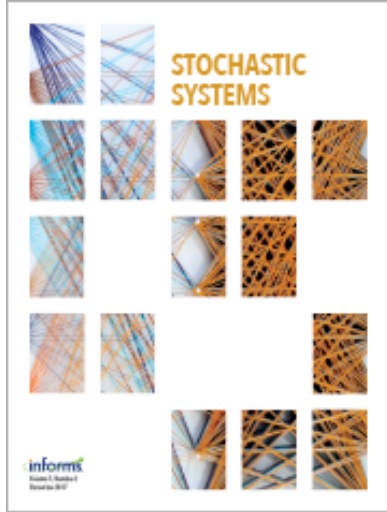




Stochastic Systems

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Asymptotic Optimality of the Speed-Aware Join-the-Shortest-Queue in the Halfin-Whitt Regime for Heterogeneous Systems

Sanidhay Bhambay, Burak Büke, Arpan Mukhopadhyay

To cite this article:

Sanidhay Bhambay, Burak Büke, Arpan Mukhopadhyay (2025) Asymptotic Optimality of the Speed-Aware Join-the-Shortest-Queue in the Halfin-Whitt Regime for Heterogeneous Systems. *Stochastic Systems* 15(2):147-193.
<https://doi.org/10.1287/stsy.2023.0054>

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “*Stochastic Systems*. Copyright © 2025 The Author(s). <https://doi.org/10.1287/stsy.2023.0054>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.”

Copyright © 2025 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Asymptotic Optimality of the Speed-Aware Join-the-Shortest-Queue in the Halfin-Whitt Regime for Heterogeneous Systems

Sanidhay Bhambay,^a Burak Büke,^b Arpan Mukhopadhyay^{c,*}

^aDurham University Business School, Durham DH1 1SL, United Kingdom; ^bSchool of Mathematics, The University of Edinburgh, Edinburgh EH9 3FD, United Kingdom; ^cDepartment of Computer Science, University of Warwick, Coventry CV4 7EZ, United Kingdom

*Corresponding author

Contact: sanidhay.bhambay@durham.ac.uk,  <https://orcid.org/0000-0002-1601-1543> (SB); B.Buke@ed.ac.uk,

 <https://orcid.org/0000-0002-6954-9144> (BB); arpan.mukhopadhyay@warwick.ac.uk,  <https://orcid.org/0000-0002-0634-1389> (AM)

Received: December 15, 2023

Revised: October 18, 2024

Accepted: January 9, 2025

Published Online in Articles in Advance:
March 31, 2025

<https://doi.org/10.1287/stsy.2023.0054>

Copyright: © 2025 The Author(s)

Abstract. The join-the-shortest-queue (JSQ) load-balancing scheme is known to minimize the average response time of jobs in homogeneous systems with identical servers. However, for *heterogeneous* systems with servers having different processing speeds, finding an optimal load balancing scheme remains an open problem for finite system sizes. Recently, for systems with heterogeneous servers, a variant of JSQ scheme, called the *speed-aware-join-the-shortest-queue* (SA-JSQ) scheme, has been shown to achieve asymptotic optimality in the fluid-scaling regime where the number of servers n tends to infinity but the normalized the arrival rate of jobs remains constant. In this paper, we show that the SA-JSQ scheme is also asymptotically optimal for heterogeneous systems in the *Halfin-Whitt* traffic regime where the normalized arrival rate scales are $1 - O(1/\sqrt{n})$. Our analysis begins by establishing that an appropriately scaled and centered version of the Markov process describing system dynamics weakly converges to a two-dimensional reflected *Ornstein-Uhlenbeck* (OU) process. We then show using *Stein's method* that the stationary distribution of the underlying Markov process converges to that of the OU process as the system size increases by establishing the validity of interchange of limits. Finally, through coupling with a suitably constructed system, we show that SA-JSQ asymptotically minimizes the diffusion-scaled total number of jobs and the diffusion-scaled number of waiting jobs in the steady state in the Halfin-Whitt regime among all policies that dispatch jobs based on queue lengths and server speeds.



Open Access Statement: This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "Stochastic Systems. Copyright © 2025 The Author(s). <https://doi.org/10.1287/stsy.2023.0054>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>."

Keywords: load balancing • speed-aware join-the-shortest queue • Halfin-Whitt regime • Stein's method • Lyapunov function • heterogeneous servers

1. Introduction

Load balancing is key to optimizing resource utilization and ensuring low delay of jobs in multiserver systems. In modern data centers, where numerous servers of various capacities coexist, load balancers play a pivotal role in distributing incoming network traffic or requests across these servers. This distribution of the load prevents any single server from being overwhelmed with jobs when there are idle servers in the system. Hence, load balancing ensures high utilization of the available resources, as well as low mean response time for the jobs.

The standard model to study the performance of load balancing schemes involves n *identical* or *homogeneous* servers, each equipped with its own queue to store pending job requests. In this setup, there is a continuous stream of jobs arriving at a rate of $n\lambda^{(n)}$. In contrast to the well-known $M/M/n$ system, there is no global queue in this model, and the dispatcher routes all incoming jobs to a server queue immediately upon arrival according to a specific load balancing scheme. This approach is especially beneficial to model the limited memory available to dispatchers in data centers for storing incoming jobs. A natural load balancing scheme to consider in this canonical setting is the *join-the-shortest-queue* (JSQ) scheme, wherein each incoming job is assigned to the server with minimum queue length with ties broken uniformly at random. The JSQ scheme was first analyzed in Winston (1977), and its optimality in terms of minimizing the mean response time of jobs was proved for Poisson

arrivals and exponential service times. This result was later extended to service time distributions having nondecreasing hazard rate by Weber (1978). The transient dynamics of the JSQ scheme has been studied by Eschenfeldt and Gamarnik (2018) in the *Halfin-Whitt regime*, introduced originally for $M/M/n$ systems by Halfin and Whitt (1981), by proving diffusion limits of the scaled system size processes when $\lambda^{(n)}$ is dependent on n , and the quantity $\sqrt{n}(1 - \lambda^{(n)})$ has a nondegenerate limit $\beta > 0$ as $n \rightarrow \infty$. Braverman (2020) showed that this diffusion limit can also be used to study the steady-state behavior of the model by proving an interchange of limits result. The rate at which the steady-state distributions of the JSQ scheme converges to the steady-state distribution of the diffusion limit was derived in Braverman (2023).

It is crucial to highlight that the effectiveness of the JSQ scheme primarily relies on the server homogeneity assumption. However, this assumption does not align with practical scenarios because data centers typically host an array of physical devices spanning multiple generations, having varying processing capabilities. Furthermore, the processing speeds of servers can diverge due to the incorporation of diverse acceleration devices like Graphics Processing Units, Field-Programmable Gate Arrays, and Application-Specific Integrated Circuits (Duato et al. 2010, Huang et al. 2016). In such heterogeneous systems, job assignment strategies like JSQ and the power-of-d-choices (Pod), which were originally designed with homogeneous systems in mind, may exhibit notably suboptimal performance (Bramson et al. 2012, Mukhopadhyay and Mazumdar 2015, Gardner et al. 2021). Additionally, it is worth noting that in finite systems, the classical JSQ scheme is recognized as nonoptimal when servers exhibit heterogeneity (Krishnan 1987, Hyytia et al. 2017). Therefore, for heterogeneous systems, speed-aware schemes that assign incoming jobs based on both the queue lengths and the server speeds are essential to minimize the average response time of jobs.

In a recent line of work by Weng et al. (2021), Bhambay and Mukhopadhyay (2022), and Rutten and Mukherjee (2023), a speed-aware variant of the JSQ scheme called the *speed-aware-join-the-shortest-queue* (SA-JSQ), also referred as *join-the-fastest-shortest-queue* (JFSQ), has been introduced for systems with n heterogeneous servers categorized into M different *types* or *pools*. In this heterogeneous model, each server pool $j \in \{1, 2, \dots, M\}$ contains $O(n)$ servers with processing speed μ_j where $\mu_1 > \mu_2 > \dots > \mu_M$. Under the SA-JSQ scheme, incoming jobs are assigned to servers with the highest speed among those with the shortest queue length. Notably, Bhambay and Mukhopadhyay (2022) prove that the SA-JSQ scheme asymptotically minimizes the average response time of jobs when $\lambda^{(n)} = \lambda \in [0, 1)$ for each n in the fluid limit. In the present work, we study the SA-JSQ load balancing scheme for heterogeneous systems in the Halfin-Whitt regime. Specifically, we study both the transient and the steady-state behavior of the SA-JSQ scheme and show that SA-JSQ is also asymptotically optimal in the Halfin-Whitt regime among all load-balancing policies that dispatch jobs based on queue lengths and server speeds. Our contributions are detailed below.

