size of the longest queue in the random static system. Using this observation, in conjunction with the stability of the random static system, it was established that the original system is stable for every $\lambda < 1$, implying that the expected number of jobs in the system remains finite for all times t.

Subsequently, it was shown that as the system size n approaches infinity, under the condition that $x^{(n)}(0)$ converges to a constant $x(0)$, the system state $x^{(n)}(.)$ converges to a deterministic system denoted as $x(.)$, where $x(.)$ is the solution to the following set of ordinary differential equations:

$$\begin{cases} \frac{dx_i}{dt} & = \lambda \left(x_{i-1}^d - x_i^d\right) - (x_i - x_{i+1}), \quad i \in \mathbb{N}, \\ x_0 & = 1. \end{cases} \tag{2.1}$$

The process $x(.)$ is referred to as the mean-field limit of the system.

Additionally, the fixed point of the mean-field limit was studied. A fixed point of the mean-field limit is a point P where $\frac{dP}{dt} = 0$. It was demonstrated that the system described by Equations (2.1) with $d \geq 2$ possesses a unique fixed point, given by

$$\begin{cases} P_i & = \lambda^{\frac{d^i - 1}{d-1}}, \quad i \in \mathbb{N}, \\ P_0 & = 1. \end{cases} \tag{2.2}$$

Furthermore, it was shown in [22] that the fixed point (2.2) is globally exponentially stable, implying that the mean-field limit described by Equations (2.1) converges exponentially fast to this fixed point as $t \to \infty$. Using this result, it was demonstrated that for $d \geq 2$, we have $\mathbb{E}\left[x_k^{(n)}(\infty)\right] \to P_k$ as $n \to \infty$, for every $k \in \mathbb{Z}_+$, where $x_k^{(n)}(\infty)$ denotes the equilibrium fraction of servers with at least k jobs. Combining this with Little's law and noting that the fixed point is bounded, it was demonstrated that the expected time a job spends in the steady-state system for $d \geq 2$ converges to

$$T_d(\lambda) = \sum_{i=1}^{\infty} \lambda^{\frac{d^i - d}{d-1}}, \tag{2.3}$$

as $n \to \infty$.

2.3 Throughput Optimality

In this section, we provide a formal definition of the system's stability region and demonstrate that the system discussed in this chapter indeed achieves the maximal stability region.

Definition 2.2. The stability region is the set of all arrival rates for which the underlying Markov process describing the time evolution of the system is positive recurrent.

Therefore, the maximal stability region corresponds to the stability region when pooling the resources of all the servers. With this definition, $\Lambda = \{\lambda(n) : 0 \leq \lambda(n) < 1\}$ is the maximal stability region that a heterogeneous network defined in Section 2.1 can achieve.

Lemma 2.3.1. *The heterogeneous system defined in Section 2.1 with the type-based JBT routing policy is stable for any value of $\lambda(n) \in \Lambda$, and hence it is throughput optimal.*

Proof. For each incoming job, let the set of all selected servers at its arrival instant, be its potential destination set. Any arbitrary set A corresponds to a potential destination set if it has exactly d_m servers of type m for all $m \in [M]$. The probability of sampling one specific potential destination set A is given by $\mathbb{P}_A = \frac{1}{\prod_{m\in[M]} \binom{n\gamma_m}{d_m}}$. Also, let $C(k)$ denote the capacity of the k^{th} server in the set $\{1, 2, \ldots, n\}$ where $C(k) \in \mathcal{C}$. We use the results of [79, Theorem 2.5] to establish the system's stability. The system is stable, if

$$\rho = \max_{B\subseteq\{1,2,\ldots,n\}} \left\{ \left(\sum_{k\in B} C(k) \right)^{-1} n\lambda(n) \sum_{A\subseteq B} \mathbb{P}_A \right\} < 1. \tag{2.4}$$

If the set B does not have at least d_m servers of type m for all $m \in [M]$, then $\mathbb{P}_A$ and consequently ρ become zero. Therefore, we denote the number of type m servers in the set B with B_m and focus on scenarios where $B_m \geq d_m$ for all $m \in [M]$. Then we can write the traffic intensity ρ in Equation (2.4) as

$$\rho = \max_{\substack{B\subseteq\{1,2,\ldots,n\} \\ d_m \leq B_m \leq n\gamma_m, \forall m\in[M]}} \left\{ \left(\sum_{k\in B} C(k) \right)^{-1} n\lambda(n) \sum_{A\subseteq B} \mathbb{P}_A \right\}. \tag{2.5}$$

Only $\prod_{m\in[M]} \binom{B_m}{d_m}$ different subsets of B correspond to a potential destination set. This gives

$$\rho = \max_{\substack{B\subseteq\{1,2,\ldots,n\} \\ d_m \leq B_m \leq n\gamma_m, \forall m\in[M]}} \left\{ \frac{n\lambda(n) \prod_{m\in[M]} \binom{B_m}{d_m}}{\sum_{m\in[M]} B_m C_m \prod_{m\in[M]} \binom{n\gamma_m}{d_m}} \right\}. \tag{2.6}$$

We show that the term $\prod_{m\in[M]}\binom{B_m}{d_m}/\sum_{m\in[M]}B_mC_m$ is increasing in each B_m. Indeed, for a fixed $l\in[M]$, let $F_{B_l}=\binom{B_l}{d_l}\prod_{m\neq l}\binom{B_m}{d_m}/\Big(\sum_{m\neq l}B_mC_m+B_lC_l\Big)$. With this definition, $F_{B_l+1}=\binom{B_l+1}{d_l}\prod_{m\neq l}\binom{B_m}{d_m}/\Big(\sum_{m\neq l}B_mC_m+(B_l+1)C_l\Big)$. We show that $\frac{F_{B_l+1}}{F_{B_l}}\geq 1$.

$$\frac{F_{B_l+1}}{F_{B_l}}=\frac{\binom{B_l+1}{d_l}\prod_{m\neq l}\binom{B_m}{d_m}\Big(\sum_{m\in[M]}B_mC_m\Big)}{\binom{B_l}{d_l}\prod_{m\neq l}\binom{B_m}{d_m}\Big(\sum_{m\neq l}B_mC_m+(B_l+1)C_l\Big)}. \tag{2.7}$$

By simplifying the expression above, we get

$$\frac{F_{B_l+1}}{F_{B_l}}=\frac{(B_l+1)\Big(\sum_{m\in[M]}B_mC_m\Big)}{(B_l+1-d_l)\Big(\sum_{m\in[M]}B_mC_m+C_l\Big)}. \tag{2.8}$$

By expanding the terms in the denominator, we have

$$\frac{F_{B_l+1}}{F_{B_l}}=\frac{(B_l+1)\Big(\sum_{m\in[M]}B_mC_m\Big)}{(B_l+1)\Big(\sum_{m\in[M]}B_mC_m\Big)+(1-d_l)C_l+B_lC_l-d_l\Big(\sum_{m\in[M]}B_mC_m\Big)}. \tag{2.9}$$

Since $d_l\geq 1$, it immediately follows that the numerator is greater than or equal to the denominator and $\frac{F_{B_l+1}}{F_{B_l}}\geq 1$. This is true for each $l\in[M]$, hence the maximum of Equation (2.6) occurs at $B_m=n\gamma_m$ for all $m\in[M]$ which gives

$$\rho=\lambda(n). \tag{2.10}$$

Equation (2.10) indicates that if $\lambda(n)\in\Lambda$, then $\rho<1$ and the system is stable. □

2.4 Mean-Field Analysis: Transient Behavior of the System

In this section, we study the behavior of the system as the number of servers increases, specifically as n approaches infinity.

For any $k\in\mathbb{Z}_+$ and $l,m\in[M]$, we define

$$\lceil k \rceil_{lm} = \max\left(\left\lceil k\frac{C_l}{C_m}\right\rceil, \alpha_l + 1\right),$$
$$\lfloor k \rfloor_{lm} = \max\left(\left\lfloor k\frac{C_l}{C_m}\right\rfloor + 1, \alpha_l + 1\right),$$

where $\lceil a \rceil$ gives the smallest integer greater than or equal to a and $\lfloor a \rfloor$ returns the greatest integer smaller than or equal to a.

We characterize the state of the system with n servers at time t using the notation $\mathbf{x}^{(n)}(t) = \left(x_{k,m}^{(n)}(t), k \in \mathbb{Z}_+, m \in [M]\right)$. In this representation, $x_{k,m}^{(n)}(t)$ represents the fraction of all $n\gamma_m$ servers that have at least k jobs in service at time t. Obviously, $\mathbf{x}^{(n)}(t)$ takes values in the space $\mathbb{U}^{(n)}$. Under the type-based JBT policy, Poisson job arrivals and exponential service times, the process $\mathbf{x}^{(n)}(\cdot)$ is a Markov process. This is attributed to the policy's dependence solely on the current system state to make routing decisions and the memoryless property of the exponential distribution. We define the generator of this Markov process in the following lemma.

Lemma 2.4.1. *Under the type-based JBT policy, the generator $\mathcal{A}^{(n)}$ of the Markov process $\mathbf{x}^{(n)}(.)$, acting on continuous functions $g \in C(\mathbb{U}^{(n)})$ is given by*

$$\begin{aligned}
\mathcal{A}^{(n)}g(\mathbf{u}) =& n\lambda(n) \sum_{m\in[M]} \sum_{k=1}^{\alpha_m+1} \left[\frac{1-(u_{\alpha_m+1,m})^{d_m}}{1-u_{\alpha_m+1,m}} (u_{k-1,m} - u_{k,m}) \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} \right. \\
&\left. \times \left(g\left(\mathbf{u} + \frac{\mathbf{e}(k,m)}{n\gamma_m}\right) - g(\mathbf{u})\right)\right] \\
&+ n\lambda(n) \sum_{m\in[M]} \sum_{k=\alpha_m+2}^{\infty} \left[((u_{k-1,m})^{d_m} - (u_{k,m})^{d_m}) \prod_{l=1}^{m-1} (u_{\lceil k-1\rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1\rfloor_{lm},l})^{d_l} \right. \\
&\left. \times \left(g\left(\mathbf{u} + \frac{\mathbf{e}(k,m)}{n\gamma_m}\right) - g(\mathbf{u})\right)\right] \\
&+ n \sum_{m\in[M]} \sum_{k=1}^{\infty} \gamma_m C_m (u_{k,m} - u_{k+1,m}) \left(g\left(\mathbf{u} - \frac{\mathbf{e}(k,m)}{n\gamma_m}\right) - g(\mathbf{u})\right),
\end{aligned} \tag{2.11}$$

where $\mathbf{u} \in \mathbb{U}^{(n)}$ and $\mathbf{e}(k,m) = (e_{k',m'}, k' \in \mathbb{Z}_+, m' \in [M])$ is the unit sequence with $e_{k,m} = 1$ and $e_{k',m'} = 0$ for all other elements.

Proof. By definition, the generator $\mathcal{A}^{(n)}$ of the Markov process $\mathbf{x}^{(n)}(\cdot)$ is given by

$$\mathcal{A}^{(n)}g(\mathbf{u}) = \sum_{\mathbf{v}\neq\mathbf{u}} r(\mathbf{u} \to \mathbf{v})(g(\mathbf{v}) - g(\mathbf{u})), \tag{2.12}$$

where $r(\mathbf{u} \to \mathbf{v})$ denotes the transition rate from state $\mathbf{u}$ to state $\mathbf{v}$. Thus it is sufficient to identify the transition rates based on the type-based JBT policy. Let $\mathbf{x}^{(n)}(t)$ be in the state $\mathbf{u} \in \mathbb{U}^{(n)}$. A new incoming job will join a type m server with exactly $k-1$ jobs, if one of the following cases happens.

1. $k - 1 \leq \alpha_m$: In this case, the job will join a type m server whose occupancy is below its threshold and has exactly $k-1$ jobs, if

   - there exists at least one such server, and
   - all selected servers of higher capacities are occupied above their thresholds. Otherwise, the job would have been routed to a server with higher capacity.

   This case happens with probability

$$\left(1 - (u_{\alpha_m+1,m})^{d_m}\right) \frac{u_{k-1,m} - u_{k,m}}{1 - u_{\alpha_m+1,m}} \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l}. \tag{2.13}$$

2. $k - 1 > \alpha_m$: In this case, the job will join a type m server whose occupancy is above its threshold and has exactly $k-1$ jobs, if

   - there exists at least one such server, and
   - all selected servers are occupied above their thresholds, otherwise the job would have been routed to some server below its threshold, and
   - $\frac{C_m}{k-1}$ is the maximum value of capacity per unfinished job. That is, for all selected servers of lower capacities $l = 1, \ldots, m-1$ with minimum occupancy q_l, $\frac{C_l}{q_l} \leq \frac{C_m}{k-1}$, and for all selected servers of types $l = m+1, \ldots, M$ with minimum occupancy q_l, we should have $\frac{C_l}{q_l} < \frac{C_m}{k-1}$.

   Combining all these together, the probability of occurrence of this event is given by

$$\left((u_{k-1,m})^{d_m} - (u_{k,m})^{d_m}\right) \prod_{l=1}^{m-1} (u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l}. \tag{2.14}$$

Also, a job will depart a type m server with exactly k jobs at the rate $\gamma_m C_m(u_{k,m} - u_{k+1,m})$. Substituting the transition rates (2.13)-(2.14) and the departure rate into the definition of the generator (2.12), we get the expression given in (2.11) for the generator $\mathcal{A}^{(n)}$. $\square$

In the next theorem, we state the weak convergence of the process $\mathbf{x}^{(n)}(.)$ to a deterministic process $\mathbf{x}(.)$.

Theorem 2.4.2. *If $\mathbf{x}^{(n)}(0) \Rightarrow \mathbf{x}_0 \in \mathbb{U}^M$ as $n \to \infty$, then the process $\mathbf{x}^{(n)}(\cdot)$ converges in distribution to a deterministic process $\mathbf{x}(\cdot)$ as $n \to \infty$. The process $\mathbf{x}(\cdot)$ lies in the space $\mathbb{U}^M$ and is the unique solution to the following set of differential equations.*

$$\begin{aligned} \mathbf{x}(0) &= \mathbf{x_0}, \\ \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t)), \end{aligned} \tag{2.15}$$

where the mapping $\mathbf{f} : \mathbb{U}^M \to \left(\mathbb{R}^{\{0,1,2,\ldots\}}\right)^M$ is given by

$$\begin{aligned} f_{0,m}(\mathbf{x}) &= 0, \quad m \in [M], \\ f_{k,m}(\mathbf{x}) &= \frac{\lambda(1 - (x_{\alpha_m+1,m})^{d_m})}{\gamma_m(1 - x_{\alpha_m+1,m})}(x_{k-1,m} - x_{k,m}) \prod_{l=m+1}^{M} (x_{\alpha_l+1,l})^{d_l} - C_m(x_{k,m} - x_{k+1,m}), \\ &\qquad m \in [M],\ 1 \le k \le \alpha_m + 1, \\ f_{k,m}(\mathbf{x}) &= \frac{\lambda}{\gamma_m}((x_{k-1,m})^{d_m} - (x_{k,m})^{d_m}) \prod_{l=1}^{m-1} (x_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (x_{\lfloor k-1 \rfloor_{lm},l})^{d_l} \\ &\qquad - C_m(x_{k,m} - x_{k+1,m}), \quad m \in [M], k > \alpha_m + 1. \end{aligned} \tag{2.16}$$

The deterministic process $\mathbf{x}(\cdot)$ is called the mean-field limit of the system.

Proof. Proof of this theorem follows from the theory of operator semigroups for Markov processes. If the corresponding operator semigroups converge and the limiting operator semigroup is Feller, then from the convergence of initial distributions, we can conclude that $\mathbf{x}^{(n)}(\cdot) \Rightarrow \mathbf{x}(\cdot)$. Details are given below.

Let $(T^{(n)}(t), t \ge 0)$ and $(T(t), t \ge 0)$ be semigroup of operators corresponding to the processes $\mathbf{x}^{(n)}(\cdot)$ and $\mathbf{x}(\cdot)$, respectively. For any continuous function $g \in C(\mathbb{U}^M)$, we have

$$\begin{aligned} T^{(n)}(t)g(\mathbf{u}) &= \mathbb{E}\left[g\left(\mathbf{x}^{(n)}(t)\right) \mid \mathbf{x}^{(n)}(0) = \mathbf{u}\right], \\ T(t)g(\mathbf{u}) &= g(\mathbf{x}(t, \mathbf{u})), \end{aligned} \tag{2.17}$$

where $\mathbf{x}(t, \mathbf{u})$ denotes the mean-field process $\mathbf{x}(t)$ starting at $\mathbf{x}(0) = \mathbf{u}$. Also, let $\mathcal{A}^{(n)}$ and $\mathcal{A}$ be the generators of operator semigroups $(T^{(n)}(t), t \geq 0)$ and $(T(t), t \geq 0)$, respectively. Then by definition, we have

$$\mathcal{A}g(\mathbf{u}) = \frac{d}{dt} g(\mathbf{x}(t, \mathbf{u}))\bigg|_{t=0}, \tag{2.18}$$

and the generator $\mathcal{A}^{(n)}$ is given in Lemma 2.4.1.

We define the space $D \subseteq C(\mathbb{U}^M)$ as the space of all continuous functions g such that $\frac{\partial g}{\partial u_{k,m}}$, $\frac{\partial^2 g}{\partial u_{k,m}^2}$ and $\frac{\partial^2 g}{\partial u_{k,m} \partial u_{k',m'}}$ exist for all $k, k' \in \mathbb{Z}_+$ and $m, m' \in [M]$, and are uniformly bounded by some finite constant K. Then, with the ℓ_2-norm for the space $\mathbb{U}^M$ and supermum norm for the space $C(\mathbb{U}^M)$, it is seen that the space D is dense in the space $C(\mathbb{U}^M)$. We claim that the mean-field process $\mathbf{x}(.)$ is in the space D that we show next.

Lemma 2.4.3. *The mean-field process $\mathbf{x}(.)$ is in the space D where for each $k, k', s \in \mathbb{Z}_+$ and $m, m', l \in [M]$, the bounds of partial derivatives are given by*

$$\begin{aligned} \left| \frac{\partial x_{s,l}(t, \mathbf{u})}{\partial u_{k,m}} \right| &\leq e^{at}, \\ \left| \frac{\partial^2 x_{s,l}(t, \mathbf{u})}{\partial u_{k,m}^2} \right|, \left| \frac{\partial^2 x_{s,l}(t, \mathbf{u})}{\partial u_{k,m} \partial u_{k',m'}} \right| &\leq \frac{b_0}{a}(e^{2at} - e^{at}), \end{aligned} \tag{2.19}$$

where $a = 2\lambda \frac{\max_m d_m}{\min_m \gamma_m} \sum_m d_m + 2 \max_m C_m$, and $b_0 = \frac{10\lambda}{\min_m \gamma_m} (\sum_m d_m)^3$.

Proof. See Appendix A.2 □

For any function $g \in D$, as $n \to \infty$ we have

$$\begin{aligned} n\gamma_m \left(g\left(\mathbf{u} + \frac{\mathbf{e}(k,m)}{n\gamma_m} \right) - g(\mathbf{u}) \right) &\to \frac{\partial g(\mathbf{u})}{\partial u_{k,m}}, \\ n\gamma_m \left(g\left(\mathbf{u} - \frac{\mathbf{e}(k,m)}{n\gamma_m} \right) - g(\mathbf{u}) \right) &\to -\frac{\partial g(\mathbf{u})}{\partial u_{k,m}}. \end{aligned} \tag{2.20}$$

From Equations (2.11) and (2.20), as $n \to \infty$, we have

$$\begin{aligned}\mathcal{A}^{(n)}g(\mathbf{u}) \to \sum_{m\in[M]} \sum_{k=1}^{\alpha_m+1} \Bigg[\Bigg(& \frac{\lambda(1-(u_{\alpha_m+1,m})^{d_m})}{\gamma_m(1-u_{\alpha_m+1,m})} (u_{k-1,m} - u_{k,m}) \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} \\ & - C_m \left(u_{k,m} - u_{k+1,m}\right) \Bigg) \frac{\partial g(\mathbf{u})}{\partial u_{k,m}} \Bigg] \\ + \sum_{m\in[M]} \sum_{k=\alpha_m+2}^{\infty} \Bigg[\Bigg(& \frac{\lambda}{\gamma_m}((u_{k-1,m})^{d_m} - (u_{k,m})^{d_m}) \prod_{l=1}^{m-1} (u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l} \\ & - C_m \left(u_{k,m} - u_{k+1,m}\right) \Bigg) \times \frac{\partial g(\mathbf{u})}{\partial u_{k,m}} \Bigg],\end{aligned} \tag{2.21}$$

which is equivalent to

$$\mathcal{A}^{(n)}g(\mathbf{u}) \to \frac{d}{dt} g(\mathbf{x}(t,\mathbf{u}))\Big|_{t=0}. \tag{2.22}$$

From Equation (2.18), we have

$$\mathcal{A}^{(n)}g(\mathbf{u}) \to \mathcal{A}g(\mathbf{u}), \qquad \forall g \in D. \tag{2.23}$$

Let $D_0 \subseteq D \subseteq C(\mathbb{U}^M)$ be the set of those functions in D that depend only on finitely many components of $\mathbf{u}$. Then D_0 is dense in D, and also in $C(\mathbb{U}^M)$. By Lemma 2.4.3, it is easy to see that $T(t)g(\mathbf{u}) \in D$, for all functions $g \in D_0$. In other words, $T(t) : D_0 \to D$, for all $t \geq 0$. From [69, Proposition 3.3, p.17], it follows that the space D is the core of generator $\mathcal{A}$. Then from Equation (2.23) and [69, Theorem 6.1, p.28], convergence of associated semigroup of operators in the space $C(\mathbb{U}^M)$ is guaranteed, and we have $T^{(n)}(t)g(\mathbf{u}) \to T(t)g(\mathbf{u})$ for every function $g \in C(\mathbb{U}^M)$ and for all $t \geq 0$. Also, it is obvious that T is a Feller semigroup. Following the arguments in the beginning of the proof and [69, Theorem 2.11, p.172], if $\mathbf{x}^{(n)}(0) \Rightarrow \mathbf{x}(0)$, then the process $\mathbf{x}^{(n)}(\cdot)$ weakly converges to the process $\mathbf{x}(\cdot)$.

Now, we show that if the initial value $\mathbf{x_0}$ of the process $\mathbf{x}(\cdot)$ lies in the space $\mathbb{U}^M$, then this process will remain in the space $\mathbb{U}^M$ for all time. Assume $x_{k,m}(t) = x_{k+1,m}(t)$ for some k, m and t. Then $f_{k,m}(\mathbf{x}(t)) \geq 0$ and $f_{k+1,m}(\mathbf{x}(t)) \leq 0$. Also, if $x_{k,m}(t)$ equals 0, then its associated time derivative $f_{k,m}(\mathbf{x}(t)) \geq 0$, and if $x_{k,m}(t)$ equals 1, then its associated time derivative $f_{k,m}(\mathbf{x}(t)) \leq 0$. Moreover, it is obvious that $x_{0,m}(t) = 1$. These conditions guarantee that $1 = x_{0,m}(t) \geq x_{k,m}(t) \geq x_{k+1,m}(t) \geq 0$ for all $k \in \mathbb{Z}_+$, $m \in [M]$ and $t \geq 0$. The summabality of the process $\mathbf{x}(\cdot)$ follows from the stability of the pre-limit process $\mathbf{x}^{(n)}(\cdot)$. Thus, the process $\mathbf{x}(\cdot)$ lies in the space $\mathbb{U}^M$. The uniqueness of this process follows by the Lipschitz property of the integral operator and the Picard-Lindelöf theorem, which is straightforward and is left out to the reader.

□

2.5 Steady-State Analysis of the System

Thus far, our focus has been on analyzing the transient dynamics of the system. In this section, we shift our focus to the steady-state behavior, which is characterized by the process $x^{(n)}(\infty)$. We establish a formal connection between this process and the fixed point of the mean-field limit, $\mathbf{P}$. Specifically, we show that the fixed point $\mathbf{P}$ exhibits global asymptotic stability, and as a result, it serves as a robust approximation for the process $x^{(n)}(\infty)$ as n approaches infinity.

2.5.1 Global Asymptotic Stability of the Fixed Point

Any fixed point $\mathbf{P}$ of the mean-field limit satisfies $\mathbf{f}(\mathbf{P}) = \mathbf{0}$. In the next theorem, we show that this fixed point is unique, and starting from any initial point, the mean-field process converges to this unique fixed point.

Theorem 2.5.1. *Let $\mathbf{x}(t, \mathbf{u})$ denote the mean-field process at time t starting from $\mathbf{u}$ and let $\mathbf{P}$ denote the fixed point of the mean-field process. If $\lambda < 1$, then $\lim_{t\to\infty} \mathbf{x}(t, \mathbf{u}) = \mathbf{P}$ for all $\mathbf{u} \in \mathbb{U}^M$. This implies the uniqueness and global asymptotic stability of the fixed point.*

Proof. The derivative of the mean-field limit given by Equation (2.16) is quasi-monotone. In other words, for each $k \in \mathbb{Z}_+$ and $m \in [M]$, $f_{k,m}(\mathbf{x})$ is non-decreasing in $x_{s,l}$ for all $s \neq k$ and $l \neq m$. From the quasi-monotonicity of the mean-field derivatives and [80, pp.70-74], if $\mathbf{u} \leq \mathbf{u}'$ for $\mathbf{u}, \mathbf{u}' \in \mathbb{U}^M$, then $\mathbf{x}(t, \mathbf{u}) \leq \mathbf{x}(t, \mathbf{u}')$ for all $t \geq 0$. This leads to the following inequality

$$\mathbf{x}\left(t, \min(\mathbf{P}, \mathbf{u})\right) \leq \mathbf{x}(t, \mathbf{u}) \leq \mathbf{x}\left(t, \max(\mathbf{P}, \mathbf{u})\right), \tag{2.24}$$

for all $t \geq 0$. Using the above inequality and the squeeze theorem, it is sufficient to show that $\lim_{t\to\infty} \mathbf{x}(t, \mathbf{u}) = \mathbf{P}$ for all $\mathbf{u} \geq \mathbf{P}$ and all $\mathbf{u} \leq \mathbf{P}$.

For each $N \in \mathbb{N}$, we define $z_N(t,\mathbf{u}) = \sum_{m\in[M]} \gamma_m \sum_{k\geq N} x_{k,m}(t,\mathbf{u})$ and $z_N(\mathbf{u}) = \sum_{m\in[M]} \gamma_m \sum_{k\geq N} u_{k,m}$. For the convenience of notation, we consider all α_m to be the same and equal to α. The general case with different α_m is similar but at the expense of notation. Referring to Equation (2.16), for $1 \leq N \leq \alpha+1$, we have

$$\begin{aligned}\frac{dz_N(t,\mathbf{u})}{dt} &= \sum_{m\in[M]} \gamma_m \sum_{k=N}^{\alpha+1} \frac{dx_{k,m}(t,\mathbf{u})}{dt} + \sum_{m\in[M]} \gamma_m \sum_{k=\alpha+2}^{\infty} \frac{dx_{k,m}(t,\mathbf{u})}{dt} \\ &= \lambda \sum_{m\in[M]} \frac{1-(x_{\alpha+1,m}(t,\mathbf{u}))^{d_m}}{1-x_{\alpha+1,m}(t,\mathbf{u})}(x_{N-1,m}(t,\mathbf{u}) - x_{\alpha+1,m}(t,\mathbf{u})) \prod_{l=m+1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} \\ &+ \lambda \prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \sum_{m\in[M]} \gamma_m C_m x_{N,m}(t,\mathbf{u}),\end{aligned} \tag{2.25}$$

and for $N > \alpha+1$, we have

$$\begin{aligned}\frac{dz_N(t,\mathbf{u})}{dt} &= \sum_{m\in[M]} \gamma_m \sum_{k=N}^{\infty} \frac{dx_{k,m}(t,\mathbf{u})}{dt} \\ &= \sum_{m\in[M]} \gamma_m \sum_{k=\alpha+2}^{\infty} \frac{dx_{k,m}(t,\mathbf{u})}{dt} - \sum_{m\in[M]} \gamma_m \sum_{k=\alpha+2}^{N-1} \frac{dx_{k,m}(t,\mathbf{u})}{dt} \\ &= \lambda \prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \sum_{m\in[M]} \gamma_m C_m x_{N,m}(t,\mathbf{u}) \\ &- \lambda \sum_{m\in[M]} \sum_{k=\alpha+2}^{N-1} ((x_{k-1,m}(t,\mathbf{u}))^{d_m} - (x_{k,m}(t,\mathbf{u}))^{d_m}) \prod_{l=1}^{m-1} (x_{\lceil k-1\rceil_{lm},l}(t,\mathbf{u}))^{d_l} \\ &\times \prod_{l=m+1}^{M} (x_{\lfloor k-1\rfloor_{lm},l}(t,\mathbf{u}))^{d_l}.\end{aligned} \tag{2.26}$$

More specifically, for $N = 1$ we have

$$\frac{dz_1(t,\mathbf{u})}{dt} = \lambda - \sum_{m\in[M]} \gamma_m C_m x_{1,m}(t,\mathbf{u}). \tag{2.27}$$

We show that $z_N(t,\mathbf{u})$ is uniformly bounded in t for all $N \in \mathbb{N}$. Since $z_N(t,\mathbf{u})$ is decreasing in N, it is enough to show that $z_1(t,\mathbf{u})$ is uniformly bounded in t. Consider the case $\mathbf{u} \leq \mathbf{P}$. From inequality (2.24), it follows that $\mathbf{x}(t,\mathbf{u}) \leq \mathbf{x}(t,\mathbf{P}) = \mathbf{P}$. This gives $z_1(t,\mathbf{u}) \leq z_1(\mathbf{P})$ for all $t \geq 0$. On the other hand, consider the case $\mathbf{u} \geq \mathbf{P}$. Then again from inequality (2.24), we have $\mathbf{x}(t,\mathbf{u}) \geq \mathbf{x}(t,\mathbf{P}) = \mathbf{P}$, or equivalently $\sum_{m\in[M]} \gamma_m C_m x_{1,m}(t,\mathbf{u}) \geq \sum_{m\in[M]} \gamma_m C_m P_{1,m}$. From Equation (2.27), we have

$$\frac{dz_1(t,\mathbf{u})}{dt} \leq \lambda - \sum_{m\in[M]} \gamma_m C_m P_{1,m} = \frac{dz_1(\mathbf{P})}{dt} = 0, \tag{2.28}$$

which means $z_1(t,\mathbf{u}) \leq z_1(\mathbf{u})$ for all $t \geq 0$. Hence, we conclude that $z_N(t,\mathbf{u})$ is uniformly bounded in t for all $N \geq 1$.

The derivative of $x_{k,m}(t,\mathbf{u})$ is bounded for all $k \in \mathbb{Z}_+$ and $m \in [M]$. Thus, the convergence $\lim_{t\to\infty} \mathbf{x}(t,\mathbf{u}) = \mathbf{P}$ holds for all $\mathbf{u} \geq \mathbf{P}$, if

$$\int_{t=0}^{\infty} (x_{k,m}(t,\mathbf{u}) - P_{k,m})\, dt < \infty, \qquad \forall k \in \mathbb{N}, m \in [M], \tag{2.29}$$

and similarly holds for all $\mathbf{u} \leq \mathbf{P}$, if

$$\int_{t=0}^{\infty} (P_{k,m} - x_{k,m}(t,\mathbf{u}))\, dt < \infty, \qquad \forall k \in \mathbb{N}, m \in [M]. \tag{2.30}$$

Two inequalities (2.29) and (2.30) are similar and we only show the case $\mathbf{u} \geq \mathbf{P}$. It is equivalent to show that

$$\int_{t=0}^{\infty} \sum_{m\in[M]} \gamma_m C_m (x_{k,m}(t,\mathbf{u}) - P_{k,m})\, dt < \infty, \qquad \forall k \in \mathbb{N}. \tag{2.31}$$

We use the induction method. First, we need to show that the inequality (2.31) is true for $k = 1$, i.e.,

$$\int_{t=0}^{\infty} \sum_{m\in[M]} \gamma_m C_m (x_{1,m}(t,\mathbf{u}) - P_{1,m})\, dt < \infty. \tag{2.32}$$

From Equation (2.27) and the fact that $\frac{dz_1(\mathbf{P})}{dt} = 0$, we have

$$\frac{dz_1(t,\mathbf{u})}{dt} = \frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_1(\mathbf{P})}{dt} = -\sum_{m\in[M]} \gamma_m C_m (x_{1,m}(t,\mathbf{u}) - P_{1,m}), \tag{2.33}$$

which gives

$$\int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m (x_{1,m}(t,\mathbf{u}) - P_{1,m})\, dt = -\int_{t=0}^{T} \frac{dz_1(t,\mathbf{u})}{dt}\, dt = z_1(\mathbf{u}) - z_1(T,\mathbf{u}). \tag{2.34}$$

$z_1(t,\mathbf{u})$ is uniformly bounded in t and the r.h.s. of the above integral is bounded, independent of T. Thus, as $T \to \infty$, we have

$$\int_{t=0}^{\infty} \sum_{m\in[M]} \gamma_m C_m (x_{1,m}(t,\mathbf{u}) - P_{1,m})\, dt < \infty. \tag{2.35}$$

Now, we show that inequality (2.31) is true for all $2 \leq k \leq \alpha+1$. Assume it holds for all $1 \leq k \leq L-1$ for some $2 \leq L \leq \alpha+1$. We show it is also true for $k = L$. From Equation (2.25), we have

$$\begin{aligned}\frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_L(t,\mathbf{u})}{dt} &= \lambda \sum_{m\in[M]} \frac{1-(x_{\alpha+1,m}(t,\mathbf{u}))^{d_m}}{1-x_{\alpha+1,m}(t,\mathbf{u})}(1-x_{L-1,m}(t,\mathbf{u})) \prod_{l=m+1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} \\ &- \sum_{m\in[M]} \gamma_m C_m (x_{1,m}(t,\mathbf{u}) - x_{L,m}(t,\mathbf{u})).\end{aligned} \tag{2.36}$$

Since $\frac{dz_1(\mathbf{P})}{dt} = \frac{dz_L(\mathbf{P})}{dt} = 0$, we have

$$\begin{aligned}\frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_L(t,\mathbf{u})}{dt} &= \frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_L(t,\mathbf{u})}{dt} - \frac{dz_1(\mathbf{P})}{dt} + \frac{dz_L(\mathbf{P})}{dt} \\ &= \lambda \sum_{m\in[M]} \frac{1-(x_{\alpha+1,m}(t,\mathbf{u}))^{d_m}}{1-x_{\alpha+1,m}(t,\mathbf{u})}(1-x_{L-1,m}(t,\mathbf{u})) \prod_{l=m+1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} \\ &- \lambda \sum_{m\in[M]} \frac{1-(P_{\alpha+1,m})^{d_m}}{1-P_{\alpha+1,m}}(1-P_{L-1,m}) \prod_{l=m+1}^{M} (P_{\alpha+1,l})^{d_l}\end{aligned}$$

$$- \sum_{m\in[M]} \gamma_m C_m(x_{1,m}(t,\mathbf{u}) - P_{1,m}) + \sum_{m\in[M]} \gamma_m C_m(x_{L,m}(t,\mathbf{u}) - P_{L,m}). \tag{2.37}$$

Taking integrals on both sides and using the fact that $x_{k,m}(t,\mathbf{u}) \geq P_{k,m}$, we have

$$\begin{aligned}\int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m(x_{L,m}(t,\mathbf{u}) - P_{L,m})\, dt \leq{}& \int_{t=0}^{T} \left(\frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_L(t,\mathbf{u})}{dt}\right) dt \\ &+ \int_{t=0}^{T} \lambda \sum_{m\in[M]} \frac{1-(x_{\alpha+1,m}(t,\mathbf{u}))^{d_m}}{1-x_{\alpha+1,m}(t,\mathbf{u})} \\ &\times (x_{L-1,m}(t,\mathbf{u}) - P_{L-1,m}) \prod_{l=m+1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l}\, dt \\ &+ \int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m(x_{1,m}(t,\mathbf{u}) - P_{1,m})\, dt,\end{aligned} \tag{2.38}$$

or equivalently

$$\begin{aligned}\int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m(x_{L,m}(t,\mathbf{u}) - P_{L,m})\, dt \leq{}& \int_{t=0}^{T} \left(\frac{dz_1(t,\mathbf{u})}{dt} - \frac{dz_L(t,\mathbf{u})}{dt}\right) dt \\ &+ \lambda \max_m d_m \int_{t=0}^{T} \sum_{m\in[M]} (x_{L-1,m}(t,\mathbf{u}) - P_{L-1,m})\, dt \\ &+ \int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m(x_{1,m}(t,\mathbf{u}) - P_{1,m})\, dt.\end{aligned} \tag{2.39}$$

$z_N(t,\mathbf{u})$ is uniformly bounded in t for all N and hence the first term on the r.h.s. is bounded, independent of T. From the assumption of the induction for $N = L-1$ and $N = 1$, it is evident that the second and third terms on the r.h.s. are also finite, independent of T. Hence, the result follows immediately as $T \to \infty$.

Finally, we need to show that inequality (2.31) is true for all $k > \alpha+1$. Assume it is true for all $1 \leq k \leq L-1$ for some $L > \alpha+1$. We show it is also true for $k = L$. From Equation (2.26) and the fact that $\frac{dz_L(\mathbf{P})}{dt} = 0$, we have

$$
\begin{aligned}
\frac{dz_L(t,\mathbf{u})}{dt} =& \frac{dz_L(t,\mathbf{u})}{dt} - \frac{dz_L(\mathbf{P})}{dt} \\
=& \lambda \left(\prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \prod_{l=1}^{M} (P_{\alpha+1,l})^{d_l} \right) - \sum_{m\in[M]} \gamma_m C_m (x_{L,m}(t,\mathbf{u}) - P_{L,m}) \\
-& \lambda \sum_{m\in[M]} \sum_{k=\alpha+2}^{L-1} ((x_{k-1,m}(t,\mathbf{u}))^{d_m} - (x_{k,m}(t,\mathbf{u}))^{d_m}) \prod_{l=1}^{m-1} (x_{\lceil k-1 \rceil_{lm},l}(t,\mathbf{u}))^{d_l} \\
\times& \prod_{l=m+1}^{M} (x_{\lfloor k-1 \rfloor_{lm},l}(t,\mathbf{u}))^{d_l} \\
+& \lambda \sum_{m\in[M]} \sum_{k=\alpha+2}^{L-1} ((P_{k-1,m})^{d_m} - (P_{k,m})^{d_m}) \prod_{l=1}^{m-1} (P_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (P_{\lfloor k-1 \rfloor_{lm},l})^{d_l}.
\end{aligned}
\tag{2.40}
$$

From the fact that $\mathbf{x}(t,\mathbf{u}) \geq \mathbf{P}$, we can rewrite the above equation as

$$
\begin{aligned}
\frac{dz_L(t,\mathbf{u})}{dt} \leq & \lambda \left(\prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \prod_{l=1}^{M} (P_{\alpha+1,l})^{d_l} \right) - \sum_{m\in[M]} \gamma_m C_m (x_{L,m}(t,\mathbf{u}) - P_{L,m}) \\
& - \lambda \sum_{m\in[M]} \sum_{k=\alpha_+2}^{L-1} ((x_{k-1,m}(t,\mathbf{u}))^{d_m} - (x_{k,m}(t,\mathbf{u}))^{d_m}) \prod_{l=1}^{m-1} (x_{\lceil k-1 \rceil_{lm},l}(t,\mathbf{u}))^{d_l} \\
& \times \prod_{l=m+1}^{M} (x_{\lfloor k-1 \rfloor_{lm},l}(t,\mathbf{u}))^{d_l} \\
& + \lambda \sum_{m\in[M]} \sum_{k=\alpha+2}^{L-1} ((x_{k-1,m}(t,\mathbf{u}))^{d_m} - (P_{k,m})^{d_m}) \prod_{l=1}^{m-1} (x_{\lceil k-1 \rceil_{lm},l})^{d_l} \\
& \times \prod_{l=m+1}^{M} (x_{\lfloor k-1 \rfloor_{lm},l})^{d_l}.
\end{aligned}
\tag{2.41}
$$

By rearranging the terms, we can write

$$
\sum_{m\in[M]} \gamma_m C_m (x_{L,m}(t,\mathbf{u}) - P_{L,m}) \leq - \frac{dz_L(t,\mathbf{u})}{dt}
$$

$$
\begin{aligned}
&+ \lambda \left(\prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \prod_{l=1}^{M} (P_{\alpha+1,l})^{d_l} \right) \\
&+ \lambda \sum_{m\in[M]} \sum_{k=\alpha+2}^{L-1} ((x_{k,m}(t,\mathbf{u}))^{d_m} - (P_{k,m})^{d_m}) \\
&\times \prod_{l=1}^{m-1} (x_{\lceil k-1 \rceil_{lm},l}(t,\mathbf{u}))^{d_l} \prod_{l=m+1}^{M} (x_{\lfloor k-1 \rfloor_{lm},l}(t,\mathbf{u}))^{d_l}, \qquad (2.42)
\end{aligned}
$$

or equivalently

$$
\begin{aligned}
\int_{t=0}^{T} \sum_{m\in[M]} \gamma_m C_m (x_{L,m}(t,\mathbf{u}) - P_{L,m})\, dt \leq{}& \int_{t=0}^{T} -\frac{dz_L(t,\mathbf{u})}{dt}\, dt \\
&+ \int_{t=0}^{T} \lambda \left(\prod_{l=1}^{M} (x_{\alpha+1,l}(t,\mathbf{u}))^{d_l} - \prod_{l=1}^{M} (P_{\alpha+1,l})^{d_l} \right) dt \\
&+ \lambda \max_{m} d_m \int_{t=0}^{T} \sum_{m\in[M]} \sum_{k=\alpha+2}^{L-1} (x_{k,m}(t,\mathbf{u}) - P_{k,m})\, dt.
\end{aligned}
\qquad (2.43)
$$

$z_L(t,\mathbf{u})$ is uniformly bounded in t which guarantees that the first term is bounded and independent of T. From this and the assumption of the induction for $k = \alpha+1, ..., L-1$, it follows that the r.h.s. of the above inequality is finite, independent of T, and as $T \to \infty$, it is immediately verified that

$$
\int_{t=0}^{\infty} \sum_{m\in[M]} \gamma_m C_m (x_{L,m}(t,\mathbf{u}) - P_{L,m})\, dt < \infty. \qquad (2.44)
$$

This completes the proof of the global asymptotic stability of the fixed point $\mathbf{P}$. □

Remark 2.2. Theorem 2.5.1 shows that for each $k \in \mathbb{Z}_+$ and $m \in [M]$, the convergence of $x_{k,m}$ to $P_{k,m}$ is faster than $t^{-\theta}$ for $\theta > 1$. Thus,

$$\begin{aligned}\|\mathbf{x}(t,\mathbf{u})-\mathbf{P}\|_2^2 &= \sum_{m\in[M]}\sum_{k\in\mathbb{Z}_+}|\mathbf{x}_{k,m}(t,\mathbf{u})-P_{k,m}|^2\\ &\leq \sum_{m\in[M]}\sum_{k\in\mathbb{Z}_+}(t^{-\theta}|u_{k,m}-P_{k,m}|)^2\\ &= t^{-2\theta}\left\|\mathbf{u}-\mathbf{P}\right\|_2^2,\end{aligned} \tag{2.45}$$

and hence,

$$\|\mathbf{x}(t,\mathbf{u})-\mathbf{P}\|_2 \leq t^{-\theta}\left\|\mathbf{u}-\mathbf{P}\right\|_2. \tag{2.46}$$

Indeed, we will show in Chapter 3 that the convergence rate of the mean-field limit to its fixed point is exponential in the homogeneous case, i.e. $M = 1$. In the heterogeneous case, the proof is much more complicated and we conjecture that it also holds for which we provide numerical evidence.

The interchange of limits for the process $\mathbf{x}^{(n)}(\cdot)$ follows from Prohorov's theorem [70] and Theorem 2.5.1. It can be readily verified that

$$\lim_{n\to\infty}\lim_{t\to\infty}\mathbf{x}^{(n)}(t) = \lim_{t\to\infty}\lim_{n\to\infty}\mathbf{x}^{(n)}(t). \tag{2.47}$$

This shows that the stationary distribution of servers, which exists and is unique due to the stability of the system, converges to the unique fixed point of the mean-field limit as $n\to\infty$. In other words,

$$\lim_{n\to\infty}\mathbf{x}^{(n)}(\infty) = \mathbf{P}. \tag{2.48}$$

Therefore, the asymptotic mean response time of jobs in the steady-state system can be calculated as follows. For a given system of size n, the average number of jobs at equilibrium is equal to $n\mathbb{E}\left[\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}x_{k,m}^{(n)}(\infty)\right]$. Let the mean sojourn time of jobs be denoted by $\mathbb{E}\left[D^{(n)}\right]$. Then by Little's law, we have

$$n\lambda(n)\mathbb{E}\left[D^{(n)}\right] = n\mathbb{E}\left[\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}x_{k,m}^{(n)}(\infty)\right]. \tag{2.49}$$

Due to the stability of the system, the above sum is finite and we have

$$\lambda(n)\mathbb{E}\left[D^{(n)}\right] = \sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty} \mathbb{E}\left[x_{k,m}^{(n)}(\infty)\right]. \tag{2.50}$$

By taking limits from both sides, we get

$$\lim_{n\to\infty} \mathbb{E}\left[D^{(n)}\right] = \frac{1}{\lambda} \sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty} P_{k,m}. \tag{2.51}$$

Hence, the asymptotic mean response time of jobs converges to the mean response time of jobs in the mean-field limit.

2.6 Conclusion

In this chapter, we studied randomized threshold-based load balancing policies designed for a class of large heterogeneous processor sharing systems. These policies exhibit remarkable adaptability, allowing for versatile configuration of threshold settings to accommodate a wide array of load balancing policies. First, we demonstrated that policies within this category achieve the maximal attainable stability region and hence are throughput optimal. Moreover, we showed that as the system size n approaches infinity, the transient empirical occupancy measure of servers converges to its deterministic mean-field limit. The mean-field limit, represented by an infinite system of ODEs, captures the system's behavior for finite times t in the system size limit. Additionally, we established that the stationary empirical distribution of system occupancy converges to the unique fixed point of the mean-field limit as n tends to infinity. The convergence offers an opportunity to approximate the complicated stationary measure of the system with the fixed point of the mean-field limit in the infinite regime enabling an analysis of the system's asymptotic mean response time in relation to that of the mean-field system.

Chapter 3

Mean-Field Fluctuations in Threshold-Based Load Balancing Policies and Finite-Sized Systems

In this chapter, we build upon the results presented in Chapter 2, with the goal of understanding how well the mean-field distribution approximates the empirical occupation measure of the system when the system size is large but finite. We continue to study large-scale heterogeneous processor sharing systems with n servers and M distinct server speeds. These servers are accessed by jobs following the type-based JBT policy, as detailed in Section 2.1. Prior research on evaluating the accuracy of mean-field approximations has predominantly focused on the stationary regime, offering error estimates exclusively during steady-state conditions. However, we study the error estimates in both transient and stationary regimes. To achieve this, we introduce a fluctuation process that represents the difference between the actual system occupancy distribution and its mean-field limit and study the limiting behavior of this fluctuation process as the system size tends to infinity. Via the use of Functional Central Limit Theorems (FCLTs), we show that as the system size n tends to infinity, the diffusion-scaled fluctuation process converges to an Ornstein-Uhlenbeck process whose drift and diffusion coefficients depend on the mean-field limit of the system. These results enable us to estimate the error of the mean-field approximations to the empirical occupancy measures associated with a system with n servers. We then use these results to show that the mean response time of jobs in a finite-sized system can be approximated by the response time given by the mean-field limit and the error is $O(\frac{1}{\sqrt{n}})$ for which the constants can be precisely calculated.

The remainder of this chapter is structured as follows. In Section 3.1, we present functional central limit theorems in the transient regime and demonstrate the convergence of the scaled fluctuation process to an OU process. In Section 3.2, we study the fluctuation process in the stationary regime and provide practical applications of the results we derive. In Section 3.3, we present numerical evidence to validate the accuracy of results. Finally, concluding remarks are found in Section 3.4.

3.1 Transient Behavior of the Fluctuation Process

In this section, we study fluctuations in the empirical distribution of system occupancy around its mean-field limit in the transient regime. We consider the system model described in Section 2.1. We introduce the diffusion-scaled fluctuation process at time t, denoted as $\mathbf{Z}^{(n)}(t)$, defined as

$$\mathbf{Z}^{(n)}(t) = \left(Z_{k,m}^{(n)}(t), k \in \mathbb{Z}_+, m \in [M]\right), \tag{3.1}$$

with each component $Z_{k,m}^{(n)}(t)$ given by

$$Z_{k,m}^{(n)}(t) = \sqrt{n\gamma_m}\left(x_{k,m}^{(n)}(t) - x_{k,m}(t)\right) \in \mathbb{R}, \tag{3.2}$$

where the processes $\mathbf{x}^{(n)}(\cdot)$ and $\mathbf{x}(\cdot)$ represent the empirical distribution of servers in a system with n servers and its mean-field limit, respectively. The mean-field limit $\mathbf{x}(\cdot)$ is established in Chapter 2, Section 2.4. To characterize the time evolution of the empirical distribution $\mathbf{x}^{(n)}(\cdot)$, we consider two sets of mutually independent unit rate standard Poisson processes, denoted as $(\mathcal{N}_{k,m}, k \in \mathbb{N}, m \in [M])$ and $(\mathcal{D}_{k,m}, k \in \mathbb{N}, m \in [M])$. Specifically, the process $\mathcal{N}_{k,m}$ represents arrivals to all type m servers with exactly $k-1$ unfinished jobs, while the process $\mathcal{D}_{k,m}$ accounts for departures from servers of type m that have exactly k processing jobs. Using arrival and departure rates derived in the proof of Lemma 2.4.1 and applying random time changes to Poisson processes as outlined in [81], we can express the time evolution of a type $m \in [M]$ server as follows.

$$x_{0,m}^{(n)}(t) = 1,$$

$$x_{k,m}^{(n)}(t) = x_{k,m}^{(n)}(0) + \frac{1}{n\gamma_m}\mathcal{N}_{k,m}\left(n\lambda(n)\int_{s=0}^{t} \frac{1-\left(x_{\alpha_m+1,m}^{(n)}(s)\right)^{d_m}}{1-x_{\alpha_m+1,m}^{(n)}(s)}\left(x_{k-1,m}^{(n)}(s) - x_{k,m}^{(n)}(s)\right)\right.$$

$$\times \prod_{l=m+1}^{M} \left(x^{(n)}_{\alpha_l+1,l}(s)\right)^{d_l} \mathbb{1}_{(k\leq\alpha_m+1)}\, ds + n\lambda(n)\int_{s=0}^{t} \left((x^{(n)}_{k-1,m}(s))^{d_m} - (x^{(n)}_{k,m}(s))^{d_m}\right)$$

$$\times \prod_{l=1}^{m-1}\left(x^{(n)}_{\lceil k-1\rceil_{lm},l}(s)\right)^{d_l}\prod_{l=m+1}^{M}\left(x^{(n)}_{\lfloor k-1\rfloor_{lm},l}(s)\right)^{d_l}\mathbb{1}_{(k>\alpha_m+1)}\,ds\Bigg)$$

$$-\frac{1}{n\gamma_m}\mathcal{D}_{k,m}\left(n\gamma_m C_m\int_{s=0}^{t} \left(x^{(n)}_{k,m}(s)-x^{(n)}_{k+1,m}(s)\right)ds\right),\quad k\in\mathbb{N}. \tag{3.3}$$

We introduce three operators W_1, W_2, and W_3, which map the space $\mathbb{U}^M$ to the space $\left(\mathbb{R}^{\{0,1,2,\dots\}}\right)^M$. For any $\mathbf{u}\in\mathbb{U}^M$ and $m\in[M]$, these operators are defined as follows.

$$\begin{aligned}
(W_1(\mathbf{u}))_{0,m} &= (W_2(\mathbf{u}))_{0,m} = (W_3(\mathbf{u}))_{0,m} = 0,\\
(W_1(\mathbf{u}))_{k,m} &= \frac{\lambda(1-(u_{\alpha_m+1,m})^{d_m})}{\gamma_m(1-u_{\alpha_m+1,m})}(u_{k-1,m}-u_{k,m})\prod_{l=m+1}^{M}(u_{\alpha_l+1,l})^{d_l},\quad 1\leq k\leq\alpha_m+1,\\
(W_1(\mathbf{u}))_{k,m} &= \frac{\lambda}{\gamma_m}((u_{k-1,m})^{d_m}-\left(u_{k,m}\right)^{d_m})\prod_{l=1}^{m-1}(u_{\lceil k-1\rceil_{lm},l})^{d_l}\prod_{l=m+1}^{M}(u_{\lfloor k-1\rfloor_{lm},l})^{d_l},\quad k>\alpha_m+1,\\
(W_2(\mathbf{u}))_{k,m} &= C_m(u_{k,m}-u_{k+1,m}),\quad k\in\mathbb{N},\\
(W_3(\mathbf{u}))_{k,m} &= \frac{\beta\sqrt{\gamma_m}}{\lambda}(W_1(\mathbf{u}))_{k,m},\quad k\in\mathbb{N}.
\end{aligned} \tag{3.4}$$

Furthermore, we define the operator $W=W_1-W_2$. This operator W corresponds to the mapping $\mathbf{f}$ that governs the dynamics of the mean-field limit, as defined in Equation (2.16). Moreover, for a given fixed $m\in[M]$, the operator $W_{1,m}:\mathbb{U}^M\to\mathbb{R}^{\{0,1,2,\dots\}}$ is expressed as

$$(W_{1,m}(\mathbf{u}))_k = (W_1(\mathbf{u}))_{k,m}\quad, k\in\mathbb{Z}_+. \tag{3.5}$$

The operators $W_{2,m}$, $W_{3,m}$, and W_m are defined analogously.

Lemma 3.1.1. *For each $m\in[M]$, the operators $W_{1,m}$, $W_{2,m}$, $W_{3,m}$, and therefore W_m, are Lipschitz with respect to both the ℓ_1-norm and the ℓ_2-norm.*

Proof. See Appendix A.3. □

Combining Equations (2.15)-(2.16) with Equations (3.2)-(3.4), we can write the fluctuation process as

$$
\begin{aligned}
Z_{k,m}^{(n)}(t) &= \sqrt{n\gamma_m}\left(x_{k,m}^{(n)}(t) - x_{k,m}(t)\right) \\
&= \sqrt{n\gamma_m}\Bigg(x_{k,m}^{(n)}(0) + \frac{1}{n\gamma_m}\mathcal{N}_{k,m}\Big(n\gamma_m\int_{s=0}^{t}\left(W_1(\mathbf{x}^{(n)}(s))\right)_{k,m}\,ds \\
&\qquad\qquad\qquad\qquad - \sqrt{n\gamma_m}\int_{s=0}^{t}\left(W_3(\mathbf{x}^{(n)}(s))\right)_{k,m}\,ds\Big) \\
&\qquad - \frac{1}{n\gamma_m}\mathcal{D}_{k,m}\left(n\gamma_m\int_{s=0}^{t}\left(W_2(\mathbf{x}^{(n)}(s))\right)_{k,m}\,ds\right) - x_{k,m}(0) \\
&\qquad - \int_{s=0}^{t}\left(W_1(\mathbf{x}(s))\right)_{k,m}\,ds + \int_{s=0}^{t}\left(W_2(\mathbf{x}(s))\right)_{k,m}\,ds\Bigg), \quad m \in [M], k \in \mathbb{Z}_+.
\end{aligned}
\tag{3.6}
$$

We write the Poisson processes in their compensated forms and reorganize the terms to obtain

$$
\begin{aligned}
Z_{k,m}^{(n)}(t) = Z_{k,m}^{(n)}(0) &+ \sqrt{n\gamma_m}\int_{s=0}^{t}\left(\left(W(\mathbf{x}^{(n)}(s))\right)_{k,m} - \left(W(\mathbf{x}(s))\right)_{k,m}\right)\,ds \\
&- \int_{s=0}^{t}\left(W_3(\mathbf{x}^{(n)}(s))\right)_{k,m}\,ds + M_{k,m}^{(n)}(t), \quad m \in [M], k \in \mathbb{Z}_+,
\end{aligned}
\tag{3.7}
$$

where $M_{k,m}^{(n)}(t)$ is a centered martingale that is uniquely characterized by its quadratic variation

$$
< M_{k,m}^{(n)} >_t = \int_{s=0}^{t}\left(\left(W_1(\mathbf{x}^{(n)}(s))\right)_{k,m} + \left(W_2(\mathbf{x}^{(n)}(s))\right)_{k,m} - \frac{1}{\sqrt{n\gamma_m}}\left(W_3(\mathbf{x}^{(n)}(s))\right)_{k,m}\right)\,ds.
\tag{3.8}
$$

For each $m \in [M]$, we group all the independent and centered martingales $M_{k,m}^{(n)}(t)$ into $\mathbf{M}_m^{(n)}(t) = \left(M_{k,m}^{(n)}(t), k \in \mathbb{Z}+\right)$. With this definition and Equation (3.5), we can reformulate Equation (3.7) as

$$
\begin{aligned}
\mathbf{Z}_m^{(n)}(t) = \mathbf{Z}_m^{(n)}(0) &+ \sqrt{n\gamma_m}\int_{s=0}^{t}\left(W_m(\mathbf{x}^{(n)}(s)) - W_m(\mathbf{x}(s))\right)\,ds - \int_{s=0}^{t}W_{3,m}(\mathbf{x}^{(n)}(s))\,ds \\
&+ \mathbf{M}_m^{(n)}(t), \quad m \in [M],
\end{aligned}
\tag{3.9}
$$

where $\mathbf{Z}_m^{(n)}(t) = \left(Z_{k,m}^{(n)}(t), k \in \mathbb{Z}_+\right)$ for the given value of m. The Stochastic Differential Equation (SDE) (3.9) describes the behavior of the fluctuation process for servers of type m in the transient regime. We claim that this fluctuation process is stochastically bounded on any finite time interval $[0, T]$ and we prove it in the subsequent lemma.

Lemma 3.1.2. *If* $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] < \infty$*, then for any finite* $T > 0$*, we have* $\limsup_{n\to\infty} \mathbb{E}\left[\sup_{0\le t\le T} \left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right] < \infty$.

Proof. First, we show that the martingale

$$\mathbf{M}^{(n)}(t) = \left(\mathbf{M}_m^{(n)}(t), m \in [M]\right) = \left(\left(M_{k,m}^{(n)}(t), k \in \mathbb{Z}_+\right), m \in [M]\right) \tag{3.10}$$

is square-integrable with respect to the ℓ_2-norm. From the definition of ℓ_2-norm and the martingale property, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{M}^{(n)}(t)\right\|_2^2\right] &= \mathbb{E}\left[\sum_{m\in[M]} \sum_{k\in\mathbb{Z}_+} |M_{k,m}^{(n)}(t)|^2\right] \\
&= \sum_{m\in[M]} \sum_{k\in\mathbb{Z}_+} \mathbb{E}\left[< M_{k,m}^{(n)} >_t\right] \\
&= \mathbb{E}\left[\left\|< \mathbf{M}^{(n)} >_t\right\|_1\right].
\end{aligned} \tag{3.11}$$

The equation above establishes a connection between the ℓ_2-norm of the martingale and the ℓ_1-norm of its quadratic variation. Utilizing the Lipschitz property of operators W_1, W_2, and W_3 with respect to the ℓ_1-norm, and taking into account that the process $\mathbf{x}^{(n)}(\cdot)$ is summable for all n, we can deduce from Equation (3.8) that the ℓ_1-norm of the quadratic variation of the martingale $\mathbf{M}^{(n)}(t)$ is finite. Therefore, the martingale $\mathbf{M}^{(n)}(t)$ is indeed square-integrable. Referring to Equation (3.9), we have

$$\begin{aligned}
\mathbf{Z}_m^{(n)}(t) = \mathbf{Z}_m^{(n)}(0) + \sqrt{n\gamma_m} \int_{s=0}^{t} \left(W_m(\mathbf{x}^{(n)}(s)) - W_m(\mathbf{x}(s))\right) ds - \int_{s=0}^{t} W_{3,m}(\mathbf{x}^{(n)}(s))\, ds \\
+ \mathbf{M}_m^{(n)}(t).
\end{aligned} \tag{3.12}$$

We take the norm of both sides and obtain

$$\left\|\mathbf{Z}_m^{(n)}(t)\right\|_2 \leq \left\|\mathbf{Z}_m^{(n)}(0)\right\|_2 + \sqrt{n\gamma_m}\int_{s=0}^{t} B_W \left\|\mathbf{x}^{(n)}(s) - \mathbf{x}(s)\right\|_2 ds + \int_{s=0}^{t} B_{W_3} \left\|\mathbf{x}^{(n)}(s)\right\|_2 ds \\ + \left\|\mathbf{M}_m^{(n)}(t)\right\|_2, \tag{3.13}$$

where B_W and B_{W_3} are the ℓ_2 Lipschitz constants associated with operators W and W_3, respectively. Employing the Cauchy-Schwartz inequality, we have

$$\left\|\mathbf{Z}_m^{(n)}(t)\right\|_2^2 \leq 4\Big(\left\|\mathbf{Z}_m^{(n)}(0)\right\|_2^2 + n\gamma_m B_W^2 t\int_{s=0}^{t} \left\|\mathbf{x}^{(n)}(s) - \mathbf{x}(s)\right\|_2^2 ds + B_{W_3}^2 t\int_{s=0}^{t} \left\|\mathbf{x}^{(n)}(s)\right\|_2^2 ds \\ + \left\|\mathbf{M}_m^{(n)}(t)\right\|_2^2\Big). \tag{3.14}$$

By the definition of ℓ_2-norm, we have

$$\left\|\mathbf{x}^{(n)}(s) - \mathbf{x}(s)\right\|_2^2 = \sum_{l\in[M]} \left\|\mathbf{x}_l^{(n)}(s) - \mathbf{x}_l(s)\right\|_2^2. \tag{3.15}$$

Substituting the above expression into inequality (3.14), we obtain

$$\left\|\mathbf{Z}_m^{(n)}(t)\right\|_2^2 \leq 4\Big(\left\|\mathbf{Z}_m^{(n)}(0)\right\|_2^2 + B_W^2 t\beta_m\int_{s=0}^{t} \left\|\mathbf{Z}^{(n)}(s)\right\|_2^2 ds + B_{W_3}^2 t\int_{s=0}^{t} \left\|\mathbf{x}^{(n)}(s)\right\|_2^2 ds \\ + \left\|\mathbf{M}_m^{(n)}(t)\right\|_2^2\Big), \tag{3.16}$$

where $\beta_m = \max_{l\in[M]} \frac{\gamma_m}{\gamma_l}$. Summing over all $m \in [M]$, we can write

$$\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2 \leq 4\Big(\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2 + B_W^2 t\beta\int_{s=0}^{t} \left\|\mathbf{Z}^{(n)}(s)\right\|_2^2 ds + MB_{W_3}^2 t\int_{s=0}^{t} \left\|\mathbf{x}^{(n)}(s)\right\|_2^2 ds \\ + \left\|\mathbf{M}^{(n)}(t)\right\|_2^2\Big), \tag{3.17}$$

where $\beta = \sum_{m\in[M]} \beta_m$. Applying Gronwall's Lemma, we obtain

$$\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2 \leq 4\left(\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2 + MB_{W_3}^2 t\int_{s=0}^{t} \left\|\mathbf{x}^{(n)}(s)\right\|_2^2 ds + \left\|\mathbf{M}^{(n)}(t)\right\|_2^2\right) e^{2\beta B_W^2 t^2}. \tag{3.18}$$

By taking the supremum of both sides, we get

$$\sup_{0\le t\le T}\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2 \le 4\Bigg(\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2 + MB_{W_3}^2T^2\sup_{0\le t\le T}\left\|\mathbf{x}^{(n)}(t)\right\|_2^2 + \sup_{0\le t\le T}\left\|\mathbf{M}^{(n)}(t)\right\|_2^2\Bigg)e^{2\beta B_W^2T^2}. \tag{3.19}$$

By Doob's Inequality,

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right] \le 4\Bigg(&\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] + MB_{W_3}^2T^2\mathbb{E}\left[\sup_{0\le t\le T}\left\|\mathbf{x}^{(n)}(t)\right\|_2^2\right]\\ &+4\mathbb{E}\left[\sum_{k\in\mathbb{Z}_+}\sum_{m\in[M]} < M_{k,m}^{(n)} >_T\right]\Bigg)e^{2\beta B_W^2T^2}.\end{aligned} \tag{3.20}$$

The ℓ_1-norm of the process $\mathbf{x}^{(n)}(t)$ is finite, in accordance with the definition of the space $\mathbb{U}^{(n)}$. Additionally, through a comparison of norms, we establish that $\left\|\mathbf{x}^{(n)}(t)\right\|_2 \le \left\|\mathbf{x}^{(n)}(t)\right\|_1$, which implies that its ℓ_2-norm is also finite. Furthermore, the martingale $\mathbf{M}^{(n)}(t)$ is square-integrable, implying that its quadratic variation lies in the L^1 space. This shows that the fluctuation process is stochastically bounded on any finite time interval $[0,T]$ and completes the proof. □

In the sequel, we define a new set of SDEs and show that as the system size increases, the fluctuation process in the transient regime converges in distribution to the solution of these SDEs. For this purpose, let the operator $L : \left(\mathbb{R}^{\{0,1,2,\ldots\}}\right)^M \to \left(\mathbb{R}^{\{0,1,2,\ldots\}}\right)^M$, be the linearization of mean-field equations (2.15)-(2.16) around a solution $\mathbf{h}(t)$ to these equations. Then we can write

$$\frac{d\mathbf{g}}{dt} = L(\mathbf{h}(t))\mathbf{g}(t), \tag{3.21}$$

where for $\mathbf{u} \in \mathbb{U}^M$ and for $\mathbf{w} \in \left(\mathbb{R}^{\{0,1,2,\ldots\}}\right)^M$, the operator $L(\mathbf{u})$ is given as follows.

$$
\begin{aligned}
&(L(\mathbf{u})\mathbf{w})_{0,m} = 0, \quad m \in [M],\\[2ex]
&(L(\mathbf{u})\mathbf{w})_{k,m} = \sum_{i=1}^{d_m-1} \frac{\lambda i}{\gamma_m} (u_{\alpha_m+1,m})^{i-1} (u_{k-1,m} - u_{k,m}) \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} w_{\alpha_m+1,m}\\
&+ \frac{\lambda(1-(u_{\alpha_m+1,m})^{d_m})}{\gamma_m(1-u_{\alpha_m+1,m})} \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} (w_{k-1,m} - w_{k,m})\\
&+ \sum_{i=m+1}^{M} \frac{\lambda d_i}{\gamma_m u_{\alpha_i+1,i}} \frac{1-(u_{\alpha_m+1,m})^{d_m}}{1-u_{\alpha_m+1,m}} (u_{k-1,m} - u_{k,m}) \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} w_{\alpha_i+1,i}\\
&- C_m(w_{k,m} - w_{k+1,m}), \quad m \in [M], 1 \le k \le \alpha_m + 1,\\[2ex]
&(L(\mathbf{u})\mathbf{w})_{k,m} = \frac{\lambda d_m}{\gamma_m}((u_{k-1,m})^{d_m-1} w_{k-1,m} - (u_{k,m})^{d_m-1} w_{k,m}) \prod_{l=1}^{m-1} (u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l}\\
&+ \sum_{i=1}^{m-1} \frac{\lambda d_i}{\gamma_m u_{\lceil k-1 \rceil_{im},i}} ((u_{k-1,m})^{d_m} - (u_{k,m})^{d_m}) \prod_{l=1}^{m-1} (u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l} w_{\lceil k-1 \rceil_{im},i}\\
&+ \sum_{i=m+1}^{M} \frac{\lambda d_i}{\gamma_m u_{\lfloor k-1 \rfloor_{im},i}} ((u_{k-1,m})^{d_m} - (u_{k,m})^{d_m}) \prod_{l=1}^{m-1} (u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l} w_{\lfloor k-1 \rfloor_{im},i}\\
&- C_m(w_{k,m} - w_{k+1,m}), \quad m \in [M], k > \alpha_m + 1.
\end{aligned}
\tag{3.22}
$$

For each $m \in [M]$, we define the operator $L(\mathbf{u})_m : \left(\mathbb{R}^{\{0,1,2,\dots\}}\right)^M \to \mathbb{R}^{\{0,1,2,\dots\}}$ such that $(L(\mathbf{u})_m\mathbf{w})_k = (L(\mathbf{u})\mathbf{w})_{k,m}$ for $k \in \mathbb{Z}_+$. With this operator in place, we are now ready to introduce the limiting set of SDEs.