1.1. Contributions

The main contributions of this paper are three-fold. (i) We present a transient analysis of a system with heterogeneous servers and dedicated queues operating in the Halfin-Whitt regime under the SA-JSQ policy. Specifically, we prove that the appropriately scaled and centered queue lengths converge to a two-dimensional reflected Ornstein-Uhlenbeck process, exhibiting a state space collapse property. (ii) We demonstrate that this limiting Ornstein-Uhlenbeck process has a stationary distribution that can be used to approximate the stationary behavior of the system as the number of servers becomes large. (iii) We establish that under the Halfin-Whitt regime the SA-JSQ policy stochastically minimizes both the diffusion-scaled total number of jobs and the diffusion-scaled number of waiting jobs in the system in the limit as $n \rightarrow \infty$. Although our transient and stationary analyzes broadly follow existing analyzes for homogeneous systems with significant technical differences introduced due to the increased complexity of the underlying state space, to the best of our knowledge, we provide the first proof of asymptotic optimality of a JSQ-type load balancing policy in the Halfin-Whitt regime for a system where the servers are heterogeneous and each server has a dedicated queue. We note that the general problem of finding optimal load balancing policies for a system of heterogeneous parallel queues is a well-known open problem in queuing theory, whereas its homogeneous counterpart was solved almost four decades ago by Winston (1977) and Weber (1978). Our result therefore can be considered as an important step toward the general understanding of efficient policies for heterogeneous systems.

1.1.1. Transient Analysis. To establish the diffusion limit for the conventional JSQ policy in homogeneous systems, Eschenfeldt and Gamarnik (2018) employ a three-step approach: (i) constructing a truncated system that differs from the original only in that the queues cannot exceed a certain bound determined by the limiting dynamics of the original system, (ii) proving the diffusion limit of the truncated system via an appropriate

Skorohod mapping, and (iii) demonstrating that the difference between the truncated system and the original system vanishes as $n \rightarrow \infty$. We generalize this approach to heterogeneous systems. However, there are several technical challenges in doing so.

First, unlike homogeneous systems, the queue lengths in different server pools scale differently, resulting in a higher-dimensional state space and necessitating a more complex Skorohod mapping to establish weak convergence. Second, the truncated system used in our proof differs from the one used in Eschenfeldt and Gamarnik (2018). Specifically, the buffer sizes are not uniform across all queues in our truncated system: Although the queues in the fastest server pool can hold at most two jobs, the queues in the remaining pools are restricted to holding no more than one job. Finally, we show that under the SA-JSQ policy, the heterogeneous system exhibits a state space collapse, where the diffusion-scaled number of idle servers in the fastest $M - 1$ pools converges to zero for all finite times (Proposition 1). This result has no direct analogue in homogeneous systems.

1.1.2. Steady-State Analysis. The proof of the *interchange of limits* is carried in two essential steps: (i) by proving that the sequence of stationary distributions of the diffusion-scaled queue length process is tight (Theorem 2) and (ii) by proving that the limiting diffusion process is (exponentially) ergodic (Theorem 3). The proofs of both theorems use Stein’s method that involves applying the generator of the underlying Markov process to a suitably chosen Lyapunov function. The key challenge in this approach is to identify a suitable Lyapunov function that will yield appropriate bounds on the expected value of the queue lengths in stationarity. The Lyapunov function is derived as the solution to a partial differential equation obtained through a generator expansion approach. In this method, the generator is approximated via a Taylor expansion, and it is shown that the error terms vanish using the gradient bounds of the solution. In a recent work, Braverman (2020) used this approach to study the behavior of the JSQ scheme for homogeneous systems in the Halfin-Whitt regime.

Even though we follow the framework in Braverman (2020) closely to prove the interchange of limits, the complexity of the underlying state space again poses several challenges in the heterogeneous setting. First, to obtain a tractable Partial Differential Equation (PDE), we need to introduce a suitable lifting operator that projects the state of the Markov process into $\mathbb{R}^2$. In the homogeneous setting, this is done naturally by only considering the scaled total number of idle servers and the scaled total number of servers with two or more jobs. In the heterogeneous setting, however, we identify that the correct way to define the lifting operator is by setting the second dimension to the scaled number of servers with two or more jobs *only* in the fastest pool as opposed to that in the entire system. Once the PDE is obtained using this lifting operator, the next challenge is to solve it to find the Lyapunov function and the gradient bounds. The heterogeneous nature of the system results in a PDE different from the one used for homogeneous systems, and this results in a notable difference in the analysis required to solve the PDE. Moreover, the resulting error terms in the generator expansion are significantly more complicated in the heterogeneous setting, and one needs to prove additional state space collapse results to derive bounds on these terms.

1.1.3. Asymptotic Optimality. Finally, we prove the asymptotic optimality of the SA-JSQ scheme in terms of stochastically minimizing the diffusion-scaled total number of jobs and the diffusion-scaled number of waiting jobs in the steady-state in the Halfin-Whitt regime (Theorem 4). To do so, we first derive a lower bound on the performance achievable by any *admissible load balancing scheme* defined as a scheme that considers only the current queue lengths and server speeds to assign each incoming job in the original system (Proposition 3). This lower bound is established by coupling the original system with a modified system where servers are not attached to queues but are free to move between queues at any time, allowing faster servers to always serve longer queues. To complete the proof of asymptotic optimality, we finally prove that the modified system described above shares the same diffusion limit as the original system operating under the SA-JSQ scheme (Proposition 4). We note that the modified system we use in our proofs differs fundamentally from the $M/M/n$ system having one central queue used in the proofs of optimality in prior works of Armony (2005) and Tezcan (2008). This difference stems from the fact that the diffusion limit of an $M/M/n$ system with a single queue is one dimensional, whereas, in the parallel queue setting, both idle servers and queues with waiting jobs can coexist yielding a two-dimensional diffusion limit.

1.2. Related Literature

The JSQ policy was shown to minimize the average response time of jobs in finite systems comprising identical servers under the assumption of Poisson arrivals and exponential service times in Winston (1977). This result was then generalized to service time distributions having nondecreasing hazard rates in Weber (1978). Further extensions covered queues with state-dependent service rates, as demonstrated in Johri (1989), and systems with finite buffers and general batch arrivals, as explored in Hordijk and Koole (1990).

Recent works, such as Mukherjee et al. (2016) and Eschenfeldt and Gamarnik (2018), analyzed the fluid and diffusion limits of the JSQ scheme. In the context of the fluid limit, Mukherjee et al. (2016) showed that under a broader class of policies including the JSQ scheme, the proportion of servers with two or more jobs tends to zero as $n \rightarrow \infty$. In the Halfin-Whitt regime, characterized by the normalized arrival rate $\lambda^{(n)}$ varying with system size n as $\lambda^{(n)} = 1 - \beta/\sqrt{n}$ for some $\beta > 0$, Eschenfeldt and Gamarnik (2018) demonstrated that the diffusion-scaled process converges to a two-dimensional reflected OU process as $n \rightarrow \infty$. Braverman (2020) showed that the many-server limit and the limit $t \rightarrow \infty$ can be interchanged and the stationary distribution of the limiting reflected OU process can be used to approximate the stationary behavior of the load balancing policy in the homogeneous setting. To do so, he showed that the steady-state proportion of idle servers, as well as servers with precisely two jobs, scale as $O(\sqrt{n})$, and the proportion of servers with more than three jobs scales as $O(1)$. Gupta and Walton (2019) analyzed the JSQ scheme in the nondegenerate slowdown (NDS) regime introduced in Atar (2012), where $\lambda^{(n)} = 1 - \beta/n$ for some fixed $\beta > 0$ and established a diffusion limit for the total customer count process.

Note that the JSQ scheme faces challenges related to communication overhead. In addressing this issue, alternative scheduling schemes, like Pod, were suggested Vvedenskaya et al. (1996) and Mitzenmacher (1996), where, as a new job arrives, $d \geq 2$ servers are chosen uniformly at random from the set of all servers, and the job is routed to the server with the shortest queue length among those d servers. The problem of communication overhead is still there in the Pod scheme if d is large. This can be eliminated with schemes relying only on the knowledge of idle servers in the system. In the join-the-idle-queue (JIQ) scheme, first introduced by Lu et al. (2011), the dispatcher assigns the incoming arrival to an idle server (if available) chosen uniformly at random. Mukherjee et al. (2016) shows an equivalence between different load-balancing schemes in the diffusion regime.