Definition 3.1. Let $\mathbf{Z}(t) = (\mathbf{Z}_m(t), m \in [M]) = ((Z_{k,m}(t), k \in \mathbb{Z}_+), m \in [M])$ represent a solution to the following set of SDEs.

$$
\mathbf{Z}_m(t) = \mathbf{Z}_m(0) + \sqrt{\gamma_m} \int_{s=0}^{t} L(\mathbf{x}(s))_m \mathbf{Z}(s)\, ds - \int_{s=0}^{t} W_{3,m}(\mathbf{x}(s))\, ds + \mathbf{M}_m(t), \qquad m \in [M],
\tag{3.23}
$$

where the process $\mathbf{x}(\cdot)$ denotes the mean-field limit, and $\mathbf{M}_m(t) = (M_{k,m}(t), k \in \mathbb{Z}_+)$ is a collection of independent, real-valued, continuous and Gaussian martingales $M_{k,m}(t)$ that are uniquely characterized by their deterministic quadratic variation

$$< M_{k,m} >_t = \int_{s=0}^{t} \left((W_1(\mathbf{x}(s)))_{k,m} + (W_2(\mathbf{x}(s)))_{k,m} \right) ds. \tag{3.24}$$

The solution to the above set of SDEs is an OU process.

The following theorem provides sufficient conditions for ensuring both the uniqueness and boundedness of the solution to the set of SDEs (3.23).

Theorem 3.1.3. *If $\mathbb{E}\left[\|\mathbf{Z}(0)\|_2^2\right] < \infty$, then SDEs (3.23) have a unique strong solution $\mathbf{Z}(t)$. Furthermore, $\mathbb{E}\left[\sup_{0 \le t \le T} \|\mathbf{Z}(t)\|_2^2\right] < \infty$ for any finite $T > 0$.*

Proof. In the first part of the proof, we establish the uniqueness of the solution. We start by showing that, for any $\mathbf{u} \in \mathbb{U}^M$ and $m \in [M]$, the operator $L(\mathbf{u})_m$ is bounded. Let us consider $\mathbf{w} \in \left(\mathbb{R}^{0,1,2,\cdots}\right)^M$. We have

$$\begin{aligned} \|L(\mathbf{u})_m \mathbf{w}\|_2^2 &= \sum_{k \in \mathbb{Z}_+} \left|(L(\mathbf{u})\mathbf{w})_{k,m}\right|^2 \\ &\le \left(4\lambda^2 \frac{\max_m d_m^2}{\min_m \gamma_m^2} \left((\alpha_M + 1)(1 + \max_m d_m^2) + 4 + \tilde{K_1} + \tilde{K_2}\right) + 16C_M^2\right) \|\mathbf{w}\|_2^2, \end{aligned} \tag{3.25}$$

where $\tilde{K_1}$ and $\tilde{K_2}$ are two positive finite constants such that for each $m \in [M]$,

$$\sum_{k=\alpha_m+2}^{\infty} \sum_{l=1}^{m-1} \left|w_{\lceil k-1 \rceil_{lm},l}\right|^2 \le \tilde{K_1} \|\mathbf{w}\|_2^2, \tag{3.26}$$

$$\sum_{k=\alpha_m+2}^{\infty} \sum_{l=m+1}^{M} \left|w_{\lfloor k-1 \rfloor_{lm},l}\right|^2 \le \tilde{K_2} \|\mathbf{w}\|_2^2. \tag{3.27}$$

Hence, for every $m \in [M]$, the operator $L(\mathbf{u})_m$ is bounded with the bound denoted as B_L, where $B_L^2 = 4\lambda^2 \frac{\max_m d_m^2}{\min_m \gamma_m^2} \left((\alpha_M + 1)(1 + \max_m d_m^2) + 4 + \tilde{K_1} + \tilde{K_2}\right) + 16C_M^2$. Using this result, we show the uniqueness of the solution to SDEs (3.23). Consider two solutions

$\mathbf{Z}^{(\mathbf{1})}(t) = \left(\mathbf{Z}_m^{(1)}(t), m \in [M]\right)$ starting from $\mathbf{Z}^{(\mathbf{1})}(0)$ and $\mathbf{Z}^{(\mathbf{2})}(t) = \left(\mathbf{Z}_m^{(2)}(t), m \in [M]\right)$ starting from $\mathbf{Z}^{(\mathbf{2})}(0)$. Then for each $m \in [M]$, we have

$$\mathbf{Z}_m^{(1)}(t) - \mathbf{Z}_m^{(2)}(t) = \mathbf{Z}_m^{(1)}(0) - \mathbf{Z}_m^{(2)}(0) + \sqrt{\gamma_m} \int_{s=0}^{t} L(\mathbf{x}(s))_m \left(\mathbf{Z}^{(1)}(s) - \mathbf{Z}^{(2)}(s)\right) ds. \quad (3.28)$$

Therefore,

$$\left\|\mathbf{Z}_m^{(1)}(t) - \mathbf{Z}_m^{(2)}(t)\right\|_2 \leq \left\|\mathbf{Z}_m^{(1)}(0) - \mathbf{Z}_m^{(2)}(0)\right\|_2 + \sqrt{\gamma_m} B_L \int_{s=0}^{t} \left\|\mathbf{Z}^{(1)}(s) - \mathbf{Z}^{(2)}(s)\right\|_2 ds. \quad (3.29)$$

From the Cauchy-Schwartz inequality, we have

$$\left\|\mathbf{Z}_m^{(1)}(t) - \mathbf{Z}_m^{(2)}(t)\right\|_2^2 \leq 2\left(\left\|\mathbf{Z}_m^{(1)}(0) - \mathbf{Z}_m^{(2)}(0)\right\|_2^2 + \gamma_m B_L^2 t \int_{s=0}^{t} \left\|\mathbf{Z}^{(1)}(s) - \mathbf{Z}^{(2)}(s)\right\|_2^2 ds\right). \quad (3.30)$$

By summing over all $m \in [M]$, we have

$$\left\|\mathbf{Z}^{(1)}(t) - \mathbf{Z}^{(2)}(t)\right\|_2^2 \leq 2\left(\left\|\mathbf{Z}^{(1)}(0) - \mathbf{Z}^{(2)}(0)\right\|_2^2 + B_L^2 t \int_{s=0}^{t} \left\|\mathbf{Z}^{(1)}(s) - \mathbf{Z}^{(2)}(s)\right\|_2^2 ds\right). \quad (3.31)$$

From Gromwall's Lemma, we have

$$\left\|\mathbf{Z}^{(1)}(t) - \mathbf{Z}^{(2)}(t)\right\|_2^2 \leq 2e^{B_L^2 t^2} \left\|\mathbf{Z}^{(1)}(0) - \mathbf{Z}^{(2)}(0)\right\|_2^2. \quad (3.32)$$

Hence,

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(1)}(t) - \mathbf{Z}^{(2)}(t)\right\|_2^2\right] \leq 2e^{B_L^2 t^2} \mathbb{E}\left[\left\|\mathbf{Z}^{(1)}(0) - \mathbf{Z}^{(2)}(0)\right\|_2^2\right]. \quad (3.33)$$

The above inequality guarantees that if the initial conditions are bounded and satisfy $\mathbf{Z}^{(1)}(0) = \mathbf{Z}^{(2)}(0)$, then the two solutions $\mathbf{Z}^{(1)}(t)$ and $\mathbf{Z}^{(2)}(t)$ are almost surely identical for every rational t. Since these solutions have continuous sample paths, we can conclude that $\mathbf{Z}^{(1)}(t) = \mathbf{Z}^{(2)}(t)$ for every $t \geq 0$.

Next, we establish the boundedness of this solution when its initial starting point is bounded. By applying the same arguments as in the proof of Lemma 3.1.2, we can demonstrate that the martingale $\mathbf{M}(t) = (\mathbf{M}_m(t), m \in [M]) = ((M_{k,m}(t), k \in \mathbb{Z}_+), m \in [M])$ is square-integrable with respect to the ℓ_2-norm.

From Equation (3.23) we have

$$\mathbf{Z}_m(t) = \mathbf{Z}_m(0) + \sqrt{\gamma_m}\int_{s=0}^{t} L(\mathbf{x}(s))_m\mathbf{Z}(s)\,ds - \int_{s=0}^{t} W_{3,m}(\mathbf{x}(s))\,ds + \mathbf{M}_m(t). \tag{3.34}$$

By taking the ℓ_2-norm of both sides, we get

$$\|\mathbf{Z}_m(t)\|_2 \leq \|\mathbf{Z}_m(0)\|_2 + \sqrt{\gamma_m}\int_{s=0}^{t} B_L\,\|\mathbf{Z}(s)\|_2\,ds + \int_{s=0}^{t} B_{W_3}\,\|\mathbf{x}(s)\|_2\,ds + \|\mathbf{M}_m(t)\|_2\,, \tag{3.35}$$

where B_L is the bound on the operator L and B_{W_3} is the Lipschitz constant associated with the operator W_3. Using the Cauchy-Schwartz inequality, we have

$$\|\mathbf{Z}_m(t)\|_2^2 \leq 4\Big(\|\mathbf{Z}_m(0)\|_2^2 + \gamma_m B_L^2 t\int_{s=0}^{t} \|\mathbf{Z}(s)\|_2^2\,ds + B_{W_3}^2 t\int_{s=0}^{t} \|\mathbf{x}(s)\|_2^2\,ds + \|\mathbf{M}_m(t)\|_2^2\Big). \tag{3.36}$$

Summing over all $m \in [M]$, we obtain

$$\|\mathbf{Z}(t)\|_2^2 \leq 4\Big(\|\mathbf{Z}(0)\|_2^2 + B_L^2 t\int_{s=0}^{t} \|\mathbf{Z}(s)\|_2^2\,ds + MB_{W_3}^2 t\int_{s=0}^{t} \|\mathbf{x}(s)\|_2^2\,ds + \|\mathbf{M}(t)\|_2^2\Big). \tag{3.37}$$

Applying Gronwall's Lemma, we have

$$\|\mathbf{Z}(t)\|_2^2 \leq 4\left(\|\mathbf{Z}(0)\|_2^2 + MB_{W_3}^2 t\int_{s=0}^{t} \|\mathbf{x}(s)\|_2^2\,ds + \|\mathbf{M}(t)\|_2^2\right) e^{2B_L^2 t^2}. \tag{3.38}$$

By taking the supremum of both sides, we obtain

$$\sup_{0\leq t\leq T}\|\mathbf{Z}(t)\|_2^2 \leq 4\left(\|\mathbf{Z}(0)\|_2^2 + MB_{W_3}^2 T^2 \sup_{0\leq t\leq T}\|\mathbf{x}(t)\|_2^2 + \sup_{0\leq t\leq T}\|\mathbf{M}(t)\|_2^2\right) e^{2B_L^2 T^2}. \tag{3.39}$$

Finally, applying Doob's Inequality gives

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\|\mathbf{Z}(t)\|_2^2\right]\leq 4\Bigg(\mathbb{E}\left[\|\mathbf{Z}(0)\|_2^2\right]+MB_{W_3}^2T^2\sup_{0\leq t\leq T}\|\mathbf{x}(t)\|_2^2 \\ +4\mathbb{E}\left[\sum_{k\in\mathbb{Z}_+}\sum_{m\in[M]}<M_{k,m}>_T\right]\Bigg)e^{2B_L^2T^2}. \tag{3.40}$$

The proof is completed by noting that the process $\mathbf{x}(t)$ is summable and the martingale $\mathbf{M}(t)$ is square-integrable. □

The following theorem states the weak convergence of the fluctuation process to a unique OU process in the transient regime.

Theorem 3.1.4. *Let $\mathbf{Z}(t)$ be the unique strong solution to the set of SDEs (3.23). Assume $\mathbf{Z}^{(n)}(0)$ converges in distribution to $\mathbf{Z}(0)$ as $n\to\infty$, then the process $\mathbf{Z}^{(n)}(t)$ also converges in distribution to the process $\mathbf{Z}(t)$.*

Proof. From the weak convergence of the initial process, we conclude the sequence $\mathbf{Z}^{(n)}(0)$ is tight i.e. $\forall\epsilon>0$ we can find a compact set K_ϵ such that $\mathbb{P}(\mathbf{Z}^{(n)}(0)\in K_\epsilon)>1-\epsilon$ and the probability of the process $\mathbf{Z}^{(n)}(0)$ lying outside the set K_ϵ is negligible. We can construct another random variable $\mathbf{x}^{(n,\epsilon)}(0)$ such that it coincides with $\mathbf{x}^{(n)}(0)$ on the set $\{\mathbf{Z}^{(n)}(0)\in K_\epsilon\}$ and outside of this set, the corresponding fluctuation process, $\mathbf{Z}^{(n,\epsilon)}(0)$, is uniformly bounded in n. Therefore, without loss of generality, we can assume the process $\mathbf{Z}^{(n)}(0)$ is uniformly bounded in n. Boundedness of the process $\mathbf{Z}^{(n)}(t)$ then follows from Lemma 3.1.2. We complete the proof by showing that Theorem 4.1 on page 354 of [69] holds. We need to show conditions (4-1) - (4-7) of this theorem hold.

Fix $m\in[M]$. Take $B_m^{(n)}(t)$ equal to $\mathbf{Z}_m^{(n)}(t)-\mathbf{Z}_m^{(n)}(0)-\mathbf{M}_m^{(n)}(t)$ and $\left(A_m^{(n)}(t)\right)_{k,s}$ as the covariation between martingales $\mathbf{M}_{k,m}^{(n)}(t)$ and $\mathbf{M}_{s,m}^{(n)}(t)$. We want to show $B_m^{(n)}(t)$ and $\left(A_m^{(n)}(t)\right)_{k,s}$ converge to $\mathbf{Z}_m(t)-\mathbf{Z}_m(0)-\mathbf{M}_m(t)$ and the covariation function between martingales $\mathbf{M}_{k,m}(t)$ and $\mathbf{M}_{s,m}(t)$, respectively.

By replacing $\mathbf{x}_m^{(n)}(t)$ with $\mathbf{x}_m(t)+\frac{\mathbf{Z}_m^{(n)}(t)}{\sqrt{n\gamma_m}}$ and from Equations (3.4)-(3.5) we have

$$W_m(\mathbf{x}^{(n)}(t)) - W_m(\mathbf{x}(t)) = L(\mathbf{x}(t))_m \left(\frac{\tilde{\mathbf{Z}}^{(n)}(t)}{\sqrt{n}}\right) + g\left(\frac{\tilde{\mathbf{Z}}^{(n)}(t)}{\sqrt{n}}\right), \tag{3.41}$$

where $\tilde{\mathbf{Z}}^{(n)}(t) = \left(\tilde{\mathbf{Z}}_l^{(n)}(t), l \in [M]\right)$ such that $\tilde{\mathbf{Z}}_l^{(n)}(t) = \frac{\mathbf{Z}_l^{(n)}(t)}{\sqrt{\gamma_l}}$ and $g\left(\frac{\tilde{\mathbf{Z}}^{(n)}(t)}{\sqrt{n}}\right)$ is $O\left(\frac{(\tilde{\mathbf{Z}}^{(n)}(t))^2}{n}\right)$.

From the definitions of functions $B_m^{(n)}(t)$ and $A_m^{(n)}(t)$, conditions (4-1)-(4-2) and (4-4)-(4-5) follow immediately. Condition (4-3) holds due to the fact that jumps of the fluctuation process are of size $\frac{1}{\sqrt{n\gamma_m}}$. Condition (4-6) follows from Equation (3.41) and Lemma 3.1.2. Finally, condition (4-7) holds due to the existence of the mean-field limit.

□

3.2 Steady-State Behavior of the Fluctuation Process

In this section, we study the asymptotic behavior of the fluctuation process in equilibrium, specifically the behavior of $\mathbf{Z}^{(n)}(\infty)$ as $n \to \infty$. From Theorem 2.5.1, we know that the mean-field process will eventually converge to its unique fixed point $\mathbf{P}$. Thus, without loss of generality we can assume that the mean-field process is located at $\mathbf{P}$ for all times. Let $\mathbf{B}(t) = (\mathbf{B}_m(t), m \in [M]) = ((B_{k,m}(t), k \in \mathbb{Z}_+), m \in [M])$ represent a collection of independent and centered Brownian motions $B_{k,m}$ such that $V_{k,m} = var(B_{k,m}(1)^2) = 2C_m(P_{k,m} - P_{k+1,m})$. Therefore, for each $m \in [M]$, the infinitesimal covariance of the process $\mathbf{B}_m(t)$ is given by $diag(\mathbf{V}_m)$, with $\mathbf{V}_m = (V_{k,m}, k \in \mathbb{Z}_+)$. From Equation (3.24) and the fact that $\mathbf{f}(\mathbf{P}) = W(\mathbf{P}) = \mathbf{0}$, it immediately follows that if the mean-field process is fixed at $\mathbf{P}$, then the processes $\mathbf{B}_m(t)$ and $\mathbf{M}_m(t)$ have the same distribution for all $m \in [M]$.

In order to study the fluctuation process in the stationary regime, a key requirement is the local exponential stability of the fixed point of the mean-field limit. The following definition and remark describe this stability and its equivalence in the linearized system.

Definition 3.2. Let $\mathbf{x}(t, \mathbf{u})$ be the mean-field process at time t starting from $\mathbf{u}$. The fixed point of the mean-field process $\mathbf{P}$, is said to be locally exponentially stable in ℓ_2-norm, if there exist some positive constants c, δ and $K < \infty$, such that

$$\|\mathbf{x}(t, \mathbf{u}) - \mathbf{P}\|_2 \leq Ke^{-\delta t} \|\mathbf{u} - \mathbf{P}\|_2, \qquad \forall \|\mathbf{u} - \mathbf{P}\|_2 \leq c. \tag{3.42}$$

Remark 3.1. The nonlinear system described by Equations (2.15)-(2.16) is locally exponentially stable at $\mathbf{P}$ if and only if the linearized system (3.21)-(3.22) around the solution $\mathbf{P}$ is exponentially stable at the origin [82]. In other words, all eigenvalues of the infinite matrix associated with $L(\mathbf{P})$ have negative real parts.

In Lemma 3.2.1, we show that when the queueing system is homogeneous ($M = 1$), the linearized mean-field system around its fixed point $\mathbf{P}$ exhibits exponential stability at the origin. This implies the local exponential stability of the fixed point. We conjecture that these results also extend to the heterogeneous case ($M > 1$), although establishing a direct proof is notably challenging. However, all cases we have considered appear to confirm this and we provide numerical evidence in Remark 3.2.

Lemma 3.2.1. *Let the queueing system be homogeneous ($M = 1$) and let $\mathbf{H}(t)$ be a solution to the linearized mean-field equations around $\mathbf{P}$, i.e.,*

$$\frac{d\mathbf{H}(t)}{dt} = L(\mathbf{P})\mathbf{H}(t). \tag{3.43}$$

Then, $\mathbf{H}(t) = \mathbf{H}(0)e^{L(\mathbf{P})t}$ and there exist $\delta > 0$ and $K < \infty$ such that

$$\left\|\mathbf{H}(t)\right\|_2 \leq Ke^{-\delta t}\left\|\mathbf{H}(0)\right\|_2. \tag{3.44}$$

Proof. The result follows if we show that $L(\mathbf{P})$ defined by Equations (3.21)-(3.22) is Hurwitz, i.e., all the eigenvalues have strictly negative real parts.

The process $\mathbf{H}(t)$ can be written as $\mathbf{H}(t)\mathbb{1}_{(k\in\Omega)} \cup \mathbf{H}(t)\mathbb{1}_{\left(k\in\bar{\Omega}\right)}$ where $\Omega = \{k \in \mathbb{Z}_+ : k \leq \alpha + 1\}$ and $\bar{\Omega} = \mathbb{Z}_+ - \Omega$. On the space Ω, the process $\mathbf{H}(t)$ evolves as a finite Markov process which can be shown to have an exponentially stable generator. However on $\bar{\Omega}$, we have

$$\begin{aligned} L(\mathbf{P})\mathbf{H}(t) &= \lambda d(P_{k-1}^{d-1}H_{k-1} - P_k^{d-1}H_k) - C(H_k - H_{k+1}) \\ &= \lambda dP_{k-1}^{d-1}H_{k-1} - (\lambda dP_k^{d-1} + C)H_k + CH_{k+1}. \end{aligned} \tag{3.45}$$

We can write the operator $L(\mathbf{P})$ in the matrix format,

$$A = \begin{bmatrix} -(\xi_1 + C) & C & 0 & 0 & \dots \\ \xi_1 & -(\xi_2 + C) & C & 0 & \dots \\ 0 & \xi_2 & -(\xi_3 + C) & C & \dots \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}, \tag{3.46}$$

where $\xi_i = \lambda d P_i^{d-1}$. The operator $L^*(\mathbf{P})$ defines a state-dependent birth-death process whose birth rates depend on the distribution $\mathbf{P}$ and death rates are constant. Note that ξ_i is monotone decreasing (associated with the tail distribution P_i). Define

$$A = (\xi_1 + C)B. \tag{3.47}$$

Then,

$$B = \begin{bmatrix} -1 & \frac{C}{\xi_1+C} & 0 & 0 & \cdots \\ \frac{\xi_1}{\xi_1+C} & -\frac{\xi_2+C}{\xi_1+C} & \frac{C}{\xi_1+C} & 0 & \cdots \\ 0 & \frac{\xi_2}{\xi_1+C} & -\frac{\xi_3+C}{\xi_1+C} & \frac{C}{\xi_1+C} & \cdots \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}. \tag{3.48}$$

Let $M = B^T + I$. Since all row sums of B^T except for the first row are zero, it follows that M is a sub-stochastic matrix. Hence applying the standard Perron-Frobenius theory to the N-W (North-West) truncation of M of order k denoted by M^k, states that the maximum eigenvalue of M^k indicated by R_k is such that $0 < R_k < 1$. Moreover from [83, Theorem 6.8, p.211], it follows that $1 > \ldots > R_k > R_{k+1} > \ldots > R \geq 0$, where R is the spectral radius of M. This implies that all the eigenvalues of $B^T = M - I$ are strictly negative from which it follows that A^T and hence A have only negative eigenvalues. Thus the operator $L(\mathbf{P})$ is exponentially stable on both spaces Ω and $\bar{\Omega}$ and we can write

$$\left\| e^{L(\mathbf{P})t}\mathbf{H}(0) \right\|_2 \leq K e^{-\delta t} \left\| \mathbf{H}(0) \right\|_2, \tag{3.49}$$

with $-\delta$ being smaller than the smallest eigenvalue of $L(\mathbf{P})$ on Ω and $\bar{\Omega}$, which are all negative. □

Remark 3.2. In Table 3.1, we provide numerical evidence to validate the Hurwitz property of the operator $L(\mathbf{P})$ in the heterogeneous case. We investigate the eigenvalues of the truncated operator $L(\mathbf{P})$ of order 100 for different system parameters. From Table 3.1, it is seen the maximum of the real part of the eigenvalues is negative in all cases. This leads us to conjecture that the result is true in the general heterogeneous case provided that the system is stable.

We define a new set of SDEs and show that if the fixed point of the mean-field limit is locally exponentially stable, the fluctuation process $\mathbf{Z}^{(n)}(\infty)$ converges to the solution of these SDEs.

Table 3.1: Maximum real part of the eigenvalues of the truncated operator $L(\mathbf{P})$ for different system parameters

$M = 2,\ \gamma_1 = \gamma_2 = 0.5,\ d_1 = d_2 = 2,\ C_1 = 2/3,\ C_2 = 4/3,\ \alpha_1 = 1,\ \alpha_2 = 2$				
λ	0.8	0.85	0.9	0.95
R_k	-0.3128	-0.2542	-0.1795	-0.0830

$M = 2,\ \gamma_1 = \gamma_2 = 0.5,\ d_1 = d_2 = 2,\ C_1 = 1/2,\ C_2 = 3/2,\ \alpha_1 = 1,\ \alpha_2 = 3$				
λ	0.8	0.85	0.9	0.95
R_k	-0.2771	-0.2344	-0.1784	-0.0977

$M = 3,\ \gamma_1 = \gamma_2 = \gamma_3 = 1/3,\ d_1 = d_2 = d_3 = 2,\ C_1 = 1/2,\ C_2 = 1,\ C_3 = 3/2$ $,\ \alpha_1 = 1,\ \alpha_2 = 2{,}\alpha_3 = 3$				
λ	0.8	0.85	0.9	0.95
R_k	-0.3452	-0.2977	-0.2366	-0.1464

Definition 3.3. Let the process $\mathbf{Z}(t) = (\mathbf{Z}_m(t), m \in [M]) = ((Z_{k,m}(t), k \in \mathbb{Z}_+), m \in [M])$ be a solution to the following set of SDEs.

$$\mathbf{Z}_m(t) = \mathbf{Z}_m(0) + \sqrt{\gamma_m} \int_{s=0}^{t} L(\mathbf{P})_m \mathbf{Z}(s)\, ds - \int_{s=0}^{t} W_{3,m}(\mathbf{P})\, ds + \mathbf{B}_m(t), \qquad m \in [M]. \tag{3.50}$$

A solution to the above set of SDEs is an OU process.

In the following theorem, we study some properties of SDEs (3.50).

Theorem 3.2.2. *Assume $\mathbb{E}\left[\|\mathbf{Z}(0)\|_2^2\right] < \infty$. Then there is a unique strong solution to the set of SDEs (3.50). This solution is given by*

$$\mathbf{Z}_m(t) = e^{\sqrt{\gamma_m} L(\mathbf{P})_m t} \mathbf{Z}_m(0) - \int_{s=0}^{t} e^{\sqrt{\gamma_m} L(\mathbf{P})_m (t-s)} W_{3,m}(\mathbf{P})\, ds + \int_{s=0}^{t} e^{\sqrt{\gamma_m} L(\mathbf{P})_m (t-s)} d\mathbf{B}_m(s), \tag{3.51}$$

for each $m \in [M]$. Additionally, $\mathbb{E}\left[\sup_{0 \le t \le T} \|\mathbf{Z}(t)\|_2^2\right] < \infty$.

Proof. The proof follows *mutatis mutandis* from the proof of Theorem 3.1.3 and hence is omitted. □

From Lemma 3.2.1 and Theorem 3.2.2, the following theorem follows immediately.

Theorem 3.2.3. *Let $\mathbf{Z}(t) = (\mathbf{Z}_m(t), m \in [M])$ be the unique solution to the set of SDEs given by (3.50). For any given $m \in [M]$, the process $\mathbf{Z}_m(t)$ converges in distribution to a Gaussian process as $t \to \infty$. This Gaussian process is uniquely determined by its mean, which is equal to $-\int_{s=0}^{\infty} e^{\sqrt{\gamma_m} L(\mathbf{P})_m s} W_{3,m}(\mathbf{P})\, ds$ and its covariance given by $\int_{s=0}^{\infty} e^{\sqrt{\gamma_m} L(\mathbf{P})_m s} diag(\mathbf{V}_m) e^{\sqrt{\gamma_m} L^*(\mathbf{P})_m s}\, ds$.*

Lemma 3.2.4. *Assume the mean-field limit is fixed at its equilibrium point $\mathbf{P}$ for all $t \geq 0$. If $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] < \infty$, then $\limsup_{n\to\infty} \sup_{t\geq 0} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right] < \infty$. Under the invariant distribution, we conclude $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] < \infty$.*

Proof. Given that the mean-field limit is located at $\mathbf{P}$, for each $m \in [M]$, we have

$$\mathbf{Z}_m^{(n)}(t) = \sqrt{n\gamma_m}\left(\mathbf{x}_m^{(n)}(t) - \mathbf{P}_m\right), \tag{3.52}$$

where $\mathbf{x}_m^{(n)}(t) = \left(x_{k,m}^{(n)}(t), k \in \mathbb{Z}_+\right)$ and $\mathbf{P}_m = (P_{k,m}, k \in \mathbb{Z}_+)$ for the specific value of m. Then we can write

$$\mathbf{x}_m^{(n)}(t) = \frac{\mathbf{Z}_m^{(n)}(t)}{\sqrt{n\gamma_m}} + \mathbf{P}_m. \tag{3.53}$$

From Equations (3.9) and (3.53) and noting that $W_m(\mathbf{P}) = \mathbf{0}$, we have

$$\mathbf{x}_m^{(n)}(t) = \mathbf{x}_m^{(n)}(0) + \int_{s=0}^{t} W_m(\mathbf{x}^{(n)}(s))\, ds - \frac{1}{\sqrt{n\gamma_m}}\int_{s=0}^{t} W_{3,m}(\mathbf{x}^{(n)}(s))\, ds + \frac{1}{\sqrt{n\gamma_m}}\mathbf{M}_m^{(n)}(t). \tag{3.54}$$

Also, let $\mathbf{x}(t, \mathbf{u})$ be the mean-field limit at time t starting from $\mathbf{u}$. Then, from Equation (2.15) and noting that the operators $\mathbf{f}$ and W are the same for each $m \in [M]$, we have

$$\mathbf{x}_m(t, \mathbf{u}) = \mathbf{u}_m + \int_{s=0}^{t} W_m(\mathbf{x}(\mathbf{s}, \mathbf{u}))\, ds, \tag{3.55}$$

where $\mathbf{x}_m(t, \mathbf{u}) = (x_{k,m}(t, \mathbf{u}), k \in \mathbb{Z}_+)$ and $\mathbf{u}_m = (u_{k,m}, k \in \mathbb{Z}_+)$ for the given m. At time $t_0 + h$, we can write the fluctuation process as

$$\mathbf{Z}_m^{(n)}(t_0 + h) = \sqrt{n\gamma_m} \left(\mathbf{x}_m^{(n)}(t_0 + h) - \mathbf{x}_m(h, \mathbf{x}^{(n)}(t_0)) + \mathbf{x}_m(h, \mathbf{x}^{(n)}(t_0)) - \mathbf{P}_m \right). \tag{3.56}$$

We define

$$\mathbf{Z}_m^{(n)}(t_0, h) = \sqrt{n\gamma_m} \left(\mathbf{x}_m^{(n)}(t_0 + h) - \mathbf{x}_m(h, \mathbf{x}^{(n)}(t_0)) \right). \tag{3.57}$$

Then we can rewrite the fluctuation process as

$$\mathbf{Z}_m^{(n)}(t_0 + h) = \mathbf{Z}_m^{(n)}(t_0, h) + \sqrt{n\gamma_m} \left(\mathbf{x}_m(h, \mathbf{x}^{(n)}(t_0)) - \mathbf{P}_m \right). \tag{3.58}$$

From Equations (2.46) and (3.58), there exist some positive $D < \infty$ and $\theta > 1$, such that

$$\left\| \mathbf{Z}_m^{(n)}(t_0 + h) \right\|_2^2 \leq 2 \left\| \mathbf{Z}_m^{(n)}(t_0, h) \right\|_2^2 + h^{-2\theta} D^2 \left\| \mathbf{Z}_m^{(n)}(t_0) \right\|_2^2. \tag{3.59}$$

By summing over all $m \in [M]$, we get

$$\left\| \mathbf{Z}^{(n)}(t_0 + h) \right\|_2^2 \leq 2 \left\| \mathbf{Z}^{(n)}(t_0, h) \right\|_2^2 + h^{-2\theta} D^2 \left\| \mathbf{Z}^{(n)}(t_0) \right\|_2^2. \tag{3.60}$$

Also, from Equation (3.55), we can rewrite Equation (3.57) as

$$\mathbf{Z}_m^{(n)}(t_0, h) = \sqrt{n\gamma_m} \left(\mathbf{x}_m^{(n)}(t_0 + h) - \mathbf{x}_m^{(n)}(t_0) \right) - \sqrt{n\gamma_m} \int_{s=0}^{h} W_m(\mathbf{x}(s, \mathbf{x}^{(n)}(t_0)))\, ds. \tag{3.61}$$

From Equation (3.54) for the term $\mathbf{x}_m^{(n)}(t_0 + h) - \mathbf{x}_m^{(n)}(t_0)$, we get

$$\begin{aligned} \mathbf{Z}_m^{(n)}(t_0, h) = \sqrt{n\gamma_m} \Bigg[& \int_{s=t_0}^{t_0+h} W_m(\mathbf{x}^{(n)}(s))\, ds - \frac{1}{\sqrt{n\gamma_m}} \int_{s=t_0}^{t_0+h} W_{3,m}(\mathbf{x}^{(n)}(s))\, ds \\ & + \frac{1}{\sqrt{n\gamma_m}} \left(\mathbf{M}_m^{(n)}(t_0 + h) - \mathbf{M}_m^{(n)}(t_0) \right) \Bigg] - \sqrt{n\gamma_m} \int_{s=0}^{h} W_m(\mathbf{x}(s, \mathbf{x}^{(n)}(t_0)))\, ds, \end{aligned} \tag{3.62}$$

or equivalently

$$\begin{aligned}\mathbf{Z}_m^{(n)}(t_0,h) = \sqrt{n\gamma_m}\int_{s=0}^{h}\left(W_m(\mathbf{x}^{(n)}(s+t_0)) - W_m(\mathbf{x}(s,\mathbf{x}^{(n)}(t_0)))\right)\,ds \\ - \int_{s=t_0}^{t_0+h} W_{3,m}(\mathbf{x}^{(n)}(s))\,ds + \left(\mathbf{M}_m^{(n)}(t_0+h) - \mathbf{M}_m^{(n)}(t_0)\right).\end{aligned} \tag{3.63}$$

By taking the ℓ_2-norm of both sides and using the Cauchy-Schwartz inequality we get

$$\begin{aligned}\left\|\mathbf{Z}_m^{(n)}(t_0,h)\right\|_2^2 \leq 3\Bigg[n\gamma_m B_W^2 h\int_{s=0}^{h}\left\|\mathbf{x}^{(n)}(s+t_0) - \mathbf{x}(s,\mathbf{x}^{(n)}(t_0))\right\|_2^2\,ds \\ + B_{W_3}^2 h\int_{s=t_0}^{t_0+h}\left\|\mathbf{x}^{(n)}(s)\right\|_2^2\,ds + \left\|\mathbf{M}_m^{(n)}(t_0+h) - \mathbf{M}_m^{(n)}(t_0)\right\|_2^2\Bigg].\end{aligned} \tag{3.64}$$

By the definition of ℓ_2-norm, we can replace the term $\left\|\mathbf{x}^{(n)}(s+t_0) - \mathbf{x}(s,\mathbf{x}^{(n)}(t_0))\right\|_2^2$ with $\sum_{l\in[M]}\left\|\mathbf{x}_l^{(n)}(s+t_0) - \mathbf{x}_l(s,\mathbf{x}^{(n)}(t_0))\right\|_2^2$, and we get

$$\begin{aligned}\left\|\mathbf{Z}_m^{(n)}(t_0,h)\right\|_2^2 \leq 3\Bigg[B_W^2 h\beta_m\int_{s=0}^{h}\left\|\mathbf{Z}^{(n)}(t_0,s)\right\|_2^2\,ds + B_{W_3}^2 h\int_{s=t_0}^{t_0+h}\left\|\mathbf{x}^{(n)}(s)\right\|_2^2\,ds \\ + \left\|\mathbf{M}_m^{(n)}(t_0+h) - \mathbf{M}_m^{(n)}(t_0)\right\|_2^2\Bigg],\end{aligned} \tag{3.65}$$

where $\beta_m = \max_{l\in[M]}\frac{\gamma_m}{\gamma_l}$. By taking sum over all $m \in [M]$, we can write

$$\begin{aligned}\left\|\mathbf{Z}^{(n)}(t_0,h)\right\|_2^2 \leq 3\Bigg[B_W^2 h\beta\int_{s=0}^{h}\left\|\mathbf{Z}^{(n)}(s,t_0)\right\|_2^2\,ds + MB_{W_3}^2 h\int_{s=t_0}^{t_0+h}\left\|\mathbf{x}^{(n)}(s)\right\|_2^2\,ds \\ + \left\|\mathbf{M}^{(n)}(t_0+h) - \mathbf{M}^{(n)}(t_0)\right\|_2^2\Bigg],\end{aligned} \tag{3.66}$$

where $\beta = \sum_{m\in[M]}\beta_m$. For any $T > 0$, by Gronwall's Lemma, we can find a positive constant K_T depending on T, such that

$$\sup_{0\le h\le T}\left\|\mathbf{Z}^{(n)}(t_0,h)\right\|_2^2 \le K_T\left(K_T+\sup_{0\le h\le T}\left\|\mathbf{M}^{(n)}(t_0+h)-\mathbf{M}^{(n)}(t_0)\right\|_2^2\right). \tag{3.67}$$

From Equations (3.60) and (3.67), we can find a positive constant S_T as a function of T such that for every $0 \le h \le T$, we have

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t_0+h)\right\|_2^2\right] \le L_T + T^{-2\theta}D^2\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t_0)\right\|_2^2\right]. \tag{3.68}$$

We choose T sufficiently large such that $T^{-2\theta}D^2 \le \epsilon < 1$. Then for all integer n and N, we have

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}((N+1)T)\right\|_2^2\right] \le L_T + \epsilon\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT)\right\|_2^2\right]. \tag{3.69}$$

We use induction to get

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT)\right\|_2^2\right] \le L_T\left(\sum_{j=1}^{N}\epsilon^{j-1}\right)+\epsilon^N\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right], \tag{3.70}$$

or

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT)\right\|_2^2\right] \le \frac{L_T}{1-\epsilon}+\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right]. \tag{3.71}$$

From Equation (3.68), we have

$$\sup_{0\le h\le T}\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT+h)\right\|_2^2\right] \le L_T + D^2\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT)\right\|_2^2\right]. \tag{3.72}$$

Substituting Equation (3.71) into (3.68), we get

$$\sup_{0\le h\le T}\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(NT+h)\right\|_2^2\right] \le L_T + D^2\left(\frac{L_T}{1-\epsilon}+\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right]\right). \tag{3.73}$$

N is an arbitrary integer, and hence the above inequality reduces to

$$\sup_{0\le h\le T}\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right] \le L_T + D^2\left(\frac{L_T}{1-\epsilon}+\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right]\right). \tag{3.74}$$

Let $\mathbf{Z}^{(n)}(\infty)$ denote the fluctuation process (3.52) in equilibrium. Then by ergodicity and Fatou's Lemma, we have

$$\mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(\infty)\right\|_2^2\right] \leq \liminf_{t\geq 0} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right] \leq \sup_{t\geq 0} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(t)\right\|_2^2\right]. \tag{3.75}$$

To show that $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(\infty)\right\|_2^2\right] < \infty$, we need to find $\mathbf{x}^{(n)}(0)$ such that $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] < \infty$. For some finite R, we choose $x_{k,m}^{(n)}(0)$ such that $-\frac{1}{2n\gamma_m} \leq x_{k,m}^{(n)}(0) - P_{k,m} \leq \frac{1}{2n\gamma_m}$ for $1 \leq k \leq R$ and $m \in [M]$. Additionally, we set $x_{k,m}^{(n)}(0) = 0$ for $k > R$ and $m \in [M]$. Consequently, $\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(0)\right\|_2^2\right] = \limsup_{n\to\infty} \sum_{m\in[M]} \frac{R}{4n\gamma_m} = 0$ and this completes the proof. □

We will now establish the weak convergence of the fluctuation process in equilibrium.