Relatively few works have explored load balancing in heterogeneous systems despite their importance in practical system design. The Pod scheme for heterogeneous systems has been studied in several papers, including Zhou et al. (2017), Mukhopadhyay and Mazumdar (2015), and Izagirre and Makowski (2014). Although Izagirre and Makowski (2014) and Zhou et al. (2017) investigated the performance of the Pod policy under light and heavy traffic conditions for finite system sizes, respectively, Mukhopadhyay and Mazumdar (2015) explored its performance in the mean-field regime. It was demonstrated that the Pod scheme exhibits a reduced stability region in heterogeneous systems due to the infrequent sampling of faster servers. Subsequent works (Mukhopadhyay et al. 2016, Gardner et al. 2021) explored variations of the Pod scheme with the aim of enhancing its performance in heterogeneous systems while preserving the maximal stability region. Stolyar (2015a) analyzed the JIQ scheme in a heterogeneous setting, revealing that the average waiting time of jobs under the JIQ scheme tends to zero in the fluid limit.

The many-server systems with heterogeneous servers have received considerable attention in the $M/M/n$ system with a single global queue setting. Armony (2005) showed that in the Halfin-Whitt regime, the fastest-server-first (FSF) policy in which each job is assigned to fastest available server is asymptotically optimal in terms of stochastically minimizing the diffusion-scaled total number of jobs in the system. Tezcan (2008) considered a similar model with heterogeneous servers where instead of a single central queue, each server pool consisting of servers of the same processing speed has a dedicated queue and an arriving job must be routed irrevocably to one of the server pools at its arrival instant. He showed that the minimum expected delay-FSF (MED-FSF) routing policy in which a job is either assigned to a queue with the minimum expected delay or to an available server having the highest speed asymptotically minimizes the stationary distribution of the diffusion-scaled total number of jobs in the system in the Halfin-Whitt regime. In both works mentioned above, the diffusion-scaled total number of jobs in the system converges to the same one-dimensional diffusion process, unlike in our system, where the limit is a two-dimensional diffusion process. In a more recent line of work, Büke and Qin (2023) focused on multiserver queuing systems characterized by heterogeneous exponential servers and renewal arrivals where service rates are random and varies for each individual server and developed a framework has been to analyze the heavy traffic limit of these queues in a random environment, employing probability measure-valued stochastic processes.

1.3. Notation

Throughout the paper we use the following general notations. We denote the set of nonnegative integers as $\mathbb{Z}_+$ and the set $\{1, \dots, M\}$ by $[M]$. We use $\bar{\mathbb{R}}$ to denote the set of extended reals, that is, $\mathbb{R} \cup \{\infty\}$, and define the sets

$$\mathbb{M}_{M,\rho} = \left\{ (b_{j,i} \in \mathbb{R}, i \in \mathbb{Z}_+, j \in [M]) : \sum_{i \in \mathbb{Z}_+} \sum_{j \in [M]} \rho^{-i} |b_{j,i}| < \infty \right\},$$

and

$$\mathbb{M}_{M,\rho}^1 = \{(b_{j,i}) \in \mathbb{M}_{M,\rho} : \text{there exists a } b_1 \in \mathbb{R} \text{ such that } b_{j,1} = b_1 \text{ for all } j \in [M]\}.$$

We endow $\bar{\mathbb{R}}$ with the order topology, where the neighborhoods of $\{\infty\}$ are the sets that contain a set $\{x > a\}$ for some $a \in \mathbb{R}$. We denote all multidimensional terms, for example, vectors, matrices, and elements of $\mathbb{M}_{M,\rho}$, with boldface letters. We also define $\mathbb{D}_{\mathbb{S}}[0, \infty)$ as the set of all cadlag functions $f : [0, \infty) \rightarrow \mathbb{S}$. We denote the uniform norm $\|\mathbf{x}\|_t = \sup_{0 \leq s \leq t} |\mathbf{x}(s)|$ for $\mathbf{x} \in \mathbb{D}_{\mathbb{S}}[0, \infty)$, where the $|\cdot|$ is considered to be the norm defined on space $\mathbb{S}$. For any $x \in \mathbb{R}$, we define $(x)_+ = \max\{x, 0\}$. We assume that all random elements used in this work are defined on the probability space $\{\tilde{\Omega}, \mathcal{F}, \mathbb{P}\}$ and $\mathbb{E}[\cdot]$ denotes the expectation with respect to the probability measure $\mathbb{P}$. We use uppercase letters to denote random elements in this probability space and corresponding lowercase letters to denote their realizations. For any stochastic process $X = (X(t), t \geq 0)$, $X(\infty)$ denotes a random variable that follows the stationary distribution of the corresponding process. We use $\Rightarrow$ to denote weak convergence and $\xrightarrow{p}$ to denote convergence in probability. For any two nonnegative real-valued functions f and g both depending on n , we write $f(n) = O(g(n))$ when $\limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} < \infty$ and $f(n) = o(g(n))$ when $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$.

2. System Model

We consider a sequence of queuing systems, where the n th system consists of n parallel servers, each with its own queue of infinite buffer size. Jobs are assumed to arrive at the n th system according to a Poisson process with rate $n\lambda^{(n)}$, where $\lambda^{(n)} < 1$ for all n . In this work, we consider the Halfin-Whitt regime, where

$$\lim_{n \rightarrow \infty} \sqrt{n}(1 - \lambda^{(n)}) = \beta, \quad (\text{A1})$$

for some constant $\beta > 0$. Upon arrival, each job is assigned to a server where it either receives service immediately (if the server is idle at that instant) or waits in the corresponding queue to be served later according to the first-come-first-served (FCFS) scheduling discipline. The servers are assumed to be *heterogeneous*; that is, there are M different server types and the service time of a job depends on the type of the server it is being served. In the n th system, there are $N_j^{(n)}$ servers of type $j \in [M]$ and the service time of a job being served by a server of type j is exponentially distributed with rate μ_j . Hence,

$$\sum_{j=1}^M N_j^{(n)} = n, \text{ for all } n \in \mathbb{Z}_+,$$

and we assume that there exists $\gamma_j > 0$ for each $j \in [M]$ such that

$$\lim_{n \rightarrow \infty} \sqrt{n} \left| \frac{N_j^{(n)}}{n} - \gamma_j \right| = 0. \quad (\text{A2})$$

Furthermore, without loss of generality we assume that the normalized system capacity is unity, that is, $\sum_{j=1}^M \mu_j \gamma_j = 1$ and the server speeds satisfy $\mu_1 > \mu_2 > \dots > \mu_M$.

The classical JSQ policy is known to yield an asymptotically optimal performance for systems with homogeneous servers. Our main interest is to analyze a modified version of the JSQ policy that we refer as the *speed-aware JSQ (SA-JSQ)* policy, and it is defined as follows.

Definition 1. Under the SA-JSQ policy, jobs are sent to a server with the minimum queue length among all the servers in the system upon arrival. Ties between servers of different types are broken by choosing the server type with the maximum speed and ties between servers of the same type are broken uniformly at random.

A load balancing scheme or policy Π is called admissible if it considers only the present state of the system, that is, the state and the queue sizes of each server, and the server speeds to dispatch jobs arriving at time $t \geq 0$. Let $\mathbf{Q}^{(n,\Pi)}(t) = (Q_{j,i}^{(n,\Pi)}(t), j \in [M], i \geq 1)$ denote the system occupancy state at time $t \geq 0$, where $Q_{j,i}^{(n,\Pi)}(t)$ is the number of type j servers with at least i jobs in the n th system under policy Π . It is clear that $Q_{j,0}^{(n,\Pi)}(t) = N_j^{(n)}$ for all $t \geq 0$ and $j \in [M]$. Moreover, it is easy to verify that the process $\mathbf{Q}^{(n,\Pi)} = (\mathbf{Q}^{(n,\Pi)}(t) : t \in \mathbb{R}_+)$ is Markov and takes values in the space $S^{(n)}$ defined as

$$S^{(n)} \triangleq \{\mathbf{s} \in \mathbb{Z}_+^{M \times \infty} : N_{j,1}^{(n)} \geq s_{j,1} \geq s_{j,2} \geq \dots \geq 0, \text{ for all } j \in [M]\}.$$

We also define the scaled process $\mathbf{Y}^{(n,\Pi)} = (\mathbf{Y}^{(n,\Pi)}(t) : t \geq 0)$, where

$$\mathbf{Y}^{(n,\Pi)}(t) = (Y_{j,i}^{(n,\Pi)}(t), j \in [M], i \geq 1)$$

with

$$Y_{j,1}^{(n,\Pi)}(t) = \frac{Q_{j,1}^{(n,\Pi)}(t) - N_j^{(n)}}{\sqrt{n}}, \quad Y_{j,i}^{(n,\Pi)}(t) = \frac{Q_{j,i}^{(n,\Pi)}(t)}{\sqrt{n}}, \quad i \geq 2, j \in [M]. \quad (2.1)$$

Moreover, we also define a process $Y_{[j_1, j_2], i}^{(n,\Pi)} = (Y_{[j_1, j_2], i}^{(n,\Pi)}(t) : t \geq 0)$, where

$$Y_{[j_1, j_2], i}^{(n,\Pi)}(t) = \sum_{j=j_1}^{j_2} Y_{j,i}^{(n,\Pi)}(t).$$

Our main focus in this work is the analysis of load-balancing systems under the SA-JSQ policy. To simplify the notation, we omit the superscript Π when we consider the state of the system under the SA-JSQ policy.