Theorem 3.2.5. *If the finite-sized system is in equilibrium, and the fixed point of its mean-field limit exhibits local exponential stability, then the process $\mathbf{Z}^{(n)}(t)$ converges in distribution to the unique solution of the set of SDEs (3.50). Additionally, the initial process $\mathbf{Z}^{(n)}(0)$ converges weakly to the invariant distribution of this process.*

Proof. From Lemma 3.2.4 and the Markov inequality, the sequence $\mathbf{Z}^{(n)}(0)$ is tight. We can now select a convergent subsequence with a square integrable limit denoted as $\mathbf{Z}^{\infty}(0)$. Applying Theorem 3.1.4 yields that the chosen subsequence converges in distribution to the unique OU process $\mathbf{Z}^{\infty}(t)$ that solves the set of SDEs (3.50) and starts at $\mathbf{Z}^{\infty}(0)$. Furthermore, from [69, Lemma 7.7, Theorem 7.8, P.131], it is established that the limit of any sequence of stationary processes remains stationary. Consequently, the subsequnce $\mathbf{Z}^{\infty}(t)$ has the stationary distribution of the unique OU process that solves the SDEs (3.50). Since this applies to every convergent subsequence, we can deduce the desired result. This concludes the proof. □

3.2.1 Applications

In this section, we study various applications of the fluctuation process. We begin with the first problem, wherein we analyze the discrepancy between the empirical distribution of servers in a finite-sized system and the mean-field distribution. We provide the order of convergence for both transient and stationary regimes. In the second problem, we focus on the accuracy of performance approximation for finite-sized systems. Specifically, we show that the stationary mean response time of jobs in the finite-sized system can be estimated by the mean response time of a system whose distribution is given by the mean-field distribution, and the accuracy of this estimation is $O(\frac{1}{\sqrt{n}})$.

Remark 3.3. We have observed that the scaled difference between $\mathbf{x}_m^{(n)}(t)$ and $\mathbf{x}_m(t)$ (representing the empirical distribution of type m servers in a system of size n and the mean-field limit of type m servers, respectively) by the scaling parameter $\sqrt{n\gamma_m}$ converges to an OU process, both in transient and stationary regimes. Applying the continuous mapping theorem with functions

$$f_1(x) = |x|,$$
$$f_2(x) = x^2,$$

to the functional central limit theorems derived in Sections 3.1 and 3.2, and taking into account that the limiting OU process possesses finite first and second moments, we conclude that the ℓ_1 difference between the empirical distribution of servers and their mean-field limit is $O(\frac{1}{\sqrt{n}})$, and the mean-squared difference is $O(\frac{1}{n})$. It is worth noting that the result for the mean-squared difference in the stationary regime and for finite-dimensional systems was previously presented in [61, 62].

In Theorem 3.2.6, we provide the order of accuracy when approximating the mean response time of jobs in the finite-sized system by its asymptotic value. We also identify the constant related to the term $O(\frac{1}{\sqrt{n}})$.

Theorem 3.2.6. *Let $\mathbb{E}\left[D^{(n)}\right]$ and $\mathbb{E}[D]$ denote the mean sojourn time of jobs in the system of size n and in the limiting system, respectively. If $\lambda(n) \in \Lambda$, then the error between these two terms is bounded by $O(\frac{1}{\sqrt{n}})$. More precisely, we have*

$$\lim_{n\to\infty} \sqrt{n}\left(\mathbb{E}\left[D^{(n)}\right] - \mathbb{E}[D]\right) = \frac{\beta}{\lambda^2}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}P_{k,m} + \frac{1}{\lambda}\sum_{m\in[M]}\sqrt{\gamma_m}\sum_{k=1}^{\infty}\nu_{k,m}, \tag{3.76}$$

where $\nu_m = (\nu_{k,m}, k \in \mathbb{Z}_+) = -\int_{s=0}^{\infty} e^{\sqrt{\gamma_m}L(\mathbf{P})_m s}W_{3,m}(\mathbf{P})\,ds$.

Proof. For a given system of size n, the average number of jobs at equilibrium is equal to $n\mathbb{E}\left[\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}x_{k,m}^{(n)}(\infty)\right]$. Let the mean sojourn time of jobs be denoted by $\mathbb{E}\left[D^{(n)}\right]$. We can apply Little's law to get

$$n\lambda(n)\mathbb{E}\left[D^{(n)}\right] = n\mathbb{E}\left[\sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty} x_{k,m}^{(n)}(\infty)\right]. \tag{3.77}$$

Due to the stability of the system, the above sum is finite and we can deduce that

$$\lambda(n)\mathbb{E}\left[D^{(n)}\right] = \sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty} \mathbb{E}\left[x_{k,m}^{(n)}(\infty)\right]. \tag{3.78}$$

From Theorem 3.2.3 and Theorem 3.2.5, we know that the diffusion limit of type m servers in the stationary regime has mean vector $\nu_m = -\int_{s=0}^{\infty} e^{\sqrt{\gamma_m}L(\mathbf{P})_m s} W_{3,m}(\mathbf{P})\, ds$. This implies

$$\mathbb{E}\left[\lim_{n\to\infty} \mathbf{Z}_m^{(n)}(\infty)\right] = \nu_m. \tag{3.79}$$

However, from Lemma 3.2.4 we have

$$\limsup_{n\to\infty} \mathbb{E}\left[\left\|\mathbf{Z}^{(n)}(\infty)\right\|_2^2\right] < \infty, \tag{3.80}$$

which enables us to interchange the order of limit and expectation in Equation (3.79). Consequently,

$$\lim_{n\to\infty} \sqrt{n\gamma_m}\left(\mathbb{E}\left[\mathbf{x}_m^{(n)}(\infty)\right] - \mathbf{P}_m\right) = \nu_m, \tag{3.81}$$

This further leads to

$$\mathbb{E}[\mathbf{x}_m^{(n)}(\infty)] = \mathbf{P}_m + \frac{\nu_m}{\sqrt{n\gamma_m}} + o(\frac{1}{\sqrt{n}}). \tag{3.82}$$

Now, from Equation (3.78), we obtain

$$\lambda(n)\mathbb{E}\left[D^{(n)}\right] = \sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty}\Big(P_{k,m} + \frac{\nu_{k,m}}{\sqrt{n\gamma_m}}\Big) + o(\frac{1}{\sqrt{n}}). \tag{3.83}$$

Since $\frac{\beta}{\lambda\sqrt{n}} < 1$, we can conclude

$$\mathbb{E}\left[D^{(n)}\right] = \frac{1}{\lambda}\left(1 + \frac{\beta}{\lambda\sqrt{n}} + o(\frac{1}{\sqrt{n}})\right) \sum_{m\in[M]} \gamma_m \sum_{k=1}^{\infty}\Big(P_{k,m} + \frac{\nu_{k,m}}{\sqrt{n\gamma_m}}\Big) + o(\frac{1}{\sqrt{n}}). \tag{3.84}$$

But we know $\frac{1}{\lambda}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}P_{k,m}=\mathbb{E}[D]$, where $\mathbb{E}[D]$ is the mean sojourn time of jobs in the limiting system. Hence

$$\mathbb{E}\left[D^{(n)}\right]=\mathbb{E}[D]+\frac{\beta}{\lambda^2\sqrt{n}}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}P_{k,m}+\frac{1}{\lambda}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}\frac{\nu_{k,m}}{\sqrt{n\gamma_m}}+o(\frac{1}{\sqrt{n}}). \tag{3.85}$$

Finally

$$\mathbb{E}\left[D^{(n)}\right]=\mathbb{E}[D]+\frac{\beta}{\lambda^2\sqrt{n}}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}P_{k,m}+\frac{1}{\lambda}\sum_{m\in[M]}\sqrt{\gamma_m}\sum_{k=1}^{\infty}\frac{\nu_{k,m}}{\sqrt{n}}+o(\frac{1}{\sqrt{n}}). \tag{3.86}$$

Or equivalently

$$\lim_{n\to\infty}\sqrt{n}\left(\mathbb{E}\left[D^{(n)}\right]-\mathbb{E}[D]\right)=\frac{\beta}{\lambda^2}\sum_{m\in[M]}\gamma_m\sum_{k=1}^{\infty}P_{k,m}+\frac{1}{\lambda}\sum_{m\in[M]}\sqrt{\gamma_m}\sum_{k=1}^{\infty}\nu_{k,m}. \tag{3.87}$$

□

Remark 3.4. In the proof above, it is observed that $O(\frac{1}{\sqrt{n}})$ terms depend on the fixed point of the mean-field limit and β associated with the arrival process. These terms are nonzero constants for $\beta\neq 0$ and are zero for $\beta=0$. This is due to the fact that the vector ν_m given by $\nu_m=-\int_{s=0}^{\infty}e^{\sqrt{\gamma_m}L(\mathbf{P})_m s}W_{3,m}(\mathbf{P})\,ds$ depends on the operator $W_{3,m}(\mathbf{P})=\frac{\beta\sqrt{\gamma_m}}{\lambda}W_{1,m}(\mathbf{P})$ defined in Equation (3.4). This guarantees that when $\beta=0$, the operator $W_{3,m}$ and hence the vector ν_m become zero. Consequently, the $O(\frac{1}{\sqrt{n}})$ term becomes zero. Therefore, we can conclude the convergence rate for $\beta=0$ is faster than $O(\frac{1}{\sqrt{n}})$.

Remark 3.5. Bounds on the higher-order moments of the difference between the average mean delay and the asymptotic delay can be obtained following the same approach.

3.3 Numerical Results

In this section, we present numerical results to validate the accuracy of our analysis. Specifically, we investigate the error between the mean response time of jobs in the system of size n and in the limiting system, numerically. We find this error for different values of n and we show it is bounded by an $O(\frac{1}{\sqrt{n}})$ term.

The mean response time of jobs in the limiting system is determined by the sum of an infinite series, as outlined in Equation (2.51). However, for practical numerical computation, we have focused on the first 500 components of $P_{k,m}$ for each $m \in [M]$. Beyond this point, these components become negligible. We computed the fixed point of the mean-field limit, denoted as $\mathbf{P}$, by solving a set of 500 nonlinear equations for each $m \in [M]$. To obtain the mean response time of jobs in a finite-sized system with n servers, we conducted simulations by averaging the response times of the initial 5×10^6 jobs entering the system. These simulations were independently repeated 100 times. The system parameters employed for these experiments were as follows: $M = 2$, $\gamma_1 = \gamma_2 = 0.5$, $d_1 = d_2 = 2$, $C_1 = 2/3$, $C_2 = 4/3$, $\alpha_1 = 1$, $\alpha_2 = 2$, $\lambda = 0.95$, and two values of $\beta = 0$ and $\beta = 1$.

The asymptotic delay is 2.4110, and simulation results are summarized in Table 3.2. When β is zero, $\lambda(n)$ is constant and independent of the system size. From Table 3.2, it is evident that $\mathbb{E}\left[D^{(n)}\right] - \mathbb{E}[D]$ decreases at a rate faster than $O(\frac{1}{\sqrt{n}})$ as n increases, and the scaled difference between the asymptotic mean response time and the mean response time of the finite-size system, with the scaling parameter $\sqrt{n}$, tends to zero. This observation is consistent with Theorem 3.2.6, since when $\beta = 0$, Equation (3.4) results in $W_{3,m}(\mathbf{P}) = 0$, which leads to a zero value for ν_m. Indeed, we can verify that in this case, the rate of convergence is $O(\frac{1}{n})$.

In Table 3.3, we present the results for the same system parameters, but with nonzero $\beta = 1$. The asymptotic delay remains unchanged at 2.4110. This is because the arrival rate to the system with an infinite number of servers is the same, regardless of the value of β. However, when the system size is finite and β is non-zero, the arrival rate of the system increases as the system size grows, as reported in Table 3.3. Consequently, the mean response times of jobs converge from below to their asymptotic delay. The scaled difference between the mean response time of jobs in the system of finite size and the asymptotic delay, with the scaling parameter $\sqrt{n}$, converges to some nonzero constant depending on β and $\mathbf{P}$, which is consistent with Theorem 3.2.6.

3.4 Conclusion

In this chapter, we assessed the accuracy of the mean-field limit for a class of large heterogeneous processor sharing systems. Our investigation centered on the fluctuations in the empirical distribution of servers' occupancy around its mean-field limit, at a diffusion scale.

Table 3.2: Difference between the mean response time of jobs in the system of finite size and in the limiting system with $\beta = 0$

n	$\lambda(n)$	$\mathbb{E}\left[D^{(n)}\right]$	$\mathbb{E}\left[D^{(n)}\right] - \mathbb{E}[D]$
50	0.95	2.6318	0.2208
100	0.95	2.5191	0.1081
200	0.95	2.4651	0.0541
300	0.95	2.4475	0.0365
400	0.95	2.4360	0.0250
600	0.95	2.4299	0.0189
800	0.95	2.4244	0.0134
1000	0.95	2.4208	0.0098
1500	0.95	2.4159	0.0049
2000	0.95	2.4142	0.0032
2500	0.95	2.4139	0.0029
4000	0.95	2.4135	0.0025

Table 3.3: Difference between the mean response time of jobs in the system of finite size and in the limiting system with $\beta \neq 0$

n	$\lambda(n)$	$\mathbb{E}\left[D^{(n)}\right]$	$\mathbb{E}\left[D^{(n)}\right] - \mathbb{E}[D]$
50	0.8086	1.9182	-0.4928
100	0.8500	1.9982	-0.4128
200	0.8793	2.0670	-0.3440
300	0.8923	2.1057	-0.3053
400	0.9000	2.1317	-0.2793
600	0.9092	2.1644	-0.2466
800	0.9146	2.1860	-0.2250
1000	0.9184	2.2038	-0.2072
1500	0.9242	2.2321	-0.1789
2000	0.9276	2.2501	-0.1609
2500	0.9300	2.2626	-0.1484
4000	0.9342	2.2812	-0.1298

Using FCLT, we showed the convergence of this process to an OU process in the transient regime. Exploiting the local exponential stability of the fixed point of the mean-field limit, we extended our analysis to derive the FCLT for the stationary regime. Using these results, we obtained error bounds for approximating the mean response time of jobs in a finite-size system, with the response time given by the mean-field limit. We established that this error diminishes at a rate of $O(\frac{1}{\sqrt{n}})$ and identified the constant associated with this term. Notably, we demonstrated that when the arrival rate remains unperturbed ($\beta = 0$), this constant becomes zero, indicating a faster convergence rate.

Chapter 4

Performance of Loss Systems with Adaptive Multiserver Jobs and Linear Speedup

In this chapter, we study large loss systems with adaptive multiserver jobs where each job or request can be split into a flexible number of sub-jobs up to a maximum limit. The number of sub-jobs a job is split into depends on the number of available servers found upon its arrival. The sub-jobs are then processed in parallel on different servers, resulting in a reduction in the original job's execution time by a factor of i when there are i sub-jobs being processed in parallel. We refer to i as the linear speedup factor. These jobs for instance can represent requests for files in a file-server system where each file is stored at multiple locations from where different parts of the file can be downloaded in parallel. We study the problem of optimal assignment of such requests when each server can process at most one sub-job at any given instant and there is no waiting room in the system. Prior research in this field often assumed that each job had access to the state information of all servers, with no system blocking. In addition, we consider limited system knowledge, where upon arrival of a job, it can only access a randomly sampled subset of servers. We analyze the steady-state system performance with full and limited system access. We develop a load balancing policy and demonstrate its effectiveness in achieving optimal average response time of jobs and zero blocking probability as the system size increases. Moreover, in cases with limited system access, we show that to achieve the same asymptotic performance results, it is necessary that the sampling size grows at a specific rate. The analysis uses Stein's and Lyapunov drift methods to establish state space collapse for large system sizes.

The remainder of this chapter is organized as follows. Section 4.1 introduces the system model, providing the foundational framework for our analysis. Section 4.2 outlines the criteria for optimality. The key results regarding the system's asymptotic optimality along with error bounds in finite-sized systems are presented in Section 4.3, with a focus on full system access in Subsection 4.3.1 and limited system access in Subsection 4.3.2. The applicability of these results to heterogeneous workloads is discussed in Section 4.4. Numerical results are provided in Section 4.5. Finally, we conclude with our closing remarks in Section 4.6.

4.1 System Model

We consider a system with a single dispatcher and n parallel servers. Each server can process only one job at any given time and there is no waiting room in the system. It is assumed that each job can be processed simultaneously at a maximum of $d \geq 1$ servers. A job's inherent processing time, i.e., the time it would take to process the job at one server, is assumed to be independent and exponentially distributed with unit mean. When a job is run in parallel on $i \in [d]$ servers, a speedup of i is obtained. That is, the job's processing time is reduced by a factor of i in comparison to its inherent service time.

Jobs arrive at the system according to a Poisson process with the rate $n\lambda(n)$, where $\lambda(n) = 1 - \beta n^{-\alpha} \geq 0$ for some $\alpha \in [0, 1)$ and $\beta > 0$. Varying the value of α enables the study of the system under different traffic regimes: (i) $\alpha = 0, \beta \in (0, 1)$ corresponds to the *stable regime*, (ii) $\alpha = 1/2$ corresponds to the *Halfin-Whitt regime*, and (iii) $\alpha \in (0, 1/2)$ (resp. $\alpha \in (1/2, 1)$) corresponds to the *sub-Halfin-Whitt (resp. super-Halfin-Whitt) regime*. When a job arrives, it must be immediately and irrevocably assigned to a subset of at most d servers where it is processed until it leaves the system. We are interested in the following two scenarios.

1. ***Full system access***: In this scenario, each job has access to the states of all the servers in the system.

2. ***Limited system access***: In this scenario, each job has access to the states of only a random subset of $k(n)$ servers chosen uniformly at random.

Note that $k(n) = n$ corresponds to the case with full system access. A job must be assigned to at most d of the $k(n)$ servers the job has access to. If none of these $k(n)$ servers are found idle upon the entry of the job, then the job is discarded or *blocked*, which corresponds to a loss.

Our objective is to design job assignment schemes that eliminate blocking and minimize the mean response time of accepted jobs at a steady state in the large system limit, as $n \to \infty$. We address this problem in each of the scenarios with full and limited system access.

Remark 4.1. A key aspect of the system is that when a job is assigned to $i \in [d]$ idle servers, it is divided into i *equal* sub-jobs or tasks. Since all sub-jobs have equal sizes and join the idle servers with the same processing speed simultaneously, their processing time is identical. This synchronized execution results in the overall job's processing time being equivalent to that of a single sub-job. Consequently, for a job with a size $\mathcal{E}$ and a maximum degree of parallelism d, the minimum attainable response time is $\frac{\mathcal{E}}{d}$.

To facilitate our analysis, we use the following notation. For each $i \in [d]$, we let $X_i(t)$ denote the number of jobs that are being processed simultaneously at i servers at time $t \geq 0$. We define $x_i(t), i \in [d]$, as $x_i(t) = X_i(t)/n$, i.e., the scaled number of jobs being processed simultaneously at i servers at time t. Clearly, under any Markovian job assignment scheme (a scheme which makes assignments based on the current state of the system), the process $x(\cdot) = (x_i(\cdot), i \in [d])$ is Markov with a unique stationary distribution. By omitting the explicit dependence on t, we denote by $x = (x_i, i \in [d])$ the state of the system distributed according to its stationary distribution. Additionally, we define the fraction of busy servers at steady-state as $q_1 = q_1(x) = \sum_{i\in[d]} ix_i$, and the fraction of idle servers at steady-state as $q_0 = q_0(x) = 1 - q_1$.

4.2 Criterion for Optimality

In this section, we develop a job assignment scheme that optimizes system performance as $n \to \infty$. To accomplish this, we first need to establish the criteria for what defines an optimal assignment policy.

Let D denote the time spent by an accepted job in the system in the steady state and P_b denote the steady-state blocking probability of a job. Since the total departure rate of jobs at steady-state is $n\mathbb{E}\left[\sum_{i\in[d]} ix_i\right] = n\mathbb{E}\left[q_1\right]$, applying the rate conservation principle and Little's law, we obtain

$$\lambda(n)(1 - P_b) = \mathbb{E}\left[q_1\right], \tag{4.1}$$

$$\lambda(n)(1 - P_b)\mathbb{E}[D] = \mathbb{E}\left[\sum_{i \in [d]} x_i\right] \tag{4.2}$$

Since a job can be processed at a maximum of d servers simultaneously, the minimum possible value for the average response time of jobs, $\mathbb{E}[D]$, according to Remark 4.1, is $1/d$. From the above equations, it is clear that for any Markovian job assignment scheme to achieve the minimum steady-state mean response time of $\frac{1}{d}$ while maintaining a zero blocking at steady-state (i.e., $P_b = 0$), $\mathbb{E}[x_i]$ for each $i \in [d]$ must satisfy:

$$\mathbb{E}[x_i] = 0, \quad \forall i \in [d-1], \tag{4.3}$$

$$\mathbb{E}[x_d] = \frac{\lambda(n)}{d}. \tag{4.4}$$

Let $x^* = \left(0, \ldots, 0, \frac{\lambda(n)}{d}\right)$. Hence, any scheme for which $\mathbb{E}\left[\|x_i - x_i^*\|^2\right] \leq \varepsilon$ for any $\varepsilon > 0$ and n sufficiently large, is asymptotically optimal. Below we define a greedy job assignment scheme that aims to achieve this.

Greedy Assignment Scheme: Under the greedy assignment scheme, upon arrival of a job, if $l \geq 1$ servers are found available among the $k(n)$ servers the job has access to, then $\min(d, l)$ of these available servers are used to process the job, i.e., under the greedy scheme all the available servers up to a maximum of d servers are used to process the incoming job.

4.3 Comparison with a Fluid Limit: Stein's Approach

In order to characterize the system's performance under the greedy scheme, we follow Stein's method to compare the dynamics of the system under the greedy scheme to that of a deterministic fluid limit. Our aim is to show for n sufficiently large, the system under the greedy scheme essentially behaves as the fluid limit.

The Fluid Limit: The deterministic fluid limit to which we compare the dynamics of the original system is defined through the following system of ODEs.

$$\dot{z}_i = -iz_i, \quad \forall i \in [d-1], \tag{4.5}$$

$$\dot{z}_d = \lambda(n) - dz_d. \tag{4.6}$$

Intuitively, in the above system, all arriving jobs find d or more available servers upon entry. It is easy to see that starting from any initial state, the above system converges to x^* as $t \to \infty$. Therefore, the deterministic system (4.5)-(4.6) is optimal in the steady state. Below we attempt to bound the distance between the steady-state performance of the original system and that of the fluid limit.

To this end, we denote by $A_i(x)$ the probability with which an incoming job is processed at $i \in \{0, 1, \ldots, d\}$ servers when the system is in state x. Specifically, $A_i(x)$ represents the probability of finding exactly i idle servers when $i \in \{0, 1, \ldots, d-1\}$, and the probability of having d or more idle servers when $i = d$. It is important to note that in this context, $A_0(x)$ corresponds to the blocking probability in state x, and $\sum_{j=0}^{d} A_j(x) = 1$.

When $k(n) = n$, by the definition of the greedy scheme, these probabilities are given by

$$A_i(x) = \begin{cases} \mathbb{1}(nq_0 = i), & \text{if } i \in \{0, 1, 2, \ldots, d-1\}, \\ \mathbb{1}(nq_0 \geq d), & \text{if } i = d. \end{cases} \tag{4.7}$$

Similarly, when $k(n) < n$, the probabilities $A_i(x)$ are given by means of a hypergeometric distribution. Specifically,

$$A_i(x) = \begin{cases} \frac{\binom{nq_0}{i}\binom{nq_1}{k(n)-i}}{\binom{n}{k(n)}}, & \text{if } i \in \{0, 1, 2, \ldots, d-1\}, \\ \sum_{l \geq d} \frac{\binom{nq_0}{l}\binom{nq_1}{k(n)-l}}{\binom{n}{k(n)}}, & \text{if } i = d. \end{cases} \tag{4.8}$$

In the lemma below, using Stein's method of generator comparison, we express the mean squared distance between x_d and x_d^* as a function of the probability $A_d(x)$ of finding at least d available servers upon entry at the steady state of the system. This expression is later used to bound the mean squared distance under different scenarios of server access.

Lemma 4.3.1. *Under the equilibrium measure of the system:*

$$\mathbb{E}\left[(\lambda(n) - dx_d)^2\right] = \frac{d}{n}\lambda(n)\mathbb{E}\left[A_d(x)\right] + \lambda(n)\mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\right]. \tag{4.9}$$

In particular, when $k(n) = n$, we have

$$\mathbb{E}\left[(\lambda(n) - dx_d)^2\right] = \frac{d}{n}\lambda(n)\mathbb{E}\left[A_d(x)\right] + \mathbb{E}\left[\mathbb{1}\left(q_1 > 1 - \frac{d}{n}\right)(\lambda(n) - dx_d)\right]. \tag{4.10}$$

Proof. We consider the Lyapunov function $V(x) = \frac{1}{2d}(\lambda(n) - dx_d)^2$. Let G be the generator of the Markov chain $x(\cdot)$. Furthermore, let L be the generator of the system of ODEs given by (4.5)-(4.6). The core of Stein's approach is to compare the drift of the function V under G to that under L. We note that under L the drift of V is given by

$$LV(x) = \frac{\partial V}{\partial x_d}(x)\dot{x}_d = -(\lambda(n) - dx_d)\dot{x}_d = -(\lambda(n) - dx_d)^2. \tag{4.11}$$

Similarly, under G the drift of V is given by

$$GV(x) = \sum_{x' \neq x} r(x, x') \left(V(x') - V(x) \right),$$

where $r(x, x')$ denotes the transition rate from state x to state x'. Given that the probability of having d or more idle servers is $A_d(x)$, the function $V(x)$ will transit from state x to state $x + \frac{e_d}{n}$ at the rate $n\lambda(n)A_d(x)$, where e_d denotes the d-dimensional unit vector with one at the d^{th} position. Additionally, the system will transit from state x to the state $x - \frac{e_d}{n}$, if one of the jobs occupying d servers departs the system. As there are nx_d jobs split into d sub-jobs, the departures occur at the rate ndx_d. Thus,

$$GV(x) = n\lambda(n)A_d(x) \left(V\left(x + \frac{1}{n}e_d\right) - V(x) \right) + ndx_d \left(V\left(x - \frac{1}{n}e_d\right) - V(x) \right). \tag{4.12}$$

Under the equilibrium measure, $\mathbb{E}\left[GV(x)\right] = 0$, and hence

$$\mathbb{E}\left[GV(x) - LV(x)\right] = \mathbb{E}\left[-LV(x)\right] = \mathbb{E}\left[(\lambda(n) - dx_d)^2\right]. \tag{4.13}$$

Therefore, to bound $\mathbb{E}\left[(\lambda(n) - dx_d)^2\right]$ it is sufficient to bound $\mathbb{E}\left[GV(x) - LV(x)\right]$. This is done as follows. Using Taylor series expansion of V and noting that $LV(x) = -\left(\frac{\partial V}{\partial x_d}\right)^2$ we have

$$\begin{aligned}\mathbb{E}\left[GV(x) - LV(x)\right] = \mathbb{E}&\left[n\lambda(n)A_d(x)\left(\frac{1}{n}\frac{\partial V}{\partial x_d}(x) + \frac{1}{2n^2}\frac{\partial^2 V}{\partial x_d^2}(\xi)\right)\right] \\ &+ \mathbb{E}\left[ndx_d\left(-\frac{1}{n}\frac{\partial V}{\partial x_d}(x) + \frac{1}{2n^2}\frac{\partial^2 V}{\partial x_d^2}(\theta)\right)\right] + \mathbb{E}\left[\left(\frac{\partial V}{\partial x_d}(x)\right)^2\right],\end{aligned} \tag{4.14}$$

where ξ and θ are d-dimensional vectors. Simplifying the RHS of the above and using $\frac{\partial^2 V}{\partial x_d^2}(y) = d$ for any vector y yields

$$\mathbb{E}\left[GV(x) - LV(x)\right] = \mathbb{E}\left[\left(\lambda(n)A_d(x) - dx_d + \frac{\partial V}{\partial x_d}(x)\right)\left(\frac{\partial V}{\partial x_d}(x)\right)\right]$$

$$+\frac{d}{2n}\mathbb{E}\left[\lambda(n)A_d(x)+dx_d\right]. \tag{4.15}$$

Jobs that occupy d servers in the system have an arrival rate of $n\lambda(n)A_d(x)$ and a departure rate of ndx_d. Since the system is in steady state, the rate conservation law applies, requiring these quantities to be equal on average. Hence $\mathbb{E}\left[dx_d\right] = \lambda(n)\mathbb{E}\left[A_d(x)\right]$. Finally, noting that $\frac{\partial V}{\partial x_d}(x) = -(\lambda(n) - dx_d)$ in the above, we have the desired result. □

The first terms on the RHS of Equations (4.9) and (4.10) are $O(\frac{1}{n})$. Therefore, to establish the convergence of x_d to x_d^*, it is sufficient to show that the second term in the above expressions scales as $O(\frac{1}{n})$. In the following sections, we establish this both when the jobs have access to all servers (i.e., $k(n) = n$) and when the jobs have access to only a subset of servers (i.e., $k(n) < n$).

4.3.1 Full System Access

We begin by considering the case when $k(n) = n$, i.e., when every arrival has access to the complete set of servers. We first show that for $k(n) = n$, the second term in Equation (4.10) is always negative. We use sample path arguments to establish this in the following lemma.

Lemma 4.3.2. *For $k(n) = n$, if the system starts at a state where $\sum_{i\in[d-1]} x_i(0) = 0$, then at all times $t \geq 0$, we have $\sum_{i\in[d-1]} x_i(t) \leq 1/n$. Furthermore, if $\lambda(n) = 1 - \beta/n^\alpha \geq 0$ for $\alpha \in [0,1)$ and $\beta > 0$, then at steady-state we have*

$$\mathbb{E}\left[\mathbb{1}\left(q_1 > 1 - \frac{d}{n}\right)(\lambda(n) - dx_d)\right] \leq 0, \tag{4.16}$$

for all sufficiently large n.

Proof. Let us denote the sum, $\sum_{i\in[d-1]} x_i$, at any state x by $S_{d-1}(x)$. If the system starts at a state where $S_{d-1}(x) = 0$, then $S_{d-1}(x)$ remains zero until the number of free servers in the system drops strictly below d since all arrivals finding d or more free servers will only increase x_d keeping the other components of x the same. Once the number of free servers drops strictly below d, the next arrival increases $S_{d-1}(x)$ from zero to at most $1/n$. Let us call this job the tagged job. If upon arrival of the tagged job, the sum $S_{d-1}(x)$ increases to $1/n$, then the system becomes fully busy after the arrival of the tagged job as the tagged job occupies all remaining servers. Hence, until the tagged job leaves the system, subsequent arrivals either find the system fully busy (hence are blocked) or find at least d servers available (which occurs if a job occupying d servers departs before the arrival occurs and the tagged job leaves the system). In either case, $S_{d-1}(x)$ remains constant at $1/n$ until the tagged job departs. When the tagged job departs, $S_{d-1}(x)$ again drops to zero. From this point onward we can apply the same chain of arguments as above. This shows that $S_{d-1}(x)$ never increases beyond $1/n$ which proves the first part of the lemma.