3. Main Results

In this section, we state our main results and their consequences. We also provide intuitions for these results. The subsequent sections are devoted to the detailed proofs of these results.

We start with presenting the state space collapse result in Proposition 1 needed to prove the convergence of the stochastic processes. The proposition establishes that the diffusion-scaled number of idle servers in the fastest $M-1$ pools remains close to zero for all large n and for all finite times. In other words, to keep track of the diffusion-scaled number of idle servers in the entire system in the limit as $n \rightarrow \infty$, it is sufficient to only keep track of the diffusion-scaled number of idle servers in the slowest pool (pool M).

Proposition 1. Suppose that (A1) and (A2) hold and $Y_{[1, M-1], 1}^{(n)}(0) \Rightarrow Y_{[1, M-1], 1}(0)$, where $Y_{[1, M-1], 1}(0)$ is a proper random variable. Then for any $\epsilon > 0$, $t_0 > 0$ and $T \geq t_0$, we have

$$\mathbb{P} \left(\sup_{s \in [t_0, T]} |Y_{[1, M-1], 1}^{(n)}(s)| > \epsilon \right) \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (3.1)$$

Moreover, if $Y_{[1, M-1], 1}^{(n)}(0) \Rightarrow 0$, then (3.1) also holds for $t_0 = 0$.

Proposition 1 states that, in the limit, almost all servers in the first $M-1$ pools should immediately become busy, even when there are idle servers in these pools at time $t = 0$. This result can be intuitively explained by the fact that the total normalized capacity available in the fastest $M-1$ pools is strictly less than one, whereas the normalized arrival rate $\lambda^{(n)}$ into the system tends to one as $n \rightarrow \infty$. Consequently, in the limit as $n \rightarrow \infty$, all servers in the fastest $M-1$ pools are used up by the SA-JSQ scheme as it prefers faster idle servers over slower ones. This can also be seen numerically in Figure 1, where we plot the diffusion-scaled idle servers in the fastest pool as a function of time for different values of n when $M=2$. It is apparent from Figure 1 that the fluctuations of the component $Y_{1,1}^{(n)}$ decreases to zero as n increases.

We now establish the limiting transient behavior of the process $Y^{(n)}$ as $n \rightarrow \infty$. To state the result, we assume that $Y_{j,1}^{(n)}(0) \Rightarrow 0$ for all $j \in [M-1]$, motivated by Proposition 1.

Figure 1. Sample Paths for $Y_{1,1}^{(n)}$ for Varying n with $M=2$, $\gamma_1 = 1/5$, $\gamma_2 = 4/5$, $\mu_1 = 20/8$, $\mu_2 = 5/8$, and $\beta=2$

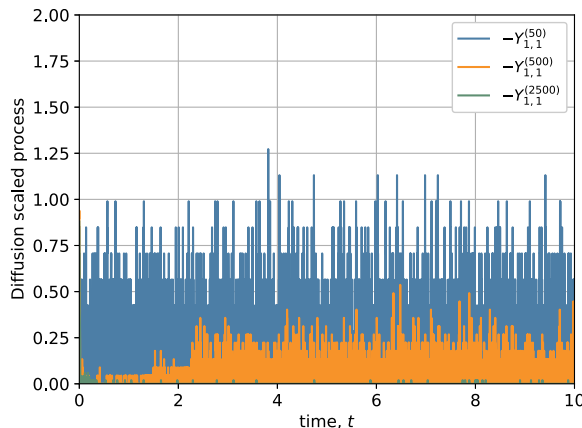
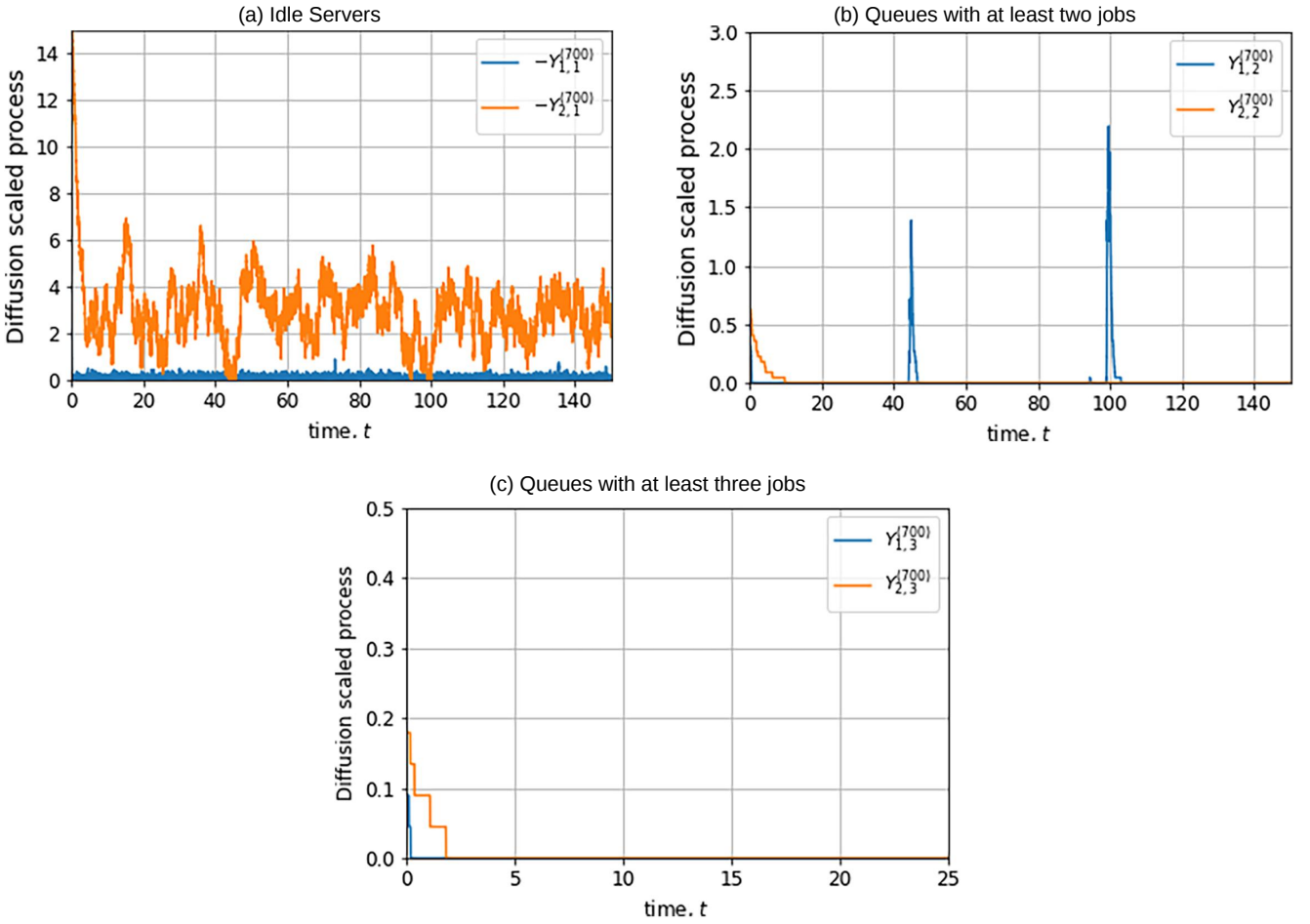


Figure 2. Sample Paths for $Y_{j,i}^{(n)}$ with $n=700, M=2, \gamma_1 = 1/5, \gamma_2 = 4/5, \mu_1 = 20/8, \mu_2 = 5/8$, and $\beta=2$