If the system starts at a state satisfying $S_{d-1}(x) > 1/n$, then it is easy to see that with non-zero probability the system enters the set of states where $S_{d-1}(x) = 0$ in finite time. However, the first part of the lemma implies that starting from states satisfying $S_{d-1}(x) = 0$, it is not possible to reach states satisfying $S_{d-1}(x) > 1/n$ and the chain always remains in states satisfying $S_{d-1}(x) \leq 1/n$. Hence, the states with $S_{d-1}(x) > 1/n$ are transient, and the states with $S_{d-1}(x) \leq 1/n$ form a single, finite communicating class. Therefore, there exists a unique invariant probability measure of the chain concentrated only on the states satisfying $S_{d-1}(x) \leq 1/n$. This implies that at steady-state we have $q_1(x) = \sum_{i\in[d]} ix_i < dx_d + dS_{d-1}(x) \leq dx_d + d/n$ with probability one. Hence, when $q_1 > 1-d/n$, we have $dx_d > 1-2d/n$. When $\lambda(n)$ is a function of n and varies as $\lambda(n) = 1-\beta/n^\alpha$ for $\alpha \in [0,1)$ and $\beta > 0$, it increases at a slower rate than $1-2d/n$. Thus, it holds that $dx_d \geq \lambda(n)$ for all n sufficiently large. This shows that $\mathbb{E}\left[\mathbb{1}\left(q_1 > 1-\frac{d}{n}\right)(\lambda(n) - dx_d)\right] \leq 0$ for all sufficiently large n. □

Combining Lemmas 4.3.1 and 4.3.2, we conclude the following.

Corollary 4.3.3. *Let $k(n) = n$ and $\lambda(n) = 1-\beta/n^\alpha \geq 0$ for $\alpha \in [0,1)$ and $\beta > 0$. Then, under the equilibrium measure, we have*

$$\mathbb{E}\left[(\lambda(n) - dx_d)^2\right] \leq \frac{d\lambda(n)}{n}, \tag{4.17}$$

for all sufficiently large n.

The above lemma and corollary establish that as the system size n increases, the higher dimensional system undergoes a state space collapse and reduces to a lower dimension. Specifically, the d-dimensional system simplifies to one dimension for large system size, where it can be fully described by only the jobs occupying d servers simultaneously. This result has important implications, as it guarantees that every accepted arrival into the system attains the minimum possible average response time in the asymptotic limit. Consequently, the system demonstrates asymptotic optimality in terms of average response time.

In the following theorem, we further demonstrate this result, along with the asymptotic optimality in terms of the blocking probability. Specifically, we establish that the system achieves the minimum average response time of $\frac{1}{d}$ and zero blocking probability as $n \to \infty$. Additionally, we present upper bounds for the performance of finite-sized systems.

Theorem 4.3.4. *Let $k(n) = n$, and $\lambda(n) = 1 - \beta/n^{\alpha} \geq 0$, where $\alpha \in [0, 1)$ and $\beta > 0$. In the steady-state regime, as $n \to \infty$, the blocking probability of the system converges to zero, and the mean response time of accepted jobs converges to $\frac{1}{d}$. Additionally, for finite-sized systems with n large enough, the error in the blocking probability is bounded by $O(\frac{1}{\sqrt{n}})$, and the error in mean response time is bounded by $O(\frac{1}{n})$.*

Proof. By the rate conservation law, we have

$$\mathbb{E}\left[q_1\right] = \lambda(n)\left(1 - P_b\right),$$

where P_b denotes the steady state blocking probability of the system. On the other hand, from the definition of $q_1 = \sum_{i \in [d]} i x_i$, we have $d x_d \leq q_1$ which leads to $\mathbb{E}\left[d x_d\right] \leq \lambda(n)\left(1 - P_b\right)$. Hence,

$$P_b^2 \leq \frac{1}{\lambda(n)^2}\left(\mathbb{E}\left[\lambda(n) - d x_d\right]\right)^2 \leq \frac{1}{\lambda(n)^2}\mathbb{E}\left[\left(\lambda(n) - d x_d\right)^2\right] \leq \frac{d}{n\lambda(n)},$$

where the second inequality follows from Jensen's inequality and the last inequality follows from Corollary 4.3.3.

Note that $\lambda(n) = 1 - \beta/n^{\alpha}$. If $\alpha = 0$ and $0 < \beta < 1$, then it readily follows that

$$P_b^2 \leq \frac{d}{n(1-\beta)}, \tag{4.18}$$

and the steady state blocking probability of the system converges to zero with an error bound of $O(\frac{1}{\sqrt{n}})$ for sufficiently large values of n. If $\alpha > 0$ and $\beta > 0$, since $\frac{\beta}{n^{\alpha}} < 1$ for large n, we have

$$P_b^2 \leq \frac{d}{n}\left(1 + \frac{\beta}{n^{\alpha}} + O(\frac{1}{n^{2\alpha}})\right) \leq \frac{d}{n} + o(\frac{1}{n}). \tag{4.19}$$

Hence, again the steady state blocking probability of the system converges to zero with an error bounded by $O(\frac{1}{\sqrt{n}})$ when n is finite.

Consider the mean response time of accepted jobs in the steady state given by $\mathbb{E}\left[D\right]$. From Little's law for the stationary regime, we have

$$\lambda(n)(1-P_b)\mathbb{E}\left[D\right] = \mathbb{E}\left[\sum_{i\in[d]} x_i\right].$$

From Lemma 4.3.2, the first $d-1$ states of the system form a single finite communicating class in steady-state, such that $\sum_{i\in[d-1]} x_i \leq 1/n$. Moreover, $dx_d \leq q_1$, due to the definition of q_1. Thus

$$\mathbb{E}\left[D\right] \leq \frac{1}{\lambda(n)(1-P_b)}\left(\frac{1}{n}+\frac{\mathbb{E}\left[q_1\right]}{d}\right).$$

But we know that $\mathbb{E}\left[q_1\right] = \lambda(n)\left(1-P_b\right)$ by the rate conservation law. Therefore,

$$\mathbb{E}\left[D\right] \leq \frac{1}{d} + \frac{1}{n\lambda(n)(1-P_b)}.$$

If $\alpha = 0$ and $0 < \beta < 1$, then from Equation (4.18), and the fact that $\sqrt{\frac{d}{n(1-\beta)}} < 1$ for large n, we have

$$\begin{aligned}\mathbb{E}\left[D\right] &\leq \frac{1}{d} + \frac{1}{n(1-\beta)}\left(1+\sqrt{\frac{d}{n(1-\beta)}} + O(\frac{1}{n})\right)\\ &\leq \frac{1}{d} + \frac{1}{n(1-\beta)} + o(\frac{1}{n}). \qquad (4.20)\end{aligned}$$

If $\alpha > 0$ and $\beta > 0$, then from Equation (4.19), and the facts $\sqrt{\frac{d}{n}} < 1$ and $\frac{\beta}{n^\alpha} < 1$, for large n, we have

$$\begin{aligned}\mathbb{E}\left[D\right] &\leq \frac{1}{d} + \frac{1}{n}\left(1+\frac{\beta}{n^\alpha} + O(\frac{1}{n^{2\alpha}})\right)\left(1+\sqrt{\frac{d}{n}+o(\frac{1}{n})} + O(\frac{1}{n})\right)\\ &\leq \frac{1}{d} + \frac{1}{n} + o(\frac{1}{n}). \qquad (4.21)\end{aligned}$$

Therefore, the asymptotic value of the mean response time of the system is $\frac{1}{d}$, with an error bound of $O(\frac{1}{n})$. □

4.3.2 Limited System Access

We now study the system with $k(n) < n$. Under this condition, the arrivals have access to a limited subset of servers of size $k(n)$, which are randomly sampled upon their arrival. We identify sufficient conditions on system parameters that guarantee the state space collapse results in the asymptotic limit. As a consequence, the system maintains the property of asymptotic optimality in terms of mean response time and blocking probability, even with limited subset access. However, the error bounds in the finite systems are different and depend on the specific characteristics of the sampled set.

In the following lemma, we derive an upper bound on the probability of an arrival not finding d idle servers in the sampled set of size $k(n)$.

Lemma 4.3.5. *If $q_1 \leq 1 - \varepsilon$ for $\varepsilon > 0$, we have*

$$1 - A_d(x) \leq d(k(n))^d \left(1 - \varepsilon\right)^{k(n)-d}.$$

Proof. From the definition of assignment policies for $k(n) \leq n$ in Equation (4.8) we have

$$1 - A_d(x) = \sum_{i=0}^{d-1} \frac{\binom{nq_0}{i}\binom{nq_1}{k(n)-i}}{\binom{n}{k(n)}} = \sum_{i=0}^{d-1} \binom{k(n)}{i} (q_0)^i (q_1)^{k(n)-i},$$

for n sufficiently large. Moreover, from the fact $q_0 \leq 1$ and the assumption $q_1 \leq 1 - \varepsilon$, we have

$$\begin{aligned} 1 - A_d(x) &\leq \sum_{i=0}^{d-1} \frac{(k(n))^i (1-\varepsilon)^{k(n)-i}}{i!} \\ &\leq d(k(n))^d (1-\varepsilon)^{k(n)-d} \end{aligned}$$

for n sufficiently large. □

Lemma 4.3.6. *If $\lambda(n) = 1 - \beta/n^\alpha \geq 0$ for $\alpha \in [0, 1)$ and $\beta > 0$, we have*

$$\begin{aligned} \mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\right] \leq \lambda(n) d(k(n))^d \left(1 - \frac{\beta}{2n^\alpha}\right)^{k(n)-d} \\ + d\ \mathbb{E}\left[(\sum_{i\in[d]} x_i) \mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]. \end{aligned} \tag{4.22}$$

Proof. We consider two cases where the fraction of busy servers exceeds the threshold $1 - \frac{\beta}{2n^\alpha}$ and when it is below that threshold. We have

$$\begin{aligned}\mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\right] =& \mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\mathbb{1}\left(q_1 \leq 1 - \frac{\beta}{2n^\alpha}\right)\right] && (4.23)\\ &+ \mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]. && (4.24)\end{aligned}$$

We bound each of the terms above in (4.23) and (4.24).

Consider the term in expression (4.23). We assume $q_1 \leq 1 - \frac{\beta}{2n^\alpha}$, otherwise this term will trivially become zero due to the indicator function. Thus, we have $q_1 \leq 1 - \varepsilon$ with $\varepsilon = \frac{\beta}{2n^\alpha} > 0$, and from Lemma 4.3.5, we conclude that

$$\mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\mathbb{1}\left(q_1 \leq 1 - \frac{\beta}{2n^\alpha}\right)\right] \leq \lambda(n)d(k(n))^d\left(1 - \frac{\beta}{2n^\alpha}\right)^{k(n)-d}. \quad (4.25)$$

Consider the second term given by expression (4.24). We have

$$\begin{aligned}&\mathbb{E}\left[(1 - A_d(x))(\lambda(n) - dx_d)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]\\ \leq& \mathbb{E}\left[(1 - A_d(x))(q_1 - \frac{\beta}{2n^\alpha} - dx_d)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]\\ \leq& \mathbb{E}\left[(1 - A_d(x))(q_1 - dx_d)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]\\ =& \mathbb{E}\left[(1 - A_d(x))(\sum_{i\in[d-1]} ix_i)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right] && (4.26)\\ \leq& d\ \mathbb{E}\left[(1 - A_d(x))(\sum_{i\in[d-1]} x_i)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right]\\ \leq& d\ \mathbb{E}\left[(\sum_{i\in[d]} x_i)\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)\right],\end{aligned}$$

where the second line follows from the indicator function $\mathbb{1}\left(q_1 > 1 - \frac{\beta}{2n^\alpha}\right)$ and the definition of $\lambda(n) = 1 - \frac{\beta}{n^\alpha}$; the fourth line follows from the definition of $q_1 = \sum_{i\in[d]} ix_i$; and the last line follows from the non-negativity of x_i. Combining all the bounds together, we get the desired result. □

In the following lemma, we establish an upper bound for expression (4.26). We introduce a new Lyapunov function $V_2(x) = \sum_{i\in[d]} x_i \ \mathbb{1}\left(q_1 > \lambda(n) + \delta\right)$ for a positive δ, and show that outside of a suitable compact set, the drift of this Lyapunov function is strictly negative. Consequently, this implies that with high probability, the function $V_2(x)$ remains within that compact set.

Intuitively, when the fraction of busy servers exceeds a threshold that tends to one as n grows, the number of accepted jobs into the system cannot increase substantially. In other words, the occurrence of two events of a significantly high number of busy servers and the acceptance of jobs into the system is highly improbable.

Lemma 4.3.7. *For any $\delta \in (0, \frac{\beta}{n^\alpha})$, define the following Lyapunov function.*

$$V_2(x) = \sum_{i\in[d]} x_i \ \mathbb{1}\left(q_1 > \lambda(n) + \delta\right). \tag{4.27}$$

If $V_2(x) \geq \kappa$ for some $\kappa > 0$, then $GV_2(x) \leq -\delta$ and $\mathbb{E}\left[V_2(x)\right] \leq \kappa + \frac{2}{n\delta}$, for all sufficiently large n.

Proof. Assume $V_2(x) \geq \kappa$ for some $\kappa > 0$. This implies

$$\mathbb{1}\left(q_1 > \lambda(n) + \delta\right) = 1, \tag{4.28}$$

and

$$q_1 > \lambda(n) + \delta. \tag{4.29}$$

Let us calculate the drift of $V_2(x)$ under G, when $V_2(x) \geq \kappa > 0$. From the definition of the generator G in Lemma 4.3.1, we have

$$GV_2(x) = \sum_{i\in[d]} n\lambda(n)A_i(x)\left(V_2\left(x + \frac{e_i}{n}\right) - V_2(x)\right) + \sum_{i\in[d]} nix_i\left(V_2\left(x - \frac{e_i}{n}\right) - V_2(x)\right),$$

where e_i denotes the d-dimensional unit vector with one at the i^{th} position. Therefore,

$$
\begin{aligned}
GV_2(x) = \sum_{i\in[d]} n\lambda(n)A_i(x)\Bigg(&\sum_{j\in[d]}\left(x_j+\frac{e_i}{n}\right)\mathbb{1}\left(q_1+\frac{i}{n}>\lambda(n)+\delta\right) \\
&-\sum_{j\in[d]} x_j\ \mathbb{1}\left(q_1>\lambda(n)+\delta\right)\Bigg) \\
+\sum_{i\in[d]} nix_i\left(\sum_{j\in[d]}\left(x_j-\frac{e_i}{n}\right)\mathbb{1}\left(q_1-\frac{i}{n}>\lambda(n)+\delta\right)-\sum_{j\in[d]} x_j\ \mathbb{1}\left(q_1>\lambda(n)+\delta\right)\right).
\end{aligned}
$$

However, under the assumption $V_2(x)\geq\kappa>0$, from Equations (4.28) and (4.29), we have

$$\mathbb{1}\left(q_1+\frac{i}{n}>\lambda(n)+\delta\right)=1,\quad \text{for any } i\in[d],$$

and

$$\mathbb{1}\left(q_1-\frac{i}{n}>\lambda(n)+\delta\right)=1,\quad \text{for any } i\in[d],\ \text{and } n \text{ sufficiently large.}$$

As a consequence, for n large enough, we have

$$
\begin{aligned}
GV_2(x) &= \sum_{i\in[d]} n\lambda(n)A_i(x)(\frac{1}{n})+\sum_{i\in[d]} nix_i(-\frac{1}{n}) \\
&= \lambda(n)\sum_{i\in[d]} A_i(x)-\sum_{i\in[d]} ix_i \\
&\leq \lambda(n)-q_1 \\
&\leq -\delta,
\end{aligned}
\tag{4.30}
$$

where the third line follows from the fact $\sum_{i\in[d]} A_i(x)\leq 1$ and the last line follows from condition (4.29). This shows that outside of the set $\{x : V_2(x)\leq\kappa, \kappa>0\}$, the function has a negative drift and completes the first part of the lemma.

For the second part, we use the results from [84, Theorem 1 - (i)] which we recall in Appendix B.1. Under the notation of of Appendix B.1, we have $p_{max}=n$ and $\nu_{max}=1/n$. Hence

$$\mathbb{E}\left[V_2(x)\right]\leq\kappa+\frac{2p_{max}(\nu_{max})^2}{\delta} \tag{4.31}$$

$$
\begin{aligned}
&\leq \kappa + \frac{2n(1/n^2)}{\delta} && (4.32)\\
&= \kappa + \frac{2}{n\delta}. && (4.33)
\end{aligned}
$$

and this completes the proof. □

Combining Lemmas 4.3.1,4.3.6 and 4.3.7, we obtain the following.

Corollary 4.3.8. *Let $\lambda(n) = 1 - \beta/n^\alpha \geq 0$ for $\alpha \in [0,1)$ and $\beta > 0$. Then, in the stationary regime, we have*

$$
\mathbb{E}\left[(\lambda(n) - dx_d)^2\right] \leq \frac{2d}{n}\lambda(n)\left(1 + \frac{2n^\alpha}{\beta}\right) + \lambda(n)^2 d(k(n))^d\left(1 - \frac{\beta}{2n^\alpha}\right)^{k(n)-d}, \qquad (4.34)
$$

for all n sufficiently large.

Proof. The result follows by choosing $\kappa = \frac{1}{n}$ and $\delta = \frac{\beta}{2n^\alpha}$ in Lemma 4.3.7 and combining with Lemmas 4.3.1 and 4.3.6. □

In the following theorem, we present the main performance bounds for systems with a finite size. We obtain sufficient conditions on the growth rate of the size of the sampling set $k(n)$, for the system to achieve asymptotic optimality.

Theorem 4.3.9. *Let $\lambda(n) = 1 - \beta/n^\alpha \geq 0$, for $\alpha \in [0,1)$ and $\beta > 0$, and $k(n) = n^\alpha \log(n)$. If $\beta > 2(\alpha(d-1)+1)$ when $\alpha > 0$, then in the steady state regime, as $n \to \infty$, the blocking probability of the system converges to zero, and the mean response time of accepted jobs converges to $\frac{1}{d}$. Furthermore, in finite-sized systems, the error in these estimations is bounded by $O(n^{-(1-\alpha)/2})$ for n large enough.*

Proof. Let P_b denote the blocking probability of the system. By the rate conservation law $\lambda(n)\,(1 - P_b) = \mathbb{E}\,[q_1] \geq \mathbb{E}\,[dx_d]$, and from the same arguments as in the proof of Theorem 4.3.4, we have

$$
\begin{aligned}
P_b^2 &\leq \frac{1}{\lambda(n)^2}\mathbb{E}\left[(\lambda(n) - dx_d)^2\right]\\
&\leq \frac{2d}{n\lambda(n)}\left(1 + \frac{2n^\alpha}{\beta}\right) + dn^{\alpha d}(\log(n))^d\left(1 - \frac{\beta}{2n^\alpha}\right)^{n^\alpha \log(n) - d}, \qquad (4.35)
\end{aligned}
$$

where the last inequality follows from Corollary 4.3.8 when $k(n) = n^\alpha \log(n)$, and n is sufficiently large. If $\alpha = 0$ and $0 < \beta < 1$, we have

$$P_b^2 \leq \frac{2d}{n(1-\beta)}\left(1+\frac{2}{\beta}\right) + o(\frac{1}{n}). \tag{4.36}$$

Thus, as $n \to \infty$, the blocking probability of the system approaches zero, and for sufficiently large values of n, it is upper bounded by $O(\frac{1}{\sqrt{n}})$

If $\alpha > 0$ and $\beta > 2(\alpha(d-1)+1)$, then from Equation (4.35) we have

$$P_b^2 \leq \frac{2d}{n\lambda(n)} + \frac{4d}{\beta\lambda(n)}n^{\alpha-1} + dn^{\alpha d}(\log(n))^d\left(1-\frac{\beta}{2n^\alpha}\right)^{n^\alpha \log(n)-d}.$$

Based on the properties of β, the second term in the above bound is the dominant term and we have

$$P_b^2 \leq \frac{4d}{\beta\lambda(n)}n^{\alpha-1} + o(n^{\alpha-1}).$$

For n large enough, $\frac{\beta}{n^\alpha} < 1$, and we have

$$\begin{aligned} P_b^2 &\leq \frac{4d}{\beta}n^{\alpha-1}\left(1+\frac{\beta}{n^\alpha}+O(\frac{1}{n^{2\alpha}})\right) + o(n^{\alpha-1}) \\ &= \frac{4d}{\beta}n^{\alpha-1} + o(n^{\alpha-1}). \end{aligned} \tag{4.37}$$

As a consequence, the blocking probability of the system approaches zero as $n \to \infty$, and for n large enough, it is upper bounded by $O(n^{-(1-\alpha)/2})$.

Let $\mathbb{E}[D]$ denote the mean response time of accepted jobs. From Little's law in the stationary regime, we have

$$\lambda(n)\,(1-P_b)\,\mathbb{E}[D] = \mathbb{E}\left[\sum_{i\in[d]} x_i\right] = \mathbb{E}\left[\sum_{i\in[d-1]} x_i\right] + \mathbb{E}[x_d].$$

For the first $d-1$ components of the system state, note that

$$\mathbb{E}\left[\sum_{i\in[d-1]} x_i\right] \leq \mathbb{E}\left[\sum_{i\in[d-1]} i x_i\right] = \mathbb{E}\left[q_1 - d x_d\right] \leq \mathbb{E}\left[\lambda(n) - d x_d\right],$$

where the equality follows from the definition of $q_1 = \sum_{i\in[d]} i x_i$; and the last inequality follows from the rate conservation law $\mathbb{E}\left[q_1\right] = \lambda(n)(1-P_b) \leq \lambda(n)$. Moreover, for the d^{th} component of the system state, we have $d x_d \leq q_1$ due to the definition of q_1 and $\mathbb{E}\left[x_d\right] \leq \frac{\mathbb{E}[q_1]}{d}$. Hence

$$\begin{aligned}\mathbb{E}\left[D\right] &\leq \frac{1}{\lambda(n)\left(1-P_b\right)}\left(\mathbb{E}\left[\lambda(n) - d x_d\right] + \frac{\mathbb{E}\left[q_1\right]}{d}\right)\\ &= \frac{1}{\lambda(n)\left(1-P_b\right)}\mathbb{E}\left[\lambda(n) - d x_d\right] + \frac{1}{d},\end{aligned}$$

where the last line follows from the rate conservation law. Therefore,

$$\begin{aligned}\left(\mathbb{E}\left[D\right] - \frac{1}{d}\right)^2 &\leq \frac{1}{\lambda(n)^2\left(1-P_b\right)^2}\left(\mathbb{E}\left[\lambda(n) - d x_d\right]\right)^2\\ &\leq \frac{1}{\lambda(n)^2\left(1-P_b\right)^2}\mathbb{E}\left[\left(\lambda(n) - d x_d\right)^2\right],\end{aligned}$$

where the second inequality follows from Jensen's inequality.

For $\alpha = 0$ and $0 < \beta < 1$, from Corollary 4.3.8 and Equation (4.36), we have

$$\mathbb{E}\left[D - \frac{1}{d}\right]^2 \leq \frac{1}{(1-\beta)^2\left(1 - 2\sqrt{\frac{2d}{n(1-\beta)}(1+\frac{2}{\beta}) + o(\frac{1}{n})}\right)}\left(\frac{2d(1-\beta)}{n}\left(1+\frac{2}{\beta}\right) + o(\frac{1}{n})\right).$$

But $2\sqrt{\frac{2d}{n(1-\beta)}(1+\frac{2}{\beta}) + o(\frac{1}{n})} < 1$, for n large enough and hence

$$\mathbb{E}\left[D - \frac{1}{d}\right]^2 \leq \frac{1}{(1-\beta)^2}\left(1 + 2\sqrt{\frac{2d}{n(1-\beta)}(1+\frac{2}{\beta}) + o(\frac{1}{n})} + O(\frac{1}{n})\right)$$

$$
\begin{aligned}
&\times \left(\frac{2d(1-\beta)}{n}\left(1+\frac{2}{\beta}\right)+o(\frac{1}{n})\right) \\
&\leq \frac{2d}{n(1-\beta)}\left(1+\frac{2}{\beta}\right)+o(\frac{1}{n}).
\end{aligned} \tag{4.38}
$$

This shows that the mean response time of accepted jobs converges to $\frac{1}{d}$ as $n \to \infty$, with an error bound in finite-sized systems of $O(\frac{1}{\sqrt{n}})$.

If $\alpha > 0$ and $\beta > 2(\alpha(d-1)+1)$, from Corrolary 4.3.8 and Equation (4.37) we have

$$
\mathbb{E}\left[D-\frac{1}{d}\right]^2 \leq \frac{1}{\left(1-\frac{2\beta}{n^\alpha}\right)\left(1-2\sqrt{\frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1})}\right)}\left(\frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1})\right).
$$

For n large enough, $\frac{2\beta}{n^\alpha} < 1$ and $2\sqrt{\frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1})} < 1$, and we have

$$
\begin{aligned}
\mathbb{E}\left[D-\frac{1}{d}\right]^2 &\leq \left(1+\frac{2\beta}{n^\alpha}+O(\frac{1}{n^{2\alpha}})\right)\left(1+2\sqrt{\frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1})}+O(n^{\alpha-1})\right) \\
&\times \left(\frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1})\right) \\
&\leq \frac{4d}{\beta}n^{\alpha-1}+o(n^{\alpha-1}).
\end{aligned} \tag{4.39}
$$

As a consequence, the mean response time of accepted jobs converges to $\frac{1}{d}$ with an error bound of $O(n^{-(1-\alpha)/2})$ for large values of n, and this completes the proof. □

4.4 Heterogeneous Workloads

So far we have considered the case with homogeneous job arrivals. The results can be extended to the heterogeneous case of multi-class arrivals where jobs of different classes correspond to different sizes and can have a different maximum degree of parallelism.

We consider a system with L classes of arrival streams denoted as $[L] = \{1, 2, \ldots, L\}$. Each arrival from class $\ell \in [L]$ follows a Poisson process with the rate $n\lambda_\ell(n)$ where $\lambda_\ell(n) = 1-\beta_\ell n^{-\alpha_\ell} \geq 0$, for some $\alpha_\ell \in [0, 1)$ and$\beta_\ell > 0$. These arrivals bring jobs where the undivided length of the job follows an exponential distribution with an average of $\frac{1}{\mu_\ell}$. It is assumed that the system has sufficient capacity to handle the incoming workload. This is formally expressed as $\rho = \sum_{\ell\in[L]} \frac{\lambda_\ell(n)}{\mu_\ell} < 1$.

The job assignment policy is such that upon the arrival of a job, $k(n)$ servers are uniformly and randomly selected from the total of n servers. Depending on the number of free servers in the sampled subset, each arrival of class ℓ is split into a maximum of d_ℓ sub-jobs and receives a speedup of i when it is divided to $i \in [d_\ell]$ parts.

The state of the system is represented by the double vector $x = ((x_{\ell,i}, \ell \in [L]), i \in [d_\ell])$, where $x_{\ell,i}$ denotes the fluid scaled number of class ℓ jobs being processed by i servers simultaneously. Additionally, the fraction of busy servers is denoted by q_1, where $q_1 := q_1(x) = \sum_{\ell\in[L]}\sum_{i\in[d_\ell]} i x_{\ell,i}$, and the fraction of idle servers is denoted by $q_0 := q_0(x) = 1 - q_1(x)$.

We show that as the system size increases, each arrival from class ℓ finds at least d_ℓ free servers to join, and hence x_{ℓ,d_ℓ} converges in probability to $\frac{\lambda_\ell(n)}{\mu_\ell d_\ell}$ for every $\ell \in [L]$. This convergence implies a state space collapse result, where the system's dimension reduces to a lower dimension. The formal statement of this result is given in Proposition 4.4.1. The proof of the proposition follows similar arguments to those in Lemma 4.3.1 and Section 4.3.2, with necessary adjustments made to address the multiple arrival streams with heterogeneous workloads.

Proposition 4.4.1. *Consider a system with L classes of arrival streams. Let $\rho = \sum_{\ell\in[L]} \frac{\lambda_\ell(n)}{\mu_\ell} < 1$. Then, under the equilibrium measure, we have*

$$\mathbb{E}\left[\sum_{\ell\in[L]}(\lambda_\ell(n) - \mu_\ell d_\ell x_{\ell,d_\ell})^2\right] \le \frac{c}{n} + \sum_{\ell\in[L]} \lambda_\ell(n)^2 d_\ell (k(n))^{d_\ell}\left(\frac{1+\rho}{2}\right)^{k(n)-d_\ell}, \tag{4.40}$$

where $c = \sum_{\ell\in[L]} \lambda_\ell(n)\mu_\ell d_{\max}\left(1 + \mu_{\max} + \frac{4\mu_{\max}}{\mu_{\min}^2(1-\rho)}\right)$, with $\mu_{\max} = \underset{\ell}{max}\ \mu_\ell$, $\mu_{\min} = \underset{\ell}{min}\ \mu_\ell$, $d_{\max} = \underset{\ell}{max}\ d_\ell$, and n is sufficiently large.

Sketch of the Proof. The proof relies on similar arguments as presented in Lemmas 4.3.1, 4.3.5 and 4.3.7. However, some adjustments are needed to accommodate the presence of multiple arrival streams as follows.

- In Lemma 4.3.1, we modify the Lyapunov function to $V(x) = \sum_{\ell\in[L]} \frac{1}{2\mu_\ell d_\ell}(\lambda_\ell(n) - \mu_\ell d_\ell x_{\ell,d_\ell})^2$, and update the system of ODEs with generator L as follows.

$$\begin{aligned} \dot{x}_{\ell,i} &= -\mu_\ell i x_{\ell,i}, \quad \forall \ell \in [L] \text{ and } i \in [d_\ell - 1], \\ \dot{x}_{\ell,d_\ell} &= \lambda_\ell(n) - \mu_\ell d_\ell x_{\ell,d_\ell}, \quad \forall \ell \in [L]. \end{aligned}$$

This results in

$$\mathbb{E}\left[\sum_{\ell\in[L]} (\lambda_\ell(n) - \mu_\ell d_\ell x_{\ell,d_\ell})^2\right] = \sum_{\ell\in[L]} \frac{\lambda_\ell(n)\mu_\ell d_\ell}{n} \mathbb{E}\left[A_{\ell,d_\ell}(x)\right] \\ + \mathbb{E}\left[\sum_{\ell\in[L]} \lambda_\ell(n)(1 - A_{\ell,d_\ell}(x))(\lambda_\ell(n) - \mu_\ell d_\ell x_{\ell,d_\ell})\right]. \quad (4.41)$$

The first term on the RHS is $O(\frac{1}{n})$, and we need to bound the second term. We consider two cases for the second term; when the fraction of busy servers q_1 is below a threshold uniformly bounded away from one, and another where q_1 is above the same threshold.

- In the first case where q_1 is below the threshold $1 - \varepsilon < 1$, we adapt the results of Lemma 4.3.5. We introduce $A_{\ell,i}(x)$, representing the probability that a class ℓ job finds $i \in \{0, 1, ..., d_\ell\}$ free servers upon arrival in state x. The probabilities $A_{\ell,i}(x)$ take a similar form as in Equation (4.8) and are defined below.

$$A_{\ell,i}(x) = \begin{cases} \frac{\binom{nq_0}{i}\binom{nq_1}{k(n)-i}}{\binom{n}{k(n)}}, & \text{if } i \in \{0, 1, 2, \ldots, d_\ell - 1\}, \\ \sum_{i\geq d_\ell} \frac{\binom{nq_0}{i}\binom{nq_1}{k(n)-i}}{\binom{n}{k(n)}}, & \text{if } i = d_\ell. \end{cases}$$

Then, for each class $\ell \in [L]$, we have

$$1 - A_{\ell,d_\ell}(x) \leq d_\ell (k(n))^{d_\ell} (1 - \varepsilon)^{k(n) - d_\ell},$$

given that $q_1 \leq 1 - \varepsilon$ for some $\varepsilon > 0$. By carefully selecting $\varepsilon = \frac{1-\rho}{2}$, we ensure that the second term on the RHS of Equation (4.41) approaches zero as $n \to \infty$.

- In the second case, where q_1 is above the same threshold $1 - \varepsilon = \frac{1+\rho}{2}$, we redefine the Lyapunov function $V_2(x)$ from Lemma 4.3.7 as

$$V_2(x) = \sum_{\ell \in [L]} \frac{1}{\mu_\ell} \sum_{i \in [d_\ell]} x_{\ell,i} \, \mathbb{1} \left(q_1 > \frac{1+\rho}{2} \right).$$

We then show that $GV_2(x) \leq -\frac{1-\rho}{2}$ when $V_2(x) \geq \kappa$ for some $\kappa > 0$. Thus, the Lyapunov function $V_2(x)$ converges to zero in probability. This implies that when the number of busy servers is sufficiently high in the system, the incoming workload to the system diminishes considerably. By manipulating terms in the second expression on the RHS of Equation (4.41), we rewrite it in terms of $V_2(x)$ and show its convergence to zero as $n \to \infty$.

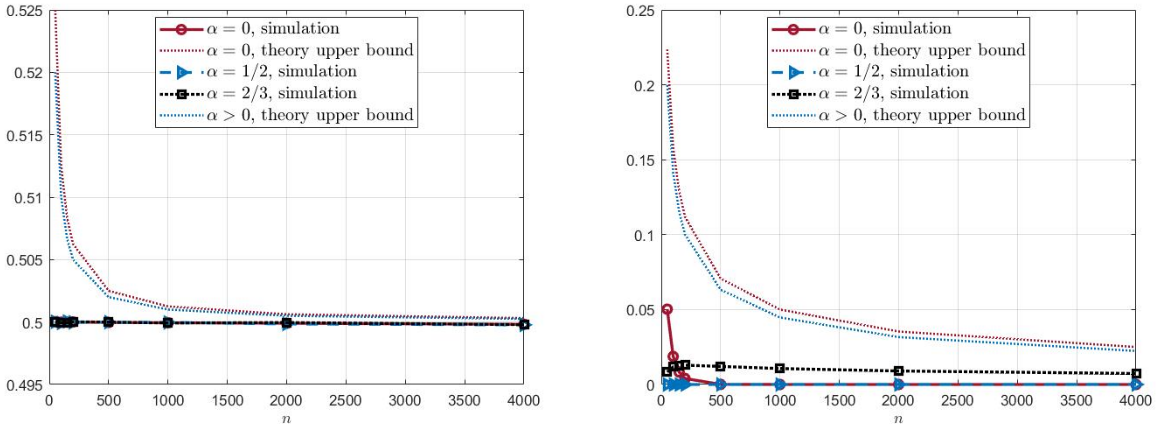
Combining all the results together, we obtain the desired bound in Proposition 4.4.1.