Theorem 1. Suppose that (A1) and (A2) hold and that there exists a random vector $\mathbf{Y}(0) = (Y_{j,i}(0), i \geq 1, j \in [M]) \in \mathbb{M}_{M,\rho}$ for some $\rho > 1$ with $Y_{j,1}(0) = 0$ for all $j \in [M-1]$, such that $\mathbf{Y}^{(n)}(0) \Rightarrow \mathbf{Y}(0)$ as $n \rightarrow \infty$. Then, we have $\mathbf{Y}^{(n)} \Rightarrow \mathbf{Y}$ as $n \rightarrow \infty$, where $\mathbf{Y} = (Y_{j,i}, i \geq 1, j \in [M])$ is the unique solution in $\mathbb{D}_{\mathbb{M}_{M,\rho}}[0, \infty)$ of the following stochastic integral equations:

$$Y_{j,1}(t) = 0, j \in \{1, \dots, M-1\}, \quad (3.2)$$

$$Y_{M,1}(t) = Y_{M,1}(0) + \sqrt{2}W(t) - \beta t - \int_0^t (\mu_M Y_{M,1}(s) - \mu_1 Y_{1,2}(s))ds + \sum_{j=2}^M \int_0^t \mu_j Y_{j,2}(s)ds - U_1(t), \quad (3.3)$$

$$Y_{1,2}(t) = Y_{1,2}(0) - \mu_1 \int_0^t (Y_{1,2}(s) - Y_{1,3}(s))ds + U_1(t), \quad (3.4)$$

$$Y_{j,2}(t) = Y_{j,2}(0) - \mu_j \int_0^t (Y_{j,2}(s) - Y_{j,3}(s))ds, j \in \{2, \dots, M\}, \quad (3.5)$$

$$Y_{j,i}(t) = Y_{j,i}(0) - \mu_j \int_0^t (Y_{j,i}(s) - Y_{j,i+1}(s))ds, i \geq 3, j \in [M], \quad (3.6)$$

where W is a standard Brownian motion, and U_1 is the unique nondecreasing and nonnegative process satisfying

$$\int_0^\infty \mathbb{1}(Y_{M,1}(t) < 0) dU_1(t) = 0. \quad (3.7)$$

Equations (3.5) and (3.6) imply that the components $Y_{j,2}$ for $j \in \{2, \dots, M\}$ and $Y_{j,i}$ for $i \geq 3, j \in [M]$ decreases to zero exponentially fast as $t \rightarrow \infty$. More precisely, the explicit solution for these components can be written as

$$Y_{j,i}(t) = e^{-t} \left\{ Y_{j,i}(0) + \sum_{k=1}^\infty \frac{1}{k!} t^k Y_{j,i+k}(0) \right\}. \quad (3.8)$$

Hence, if $Y_{j,2}(0) = 0$ for $j \in \{2, \dots, M\}$ and $Y_{j,i}(0) = 0$ for $j \in [M]$ and $i \geq 3$, then $Y_{M,1}$ and $Y_{1,2}$ are the only nonzero components of the limiting diffusion process $\mathbf{Y}$, and Equations (3.3) and (3.4) reduce to

$$Y_{M,1}(t) = Y_{M,1}(0) + \sqrt{2}W(t) - \beta t + \int_0^t (\mu_1 Y_{1,2}(s) - \mu_M Y_{M,1}(s)) ds - U_1(t), \quad (3.9)$$

$$Y_{1,2}(t) = Y_{1,2}(0) - \mu_1 \int_0^t Y_{1,2}(s) ds + U_1(t). \quad (3.10)$$

Equations (3.9) and (3.10) define a two-dimensional reflected OU process where the regulator process U_1 ensures that $Y_{M,1}$ remains nonpositive and $Y_{1,2}$ remains nonnegative at all times. Furthermore, we note that $Y_{1,2}$ can only increase when $Y_{M,1}$ is zero (see Figure 2), a feature of any JSQ-type policy that ensures that jobs are sent to queues having an existing job only when there are no idle servers in the system. Furthermore, both $Y_{M,1}$ and $Y_{1,2}$ can be nonzero at the same time. This is because even in the limit it can happen that jobs wait in queues belonging to the fastest pool when there are idle servers in the slowest server pool. This is a fundamental drawback of having a separate queue for each server whose speed does not scale with n . In contrast, for systems where each queue is served by $O(n)$ servers of constant speed, in the limit as $n \rightarrow \infty$, no job has to wait in a queue while there are idle servers available. This yields a one dimensional diffusion limit as in Armony (2005) and Tezcan (2008).

Furthermore, the presence of the term $\mu_1 Y_{1,2}(s)$ in (3.9) can be intuitively explained as follows. In the Halfin-Whitt limit each departure from the fastest $M-1$ pools is immediately compensated by an arrival because in the limit, the arrival rate of jobs is strictly above the rate at which the servers in the fastest $M-1$ pool can process jobs. This means that the moment a job leaves a queue with two jobs in the fastest pool a new job arrives that must be routed to an idle queue of the slowest pool because no idle queue exists in the fastest $M-1$ pools in the diffusion limit.

The stability of a load-balancing system operating under the SA-JSQ scheme for all $\lambda^{(n)} < 1$ was established in Bhambay and Mukhopadhyay (2022). In our subsequent results, we investigate the steady-state characteristics of the process $\mathbf{Y}^{(n)}$. We first establish the tightness of the sequence $(\mathbf{Y}^{(n)}(\infty))_{n \in \mathbb{Z}_+}$ in Theorem 2.

Theorem 2. Suppose that (A1) and (A2) hold. Then, there exist positive constants C_1 , C_2 , and C_3 such that for all sufficiently large n the following inequalities hold:

$$\mathbb{E}[|Y_{j,1}^{(n)}(\infty)|] < C_1 n^{-1/2} \text{ for all } 1 \leq j \leq M-1, \quad (3.11)$$

$$\mathbb{E}[|Y_{M,1}^{(n)}(\infty)|] \leq C_2, \quad (3.12)$$

$$\mathbb{E}[|Y_{1,2}^{(n)}(\infty)|] \leq C_3. \quad (3.13)$$

Furthermore, for any $a \in (1/2, 1)$ there exists a constant $C_4(a)$ such that for all sufficiently large n we have

$$\mathbb{E}[|Y_{j,i}^{(n)}(\infty)|] \leq C_4(a) n^{-(a-1/2)} \text{ for all } 2 \leq j \leq M, i = 2 \text{ and for all } 1 \leq j \leq M, i \geq 3. \quad (3.14)$$

Theorem 2 implies that the expected number of idle servers in the fastest $M-1$ pools are bounded above by constants independent of n . The expected number of servers with at least two jobs in pool 1 and the expected number of idle servers in pool M are $O(\sqrt{n})$. Moreover, the expected number of servers with at least three jobs in all pools and the expected number of servers with at least two jobs in the slowest $M-1$ pools all approaches to zero faster than $n^{-(a-1/2)}$ for any $1/2 < a < 1$. As mentioned before, these results can be intuitively explained by the facts that (i) the spare capacity $n(1 - \lambda^{(n)})$ in the Halfin-Whitt regime scales as $O(\sqrt{n})$, and (ii) this spare capacity is mostly available through the idle servers in the slowest pool because the SA-JSQ scheme prefers faster servers over slower ones whenever they have the same queue lengths. By a simple application of the Markov inequality, it follows from the above theorem that the sequence $(\mathbf{Y}^{(n)}(\infty))_{n \in \mathbb{Z}_+}$ is tight.

Remark 1. From Theorem 2 we can compute a bound on the probability with which a job has to wait in a queue before being processed in the steady state. This probability is given by $\mathbb{P}(\sum_{j \in [M]} Q_{j,1}^{(n)}(\infty) = n) \leq \mathbb{P}(Q_{M,1}^{(n)}(\infty) = N_M^{(n)}) \leq \mathbb{E}[Q_{M,1}^{(n)}]/N_M^{(n)} = O(1/\sqrt{n})$, where the last equality follows by (3.12) and (A2). Thus, the stationary waiting probability of a job goes to zero at rate $O(1/\sqrt{n})$.

Next, we establish the positive recurrence of the limiting diffusion process $\mathbf{Y}$. This is the final ingredient needed to establish the interchange of limits as stated in Corollary 1.