4.5 Numerical Results

In this section, we present simulation results to validate the accuracy of our analysis. We conducted all simulation experiments independently, repeating them 100 times. At each iteration, we measured system performance metrics for the first 5×10^6 jobs arriving at the system.

Figure 4.1 illustrates the simulation results when jobs have access to the full state of the system, i.e., when $k(n) = n$. The maximum degree of parallelism is considered to be $d = 2$. Different values of $\lambda(n)$ are considered including $(i)\alpha = 0, \beta = 0.2$, (ii) $\alpha = 1/2, \beta = 5$, and (iii)$\alpha = 2/3, \beta = 5$. Figure 4.1(a) displays the mean response time of accepted jobs obtained through simulations, alongside the theoretical upper bounds derived in Theorem 4.3.4. This figure highlights a clear convergence of the average response time to the theoretically computed value of $1/d$ as the system size n increases. Moreover, the upper bounds derived are indeed tight in the asymptotic regime and the convergence occurs at a rate of $O(\frac{1}{n})$. Figure 4.1(b) presents the blocking probability of the system, along with the theoretical upper bounds. This figure shows that the rate at which the blocking potability approaches zero exceeds $O(\frac{1}{\sqrt{n}})$, as the system size n increases. In fact, upon closer inspection of the simulation results, it becomes evident that the rate of convergence for the blocking probability when $k(n) = n$ also is $O(\frac{1}{n})$.

Figure 4.2, illustrates the system performance metrics for the same heavy traffic parameters when the jobs have access to only a sampled subset of servers upon their arrival.



(a) Mean response time of accepted jobs

(b) Blocking probability of the system

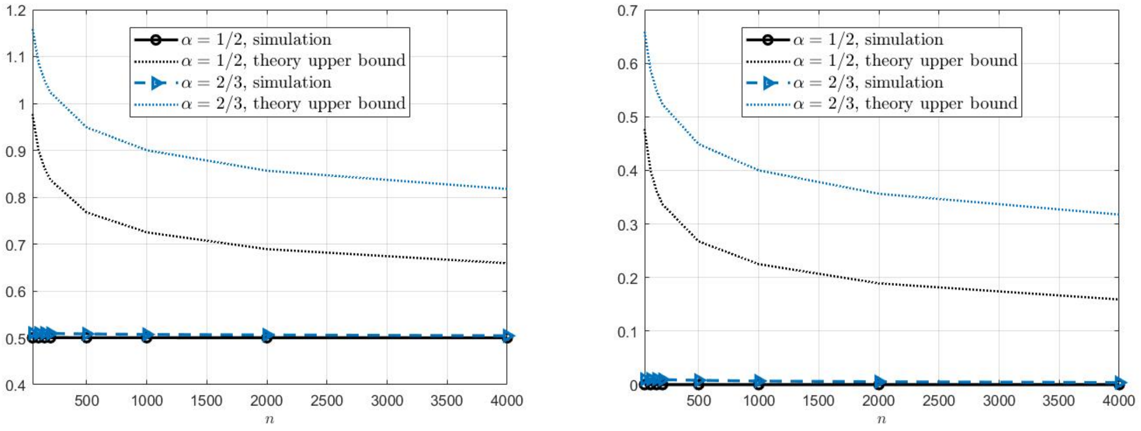
Figure 4.1: System performance metrics for different system sizes n, sampling size $k(n) = n$, degree of parallelism $d = 2$, and various values of load parameters (i)$\alpha = 0, \beta = 0.2$, (ii)$\alpha = 1/2, \beta = 5$, (iii)$\alpha = 2/3, \beta = 5$.

The sampling size $k(n)$ is chosen to be $k(n) = \lceil n^{\alpha} \log(n) \rceil$. Figures 4.2(a) and 4.2(b) illustrate the mean response time of accepted jobs and the blocking probability of the system, respectively. Additionally, upper bounds on system performance metrics are provided. It is observed that the upper bounds are not tight in the partial system access scenario.

The simulation results align consistently with the theoretical results presented in Theorems 4.3.4 and 4.3.9, where it is proven that all arrivals are accepted to the system and distributed across d servers. Moreover, as the graphs indicate, when the sampling size $k(n)$ equals the total number of servers n, the upper bound on the average response time is tight, demonstrating a convergence rate of $O(\frac{1}{n})$.

4.6 Conclusion

In this chapter, we studied the performance of adaptive multiserver-job systems with n servers. The system load was modeled as $1 - \beta n^{-\alpha} \geq 0$, where $\alpha \in [0, 1)$ and $\beta > 0$. Upon arrival, each incoming job had the flexibility to split into a maximum of d smaller tasks, based on the system state and available resources. We showed that when arriving jobs had complete knowledge of the system state and all servers were accessible, the system



(a) Mean response time of accepted jobs

(b) Blocking probability of the system

Figure 4.2: System performance metrics for different system sizes n, sampling size $k(n) = \lceil n^\alpha \log(n) \rceil$, degree of parallelism $d = 2$, and various values of load parameters (i)$\alpha = 1/2, \beta = 5$, (ii)$\alpha = 2/3, \beta = 5$

achieved the optimal average response time of $1/d$ and zero blocking probability in the limit as the system size approached infinity. We further characterized the error bounds in finite-sized systems, demonstrating a bound of $O(1/\sqrt{n})$ for the blocking probability and $O(1/n)$ for the mean response time. When arrivals had partial knowledge of the system state, sampled upon their arrival, similar optimal performance measures were attainable. A necessary condition was that the sampling size must grow at rate $\omega\left(n^\alpha\right)$. In this case, the error bound was established as $O(n^{-(1-\alpha)/2})$ for both the mean delay and blocking probability.

Chapter 5

Performance of Loss Systems with Adaptive Multiserver Jobs and Sublinear Speedup

In this chapter, we expand upon the analysis from the previous chapter to study systems in which the speedup in job processing time is no longer linear, meaning that it does not directly correspond to the number of servers used to process the job. This nonlinearity is a result of the additional overheads that arise in real-world scenarios involving parallel processing. For example, consider a file-server system where requests are made to access files stored redundantly across multiple locations, allowing for simultaneous downloads of different file segments. This introduces challenges such as managing chunk sequence numbers and file reconstruction. These complexities result in job processing speedup being a sublinear function of the number of servers employed. In particular, we focus on scenarios where the speedup function exhibits the characteristics of being strictly increasing, sublinear, and concave concerning the number of servers used for job processing. Within the same system model as in the previous chapter, we characterize the optimal state of the system and show that it is at most two-dimensional. A key observation from this optimal solution is that, under heavy traffic, parallelization no longer offers performance advantages. In such cases, the optimal approach is to assign jobs to individual servers. We introduce a probabilistic load balancing policy and demonstrate that, as the system size increases, the system converges to this established optimal state. As a result, it is asymptotically optimal in terms of average job response time and blocking probability.

The subsequent sections of this chapter are structured as follows. The criteria for optimality are outlined in Section 5.1. The main results regarding the system's asymptotic optimality are presented in Section 5.2, including an analysis of state space collapsing to one dimension, addressed in Subsection 5.2.1, and state space collapsing to two dimensions, discussed in Subsection 5.2.2. Numerical insights are provided in Section 5.3. We draw our conclusions in Section 5.4.

5.1 Criterion for Optimality

As in our prior study in Chapter 4, the initial step requires the establishment of criteria for optimality. We begin by introducing the concept of a speedup function, which quantifies the acceleration in job processing time when executed across multiple servers concurrently. We assume that when a job is parallelized across $i \in [d]$ servers, it receives a speedup of s_i where s_i is a sublinear, strictly increasing and concave function of i satisfying $1 = s_1 < s_2 < \ldots < s_d$, and $s_i/i \geq s_{i+1}/(i+1)$ for all $i \in [d-1]$. The goal is to design job assignment schemes that eliminate the occurrence of job blocking and minimize the average response time of accepted jobs in the steady state system as $n \to \infty$. However, identifying the optimal dynamics of this system is not clear. To explain, let us consider the rate conservation principle and Little's law as follows.

$$\lambda(n)(1-P_b) = \mathbb{E}\left[\sum_{i\in[d]} s_i x_i\right], \tag{5.1}$$

$$\lambda(n)(1-P_b)\mathbb{E}\left[D\right] = \mathbb{E}\left[\sum_{i\in[d]} x_i\right]. \tag{5.2}$$

An ideal optimal system state can be x^*, characterized by an expected value $\mathbb{E}\left[x^*\right] = (0, 0, \ldots, \lambda(n)/s_d)$. This state achieves an average response time of jobs equal to $1/s_d$ which is the smallest attainable value based on Remark 4.1, and ensures zero blocking probability. However, the proposed system state x^*, may not always be feasible due to the physical constraints of the system. Specifically, consider the fraction of busy servers q_1. We find that

$$\mathbb{E}\left[q_1\right] = \sum_{i\in[d]} i\mathbb{E}\left[x_i\right] = \frac{d}{s_d}\lambda(n).$$

The above quantity must always be bounded by one as the fraction of busy servers cannot exceed the total number of servers in the system. However, given that $s_d < d$ (which is the case unless we encounter a scenario in which $s_i = i$ for all $i \in [d]$, effectively reverting to the linear speedup case), and depending on the specific value of $\lambda(n)$, it may not be possible to satisfy the physical constraint associated with q_1.

To this end, we formulate an optimization problem that aims to minimize the steady-state mean response time of the jobs subject to the constraint that the steady-state blocking probability is zero and the physical limitations of the system are satisfied. Therefore, to achieve the optimal system performance, the expected system state $\mathbb{E}[x]$ should be the solution to the optimization problem presented below.

$$\begin{aligned} \underset{y=(y_i, i\in[d])\in R_+^d}{\text{minimize}} \quad & \sum_{i\in[d]} y_i \\ \text{subject to} \quad & \sum_{i\in[d]} s_i y_i = \lambda(n), \\ & \sum_{i\in[d]} i y_i \leq 1, \end{aligned} \tag{5.3}$$

where the objective function corresponds to minimizing the total expected number of jobs in steady-state (note that by (5.2) this sum is proportional to the mean response time of accepted jobs when $P_b = 0$), the first constraint corresponds to the zero blocking condition obtained by setting $P_b = 0$ in (5.1), and the last constraint comes from the fact that the total expected number of occupied servers cannot exceed the total number of servers in the system. Note that $\lambda(n) \leq 1$ and $s_1 = 1$ guarantee the existence of an optimal solution to (5.3) since the d-dimensional vector $y = (\lambda(n), 0, \ldots, 0)$ is always a feasible solution. Let $y^* = (y_i^*, i \in [d])$ denote an optimal solution to the above optimization problem. In the proposition below, we characterize the optimal solution y^* of (5.3) as a function of the normalized arrival rate $\lambda(n)$ and the speed-up function.

Proposition 5.1.1. *Let s_i be a strictly increasing, concave function of i satisfying*

$$1 = \frac{s_1}{1} \geq \frac{s_2}{2} \geq \ldots \geq \frac{s_d}{d}. \tag{5.4}$$

The following results hold.

(i) If $\lambda(n) \leq \frac{s_d}{d}$, then the optimal solution is unique and is given by $y^* = (0, 0, \ldots, 0, \frac{\lambda(n)}{s_d})$.

(ii) If $\lambda(n) = s_i/i$ for some $i \in [d]$, then the optimal solution is unique and satisfies $y_{i_1}^* = \lambda(n)/s_{i_1}$ and $y_j^* = 0$ for all $j \neq i_1$, where $i_1 = \max\{i \in [d] : \lambda(n) = s_i/i\}$.

(iii) If $\lambda(n) \in (\frac{s_{i+1}}{i+1}, \frac{s_i}{i})$ for some $i \in [d-1]$ satisfying $\frac{s_{i+1}}{i+1} < \frac{s_i}{i}$, then an optimal solution to (5.3) can be obtained by setting

$$y_i^* = \frac{\frac{1}{i}\left(\lambda(n) - \frac{s_{i+1}}{i+1}\right)}{\frac{s_i}{i} - \frac{s_{i+1}}{i+1}}, \tag{5.5}$$

$$y_{i+1}^* = \frac{\frac{1}{i+1}\left(\frac{s_i}{i} - \lambda(n)\right)}{\frac{s_i}{i} - \frac{s_{i+1}}{i+1}}, \tag{5.6}$$

$$y_j^* = 0, \forall j \notin \{i, i+1\}. \tag{5.7}$$

Furthermore, the above solution is unique if $s_i - s_{i-1} > s_{i+1} - s_i > s_{i+2} - s_{i+1}$.

Proof. The Lagrangian function is given by

$$\mathcal{L}\left(y, \nu, \theta_0, \theta_1, \ldots, \theta_d\right) = \sum_{i\in[d]} y_i + \nu\left(\sum_{i\in[d]} s_i y_i - \lambda(n)\right) + \theta_0(\sum_{i\in[d]} i y_i - 1) - \sum_{i\in[d]} \theta_i y_i,$$

where ν, $\theta_0 \geq 0$, and $\theta_i \geq 0$ for $i \in [d]$ represent the Lagrange multipliers corresponding to their respective constraints. Slater's condition for strong duality holds since $y = (\lambda(n), 0, \ldots, 0)$ is a feasible solution for all $\lambda(n) \leq 1$. Consequently, any primal optimal solution $y = (y_1, \ldots, y_d)$ and dual optimal solution $(\nu, \theta_0, \theta_1, \ldots, \theta_d)$ must satisfy the Karush-Kuhn-Tucker (KKT) conditions outlined below.

$$\frac{\partial \mathcal{L}}{\partial y_i} = 1 + \nu s_i + \theta_0 i - \theta_i = 0, \forall i \in [d], \qquad \text{(Stationarity)} \tag{5.8}$$

$$\theta_0 \geq 0, \quad \theta_i \geq 0, \forall i \in [d], \qquad \text{(Dual Feasibility)} \tag{5.9}$$

$$\theta_0\left(\sum_{i\in[d]} i y_i - 1\right) = 0, \quad \theta_i y_i = 0, \forall i \in [d], \qquad \text{(Complimentary Slackness)} \tag{5.10}$$

$$\sum_{i\in[d]} s_i y_i = \lambda(n), \quad \sum_{i\in[d]} i y_i \leq 1, \quad y_i \geq 0, \forall i \in [d]. \quad \text{(Primal Feasibility)} \tag{5.11}$$

From the primal feasibility constraint (5.11): $\sum_{i\in[d]} s_i y_i = \lambda(n) > 0$, it is necessary that $y_i > 0$ for at least one $i \in [d]$. Then, the condition (5.10): $\theta_i y_i = 0$ requires the corresponding Lagrange multiplier θ_i to be set to zero.

Let $\theta_i = 0$ for $K \geq 1$ distinct indices of $i \in \{i_1, i_2, \ldots, i_K\} \subseteq [d]$ with $i_1 < i_2 < \ldots < i_K$, and $\theta_i > 0$ for all other indices. Equation (5.8) gives rise to the relationship below.

$$1 + \nu s_{i_k} + \theta_0 i_k = \theta_{i_k} = 0, \quad \forall k \in [K]. \tag{5.12}$$

This results in

$$-\frac{\theta_0}{\nu} = \frac{s_{i_{k_2}} - s_{i_{k_1}}}{i_{k_2} - i_{k_1}}, \quad \forall k_1 < k_2,\ k_1, k_2 \in [K]. \tag{5.13}$$

We show that all indices $\{i_1, i_2, \dots, i_K\}$ are consecutive. Assume it is not true, and $i_k + 1 < i_{k+1}$ for some $k \in [K]$. Then, (i_k, i_{k+1}) is a non-empty set and for any $i \in (i_k, i_{k+1})$, we must have $\theta_i > 0$. Consequently, Equation (5.8) for i and i_k yields

$$1 + \nu s_i + \theta_0 i = \theta_i > 0, \quad \forall i \in (i_k, i_{k+1}) \tag{5.14}$$

$$1 + \nu s_{i_k} + \theta_0 i_k = \theta_{i_k} = 0. \tag{5.15}$$

The above equations result in

$$-\frac{\theta_0}{\nu} > \frac{s_i - s_{i_k}}{i - i_k}, \quad \forall i \in (i_k, i_{k+1}). \tag{5.16}$$

Combined with Equation (5.13), we get

$$\frac{s_{i_{k+1}} - s_{i_k}}{i_{k+1} - i_k} > \frac{s_i - s_{i_k}}{i - i_k}, \quad \forall i \in (i_k, i_{k+1}), \tag{5.17}$$

which is equivalent to

$$s_i < \left(1 - \frac{i - i_k}{i_{k+1} - i_k}\right) s_{i_k} + \frac{i - i_k}{i_{k+1} - i_k} s_{i_{k+1}}, \quad \forall i \in (i_k, i_{k+1}). \tag{5.18}$$

Note that $\frac{i-i_k}{i_{k+1}-i_k} \in (0,1)$ for any i in (i_k, i_{k+1}). Thus, the above inequality violates the concavity of the speedup function and is not feasible. Therefore, $i_k + 1 < i_{k+1}$ cannot happen for any $k \in [K]$, and we have

$$i_{k+1} = i_k + 1, \quad \forall k \in [K-1]. \tag{5.19}$$

From Equation (5.13), we conclude that

$$s_{i_{k+1}} - s_{i_k} = -\frac{\theta_0}{\nu} = \Delta, \quad \forall k \in [K-1], \tag{5.20}$$

where Δ is a positive constant independent of i and k. We consider two cases: one where $\theta_i = 0$ for at least two distinct indices of i, i.e., $K \geq 2$, and the other where $\theta_i = 0$ for a single index i, i.e., $K = 1$.

1. $\theta_i = 0$ **for at least two distinct indices of** $i \in [d]$**:**

   If $\theta_i = 0$ for more than one $i \in [d]$, then θ_0 must be non-zero. Otherwise, Equation (5.8) leads to $\nu = -1/s_i$ for those $i \in [d]$ where $\theta_i = 0$, contradicting the strictly increasing property of s_i. Thus, $\theta_0 > 0$, and Equation (5.10): $\theta_0 \left(\sum_{i\in[d]} iy_i - 1\right) = 0$, implies $\sum_{i\in[d]} iy_i = 1$. Since y_i is non-zero only for $i \in \{i_1, i_2, \ldots, i_K\}$, this leads to a system of equations

$$\sum_{k=1}^{K} s_{i_k} y_{i_k} = \lambda(n), \tag{5.21}$$

$$\sum_{k=1}^{K} i_k y_{i_k} = 1. \tag{5.22}$$

   However, from Equations (5.19) and (5.20), we have $i_k = i_1 + (k-1)$ and $s_{i_k} = s_{i_1} + (k-1)\Delta$ for any $k \in [K]$. As a result, we can rewrite the above system as

$$s_{i_1} \sum_{k=1}^{K} y_{i_k} + \Delta \sum_{k=2}^{K} (k-1) y_{i_k} = \lambda(n), \tag{5.23}$$

$$i_1 \sum_{k=1}^{K} y_{i_k} + \sum_{k=2}^{K} (k-1) y_{i_k} = 1. \tag{5.24}$$

   Solving these equations results in

$$(s_{i_1} - \Delta i_1) \sum_{k=1}^{K} y_{i_k} = \lambda(n) - \Delta, \tag{5.25}$$

$$\sum_{k=1}^{K} (k-1) y_{i_k} = 1 - i_1 \sum_{k=1}^{K} y_{i_k}. \tag{5.26}$$

   We consider two sub-cases: one where $\Delta = \lambda(n)$ and another where $\Delta \neq \lambda(n)$.

- Sub-Case 1, $\Delta = \lambda(n)$: In this sub-case, Equation (5.25) implies that

  $$(s_{i_1} - \Delta i_1) \sum_{k=1}^{K} y_{i_k} = 0.$$

  Since $y_i > 0$ for all $i \in \{i_1, \ldots, i_K\}$, it follows that $s_{i_1} - \Delta i_1 = 0$. However, from Equation (5.12) and the definition of Δ in Equation (5.20), we derive

  $$s_{i_K} - \Delta i_K = s_{i_{K-1}} - \Delta i_{K-1} = \ldots = s_{i_1} - \Delta i_1 = 0,$$

  which leads to $\frac{s_{i_k}}{i_k} = \Delta = \lambda(n)$, for all $k \in [K]$. Substituting into Equation (5.12) leads to

  $$\nu\lambda(n) + \theta_0 = -\frac{1}{i_k}, \quad \forall k \in [K].$$

  This contradicts the possibility of having multiple values of i_k that satisfy the above equation, rendering this sub-case infeasible.

- Sub-Case 2, $\Delta \neq \lambda(n)$: In this sub-case, Equations (5.25)-(5.26) imply that

  $$\sum_{k=1}^{K} y_{i_k} = \frac{\lambda(n) - \Delta}{s_{i_1} - \Delta i_1}, \tag{5.27}$$

  $$\sum_{k=2}^{K} (k-1) y_{i_k} = 1 - i_1 \sum_{k=1}^{K} y_{i_k} = \frac{s_{i_1} - \lambda(n) i_1}{s_{i_1} - \Delta i_1}. \tag{5.28}$$

  Since $\sum_{k=2}^{K}(k-1)y_{i_k} = 1 - i_1 \sum_{k=1}^{K} y_{i_k} < 1$, and from Equation (5.28), we conclude that $\lambda(n) > \Delta$. This, alongside the condition that $\sum_{k=1}^{K} y_{i_k} > 0$ and Equation (5.27), imply that $s_{i_1} - \Delta i_1 > 0$. Additionally, we have

  $$s_{i_K} - \Delta i_K = s_{i_{K-1}} - \Delta i_{K-1} = \ldots = s_{i_1} - \Delta i_1 > 0,$$

  and from the definition of Δ in (5.20), we get

  $$\frac{s_{i_k}}{i_k} > \Delta = s_{i_k+1} - s_{i_k}, \quad \forall k \in [K-1]. \tag{5.29}$$

This condition leads to

$$\frac{s_{i_1}}{i_1} > \frac{s_{i_2}}{i_2} > \ldots > \frac{s_{i_K}}{i_K}. \tag{5.30}$$

Under the above condition and considering $\frac{s_{i_K}}{i_K} \leq \lambda(n) \leq \frac{s_{i_1}}{i_1}$, the objective function simplifies to $\sum_{k=1}^{K} y_{i_k} = \frac{\lambda(n)-\Delta}{s_{i_1}-\Delta i_1}$. A feasible solution to achieve this objective function when $\frac{s_{i_{k+1}}}{i_{k+1}} \leq \lambda(n) \leq \frac{s_{i_k}}{i_k}$ is given by

$$y_{i_k} = \frac{\frac{1}{i_k}\left(\lambda(n) - \frac{s_{i_{k+1}}}{i_{k+1}}\right)}{\frac{s_{i_k}}{i_k} - \frac{s_{i_{k+1}}}{i_{k+1}}}, \quad y_{i_{k+1}} = \frac{\frac{1}{i_{k+1}}\left(\frac{s_{i_k}}{i_k} - \lambda(n)\right)}{\frac{s_{i_k}}{i_k} - \frac{s_{i_{k+1}}}{i_{k+1}}}, \quad y_i = 0, \forall i \notin \{i_k, i_{k+1}\}.$$

From Equation (5.13), it is seen that when $s_{i_k} - s_{i_{k-1}} > s_{i_{k+1}} - s_{i_k} > s_{i_{k+2}} - s_{i_{k+1}}$, then $k = 2$ is the only feasible option and the above solution is unique.

2. **$\theta_i = 0$ for exactly one index i:**
Let i_1 be the index for which $\theta_i = 0$. Therefore, $\theta_i > 0$ for all $i \neq i_1$, and the condition (5.10): $\theta_i y_i = 0$ implies $y_i = 0$ for all $i \neq i_1$. As a result, from primal feasibility constraint (5.11): $\sum_{i\in[d]} s_i y_i = \lambda(n)$, we get

$$y_{i_1} = \frac{\lambda(n)}{s_{i_1}}, \quad y_i = 0, \forall i \neq i_1.$$

Two sub-cases emerge based on whether $\lambda(n) < s_{i_1}/i_1$ or $\lambda(n) = s_{i_1}/i_1$.

- Sub-Case 1, $\lambda(n) < s_{i_1}/i_1$: In this sub-case, the condition (5.11): $\sum_{i\in[d]} i y_i = i_1 y_{i_1} < 1$ requires $\theta_0 = 0$. Consequently, from Equation (5.8) we obtain

$$\begin{aligned} 1 + \nu s_{i_1} &= \theta_{i_1} = 0 \\ 1 + \nu s_i &= \theta_i > 0, \quad \forall i \neq i_1. \end{aligned}$$

  The above inequalities imply $s_i < s_{i_1}$ for all $i \neq i_1$. Given the strictly increasing property of s_i, it follows that $i_1 = d$ and $y^* = (0, \ldots, 0, \lambda(n)/s_d)$.

- Sub-Case 2, $\lambda(n) = s_{i_1}/i_1$: In this sub-case, the only non-zero component y_{i_1} is given by

$$y_{i_1} = \frac{\lambda(n)}{s_{i_1}} = \frac{1}{i_1}.$$

With this value, the objective function becomes $1/i_1$ and is minimized at $i = \max\{i_1 : \lambda(n) = s_{i_1}/i_1\} = \max\{i_1 : s_{i_1} = \lambda(n)i_1\}$. Hence,

$$y^* = (0, \ldots, 0, \lambda(n)/s_i, 0, \ldots, 0)\,,$$

where i corresponds to the given maximum index.

This completes the analysis of all possible cases and sub-cases. □

Based on Proposition 5.1.1, it is evident that the optimal solution y^* is characterized by having at most two nonzero components. We introduce the set $I^* = \{i \in [d] : y_i^* > 0\}$, representing the indices of these nonzero components. Therefore, $|I^*|= 1$, or 2. A job assignment scheme aligning the system's state x with the conditions $\mathbb{E}[x_i] = y_i^*$ for all $i \in I^*$ and $\mathbb{E}[x_i] = 0$ for all other indices replicates the optimal behavior and leads to asymptotic optimality in the steady state. We proceed to present such a scheme.

Probabilistic Assignment Scheme: In the probabilistic assignment scheme, when an incoming job arrives, it is divided into $i \in [d]$ sub-jobs with probabilities p_i. These sub-jobs are assigned to i available servers, provided there are sufficient free servers within the system. In the event of an insufficient number of available servers, the job becomes blocked. The probabilities p_i are determined as follows.

$$p_i = \frac{s_i y_i^*}{\lambda(n)}, \quad i \in [d]. \tag{5.31}$$

Note that $\sum_{i\in[d]} p_i = \sum_{i\in I^*} p_i = 1$ due to the first constraint of the optimization problem. Moreover, under the proposed assignment scheme, the probabilities $A_i(x)$ as the probability with which an incoming job is processed at $i \in [d]$ servers when the system is in state x is given by

$$A_i(x) = \begin{cases} p_i \mathbb{1}\left(nq_0 \geq i\right), & i \in I^*, \\ 0, & i \in [d] - I^*. \end{cases} \tag{5.32}$$

Remark 5.1. An alternative approach to job assignment involves simplifying the system's state to a single dimension. An example of such a scheme is the I-greedy assignment scheme, where $I := \max\{i \in [d] : \lambda(n) \leq s_i/i\}$. Under this scheme, all available servers are used up to a maximum of I servers. In cases where no free server is available, a job is subjected to blocking. The selection of the parameter I depends on both the normalized arrival rate of the system and the specific characteristics of the speedup function. This choice is essentially grounded in the optimality criterion established in Proposition 5.1.1 such that a zero blocking probability is required. The analytical approach to this policy follows a similar methodology to that of the greedy job assignment scheme, with necessary adjustments required to accommodate the introduction of the new threshold level I and the nonlinearity associated with the speedup function. Under this policy a zero blocking probability and an average response time of $\frac{1}{s_I}$ are achieved, with the analysis details left to the astute reader. The discussion provided here serves as supplementary information and in the subsequent sections of this thesis, we will place our emphasis on the probabilistic routing scheme.

Remark 5.2. Proposition 5.1.1 provides a key result regarding job parallelization under heavy traffic conditions. In this context, the normalized arrival rate is represented as $\lambda(n) = 1 - \beta n^{-\alpha}$, where $\alpha > 0$ and $\beta > 0$. This implies that $1 = \lim_{n\to\infty} \lambda(n) = s_{i_1}/i_1$, where $i_1 = \max\{i \in [d], s_i/i = 1\}$. If $i_1 > 1$, it is implied that $s_i = i$ for all $i \in [i_1]$, leading to a system with linear speedup functions. Therefore, in the context of sublinear functions within the asymptotic regime and under heavy traffic conditions, we conclude that $i_1 = 1$, and job parallelization fails to improve system performance. Consequently, the optimal strategy in these scenarios involves allocating all incoming jobs to individual servers to get a zero blocking probability in the system.

5.2 Comparison with a Fluid Limit: Stein's Approach

In order to analyze the system's performance under the probabilistic job assignment scheme, we introduce a simple dynamical system and characterize the difference between the generator of the original Markov process $x(\cdot)$ and the generator of this dynamical system.

The Fluid Limit: The simple dynamical system that we want the original steady-state system to converge to its fixed point is described by a set of ODEs defined below.

$$\dot{z}_i = s_i y_i^* - s_i z_i, \quad i \in [d], \tag{5.33}$$

where y^* is the solution to the optimization problem in Proposition 5.1.1. It is important to note that the fixed point of this system corresponds to y^*.

In the following lemma, we evaluate the accuracy of representing the system state x by the fixed point of the fluid model (5.33). Based on the assignment probabilities described in (5.32), it is evident that the arrival rate of jobs being split into $i \in [d] - I^*$ sub-jobs is zero. Given that the system is in a stationary state, through sample path arguments, we conclude that these jobs will eventually vanish in the stationary state, leaving only the jobs occupying $i \in I^*$ sub-jobs. Consequently, the state space collapses to $|I^*|$ dimensions and it is sufficient to study x_i only for $i \in I^*$. More specifically, the objective is to show that as n approaches infinity, $\mathbb{E}\left[x_i\right]$ converges in probability to $\lim_{n\to\infty} y_i^*$ for $i \in I^*$.

Lemma 5.2.1. *Under the equilibrium measure of the system, we have*

$$\mathbb{E}\left[\sum_{i\in I^*} c_i s_i \left(x_i - y_i^*\right)^2\right] = \frac{\lambda(n)}{n}\sum_{i\in I^*} c_i \mathbb{E}\left[A_i(x)\right] + \sum_{i\in I^*}\mathbb{E}\left[c_i s_i y_i^* \mathbb{1}\left(q_1 > 1 - \frac{i}{n}\right)\left(y_i^* - x_i\right)\right], \tag{5.34}$$

where $c_i > 0$ for $i \in I^*$.

Proof. We introduce a Lyapunov function $V(x) = \sum_{i\in I^*} \frac{c_i}{2}\left(x_i - y_i^*\right)^2$, where c_i is a strictly positive constant for every $i \in I^*$. We compare the drift of this function under G, the generator of the Markov chain $x(\cdot)$, to that under L, the generator of the system of ODEs given by (5.33). We note that under L the drift of V is given by

$$LV(x) = \sum_{i\in I^*}\frac{\partial V}{\partial x_i}(x)\dot{x}_i = \sum_{i\in I^*} c_i\left(x_i - y_i^*\right) s_i\left(y_i^* - x_i\right) = -\sum_{i\in I^*} c_i s_i\left(x_i - y_i^*\right)^2. \tag{5.35}$$

Following similar arguments as in the proof of Lemma 4.3.1 for the generator G, we have

$$GV(x) = \sum_{i\in I^*} n\lambda(n)A_i(x)\left(V\left(x + \frac{1}{n}e_i\right) - V(x)\right) + ns_i x_i\left(V\left(x - \frac{1}{n}e_i\right) - V(x)\right), \tag{5.36}$$

where e_i denotes the d-dimensional unit vector with a value of one at the i^{th} position. In the steady state, the expectation of the drift of any suitable function V under G is zero, which can be expressed as

$$\mathbb{E}\left[GV(x)\right] = 0. \tag{5.37}$$

From Equation (5.35), we can rewrite the above equation as

$$\mathbb{E}\left[GV(x) - LV(x)\right] = \mathbb{E}\left[-LV(x)\right] = \mathbb{E}\left[\sum_{i\in I^*} c_i s_i\left(x_i - y_i^*\right)^2\right]. \tag{5.38}$$

The above equation provides a means to establish bounds on $\mathbb{E}\left[\sum_{i\in I^*} c_i s_i \left(x_i - y_i^*\right)^2\right]$ by comparing the drift of the function V under G to that under L. Using the Taylor series expansion of V and combining Equation (5.35) and (5.36), we have

$$\begin{aligned}\mathbb{E}\left[GV(x) - LV(x)\right] = &\sum_{i\in I^*} \mathbb{E}\left[n\lambda(n)A_i(x)\left(\frac{1}{n}\frac{\partial V}{\partial x_i}(x) + \frac{1}{2n^2}\frac{\partial^2 V}{\partial x_i^2}(\xi)\right)\right] \\ &+ \sum_{i\in I^*} \mathbb{E}\left[ns_i x_i\left(-\frac{1}{n}\frac{\partial V}{\partial x_i}(x) + \frac{1}{2n^2}\frac{\partial^2 V}{\partial x_i^2}(\theta)\right) - \frac{\partial V}{\partial x_i}(x)\dot{x}_i\right], \qquad (5.39)\end{aligned}$$

where ξ and θ are d-dimensional vectors. Simplifying the RHS of the above and using the fact that $\frac{\partial^2 V}{\partial x_i^2}(y) = c_i$ for any vector y, we get

$$\begin{aligned}\mathbb{E}\left[GV(x) - LV(x)\right] = &\sum_{i\in I^*} \mathbb{E}\left[\left(\lambda(n)A_i(x) - s_i x_i - \dot{x}_i\right)\frac{\partial V}{\partial x_i}(x)\right] \\ &+ \frac{1}{2n}\sum_{i\in I^*} c_i \mathbb{E}\left[\lambda(n)A_i(x) + s_i x_i\right]. \qquad (5.40)\end{aligned}$$

From Equation (5.33), and using the form of $A_i(x) = p_i \mathbb{1}\left(nq_0 \geq i\right) = \frac{s_i y_i^*}{\lambda(n)}\mathbb{1}\left(nq_0 \geq i\right)$, we can rewrite the above as

$$\begin{aligned}\mathbb{E}\left[GV(x) - LV(x)\right] = &\sum_{i\in I^*} \mathbb{E}\left[s_i y_i^*\left(\mathbb{1}\left(nq_0 \geq i\right) - 1\right)\frac{\partial V}{\partial x_i}(x)\right] \\ &+ \frac{1}{2n}\sum_{i\in I^*} c_i \mathbb{E}\left[\lambda(n)A_i(x) + s_i x_i\right]. \qquad (5.41)\end{aligned}$$

Finally, from $\frac{\partial V}{\partial x_i}(x) = c_i\left(x_i - y_i^*\right)$, and the equality $\mathbb{E}\left[s_i x_i\right] = \lambda(n)\mathbb{E}\left[A_i(x)\right]$, we arrive at the desired result. This last equality stems from the system being in the stationary regime, where the arrival rate and departure rate of each class of jobs that occupy i servers simultaneously should be equal on average for each value of $i \in [d]$. This completes the proof. □

To establish the convergence of x_i to y_i^* for $i \in I^*$, it is sufficient to show that the second term on the right-hand side of Equation (5.34) converges to zero as $n \to \infty$. In the following sections, we provide a detailed analysis of this result for both scenarios when $|I^*|= 1$ and when $|I^*|= 2$.

5.2.1 State Space Collapse to One Dimension

Here, we study the system when $|I^*|= 1$, implying that the optimal solution y^* comprises only one nonzero component. According to Proposition 5.1.1, this scenario occurs when the arrival rate satisfies $\lambda(n) \leq \frac{s_d}{d}$ or $\lambda(n) = \frac{s_i}{i}$ for a specific $i \in [d]$. We show that in this particular setting, the system achieves asymptotic optimality. More precisely, it achieves both a blocking probability of zero and the minimum achievable average response time for accepted jobs as n approaches infinity. Additionally, we provide upper bounds on system performance for finite system sizes.

Lemma 5.2.2. *Let $I^* = \{i_1\}$ for some $i_1 \in [d]$. Then, under the equilibrium measure of the system we have*

$$\sum_{i\in I^*} \mathbb{E}\left[i\mathbb{1}\left(q_1 > 1 - \frac{i}{n}\right)(y_i^* - x_i)\right] \leq \frac{i_1}{n}. \tag{5.42}$$

Proof. Noting that $I^* = \{i_1\}$, we arrive at the following equation.

$$\sum_{i\in I^*} \mathbb{E}\left[i\mathbb{1}\left(q_1 > 1 - \frac{i}{n}\right)(y_i^* - x_i)\right] = \mathbb{E}\left[i_1\mathbb{1}\left(q_1 > 1 - \frac{i_1}{n}\right)\left(y_{i_1}^* - x_{i_1}\right)\right]. \tag{5.43}$$

Additionally, when $I^* = \{i_1\}$, as discussed earlier, it follows that the probabilities $p_i(x)$ are zero for all $i \neq i_1$, while $p_{i_1} = 1$. Therefore, by sample path arguments, we conclude that in the stationary regime, $x_i = 0$ for $i \neq i_1$. This implies $q_1 = \sum_{i\in[d]} ix_i = i_1x_{i_1}$. Substituting this into Equation (5.43), we obtain

$$\begin{aligned}
\sum_{i\in I^*} \mathbb{E}\left[i\mathbb{1}\left(q_1 > 1 - \frac{i}{n}\right)(y_i^* - x_i)\right] &= \mathbb{E}\left[\mathbb{1}\left(q_1 > 1 - \frac{i_1}{n}\right)\left(i_1y_{i_1}^* - q_1\right)\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left(q_1 > 1 - \frac{i_1}{n}\right)(1 - q_1)\right] \\
&\leq \frac{i_1}{n},
\end{aligned} \tag{5.44}$$

where the second line follows from the constraints of the optimization problem, which specify that $\sum_{i\in[d]} iy_i^* \leq 1$. This implies $i_1y_{i_1}^* \leq 1$. The last line is a consequence of the indicator function. □

Combining Lemmas 5.2.1 and 5.2.2, we get the following corollary.

Corollary 5.2.3. *Let $I^* = \{i_1\}$ for some $i_1 \in [d]$. Under the equilibrium measure of the system we have*

$$\mathbb{E}\left[\sum_{i \in I^*} \frac{i}{y_i^*} (x_i - y_i^*)^2\right] \leq 2\frac{i_1}{n}. \tag{5.45}$$

Proof. The proof is straightforward by selecting $c_i = \frac{i}{s_i y_i^*}$ for $i \in I^*$ in Lemma 5.2.1, and noting that $y_{i_1}^* = \frac{\lambda(n)}{s_{i_1}}$ when $I^* = \{i_1\}$. □

In the following theorem, we show that the system achieves asymptotic optimality in terms of the average response time of accepted jobs and the blocking probability. Additionally, we present upper bounds on the system's performance for finite-sized configurations.

Theorem 5.2.4. *Let $I^* = \{i_1\}$ for some $i_1 \in [d]$. Then, the steady-state blocking probability of the system converges to zero with an error bound of $O(\frac{1}{\sqrt{n}})$, while the steady-state mean response time of accepted jobs remains at $\frac{1}{s_{i_1}}$ for sufficiently large values of n.*

Proof. The proof follows the same reasoning as the proof of Theorem 4.3.4. The detailed analysis is given in Appendix A.4. □

5.2.2 State Space Collapse to Two Dimensions

We now consider the scenario where the normalized arrival rate $\lambda(n)$ satisfies $\lambda(n) \in \left(\frac{s_{i_1+1}}{i_1+1}, \frac{s_{i_1}}{i_1}\right)$ for some $i_1 \in [d-1]$. Consequently, as per Proposition 5.1.1, we conclude that the optimal solution y^* consists of two nonzero components, and $I^* = \{i_1, i_1+1\}$. We present similar asymptotic optimality results as in the previous section.

Lemma 5.2.5. *Let $I^* = \{i_1, i_2\}$, where $i_2 = i_1 + 1$ and $i_1 \in [d-1]$. Under the equilibrium measure of the system, we have*

$$\sum_{i \in I^*} \mathbb{E}\left[i\mathbb{1}\left(q_1 > 1 - \frac{i}{n}\right)(y_i^* - x_i)\right] \leq \frac{i_1}{n}$$

$$+\mathbb{E}\left[\mathbb{1}\left(q_1 = 1-\frac{i_1}{n}, r_1 \le \lambda(n)+\delta\right)\left(i_2 y_{i_2}^* - i_2 x_{i_2}\right)\right]$$
$$+ i_2\mathbb{E}\left[\mathbb{1}\left(r_1 > \lambda(n)+\delta\right)\left(x_{i_1}+x_{i_2}\right)\right], \tag{5.46}$$

where $r_1 = \sum_{i\in I^*} s_i x_i$ denotes the departure rate of the system, $\delta \in (0, 1-\lambda(n))$ and n is sufficiently large.

Proof. Since $I^* = \{i_1, i_2\}$, we have

$$\begin{aligned}\sum_{i\in I^*}\mathbb{E}\left[i\mathbb{1}\left(q_1 > 1-\frac{i}{n}\right)\left(y_i^* - x_i\right)\right] &= \mathbb{E}\left[i_1\mathbb{1}\left(q_1 > 1-\frac{i_1}{n}\right)\left(y_{i_1}^* - x_{i_1}\right)\right] \\ &\quad + \mathbb{E}\left[i_2\mathbb{1}\left(q_1 > 1-\frac{i_2}{n}\right)\left(y_{i_2}^* - x_{i_2}\right)\right] \\ &= \mathbb{E}\left[\mathbb{1}\left(q_1 > 1-\frac{i_1}{n}\right)\left(i_1 y_{i_1}^* + i_2 y_{i_2}^* - i_1 x_{i_1} - i_2 x_{i_2}\right)\right] \qquad (5.47) \\ &\quad + \mathbb{E}\left[\mathbb{1}\left(1-\frac{i_2}{n} < q_1 \le 1-\frac{i_1}{n}\right)\left(i_2 y_{i_2}^* - i_2 x_{i_2}\right)\right] \qquad (5.48)\end{aligned}$$

We consider each of the terms (5.47) and (5.48) independently and establish suitable bounds for each.

Let us start with expression (5.47). Considering the optimal solution y^*, it is clear that $i_1 y_{i_1}^* + i_2 y_{i_2}^* = 1$. Furthermore, through sample path analysis, we can deduce that $x_i = 0$ for all $i \notin I^*$ in the stationary regime. Therefore, we can simplify the expression for q_1, reducing it to $q_1 = i_1 x_{i_1} + i_2 x_{i_2}$. Hence,

$$\mathbb{E}\left[\mathbb{1}\left(q_1 > 1-\frac{i_1}{n}\right)\left(i_1 y_{i_1}^* + i_2 y_{i_2}^* - i_1 x_{i_1} - i_2 x_{i_2}\right)\right] = \mathbb{E}\left[\mathbb{1}\left(q_1 > 1-\frac{i_1}{n}\right)\left(1-q_1\right)\right] \le \frac{i_1}{n}, \tag{5.49}$$

where the inequality follows from the indicator function.

Now, let us examine expression (5.48). According to the lemma statement, we have $i_2 = i_1 + 1$. Therefore, the indicator function $1 - \frac{i_1+1}{n} < q_1 \le 1 - \frac{i_1}{n}$ ensures that $q_1 = 1 - \frac{i_1}{n}$. Consequently,

$$\mathbb{E}\left[\mathbb{1}\left(1-\frac{i_2}{n}<q_1\leq 1-\frac{i_1}{n}\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right]=\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n}\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right]. \tag{5.50}$$

We analyze the term (5.50) under two distinct conditions related to the departure rate of the system: when it is either less than the arrival rate or greater than it. To facilitate this analysis, we define r_1 as the departure rate of the system, given by $r_1=\sum_{i\in I^*}s_ix_i$. For any $\delta\in(0,1-\lambda(n))$, we can write

$$\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n}\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right]=\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1\leq\lambda(n)+\delta\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right] \tag{5.51}$$

$$+\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1>\lambda(n)+\delta\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right] \tag{5.52}$$

As per Proposition 5.1.1, the optimal solution $y^*_{i_2}$ is determined by the following.

$$y^*_{i_2}=\frac{\frac{1}{i_2}\left(\frac{s_{i_1}}{i_1}-\lambda(n)\right)}{\frac{s_{i_1}}{i_1}-\frac{s_{i_2}}{i_2}}. \tag{5.53}$$

The set $I^*=\{i_1,i_2\}$ is only applicable when $\frac{s_{i_2}}{i_2}<\lambda(n)<\frac{s_{i_1}}{i_1}$. Consequently, we can infer that $y^*_{i_2}<\frac{1}{i_2}$ or equivalently $i_2y^*_{i_2}<1$. On the other hand, the fraction of busy servers is characterized by $q_1=1-\frac{i_1}{n}$. Therefore, by choosing a n sufficiently large, we ensure that $i_2y^*_{i_2}\leq q_1$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1>\lambda(n)+\delta\right)\left(i_2y^*_{i_2}-i_2x_{i_2}\right)\right]\\
&\leq\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1>\lambda(n)+\delta\right)\left(q_1-i_2x_{i_2}\right)\right]\\
&\leq\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1>\lambda(n)+\delta\right)q_1\right]\\
&\leq i_2\mathbb{E}\left[\mathbb{1}\left(q_1=1-\frac{i_1}{n},r_1>\lambda(n)+\delta\right)\sum_{i\in I^*}x_i\right]
\end{aligned}$$

$$\leq i_2\mathbb{E}\left[\mathbb{1}\left(r_1 > \lambda(n) + \delta\right)\sum_{i\in I^*} x_i\right], \tag{5.54}$$

where the second line follows from the inequality $i_2 y^*_{i_2} \leq q_1$ and the fourth line follows directly from the definition of q_1, which is given as $q_1 = \sum_{i\in I^*} ix_i$.

Combining (5.49), (5.51) and (5.54) we get the desired result. □

In the following lemmas, we further bound the terms established in Lemma 5.2.5.

Lemma 5.2.6. *Let $I^* = \{i_1, i_2\}$, where $i_2 = i_1 + 1$ and $i_1 \in [d-1]$. Under the equilibrium measure of the system, we have*

$$\mathbb{E}\left[\mathbb{1}\left(q_1 = 1 - \frac{i_1}{n}, r_1 \leq \lambda(n) + \delta\right)\left(i_2 y^*_{i_2} - i_2 x_{i_2}\right)\right] \leq \frac{s_{i_1}}{n\left(\frac{s_{i_1}}{i_1} - \frac{s_{i_2}}{i_2}\right)} + \frac{\delta}{\frac{s_{i_1}}{i_1} - \frac{s_{i_2}}{i_2}}, \tag{5.55}$$

where $r_1 = \sum_{i\in I^*} s_i x_i$ denotes the departure rate of the system and $\delta \in (0, 1-\lambda(n))$.

Proof. Given that $I^* = \{i_1, i_2\}$ and the system is in the stationary regime, we can employ similar reasoning as in previous lemmas to determine that the fraction of busy servers is defined as $q_1 = i_1 x_{i_1} + i_2 x_{i_2}$. Consequently, from the indicator function, we can express the following system of equations.

$$i_1 x_{i_1} + i_2 x_{i_2} = 1 - \frac{i_1}{n}. \tag{5.56}$$

$$s_{i_1} x_{i_1} + s_{i_2} x_{i_2} \leq \lambda(n) + \delta. \tag{5.57}$$

By solving the equation for x_{i_1} and substituting in the inequality, we obtain

$$x_{i_1} = \frac{1}{i_1} - \frac{1}{n} - \frac{i_2}{i_1} x_{i_2}, \tag{5.58}$$

$$\frac{s_{i_1}}{i_1} - \frac{s_{i_1}}{n} - \frac{s_{i_1} i_2}{i_1} x_{i_2} + s_{i_2} x_{i_2} \leq \lambda(n) + \delta. \tag{5.59}$$

We rewrite inequality (5.59) as

$$\left(\frac{s_{i_1}}{i_1} - \lambda(n)\right) - i_2 x_{i_2}\left(\frac{s_{i_1}}{i_1} - \frac{s_{i_2}}{i_2}\right) \leq \frac{s_{i_1}}{n} + \delta. \tag{5.60}$$

Noting that $y_{i_2}^* = \frac{\frac{1}{i_2}\left(\frac{s_{i_1}}{i_1}-\lambda(n)\right)}{\frac{s_{i_1}}{i_1}-\frac{s_{i_2}}{i_2}}$, the result follows.

□

In the following lemma, we introduce a new Lyapunov function, denoted as $V_2(x) = \mathbb{1}\left(\sum_{i\in I^*} s_i x_i > \lambda(n)+\delta\right)\sum_{i\in I^*} x_i$, where δ is a positive constant. This function provides insights into scenarios where the system's departure rate is sufficiently high, resulting in most servers being occupied, and thus, limiting the acceptance of new jobs. More precisely, we show that outside of a suitable compact set, the drift of this Lyapunov function is negative. This observation implies that, with high probability, the function $V_2(x)$ remains in that compact set.

Lemma 5.2.7. *Let $\lambda(n) = 1-\beta/n^\alpha \geq 0$ for $\alpha \in [0,1)$ and $\beta > 0$. For any $\delta \in (0, \beta/n^\alpha)$, we define the following Lyapunov function.*

$$V_2(x) = \mathbb{1}\left(\sum_{i\in I^*} s_i x_i > \lambda(n)+\delta\right)\sum_{i\in I^*} x_i. \tag{5.61}$$

If $V_2(x) \geq \kappa$ for some $\kappa > 0$, then $GV_2(x) \leq -\delta$, and furthermore, $\mathbb{E}\left[V_2(x)\right] \leq \kappa + \frac{2}{n\delta}$, for all n sufficiently large.

Proof. The proof follows from the same arguments as in the proof of Lemma 4.3.7 and hence is omitted here. □

Through the combination of Lemma 5.2.1 and Lemmas 5.2.5- 5.2.7, we derive the following.

Corollary 5.2.8. *Let s_i be a strictly increasing, concave function of i satisfying $1 = \frac{s_1}{1} \geq \frac{s_2}{2} \geq \ldots \geq \frac{s_d}{d}$, and assume $\lambda(n) = 1-\beta/n^\alpha \geq 0$ for some $\alpha \in [0,1)$ and $\beta > 0$. If $\lambda(n) \in \left(\frac{s_{i_1+1}}{i_1+1}, \frac{s_{i_1}}{i_1}\right)$ for some $i_1 \in [d-1]$, then under the equilibrium measure of the system we have*

$$\mathbb{E}\left[\sum_{i\in I^*} \frac{i}{y_i^*}\left(x_i - y_i^*\right)^2\right] \leq \frac{c}{n} + \frac{\beta}{2\left(\frac{s_{i_1}}{i_1} - \frac{s_{i_1+1}}{i_1+1}\right)} n^{-(\alpha+\varepsilon)} + \frac{4(i_1+1)}{\beta} n^{\varepsilon-(1-\alpha)}, \tag{5.62}$$

where $c = \lambda(n)\sum_{i\in I^*}\frac{i}{s_i y_i^*} + 2i_1 + 1 + \frac{s_{i_1}}{\frac{s_{i_1}}{i_1}-\frac{s_{i_1+1}}{i_1+1}}$, $\varepsilon \in (0, 1-\alpha)$ and n is sufficiently large.

Proof. The result follows by choosing $c_i = \frac{i}{s_i y_i^*}$ for $i \in I^*$ in Lemma 5.2.1, $\delta = \frac{\beta}{2n^{\alpha+\varepsilon}}$ for some $\varepsilon \in (0, 1-\alpha)$ and $\kappa = \frac{1}{n}$ in Lemmas 5.2.6 and 5.2.7, and combining with Lemma 5.2.5. □

In the following theorem, we establish the asymptotic optimality of the system.

Theorem 5.2.9. *Let s_i be a strictly increasing, concave function of i satisfying $1 = \frac{s_1}{1} \geq \frac{s_2}{2} \geq \ldots \geq \frac{s_d}{d}$. Assume that $\lambda(n) = 1 - \beta/n^\alpha \geq 0$ for $\alpha \in [0, 1)$ and $\beta > 0$. If $\lambda(n)$ lies in the interval $\left(\frac{s_{i_1+1}}{i_1+1}, \frac{s_{i_1}}{i_1}\right)$ for some $i_1 \in [d-1]$, then the steady-state blocking probability of the system, denoted by P_b, converges to zero as $n \to \infty$. Additionally, the steady-state mean response time of jobs, denoted by $\mathbb{E}[D]$, converges to its minimum possible value as $n \to \infty$. In other words,*

$$\lim_{n\to\infty} P_b = 0, \tag{5.63}$$

$$\lim_{n\to\infty} \mathbb{E}[D] = \frac{y_{i_1}^* + y_{i_1+1}^*}{\sigma} = \frac{1 - \frac{s_{i_1+1} - s_{i_1}}{\sigma}}{s_{i_1} - i_1 \left(s_{i_1+1} - s_{i_1}\right)}, \tag{5.64}$$

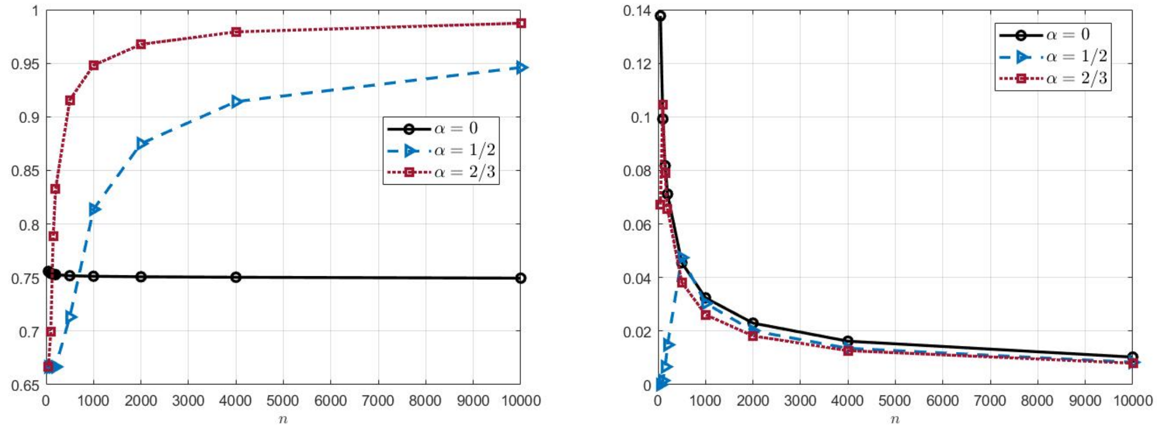
where $\sigma = \lim_{n\to\infty} \lambda(n)$.

Proof. The proof follows from Corollary 5.2.8, considering the rate conservation principle and Little's law. □

5.3 Numerical Results

In this section, we present the results of our simulations conducted under the conditions of a strictly increasing, sublinear, and concave speedup function. The simulations were repeated 100 times, each time simulating the arrival of the first 5×10^6 jobs to the system. The system was configured with the following parameters $d = 2$, $s_1 = 1$, $s_2 = 1.5$ and various traffic regimes including (i)$\alpha = 0, \beta = 0.2$, (ii) $\alpha = 1/2, \beta = 5$ and (iii)$\alpha = 2/3, \beta = 5$.

Figure 5.1 presents the system's performance metrics when the probabilistic job assignment scheme is employed. In Figure 5.1(a), the mean response time of accepted jobs is illustrated, while Figure 5.1(b) depicts the blocking probability of the system. Across all traffic regimes, it is observed that the blocking probability of the system consistently converges to zero as the system size n increases. Furthermore, when the system operates under heavy traffic conditions, i.e., $\alpha > 0$, the optimal solution y^* asymptotically converges to



(a) Mean response time of accepted

(b) Blocking probability of the system

Figure 5.1: System performance metrics for different system sizes n, degree of parallelism $d = 2$, speedups $s_1 = 1$, $s_2 = 1.5$ and various values of load parameters (i)$\alpha = 0, \beta = 0.2$, (ii)$\alpha = 1/2, \beta = 5$, and (iii)$\alpha = 2/3, \beta = 5$.

the value $(1, 0)$ as the system size n tends to infinity. This convergence results in the mean response time of accepted jobs converging to 1, as demonstrated by the graphs. Conversely, when the traffic conditions are such that $\alpha = 0$ and $\beta = 0.2$, the optimal solution is given by $y^* = (0.2, 0.4)$, and the mean response time of jobs converges to $\sum y_i^*/(1-\beta) = 0.75$ as depicted in the graph. These empirical results align with the theoretical findings derived in Theorems 5.2.4 and 5.2.9, underscoring the robustness of our analysis in various traffic scenarios.

5.4 Conclusion

In this chapter, our focus was on adaptive multi-server job systems, where the concept of a perfect speedup in job processing time was not applicable. Our objective was to optimize system performance in the asymptotic limit, across varying traffic regimes. When the speedup function was strictly increasing, sublinear, and concave with respect to the number of servers assigned to job processing, we identified the optimal system state and showed that it is at most two-dimensional. This optimal state underscored a critical observation: Under heavy traffic conditions, the advantages of parallelizing jobs across multiple servers diminish as the system size increases, and the optimal approach is to assign each job to an individual server. Moreover, this optimal solution enabled us to represent the optimal system with a simplified, at most two-dimensional fluid model, offering a powerful tool for comprehending and predicting system performance. We devised a probabilistic job assignment scheme and showed that in the large system size limit, the original system can be approximated by this fluid model, resulting in zero blocking probability and the minimum average response time of jobs.

Chapter 6

Conclusions and Future Work

We conclude with a summary of the contributions of this thesis and discuss the extensions and issues that can be addressed in future work.

6.1 Summary

In this thesis, we have focused on the design and analysis of randomized dynamic load balancing policies in large-scale networks arising in applications such as Infrastructure-as-a-Service (IaaS) clouds. These policies play a crucial role in efficiently allocating incoming user requests to servers, based on job assignment schemes that are adapted to the current state of a sampled subset of servers. The objective is to optimize system performance in terms of specific metrics such as the mean response time of jobs and the probability of loss. Due to complex interdependencies among servers which is caused by incorporating system state into routing decisions and the large size of these systems, the exact analysis of these systems is intractable. To address this challenge, we have employed relevant limit theorems, including mean-field techniques and Stein's approach, to study the system as its size converges to infinity. Additionally, we have evaluated the accuracy of these limits when applied to systems of finite size using FCLTs and Stein's method.

In Chapter2, our focus was on the analysis of large heterogeneous processor sharing systems, where servers varied in processing speeds and were categorized into distinct groups based on their speeds. We introduced a randomized threshold-based load balancing policy in which servers were sampled from each different group to incorporate the inherent heterogeneity in the system. Job assignments were made then based on the sampled servers' occupancy, processing speeds, and predetermined server-specific thresholds. We showed that this policy ensures system stability for any arrival rate strictly below the system's total capacity, making it a throughput-optimal system. Under certain assumptions on the initial empirical random measure of the system, we showed that as the system size tends to infinity, the empirical occupancy measure converges to a deterministic limit, characterized by a set of ODEs. This deterministic system, known as the mean-field limit, offers a reliable approximation for the system's empirical measure over finite time periods. Furthermore, we established the global asymptotic stability of the fixed point of the mean-field limit, relying on the monotonicity properties of the system of ODEs. Combining these results, we showed that the empirical stationary distribution of the system converges to this unique fixed point in the system size limit. Consequently, we demonstrated that the asymptotic steady-state mean response time of jobs corresponds to the response time of the mean-field limit when it reaches its fixed point. Notably, these results proved to be independent of the specific threshold values in the system, rendering them applicable to any load balancing policy that conforms to the proposed framework.

In Chapter 3, we extended the results from Chapter 2 to evaluate the accuracy of representing the empirical measure of a finite-size system by its mean-field limit. Our focus was on large heterogeneous processor-sharing systems operating under two distinct regimes including the heavy traffic Halfin-Whitt regime and a sub-critical regime where the arrival rate always remains below the system's total capacity. We introduced the fluctuation process as the difference between the system's empirical measure and its mean-field limit. Using FCLTs, we showed that as the system size approaches infinity, the diffusion-scaled fluctuation process converges to an OU process whose drift and diffusion coefficients depend on the traffic regime and the mean-field limit. Furthermore, using that the mean-field limit of the system exhibited local exponential stability at its unique fixed point, we demonstrated analogous convergence results for the fluctuation process in the stationary regime. Based on these outcomes, we established the rates at which the empirical measure of the system approached its limit, both in transient and stationary regimes. This rate scaled proportionally with the inverse square root of the system size in the Halfin-Whitt traffic regime and exhibited even faster convergence in the sub-critical regime. Additionally, we showed that the average response time of jobs in the finite-size system followed a similar rate of convergence toward its asymptotic value.

In Chapter 4, we studied large loss systems with adaptive multiserver jobs. These jobs had the flexibility to run on a variable number of servers, up to a maximum of d servers. A key assumption was that the job's processing time decreased proportionally with the number of servers allocated to it. Within this context, we established a criterion for the asymptotic optimality of the steady-state system in terms of the mean response time of jobs and the blocking probability. To achieve this, we introduced a greedy job assignment scheme where a random subset of servers was sampled, and the number of servers assigned to process the job depended on the available servers within this sampled subset. We demonstrated that in scenarios where incoming jobs possessed complete knowledge of the system state and all servers were accessible, the steady-state system approached the optimal average response time of $1/d$ and zero blocking probability as the system size grew to infinity. Furthermore, we quantified error bounds for these values in finite-size systems as functions of the number of servers. Moreover, we studied cases where arrivals had partial knowledge of the system state, determined upon their arrival. We showed that the system achieves similar optimal performance measures. However, achieving this required that the sampling size increased at a rate exceeding n^{α}, where α represented the rate at which the arrival rate approached its critical value and n was the number of servers. Additionally, we extended these results to systems with multiple arrival streams, each characterized by varying job sizes and degrees of parallelization. Remarkably, the same asymptotic optimality results held true. Our analysis covered a range of traffic regimes, including the mean-field regime and the (sub- and super-) Halfin-Whitt regimes.

In Chapter 5, we expanded on the results from the previous chapter, addressing scenarios where a job's processing time does not exhibit a linear relationship with the number of servers assigned to it. Specifically, we considered speedup functions that were strictly increasing, concave, and sublinear with respect to the number of servers allocated to a job. Firstly, we established the optimal behavior for such systems. We introduced a probabilistic job assignment scheme and showed that as the system size approaches infinity, the steady-state system achieves the optimal system performance as defined by zero blocking probability and the minimal achievable average response time for accepted jobs. A key result is that in heavy traffic regimes, under these conditions for the speedup function, job parallelization does not enhance system performance. Instead, the optimal approach in heavy traffic involves assigning each job to individual servers.

6.2 Future Work

Following the results presented in this dissertation, further research can be carried out to address the following issues.

- An open problem in Chapter 3 is demonstrating the local exponential stability of the mean-field limit at its fixed point, specifically in the context of heterogeneous systems. While we delivered rigorous proof in the homogeneous case, addressing this question in the heterogeneous scenario remains an open challenge. Establishing this property enhances the robustness and quality of this study.

- In Chapters 4 and 5, we considered adaptive multiserver jobs, where the length of an undivided job follows an exponential distribution with a mean of one. We established the asymptotic optimality of the system under various load balancing policies, including the greedy assignment scheme for linear speed functions, and the I-greedy and probabilistic assignment schemes for sublinear speedup functions. An interesting question arises: Do the results carry over to the case of non-exponential job lengths, i.e., are such state-dependent multi-rate models insensitive to the job length distribution? We conjecture that this is indeed true and we provide numerical evidence to support the claim below. The proof of this conjecture presents an open challenge, requiring the construction of Markov processes on continuous state spaces and analysis of their corresponding generators.

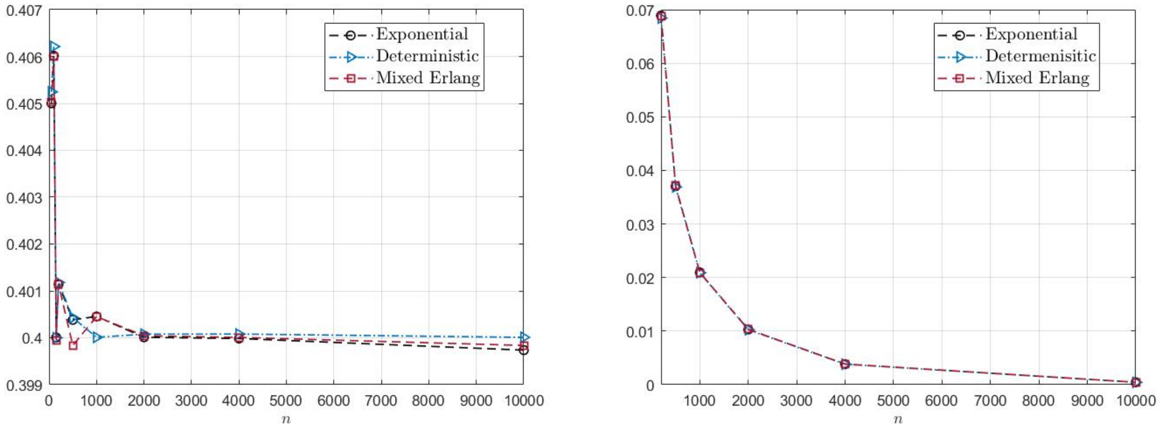
  In Figure 6.1, we provide numerical evidence to support the insensitivity of the system, focusing on the scenario where jobs receive sublinear speedup and the I-greedy assignment policy is applied (as discussed in Remark 5.1). We consider three different job length distributions: the exponential distribution, the deterministic distribution, and the Mixed-Erlang distribution, all having the same unit mean. The Mixed-Erlang distribution comprises sums of independent exponentially distributed random variables (referred to as an Erlang distribution) where the number of exponential phases is equal to $i \in \{1, 2, \ldots, N\}$ with probability p_i, and $\sum_{i=1}^{N} p_i = 1$. Each exponential phase is assumed to have a mean of $\frac{1}{\mu_p}$. Consequently, we have

  $$\frac{\sum_{i=1}^{N} i p_i}{\mu_p} = 1.$$

  We calculate μ_p based on the given values of p_i. Our system parameters are as follows: $\alpha = 0, \beta = 0.2, d = 4, s_1 = 1, s_2 = 1.8, s_3 = 2.5, s_4 = 3, N = 2, p_1 = 0.4$, and $p_2 = 0.6$. With these parameters, following Remark 5.1, we find that $I = 3$. Consequently, the mean response time of jobs converges to $\frac{1}{s_I} = \frac{1}{2.5} = 0.4$.

  The graphs illustrate that the system's performance is insensitive to the exact distribution of job lengths and remains asymptotically optimal for all job size distributions.

- In Chapter 5, we considered a probabilistic assignment scheme in which each incoming job is split into $i \in [d]$ sub-jobs with probabilities p_i. If there aren't enough available servers, the job is blocked. An adjusted version of this policy involves splitting a job into $i \in [d]$ sub-jobs and in the event of an insufficient number of available



(a) Mean response time of accepted jobs

(b) Blocking probability of the system

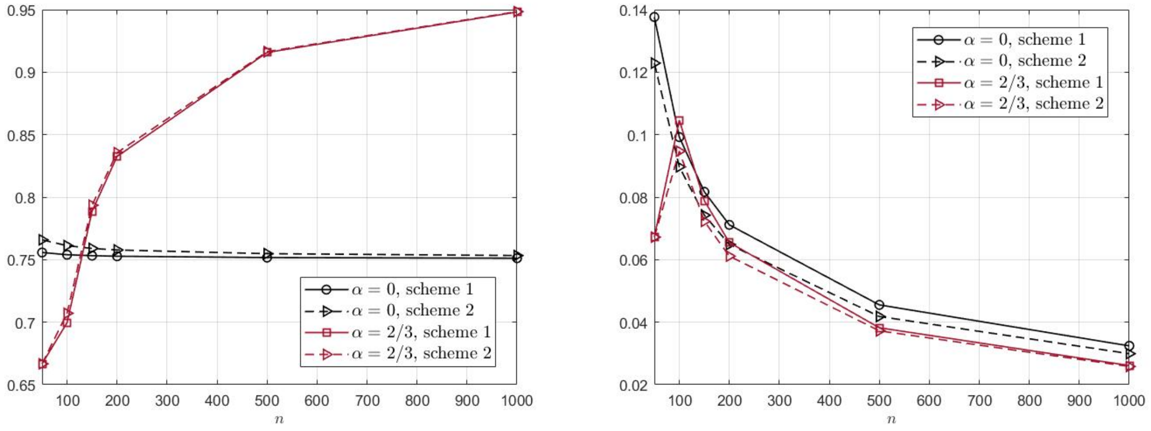
Figure 6.1: System performance metrics for different system sizes n, sampling size $k(n) = n$, arrival rate parameters $\alpha = 0, \beta = 0.2$, degree of parallelism $d = 4$, speedups $s_1 = 1,\ s_2 = 1.8,\ s_3 = 2.5,\ s_4 = 3$ and various job length distributions.

servers, allowing the job to occupy any available servers left in the system. The job gets blocked only when the system is full. This modified policy results in updated assignment probabilities, given by

$$A_i(x) = \mathbb{1}\left(nq_0 \geq i\right) p_i + \mathbb{1}\left(nq_0 = i\right) \sum_{j>i} p_j.$$

Let us refer to the probabilistic policy employed in Chapter 5 as "scheme 1", and the modified version introduced here as "scheme 2". Figure 6.2 provides numerical evidence comparing these two policies, demonstrating that scheme 2 is also asymptotically optimal and offers better results in terms of blocking probability in finite-sized systems. Studying this modified policy further, along with analyzing its asymptotic behavior and performance characteristics, could be a valuable avenue for future research.