Theorem 3. *The diffusion process $\mathbf{Y}$ defined in Theorem 1 is positive recurrent.*

Corollary 1. *Let $\pi^{(n)}$ be the stationary distribution of the process $\mathbf{Y}^{(n)}$ and let π denote the stationary distribution of $\mathbf{Y}$. Then, $\pi^{(n)} \Rightarrow \pi$.*

Hence, by Corollary 1, we can conclude that the stationary distribution of the diffusion-scaled queue length process $\mathbf{Y}^{(n)}$ can be well approximated by that of the limiting diffusion process $\mathbf{Y}$ for large n .

In our final result, we focus on the asymptotic optimality of the SA-JSQ scheme in the Halfin-Whitt regime. Note that optimality can be defined in various ways. In the present paper, we are interested in asymptotic optimality, which is defined in terms of minimizing the steady-state distribution of the diffusion-scaled total number of jobs and the diffusion-scaled number of waiting jobs in the system as in Armony (2005) and Tezcan (2008). To define a similar notion of optimality we define

$$Y_{+1}^{(n)}(t) = \frac{\sum_{j \in [M]} \sum_{i \geq 1} Q_{j,i}^{(n)}(t) - n}{\sqrt{n}}, \quad \text{and} \quad Y_{+2}^{(n)}(t) = \sum_{j \in [M]} \sum_{i \geq 2} \frac{Q_{j,i}^{(n)}(t)}{\sqrt{n}} \quad (3.15)$$

to be, respectively, the diffusion-scaled total number of customers in the system and the diffusion-scaled number of waiting jobs in the system at time $t \in [0, \infty]$ under the SA-JSQ policy. The same quantities under a stable admissible policy Π are denoted by $Y_{+1}^{(n, \Pi)}(t)$ and $Y_{+2}^{(n, \Pi)}(t)$, respectively. We are now ready to state our asymptotic optimality result.

Theorem 4. *Assume (A1) and (A2) hold. Then, for any stable admissible policy Π , we have*

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_{+1}^{(n)}(\infty) > y) \leq \liminf_{n \rightarrow \infty} \mathbb{P}(Y_{+1}^{(n, \Pi)}(\infty) > y), \quad (3.16)$$

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_{+2}^{(n)}(\infty) > y) \leq \liminf_{n \rightarrow \infty} \mathbb{P}(Y_{+2}^{(n, \Pi)}(\infty) > y), \quad (3.17)$$

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

Theorem 4 ensures that the SA-JSQ scheme asymptotically minimizes the steady-state distribution of the diffusion-scaled total number of jobs and the diffusion-scaled number of waiting jobs in the system in the Halfin-Whitt regime.

The remainder of the paper is organized as follows. The proof of the state space collapse result, stated in Proposition 1, is provided in Section 4. Section 5 is dedicated to the proof of Theorem 1. Theorem 2 is proved in Section 6. One essential component of the proof in Section 6 is finding a suitable Lyapunov function. We do this in Section 6.1. In Section 7, we prove the asymptotic optimality of the SA-JSQ scheme, as stated in Theorem 4. Finally, we conclude the paper in Section 8.

4. State Space Collapse: Proof of Proposition 1

In this section, we establish the state space collapse result in load-balancing systems under the SA-JSQ scheme. To prove Proposition 1, we first express the evolution of $Y_{[1, M-1], 1}^{(n)}$ in terms of arrival and departure processes. We then examine the system's behavior in two parts: first, from $t = 0$ up to the point where $Y_{[1, M-1], 1}^{(n)}$ reaches zero (defined as $\theta_0^{(n)}$ below), and second, the behavior after hitting zero. We show that once $Y_{[1, M-1], 1}^{(n)}$ hits zero, it stays very close to zero, that is, the probability that it will deviate by an amount ϵ approaches 0 as $n \rightarrow \infty$. Finally, we show that the time $\theta_0^{(n)}$ to hit zero also converges to zero in probability. The following lemma from Pang et al. (2007) is key in our proofs of Proposition 1 and Theorem 1.

Lemma 1 (Lemma 5.8 in Pang et al. 2007). *Suppose that, for each $n \geq 1$, $M_n = (M_n(t) : t \geq 0)$ is a square-integrable martingale (with respect to a specified filtration) with predictable quadratic variation $\langle M_n \rangle = (\langle M_n \rangle(t) : t \geq 0)$, that is, $M_n^2 - \langle M_n \rangle = (M_n^2(t) - \langle M_n \rangle(t) : t \geq 0)$ is a martingale by the Doob-Meyer decomposition. If the sequence of random variables $\{\langle M_n \rangle(T)\}_{n \geq 1}$ is tight for each $T \geq 0$, then the sequence of stochastic processes $\{M_n : n \geq 1\}$ is stochastically bounded in $\mathbb{D}_{\mathbb{R}}[0, \infty)$ that is, the sequence $\{\sup_{0 \leq t \leq T} |M_n(t)|\}_{n \geq 1}$ is tight for every $T \geq 0$.*

Proof of Proposition 1. We first define

$$\theta_0^{(n)} = \inf\{t \geq 0 : Y_{[1, M-1], 1}^{(n)}(t) = 0\}.$$

Now, we can write (3.1) as

$$\begin{aligned} \mathbb{P}\left(\sup_{s \in [t_0, T]} |Y_{[1, M-1], 1}^{(n)}(s)| > \epsilon\right) &= \mathbb{P}\left(\sup_{s \in [t_0, T]} -Y_{[1, M-1], 1}^{(n)}(s) > \epsilon, \theta_0^{(n)} \leq t_0\right) + \mathbb{P}\left(\sup_{s \in [t_0, T]} -Y_{[1, M-1], 1}^{(n)}(s) > \epsilon, \theta_0^{(n)} > t_0\right) \\ &\leq \mathbb{P}\left(\sup_{s \in [t_0, T]} -Y_{[1, M-1], 1}^{(n)}(s) > \epsilon, \theta_0^{(n)} \leq t_0\right) + \mathbb{P}(\theta_0^{(n)} > t_0) \\ &\leq \mathbb{P}\left(\sup_{s \in [\theta_0^{(n)}, T]} -Y_{[1, M-1], 1}^{(n)}(s) > \epsilon\right) + \mathbb{P}(\theta_0^{(n)} > t_0). \end{aligned} \quad (4.1)$$

To prove (3.1), it is sufficient to show that each term on the right-hand side of (4.1) converges to zero as $n \rightarrow \infty$. For a given state $\mathbf{q} \in S^{(n)}$, the process $Y_{[1, M-1], 1}^{(n)}$ decreases by $1/\sqrt{n}$ with rate $\sum_{j \in [M-1]} \mu_j(q_{j,1} - q_{j,2})$, and it increases by $1/\sqrt{n}$ with rate $n\lambda^{(n)} \mathbb{1}(y_{[1, M-1], 1} < 0)$. Let A and D be unit rate Poisson processes and define square integrable martingales

$$\hat{M}_A^{(n)}(t) = \frac{A(n\lambda^{(n)}t) - n\lambda^{(n)}t}{\sqrt{n}}, \quad (4.2)$$

$$\hat{M}_D^{(n)}(t) = \frac{1}{\sqrt{n}} \left\{ D\left(\sum_{j \in [M-1]} \mu_j N_j^{(n)} t\right) - \sum_{j \in [M-1]} \mu_j N_j^{(n)} t \right\}. \quad (4.3)$$

For any $0 \leq t \leq \theta_0^{(n)}$, we can write

$$\begin{aligned} -Y_{[1, M-1], 1}^{(n)}(t) &= -Y_{[1, M-1], 1}^{(n)}(0) - \frac{A\left(n\lambda^{(n)} \int_0^t \mathbb{1}(Y_{[1, M-1], 1}^{(n)}(s) < 0) ds\right)}{\sqrt{n}} + \frac{1}{\sqrt{n}} D\left[\int_0^t \sum_{j \in [M-1]} \mu_j (Q_{j,1}^{(n)}(s) - Q_{j,2}^{(n)}(s)) ds\right] \\ &= -Y_{[1, M-1], 1}^{(n)}(0) - \frac{A(n\lambda^{(n)}t)}{\sqrt{n}} + \frac{1}{\sqrt{n}} D\left[\int_0^t \sum_{j \in [M-1]} \mu_j (Q_{j,1}^{(n)}(s) - Q_{j,2}^{(n)}(s)) ds\right] \\ &\leq -Y_{[1, M-1], 1}^{(n)}(0) - \frac{A(n\lambda^{(n)}t)}{\sqrt{n}} + \frac{1}{\sqrt{n}} D\left(\sum_{j \in [M-1]} \mu_j N_j^{(n)} t\right) \\ &= -Y_{[1, M-1], 1}^{(n)}(0) - \hat{M}_A^{(n)}(t) - \frac{n\lambda^{(n)}t}{\sqrt{n}} + \hat{M}_D^{(n)}(t) + \frac{\sum_{j \in [M-1]} \mu_j N_j^{(n)} t}{\sqrt{n}} \\ &= -Y_{[1, M-1], 1}^{(n)}(0) - \hat{M}_A^{(n)}(t) + \hat{M}_D^{(n)}(t) - c(t, n), \end{aligned} \quad (4.4)$$

where

$$c(t, n) = \sqrt{n}t \left(\lambda^{(n)} - n^{-1} \sum_{j \in [M-1]} N_j^{(n)} \mu_j \right). \quad (4.5)$$

Here, the second equality follows as for any $0 \leq s \leq \theta_0^{(n)}$, we have $\mathbb{1}(Y_{[1, M-1], 1}^{(n)}(s) < 0) = 1$, and the inequality follows from the fact that $Q_{j,1}^{(n)} \leq N_j^{(n)}$. Assumptions (A1) and (A2) imply that $\lambda^{(n)} \rightarrow 1$ and $n^{-1} \sum_{j \in [M-1]} N_j^{(n)} \mu_j \rightarrow 1 - \gamma_M \mu_M$, and hence, we have $c(t, n) = O(\sqrt{n})$ for any $t \geq 0$. For large enough n , we observe that for any $t_0 > 0$, we can write

$$\begin{aligned} \mathbb{P}(\theta_0^{(n)} > t_0) &\leq \mathbb{P}\left(\sup_{0 \leq t \leq t_0} Y_{[1, M-1], 1}^{(n)}(t) < 0\right) \\ &\leq \mathbb{P}\left(\sup_{0 \leq t \leq t_0} Y_{[1, M-1], 1}^{(n)}(0) + \hat{M}_A^{(n)}(t) - \hat{M}_D^{(n)}(t) + c(t, n) < 0\right) \\ &\leq \mathbb{P}(Y_{[1, M-1], 1}^{(n)}(0) + \hat{M}_A^{(n)}(t_0) - \hat{M}_D^{(n)}(t_0) < -c(t_0, n)) \\ &\leq \mathbb{P}(|Y_{[1, M-1], 1}^{(n)}(0)| \geq c(t_0, n)/3) + \mathbb{P}(|\hat{M}_A^{(n)}(t_0)| \geq c(t_0, n)/3) + \mathbb{P}(|\hat{M}_D^{(n)}(t_0)| \geq c(t_0, n)/3), \end{aligned} \quad (4.6)$$

where the second inequality follows from (4.4), and the last inequality follows from the union bound.

The optional quadratic variations $\langle \hat{M}_A^{(n)} \rangle(t) = \lambda^{(n)}t$ and

$$\langle \hat{M}_D^{(n)} \rangle(t) = \sum_{j \in [M-1]} \mu_j N_j^{(n)} t / n$$

are stochastically bounded due to Assumptions (A1) and (A2). Lemma 1 implies that the martingales $\hat{M}_A^{(n)}$ and $\hat{M}_D^{(n)}$ are also stochastically bounded. As $n \rightarrow \infty$, the sequence $\{Y_{[1,M-1],1}^{(n)}(0)\}_{n \geq 1}$ is weakly convergent. Furthermore, since $c(t_0, n) \rightarrow \infty$, each term on the right-hand side of (4.6) converges to zero.

Next, we define

$$\begin{aligned} \theta_\epsilon^{n,1} &= \inf\{s > \theta_0^{(n)} : -Y_{[1,M-1],1}^{(n)}(s) > \epsilon\}, \\ \theta_\epsilon^{n,2} &= \sup\{\theta_0^{(n)} \leq s < \theta_\epsilon^{n,1} : -Y_{[1,M-1],1}^{(n)}(s) < \epsilon/2\}. \end{aligned}$$

Now, using a similar reasoning as in (4.4), we can write

$$\begin{aligned} \mathbb{P}\left(\sup_{s \in [\theta_0^{(n)}, T]} -Y_{[1,M-1],1}^{(n)}(s) > \epsilon\right) &\leq \mathbb{P}(\theta_0^{(n)} < \theta_\epsilon^{n,2} < \theta_\epsilon^{n,1} < T) \\ &\leq \mathbb{P}\left(\sup_{\theta_\epsilon^{n,2} \leq t \leq T} \frac{1}{\sqrt{n}} D\left(\sum_{j \in [M-1]} \mu_j N_j^{(n)}(t - \theta_\epsilon^{n,2})\right) - \frac{A(n\lambda^{(n)}(t - \theta_\epsilon^{n,2}))}{\sqrt{n}} > \epsilon/2\right) \\ &\leq \mathbb{P}\left(\sup_{\theta_\epsilon^{n,2} \leq t \leq T} \hat{M}_D^{(n)}(t - \theta_\epsilon^{n,2}) - \hat{M}_A^{(n)}(t - \theta_\epsilon^{n,2}) - c(t - \theta_\epsilon^{n,2}, n) > \epsilon/2\right), \end{aligned} \quad (4.7)$$

where the second inequality follows because the term in the supremum is the change in the process $-Y_{[1,M-1],1}^{(n)}$ from time $\theta_\epsilon^{n,2}$ to time t . Again, using the fact that the martingales $\hat{M}_A^{(n)}$, $\hat{M}_D^{(n)}$ are stochastically bounded and $c(t - \theta_\epsilon^{n,2}, n)$ diverges to ∞ , the right-hand side converges to zero. When $Y_{[1,M-1],1}^{(n)}(0) \xrightarrow{p} 0$, the second claim of the proposition can be proved using the same argument by writing

$$\mathbb{P}\left(\sup_{s \in [0, T]} -Y_{[1,M-1],1}^{(n)}(s) > \epsilon\right) \leq \mathbb{P}(-Y_{[1,M-1],1}^{(n)}(0) > \epsilon/2) + \mathbb{P}(0 < \theta_\epsilon^{n,1} < T),$$

and replacing $\theta_\epsilon^{n,2}$ with zero in (4.7). This completes the proof. $\square$

5. Transient Analysis: Proof of Theorem 1

The proof of Theorem 1 relies on the continuous mapping theorem by showing that the system of integral equations in Proposition 2 defines a continuous mapping on $\mathbb{D}_{\mathbb{M}_M, \rho}[0, \infty)$. We first demonstrate that these equations possess a unique solution and hence correspond to a well-defined mapping. The proof of Proposition 2 can be found in Appendix A. As the next step, we introduce a truncated system and formulate the evolution equations in terms of martingales specific to this truncated system. We establish the convergence of the martingale components appearing in the evolution equations for this truncated system, and finally, we prove that both the truncated and untruncated systems exhibit similar behavior in the limit.

Proposition 2. Consider the following set of integral equations:

$$y_1(t) = b_1 + z_1(t) - \int_0^t (\mu_M y_1(s) - \mu_1 y_{1,2}(s)) ds + \sum_{j=2}^M \mu_j \int_0^t y_{j,2}(s) ds - u_1(t), \quad (5.1)$$

$$y_{1,2}(t) = b_{1,2} + z_{1,2}(t) - \int_0^t \mu_1 (y_{1,2}(s) - y_{1,3}(s)) ds + u_1(t) - u_2(t), \quad (5.2)$$

$$y_{j,2}(t) = b_{j,2} + z_{j,2}(t) - \int_0^t \mu_j (y_{j,2}(s) - y_{j,3}(s)) ds, \quad \forall j \in \{2, \dots, M\}, \quad (5.3)$$

$$y_{j,i}(t) = b_{j,i} + z_{j,i}(t) - \int_0^t \mu_j (y_{j,i}(s) - y_{j,i+1}(s)) ds, \quad i \geq 3, j \in [M], \quad (5.4)$$

$$y_1(t) \leq 0, 0 \leq y_{1,2}(t) \leq B, y_{j,i}(t) \geq 0, \text{ for all } i \geq 2, j \in [M], \quad (5.5)$$

where u_1 and u_2 are nondecreasing and nonnegative functions such that

$$\int_0^\infty \mathbb{1}(y_1(t) < 0) du_1(t) = 0, \quad (5.6)$$

$$\int_0^\infty \mathbb{1}(y_{1,2}(t) < B) du_2(t) = 0. \quad (5.7)$$

Given $\mathbf{b} \in \mathbb{M}_{M,\rho}^1$ and $\mathbf{z} \in \mathbb{D}_{\mathbb{M}_{M,\rho}^1}[0, \infty)$ the integral Equations (5.1)–(5.3) have a unique solution $(\mathbf{y}, \mathbf{u}) \in \mathbb{D}_{\mathbb{M}_{M,\rho}^1}[0, \infty) \times \mathbb{D}_{\mathbb{R}_+^2}[0, \infty)$. Moreover, there exists a well-defined function $(f, g) : \mathbb{R}_+ \times \mathbb{M}_{M,\rho}^1 \times \mathbb{D}_{\mathbb{M}_{M,\rho}^1}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{M}_{M,\rho}^1}[0, \infty) \times \mathbb{D}_{\mathbb{R}_+^2}[0, \infty)$ that maps $(B, \mathbf{b}, \mathbf{z})$ in to $\mathbf{y} = f(B, \mathbf{b}, \mathbf{z})$ and $\mathbf{u} = g(B, \mathbf{b}, \mathbf{z})$. Furthermore, the function (f, g) is continuous on $\mathbb{R}_+ \times \mathbb{M}_{M,\rho}^1 \times \mathbb{D}_{\mathbb{M}_{M,\rho}^1}[0, \infty)$. Finally, $\mathbf{y}$ being continuous implies that $\mathbf{x}$ and $\mathbf{u}$ are also continuous.