- The model we have introduced in Chapters 4 and 5 assumes that each incoming user request is compatible with all the servers in the system. However, in scenarios like file retrieval systems, where an original file is recovered from multiple data chunks stored in parallel, jobs can only be assigned to servers if they meet specific compat-



(a) Mean response time of accepted jobs

(b) Blocking probability of the system

Figure 6.2: System performance metrics for different system sizes n, sampling size $k(n) = n$, degree of parallelism $d = 2$, speedups $s_1 = 1,\ s_2 = 1.5$ and different arrival rate parameters (i)$\alpha = 0, \beta = 0.2$, (ii)$\alpha = 2/3, \beta = 5$.

ibility criteria, such as having the required data. Incorporating these compatibility restrictions into the system model remains an open challenge. An intriguing question that emerges is: how many of the servers should be compatible with a specific type of arrival to ensure that the system achieves asymptotic optimality? This problem presents an avenue for further research.

- In Chapter 5, we introduced two job assignment schemes: the I-greedy and the probabilistic scheme. Both of these policies rely on having knowledge of the arrival rate $\lambda(n)$ and system parameters $s_i,\ i \in [d]$. However, in practice, the exact speed-up function and the precise region in which $\lambda(n)$ lies may not be readily known. An alternative approach is to employ adaptive policies that eliminate the need for detailed system parameter knowledge. These adaptive policies determine the arrival rate region by monitoring job blocking occurrences. We present two such adaptive policies.

  (i) In the first adaptive policy, the initial degree of parallelism is set to one. At each arrival instance, we increment the degree of parallelism by one, with a maximum value of d. However, once the first job blocking event occurs, the degree of parallelism is reset to one, and the process begins again.

(ii) In the second policy, instead of increasing the degree of parallelism with each job arrival, we begin with the maximum degree of parallelism, which is d. If blocking occurs, we reduce the degree of parallelism by one until there is no blocking. We conjecture that this policy achieves a similar level of performance as the I-greedy policy.

Importantly, these policies do not require a priori knowledge of the arrival rate and speedup function and align well with practical applications. The analysis of these policies remains an open problem, leading to intriguing questions regarding the system's performance when employing such policies.

References

[1] T. G. Dietterich. Machine Learning for Sequential Data: A Review. In T. Caelli, A. Amin, R. P. W. Duin, D. de Ridder, and M. Kamel, editors, *Structural, Syntactic, and Statistical Pattern Recognition SSPR/SPR*, Berlin, Heidelberg, 2002. Springer.

[2] V. Singh, S.-S. Chen, M. Singhania, B. Nanavati, A. kumar kar, and A. Gupta. How Are Reinforcement Learning and Deep Learning Algorithms Used for Big Data Based Decision Making in Financial Industries–A Review and Research Agenda. *International Journal of Information Management Data Insights*, 2(2), 2022.

[3] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes. Large-Scale Cluster Management at Google with Borg. In *Proceedings of the Tenth European Conference on Computer Systems*, EuroSys '15, New York, NY, USA, 2015. Association for Computing Machinery.

[4] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[5] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA, November 2016. USENIX Association.

[6] K. Lee, N. B. Shah, L. Huang, and RamchandranKannan. The MDS Queue: Analysing the Latency Performance of Erasure Codes. *IEEE Transactions on Information Theory*, 63(5):2822–2842, 2017.

[7] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's Highly Available Key-Value Store. *SIGOPS Operating Systems Review*, 41(6):205–220, 2007.

[8] J. Yi, F. Maghoul, and J. Pedersen. Deciphering Mobile Search Patterns: A Study of Yahoo! Mobile Search Queries. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 257–266, New York, NY, USA, 2008. Association for Computing Machinery.

[9] W. Winston. Optimality of the Shortest Line Discipline. *Journal of Applied Probability*, 14(1):181–189, 1977.

[10] V. Gupta, M. Harchol Balter, K. Sigman, and W. Whitt. Analysis of Join-The-Shortest-Queue Routing for Web Server Farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

[11] L. Kleinrock. Time-Shared Systems: A Theoretical Treatment. *Journal of the ACM*, 14(2):242–261, 1967.

[12] R. Schassberger. A New Approach to the M / G / 1 Processor-Sharing Queue. *Advances in Applied Probability*, 16(1):202–213, 1984.

[13] S. F. Yashkov and A. S. Yashkova. Processor Sharing: A Survey of the Mathematical Theory. *Automation and Remote Control*, 68:1662–1731, 2007.

[14] P. H. Brill and L. Green. Queues in Which Customers Receive Simultaneous Service from a Random Number of Servers: A System Point Approach. *Management Science*, 30(1):51–68, 1984.

[15] W. Cirne and F. Berman. Using Moldability to Improve the Performance of Supercomputer Jobs. *Journal of Parallel and Distributed Computing*, 62(10):1571–1601, 2002.

[16] Srinivasan, Krishnamoorthy, and Sadayappan. A Robust Scheduling Technology for Moldable Scheduling of Parallel Jobs. In *2003 Proceedings IEEE International Conference on Cluster Computing*, pages 92–99, Hong Kong, China, 2003. IEEE.

[17] T. Özden, T. Beringer, A. Mazaheri, H. M. Fard, and F. Wolf. ElastiSim: A Batch-System Simulator for Malleable Workloads. In *Proceedings of the 51st International Conference on Parallel Processing*, ICPP '22, New York, NY, USA, 2023. Association for Computing Machinery.

[18] P. Sanders and D. Schreiber. Decentralized Online Scheduling of Malleable NP-hard Jobs. In J. Cano and P. Trinder, editors, *Euro-Par 2022: Parallel Processing*, Cham, 2022. Springer.

[19] F. A. Haight. Two Queues in Parallel. *Biometrika*, 45(3-4):401–410, 1958.

[20] J. F. C. Kingman. Two Similar Queues in Parallel. *The Annals of Mathematical Statistics*, 32(4):1314–1323, 1961.

[21] A. Ephremides, P. Varaiya, and J. Walrand. A Simple Dynamic Routing Problem. *IEEE Transactions on Automatic Control*, 25(4):690–693, 1980.

[22] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California, Berkeley, 1991.

[23] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing System with Selection of the Shortest of Two Queues: An Asymptotic Approach. *Problemy Peredachi Informatsii*, 32(1):20—34, 1996.

[24] M. Bramson, Y. Lu, and B. Prabhakar. Randomized Load Balancing with General Service Time Distributions. *SIGMETRICS Performance Evaluation Review*, 38(1):275–286, 2010.

[25] 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. *SIGMETRICS Performance Evaluation Review*, 46(3):54–55, 2019.

[26] A. Mukhopadhyay and R. R. Mazumdar. Rate-Based Randomized Routing in Large Heterogeneous Processor Sharing Systems. In *2014 26th International Teletraffic Congress (ITC)*, pages 1–9. IEEE, 2014.

[27] 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*, 3(2):116–126, 2016.

[28] A. Mukhopadhyay, A. Karthik, and R. R. Mazumdar. Randomized Assignment of Jobs to Servers in Heterogeneous Clusters of Shared Servers for Low Delay. *Stochic Systems*, 6(1):90–131, 2016.

[29] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia. Delay, Memory, and Messaging Tradeoffs in Distributed Service Systems. *SIGMETRICS Performance Evaluation Review*, 44(1):1–12, 2016.

[30] D. Mukherjee, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting. Universality of Power-of-d Load Balancing in Many-Server Systems. *Stochastic Systems*, 8(4):265–292, 2018.

[31] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg. Join-Idle-Queue: A Novel Load Balancing Algorithm for Dynamically Scalable Web Services. *Performance Evaluation*, 68(11):1056–1071, 2011.

[32] M. Mitzenmacher. Analyzing Distributed Join-Idle-Queue: A Fluid Limit Approach. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 312–318, 2016.

[33] A. L. Stolyar. Pull-Based Load Distribution among Heterogeneous Parallel Servers: the Case of Multiple Routers. *Queueing Systems*, 85:31–65, 2017.

[34] S. Foss and A. L. Stolyar. Large-Scale Join-Idle-Queue System with General Service Times. *Journal of Applied Probability*, 54(4):995–1007, 2017.

[35] X. Zhou, J. Tan, and N. Shroff. Heavy-Traffic Delay Optimality in Pull-Based Load Balancing Systems: Necessary and Sufficient Conditions. *Proceeedings of the ACM Measurement and Analysis of Computing Systems*, 2(3), 2018.

[36] X. Zhou, F. Wu, J. Tan, Y. Sun, and N. Shroff. Designing Low-Complexity Heavy-Traffic Delay-Optimal Load Balancing Schemes: Theory to Algorithms. *Proceedings of the ACM on Measurement and Analysis of Computing Sytems*, 1(2):1–30, 2017.

[37] T. Vasantam and R. R. Mazumdar. On Occupancy Based Randomized Routing Schemes in Large Systems of Shared Servers. In *2018 30th International Teletraffic Congress (ITC 30)*, Vienna, Austria, 2018. ITC Press.

[38] I. A. Horváth, Z. Scully, and B. Van Houdt. Mean Field Analysis of Join-Below-Threshold Load Balancing for Resource Sharing Servers. *Proceedings of the ACM on Measurement and Analysis of Computing Sytems*, 3(3):1–21, 2019.

[39] L. Flatto and S. Hahn. Two Parallel Queues Created by Arrivals with Two Demands I. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

[40] F. Baccelli, A. M. Makowski, and A. Shwartz. The Fork-Join Queue and Related Systems with Synchronization Constraints: Stochastic Ordering and Computable Bounds. *Advances in Applied Probability*, 21(3):629–660, 1989.

[41] A. Rizk, F. Poloczek, and F. Ciucu. Stochastic Bounds in Fork–Join Queueing Systems under Full and Partial Mapping. *Queueing Systems*, 83(3-4):261–291, 2016.

[42] A. Thomasian. Analysis of Fork/Join and Related Queueing Systems. *ACM Computing Surveys*, 47(2):1–71, 2014.

[43] R. Jinan, G. Gautam, P. Parag, and V. Aggarwal. Asymptotic Analysis of Probabilistic Scheduling for Erasure-Coded Heterogeneous Systems. *SIGMETRICS Performance Evaluation Review*, 50(4):8–10, 2023.

[44] W. Wang, M. Harchol-Balter, H. Jiang, A. Scheller-Wolf, and R. Srikant. Delay Asymptotics and Bounds for Multitask Parallel Jobs. *Queueing Systems*, 46(3):207–239, 2019.

[45] W. Weng and W. Wang. Achieving Zero Asymptotic Queueing Delay for Parallel Jobs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):36, 2021.

[46] D. Filippopoulos and H. Karatza. An M/M/2 Parallel System Model with Pure Space Sharing among Rigid Jobs. *Mathematical and Computer Modelling*, 45(5):491–530, 2007.

[47] W. Wang, Q. Xie, and M. Harchol-Balter. Zero Queueing for Multi-Server Jobs. *SIGMETRICS Performance Evaluation Review*, 49(1):13–14, 2022.

[48] Y. Hong and W. Wang. Sharp Waiting-Time Bounds for Multiserver Jobs. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, MobiHoc '22, New York, NY, USA, 2022. Association for Computing Machinery.

[49] I. Grosof, Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. Optimal Scheduling in the Multiserver-Job Model under Heavy Traffic. *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 6(3), 2022.

[50] W. Whitt. Blocking when Service is Required from Several Facilities Simultaneously. *AT&T Technical Journal*, 64(8):1807–1856, 1985.

[51] J. Kaufman. Blocking in a Shared Resource Environment. *IEEE Transactions on Communications*, 29(10):1474–1481, 1981.

[52] M. D. Hill and M. R. Marty. Amdahl's Law in the Multicore Era. *Computer*, 41(7):33–38, 2008.

[53] B. Berg, J.-P. Dorsman, and M. Harchol-Balter. Towards Optimality in Parallel Job Scheduling. *SIGMETRICS Performance Evaluation Review*, 46(1):116–118, 2018.

[54] B. Berg, M. Harchol-Balter, B. Moseley, W. Wang, and J. Whitehouse. Optimal Resource Allocation for Elastic and Inelastic Jobs. In *Proceedings of the 32nd ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '20, page 75–87, New York, NY, USA, 2020. Association for Computing Machinery.

[55] I. Grosof, M. Harchol-Balter, and A. Scheller-Wolf. WCFS: A New Framework for Analyzing Multiserver Systems. *Queueing Systems*, 102(1-2):143–174, 2022.

[56] L. Bortolussi and R. A. Hayden. Bounds on the Deviation of Discrete-Time Markov Chains from their Mean-Field Model. *Performance Evaluation*, 70(10):736–749, 2013.

[57] C. Graham. Functional Central Limit Theorems for a Large Network in which Customers Join the Shortest of Several Queues. *Probability Theory and Related Fields*, 131:97–120, 2005.

[58] T. Vasantam and R. R. Mazumdar. Fluctuations around the Mean-Field for a Large Scale Erlang Loss System under the SQ(d) Load Balancing. In *2019 31st International Teletraffic Congress (ITC 31)*, Budapest, Hungary, Hungary, 2019. IEEE.

[59] T. Vasantam and R. R. Mazumdar. Sensitivity of Mean-Field Fluctuations in Erlang Loss Models with Randomized Routing. *Journal of Applied Probability*, 58(2):428–448, 2021.

[60] S. Halfin and W. Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations research*, 29(3):567–588, 1981.

[61] L. Ying. On the Approximation Error of Mean-Field Models. *SIGMETRICS Performance Evaluation Review*, 44(1):285–297, 2016.

[62] L. Ying. Stein's Method for Mean-Field Approximations in Light and Heavy Traffic Regimes. *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 1(1):1–27, 2017.

[63] N. Gast. Expected Values Estimated via Mean-Field Approximation are 1/N-Accurate. *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 1(1):1–26, 2017.

[64] N. Gast and B. Van Houdt. A Refined Mean Field Approximation. *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 1(2):1–28, 2017.

[65] N. Gast, L. Bortolussi, and M. Tribastone. Size Expansions of Mean Field Approximation: Transient and Steady-State Analysis. *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 46(3):25–26, 2019.

[66] Hairi, X. Liu, and L. Ying. Beyond Scaling: Calculable Error Bounds of the Power-of-Two-Choices Mean-Field Model in Heavy-Traffic. In *Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '21)*, New York, NY, USA, 2021. Association for Computing Machinery.

[67] S. Allmeier and N. Gast. Mean Field and Refined Mean Field Approximations for Heterogeneous Systems: It Works! *Proceedings of the ACM Measurement and Analysis of Computing Systems*, 6(1):1–43, 2022.

[68] M. Kac. Foundations of Kinetic Theory. In *Proceedings of The Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 171–197, 1956.

[69] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence.* John Wiley & Sons, New York, 1986.

[70] P. Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, New York, 1999.

[71] C. Stein. A Bound for the Error in the Normal Approximation to the Distribution of a Sum of Dependent Random Variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 583–603. University of California Press, Berkeley, CA, 1972.

[72] C. Stein. *Approximate Computation of Expectations*, volume 7 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[73] A. D. Barbour. Stein's Method and Poisson Process Convergence. *Journal of Applied Probability*, 25(A):175–184, 1988.

[74] A. L. Stolyar. Tightness of Stationary Distributions of a Flexible-Server System in the Halfin-Whitt Asymptotic Regime. *Stochastic Systems*, 5(2):239–267, 2015.

[75] A. Braverman, J.G. Dai, and J. Feng. Stein's Method for Steady-State Diffusion Approximations: An Introduction Through the Erlang-A and Erlang-C Models. *Stochastic Systems*, 6(2):301–366, 2017.

[76] A. Braverman. Steady-State Analysis of the Join-the-Shortest-Queue Model in the Halfin–Whitt Regime. *Mathematics of Operations Research*, 45(3):1069–1103, 2020.

[77] S. Ghanbarian and R. R. Mazumdar. Mean-Field Fluctuations at Diffusion Scale in Threshold-Based Randomized Routing for Processor Sharing Systems and Applications. *Stochastic Models*, pages 1–46, 2023.

[78] S. Ghanbarian, A. Mukhopadhyay, F. M. Guillemin, and R. R. Mazumdar. On the Performance of Large Loss Systems with Adaptive Multiserver Jobs. *Under Review (arXiv preprint arXiv:2309.00060)*, 2023.

[79] S. Foss and N. Chernova. On the Stability of a Partially Accessible Multi-Station Queue with State-Dependent Routing. *Queueing Systems*, 29:55–73, 1998.

[80] K. Deimling. *Ordinary Differential Equations in Banach Spaces.* Springer, Heidelberg, 1977.

[81] G. Pang, R. Talreja, and W. Whitt. Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*, 4:193 – 267, 2007.

[82] H. K. Khalil. *Nonlinear systems.* Prentice-Hall, New Jersey, 2002.

[83] E. Seneta. *Non-Negative Matrices and Markov Chains.* Springer, New York, 2006.

[84] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. Performance of Multiclass Markovian Queueing Networks via Piecewise Linear Lyapunov Functions. *Annals of Applied Probability*, 11(4):1384–1428, 2001.

[85] J.B. Martin and Y. M. Suhov. Fast Jackson networks. *The Annals of Applied Probability*, 9(3):854 – 870, 1999.

APPENDICES

Appendix A

Proofs

In this section, we provide some technical proofs not included in the main text.

A.1 Proof of Compactness of the Space $\mathbb{U}^M$

Since the Cartesian product of compact spaces is compact, it is sufficient to show that the space $\mathbb{U}$ is compact under the ℓ_2-norm. Consider a sequence of elements $\{\mathbf{u}^n\}_{n=1}^{\infty}$ where for each n,

$$\mathbf{u}^n = (u_0^n, u_1^n, u_2^n, ..., u_k^n, ...) \in \mathbb{U}. \tag{A.1}$$

For a fixed $k \in \mathbb{Z}_+$, the sequence $\{u_k^n\}_{n=1}^{\infty}$ is a bounded real sequence. Thus by the Bolzano-Weierstrass theorem, it has a convergent subsequence in $\mathbb{R}$. By the process of diagonalization, we can find a subsequence $\{\mathbf{u}^{n_l}\}_{l=1}^{\infty}$ such that for some $\tilde{\mathbf{u}} = (\tilde{u}_0, \tilde{u}_1, \tilde{u}_2, ..., \tilde{u}_k, ...)$,

$$\lim_{l \to \infty} u_k^{n_l} = \tilde{u}_k, \quad \forall k \in \mathbb{Z}_+. \tag{A.2}$$

We show that the limit $\tilde{\mathbf{u}}$ lies in the space $\mathbb{U}$ and also the subsequence $\{\mathbf{u}^{n_l}\}_{l=1}^{\infty}$ converges to this limit under the ℓ_2-norm. From Equation (A.2), it is obvious that $\tilde{u}_0 = 1$, and $\tilde{u}_k \geq \tilde{u}_{k+1}$ for every index k. Also,

$$\sum_{k\in\mathbb{Z}_+}|\tilde{u}_k| = \sum_{k\in\mathbb{Z}_+}\lim_{l\to\infty}|u_k^{n_l}| = \lim_{l\to\infty}\sum_{k\in\mathbb{Z}_+}|u_k^{n_l}| < \infty. \tag{A.3}$$

These conditions guarantee that $\tilde{\mathbf{u}} \in \mathbb{U}$. Moreover, from Equation (A.2), we can find finite $N \in \mathbb{N}$, such that

$$|u_k^{n_l} - \tilde{u}_k| < \sqrt{\frac{\epsilon}{N}}, \quad \forall l \geq N, \forall k \in \mathbb{Z}_+. \tag{A.4}$$

We choose l sufficiently large such that

$$\sum_{k=1}^{N}|u_k^{n_l} - \tilde{u}_k|^2 \leq N \sup_k |u_k^{n_l} - \tilde{u}_k|^2 < \epsilon. \tag{A.5}$$

By letting N to go to infinity, we get

$$\sum_{k=1}^{\infty}|u_k^{n_l} - \tilde{u}_k|^2 < \epsilon. \tag{A.6}$$

The above inequality holds for every positive ϵ, consequently the ℓ_2 convergence holds. Thus we have shown that every sequence of elements in the space $\mathbb{U}$ has a convergent subsequence with the limit in the same space $\mathbb{U}$ and this completes the proof.

A.2 Proof of Lemma 2.4.3

For a fixed $k \in \mathbb{Z}_+$ and $m \in [M]$, we define $\mathbf{x}'(t)$ as the partial derivative of the process $\mathbf{x}(t)$ with respect to $u_{k,m}$. By differentiating Equation (2.16) with respect to $u_{k,m}$, we get

$$\begin{aligned}\frac{dx'_{s,l}}{dt} &= \sum_{i=1}^{d_l-1}\frac{\lambda i}{\gamma_l}(x_{\alpha_l+1,l})^{i-1}x'_{\alpha_l+1,l}(x_{s-1,l} - x_{s,l})\prod_{j=l+1}^{M}(x_{\alpha_j+1,j})^{d_j}\\ &+ \frac{\lambda(1-(x_{\alpha_l+1,l})^{d_l})}{\gamma_l(1-x_{\alpha_l+1,l})}(x'_{s-1,l} - x'_{s,l})\prod_{j=l+1}^{M}(x_{\alpha_j+1,j})^{d_j}\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=l+1}^{M} \frac{\lambda d_i}{\gamma_l x_{\alpha_i+1,i}} \frac{1-(x_{\alpha_l+1,l})^{d_l}}{1-x_{\alpha_l+1,l}} (x_{s-1,l}-x_{s,l}) \prod_{j=l+1}^{M} (x_{\alpha_j+1,j})^{d_j} x'_{\alpha_i+1,i} \\
&- C_l(x'_{s,l}-x'_{s+1,l}), \quad l \in [M], 1 \leq s \leq \alpha_l+1, \qquad \text{(A.7)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{dx'_{s,l}}{dt} &= \frac{\lambda d_l}{\gamma_l}((x_{s-1,l})^{d_l-1}x'_{s-1,l}-(x_{s,l})^{d_l-1}x'_{s,l}) \prod_{j=1}^{l-1}(x_{\lceil s-1\rceil_{jl},j})^{d_j} \prod_{j=l+1}^{M}(x_{\lfloor s-1\rfloor_{jl},j})^{d_j} \\
&+\sum_{i=1}^{l-1} \frac{\lambda d_i}{\gamma_l x_{\lceil s-1\rceil_{il},i}}((x_{s-1,l})^{d_l}-(x_{s,l})^{d_l}) \prod_{j=1}^{l-1}(x_{\lceil s-1\rceil_{jl},j})^{d_j} \prod_{j=l+1}^{M}(x_{\lfloor s-1\rfloor_{jl},j})^{d_j} x'_{\lceil s-1\rceil_{il},i} \\
&+\sum_{i=l+1}^{M} \frac{\lambda d_i}{\gamma_l x_{\lfloor s-1\rfloor_{il},i}}((x_{s-1,l})^{d_l}-(x_{s,l})^{d_l}) \prod_{j=1}^{l-1}(x_{\lceil s-1\rceil_{jl},j})^{d_j} \prod_{j=l+1}^{M}(x_{\lfloor s-1\rfloor_{jl},j})^{d_j} x'_{\lfloor s-1\rfloor_{il},i} \\
&- C_l(x'_{s,l}-x'_{s+1,l}), \quad l \in [M], s > \alpha_l+1. \qquad \text{(A.8)}
\end{aligned}
$$

For the simplicity of notation, we have dropped the time argument t in the above equations. Also at $t=0$, we have

$$
x'_{s,l}(0) = \begin{cases} 1 & , s=k, l=m \\ 0 & , otherwise \end{cases} \qquad \text{(A.9)}
$$

From [85, Lemma 3.1] with $a = 2\lambda \frac{\max_m d_m}{\min_m \gamma_m} \sum_m d_m + 2\max_m C_m$, $b_0 = 0$, and $c = 1$, we get the upper bound for the first partial derivative. If we differentiate $\frac{dx'_{s,l}}{dt}$ one more time with respect to $u_{k,m}$ or $u_{k',m'}$, we can use the same Lemma 3.1 of [85] to get the upper bound for the second partial derivative.

A.3 Proof of Lemma 3.1.1

We provide results for the operator $W_{1,m}$ for a particular fixed value of $m \in [M]$, employing the ℓ_2-norm. Similar results for other cases can be derived analogously. Using Equations (3.4)-(3.5) and considering the definition of ℓ_2-norm, we can establish the following relationship for any pair of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{U}^M$

$$
\|W_{1,m}(\mathbf{u}) - W_{1,m}(\mathbf{v})\|_2^2 = \sum_{k \in \mathbb{Z}_+} |(W_1(\mathbf{u}))_{k,m} - (W_1(\mathbf{v}))_{k,m}|^2
$$

$$\begin{aligned}
&= \sum_{k=1}^{\alpha_m+1} \left| \frac{\lambda(1-(u_{\alpha_m+1,m})^{d_m})}{\gamma_m(1-u_{\alpha_m+1,m})}(u_{k-1,m}-u_{k,m}) \prod_{l=m+1}^{M} (u_{\alpha_l+1,l})^{d_l} \right. \\
&\qquad \left. - \frac{\lambda(1-(v_{\alpha_m+1,m})^{d_m})}{\gamma_m(1-v_{\alpha_m+1,m})}(v_{k-1,m}-v_{k,m}) \prod_{l=m+1}^{M} (v_{\alpha_l+1,l})^{d_l} \right|^2 \\
&+ \sum_{k=\alpha_m+2}^{\infty} \left| \frac{\lambda}{\gamma_m}((u_{k-1,m})^{d_m}-(u_{k,m})^{d_m}) \prod_{l=1}^{m-1}(u_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (u_{\lfloor k-1 \rfloor_{lm},l})^{d_l} \right. \\
&\qquad \left. - \frac{\lambda}{\gamma_m}((v_{k-1,m})^{d_m}-(v_{k,m})^{d_m}) \prod_{l=1}^{m-1}(v_{\lceil k-1 \rceil_{lm},l})^{d_l} \prod_{l=m+1}^{M} (v_{\lfloor k-1 \rfloor_{lm},l})^{d_l} \right|^2 . \qquad \text{(A.10)}
\end{aligned}$$

From the inequality $|a_1b_1^p - a_2b_2^p| \leq |a_1-a_2| + p|b_1-b_2|$ for a_1, a_2, b_1 and $b_2 \in [0,1]$, we get

$$\begin{aligned}
\left\| W_{1,m}(\mathbf{u}) - W_{1,m}(\mathbf{v}) \right\|_2^2 \leq{}& 3\frac{\lambda^2 d_m^2}{\gamma_m^2} \sum_{k=1}^{\alpha_m+1} \left[\left| \sum_{l=m}^{M} d_m(u_{\alpha_l+1,l} - v_{\alpha_l+1,l}) \right|^2 \right. \\
&\qquad \left. + \left| u_{k-1,m} - v_{k-1,m} \right|^2 + \left| u_{k,m} - v_{k,m} \right|^2 \right] \\
&+ 2\frac{\lambda^2}{\gamma_m^2} \sum_{k=\alpha_m+2}^{\infty} \left[\left| \sum_{l=1}^{m} d_l(u_{\lceil k-1 \rceil_{lm},l} - v_{\lceil k-1 \rceil_{lm},l}) \right|^2 \right. \\
&\qquad \left. + \left| \sum_{l=m}^{M} d_l(u_{\lfloor k-1 \rfloor_{lm},l} - v_{\lfloor k-1 \rfloor_{lm},l}) \right|^2 \right] \\
\leq{}& \lambda^2 \frac{\max_l d_l^2}{\min_l \gamma_l^2} (3M \max_l d_l^2(\alpha_M+1) + 6 + 2MK_1 + 2MK_2) \\
&\times \left\| \mathbf{u} - \mathbf{v} \right\|_2^2 , \qquad \text{(A.11)}
\end{aligned}$$

where K_1 and K_2 are some positive finite constants such that for each $m \in [M]$,

$$\sum_{k=\alpha_m+2}^{\infty} \sum_{l=1}^{m} \left| u_{\lceil k-1 \rceil_{lm},l} - v_{\lceil k-1 \rceil_{lm},l} \right|^2 \leq K_1 \left\| \mathbf{u} - \mathbf{v} \right\|_2^2 , \qquad \text{(A.12)}$$

$$\sum_{k=\alpha_m+2}^{\infty}\sum_{l=m}^{M}\left|u_{\lfloor k-1\rfloor_{lm},l}-v_{\lfloor k-1\rfloor_{lm},l}\right|^2\leq K_2\left\|\mathbf{u}-\mathbf{v}\right\|_2^2. \tag{A.13}$$

So we can write $\left\|W_{1,m}(\mathbf{u})-W_{1,m}(\mathbf{v})\right\|_2\leq B_{W_1}\left\|\mathbf{u}-\mathbf{v}\right\|_2$, where

$$B_{W_1}^2=\lambda^2\frac{\max_l d_l^2}{\min_l\gamma_l^2}(3M\max_l d_l^2(\alpha_M+1)+6+2MK_1+2MK_2). \tag{A.14}$$

This completes the Lipschitz property for the operator $W_{1,m}$ with respect to the ℓ_2-norm.

A.4 Proof of Theorem 5.2.4

According to the rate conservation law, we can express

$$\mathbb{E}\left[\sum_{i\in[d]}s_ix_i\right]=\lambda(n)\left(1-P_b\right), \tag{A.15}$$

where P_b represents the steady-state blocking probability of the system. Since $x_i=0$ for every $i\notin I^*$ and $I^*=\{i_1\}$, we deduce

$$\mathbb{E}\left[s_{i_1}x_{i_1}\right]=\lambda(n)\left(1-P_b\right). \tag{A.16}$$

Therefore, we have

$$\begin{aligned}P_b^2&\leq\frac{1}{\lambda(n)^2}\left(\mathbb{E}\left[\lambda(n)-s_{i_1}x_{i_1}\right]\right)^2\\&\leq\frac{1}{\lambda(n)^2}\mathbb{E}\left[\left(\lambda(n)-s_{i_1}x_{i_1}\right)^2\right]\\&\leq\frac{s_{i_1}^2}{\lambda(n)^2}\mathbb{E}\left[\left(y_{i_1}^*-x_{i_1}\right)^2\right],\end{aligned} \tag{A.17}$$

where the second inequality follows from Jensen's inequality, and the last inequality stems from the fact that $y_{i_1}^*=\frac{\lambda(n)}{s_{i_1}}$ when $I^*=\{i_1\}$. From Corollary 5.2.3, we derive

$$P_b^2 \le \frac{s_{i_1}^2}{\lambda(n)^2}\,\frac{2i_1}{n}\,\frac{y_{i_1}^*}{i_1} = 2\frac{s_{i_1}}{n\lambda(n)}. \tag{A.18}$$

Again, the last equality follows from the relationship $y_{i_1}^* = \frac{\lambda(n)}{s_{i_1}}$. Note that $\lambda(n) = 1 - \beta/n^{\alpha}$. If $\alpha = 0$ and $0 < \beta < 1$, it readily follows that

$$P_b^2 \le 2\frac{s_{i_1}}{n(1-\beta)}. \tag{A.19}$$

Consequently, the steady-state blocking probability of the system converges to zero with an upper bound of $O(\frac{1}{\sqrt{n}})$. If $\alpha > 0$ and $\beta > 0$, as $\frac{\beta}{n^\alpha} < 1$ for sufficiently large n, we arrive at

$$P_b^2 \le 2\frac{s_{i_1}}{n}\left(1 + \frac{\beta}{n^\alpha} + O(\frac{1}{n^{2\alpha}})\right) \le 2\frac{s_{i_1}}{n} + o(\frac{1}{n}). \tag{A.20}$$

Hence, once again, the steady-state blocking probability of the system converges to zero with an error bound of $O(\frac{1}{\sqrt{n}})$.

Now, let us consider the mean response time of accepted jobs in the stationary regime, denoted as $\mathbb{E}[D]$. According to Little's law, we have

$$\lambda(n)(1-P_b)\mathbb{E}[D] = \mathbb{E}\left[\sum_{i\in[d]} x_i\right] = \mathbb{E}[x_{i_1}]. \tag{A.21}$$

The last equality follows from the fact that only the i_1^{th} component remains in the steady state, as the other components have an arrival rate of zero. We already know that $\mathbb{E}[s_{i_1}x_{i_1}] = \lambda(n)(1-P_b)$ according to the rate conservation law. Therefore,

$$\mathbb{E}[D] = \frac{1}{s_{i_1}}. \tag{A.22}$$

This equation shows that the mean response time of accepted jobs is equal to $\frac{1}{s_{i_1}}$ for sufficiently large n.

Appendix B

Useful Results

We recall in this section Theorem 1 of [84]. This result is used in the proof of Lemmas 4.3.7 and 5.2.7.

B.1 Tail Bounds

Consider a continuous time Markov chain $X(t)$ which takes values in some countable set $\mathcal{X}$, with a stationary probability distribution π. For any two vectors $x, x^{'} \in \mathcal{X}$, let $p(x, x^{'})$ denote the transition probability from state x to state $x^{'}$. For a given function $\Phi : \mathcal{X} \to \mathbb{R}^{+}$ such that $\mathbb{E}_{\pi}\left[\Phi(X(t))\right] < \infty$, let

$$\nu_{max} = \sup_{x,x^{'} \in X, p(x,x^{'})>0} |\Phi(x') - \Phi(x)| < \infty$$

Namely, ν_{max} is the largest possible change of the function Φ during an arbitrary transition. Also, let

$$p_{max} = \sup_{x \in X} \sum_{x^{'} \in \mathcal{X}, \Phi(x) < \Phi(x^{'})} p(x, x^{'}) < \infty.$$

Namely, $p_{max}$is the tight upper bound on the probability that the value of Φ is increasing during an arbitrary transition.

(i) If there exists a Lyapunov function Φ such that for any $x \in \mathcal{X}$ with $\Phi(x) > B$,

$$G\Phi(x) = \sum_{x^{'} \neq x} p(x, x^{'}) \Big(\Phi(x^{'}) - \Phi(x) \Big) \leq -\gamma,$$

for some $\gamma > 0$ and $B \geq 0$, then for any $m = 0, 1, 2, ...,$

$$\mathbb{P}_{\pi}\left(\Phi(X(t)) > B + 2\nu_{max} m\right) \leq \left(\frac{p_{max}\nu_{max}}{p_{max}\nu_{max} + \gamma}\right)^{m+1}.$$

As a result,

$$\mathbb{E}_{\pi}\left[\Phi(X(t))\right] \leq B + \frac{2p_{max}(\nu_{max})^2}{\gamma}.$$