Next, we consider a truncated system where the queues with more than two jobs are not allowed to form in pool 1 and no queue with more than one job will form in pools $\{2, \dots, M\}$, even though there might be such queues initially present at time 0. We first analyze this system and then show that the truncated system and the original system are indistinguishable in the limit. For any $t \geq 0$, let $\hat{\mathbf{Q}}^{(n)}(t) = (\hat{Q}_{j,i}^{(n)}(t), i \geq 1, j \in [M])$, where $\hat{Q}_{j,i}^{(n)}(t)$ is the number of type j servers with at least i jobs in the truncated system. For all $t \geq 0$, we can represent the dynamics of the truncated system as

$$\hat{Q}_{1,1}^{(n)}(t) = \hat{Q}_{1,1}^{(n)}(0) + A(n\lambda^{(n)}t) - D_{1,1} \left(\mu_1 \int_0^t (\hat{Q}_{1,1}^{(n)}(s) - \hat{Q}_{1,2}^{(n)}(s)) ds \right) - \hat{U}_{1,1}^{(n)}(t), \quad (5.8)$$

$$\hat{Q}_{j,1}^{(n)}(t) = \hat{Q}_{j,1}^{(n)}(0) + \hat{U}_{j,1}^{(n)}(t) - D_{j,1} \left(\mu_j \int_0^t (\hat{Q}_{j,1}^{(n)}(s) - \hat{Q}_{j,2}^{(n)}(s)) ds \right), \quad 2 \leq j \leq M, \quad (5.9)$$

$$\hat{Q}_{1,2}^{(n)}(t) = \hat{Q}_{1,2}^{(n)}(0) + \hat{U}_{1,2}^{(n)}(t) - D_{1,2} \left(\mu_1 \int_0^t (\hat{Q}_{1,2}^{(n)}(s) - \hat{Q}_{1,3}^{(n)}(s)) ds \right), \quad (5.10)$$

$$\hat{Q}_{j,2}^{(n)}(t) = \hat{Q}_{j,2}^{(n)}(0) - D_{j,2} \left(\mu_j \int_0^t (\hat{Q}_{j,2}^{(n)}(s) - \hat{Q}_{j,3}^{(n)}(s)) ds \right), \quad j \in \{2, \dots, M\}, \quad (5.11)$$

$$\hat{Q}_{j,i}^{(n)}(t) = \hat{Q}_{j,i}^{(n)}(0) - D_{j,i} \left(\mu_j \int_0^t (\hat{Q}_{j,i}^{(n)}(s) - \hat{Q}_{j,i+1}^{(n)}(s)) ds \right), \quad i \geq 3, j \in [M], \quad (5.12)$$

where A and $D_{j,i}$ are mutually independent unit-rate Poisson processes and

$$\hat{U}_{1,1}^{(n)}(t) = \int_0^t \mathbb{1}(\hat{Q}_{1,1}^{(n)}(s) = N_1^{(n)}) dA(n\lambda^{(n)}s), \quad (5.13)$$

$$\hat{U}_{j,1}^{(n)}(t) = \int_0^t \mathbb{1}(\hat{Q}_{j,1}^{(n)}(s) = N_j^{(n)}, \forall l \in [j-1], \hat{Q}_{j,l}^{(n)}(s) < N_j^{(n)}) dA(n\lambda^{(n)}s), \quad 2 \leq j \leq M, \quad (5.14)$$

$$\hat{U}_{1,2}^{(n)}(t) = \int_0^t \mathbb{1}(\hat{Q}_{1,1}^{(n)}(s) = N_1^{(n)}, \forall j \in [M], \hat{Q}_{1,2}^{(n)}(s) < N_1^{(n)}) dA(n\lambda^{(n)}s). \quad (5.15)$$

It is easy to see that $\hat{U}_{1,1}^{(n)}(t)$ counts the number of arrivals that see all pool 1 servers busy upon arrival within the time interval $[0, t]$. Similarly, $\hat{U}_{j,1}^{(n)}(t)$ for $2 \leq j \leq M$ counts the number of arrivals that see all servers in pools in $[j-1]$ busy but finds at least one idle server in pool j within the time interval $[0, t]$. Finally, $\hat{U}_{1,2}^{(n)}(t)$ counts the number of arrivals that, upon arrival, sees all servers in the system busy but find some servers in pool 1 with only one job within time interval $[0, t]$.

Similar to (2.1), we define the diffusion-scaled version of the process $\hat{\mathbf{Q}}^{(n)}$ as

$$\hat{Y}_{j,1}^{(n)}(t) = \frac{\hat{Q}_{j,1}^{(n)}(t) - N_j^{(n)}}{\sqrt{n}}, \quad \hat{Y}_{j,i}^{(n)}(t) = \frac{\hat{Q}_{j,i}^{(n)}(t)}{\sqrt{n}}, \quad i \geq 2, j \in [M], \quad \text{and} \quad \hat{Y}_{[1,M],1}^{(n)}(t) = \sum_{j \in [M]} \hat{Y}_{j,1}^{(n)}(t).$$

In addition to $\hat{M}_A^{(n)}$ (defined in (4.2)) and $\hat{M}_D^{(n)}$, for any $i \geq 1$ and $j \in [M]$, we also define following martingales:

$$\hat{M}_{D,j,i}^{(n)}(t) = \frac{1}{\sqrt{n}} D_{j,i} \left(\mu_j \int_0^t (\hat{Q}_{j,i}^{(n)}(s) - \hat{Q}_{j,i+1}^{(n)}(s)) ds \right) - \frac{1}{\sqrt{n}} \int_0^t \mu_j (\hat{Q}_{j,i}^{(n)}(s) - \hat{Q}_{j,i+1}^{(n)}(s)) ds. \quad (5.16)$$

The martingale $\hat{M}_{D,j,i}^{(n)}(t)$ are square integrable martingale (lemma 3.2 in Pang et al. (2007)) with optional and predictable quadratic variations

$$[\hat{M}_{D,j,i}^{(n)}](t) = \frac{1}{n} D_{j,i} \left(\mu_j \int_0^t (\hat{Q}_{j,i}^{(n)}(s) - \hat{Q}_{j,i+1}^{(n)}(s)) ds \right)$$

$$\langle \hat{M}_{D,j,i}^{(n)} \rangle(t) = \frac{1}{n} \mu_j \int_0^t (\hat{Q}_{j,i}^{(n)}(s) - \hat{Q}_{j,i+1}^{(n)}(s)) ds.$$

Combining (5.8), and (5.9), we have

$$\begin{aligned} \hat{Y}_{[1,M],1}^{(n)}(t) &= \hat{Y}_{[1,M],1}^{(n)}(0) + \hat{M}_A^{(n)}(t) - \sum_{j \in [M]} \hat{M}_{D,j,1}^{(n)}(t) - \sqrt{n}(1 - \lambda^{(n)})t \\ &\quad - \sum_{j \in [M]} \int_0^t \mu_j (\hat{Y}_{j,1}^{(n)}(s) - \hat{Y}_{j,2}^{(n)}(s)) ds - \frac{\hat{U}_{1,1}^{(n)}(t)}{\sqrt{n}} + \frac{\sum_{j=2}^M \hat{U}_{j,1}^{(n)}(t)}{\sqrt{n}}. \end{aligned}$$

Slightly rearranging the terms to obtain the form in (5.1), we get

$$\begin{aligned} \hat{Y}_{[1,M],1}^{(n)}(t) &= \hat{Y}_{[1,M],1}^{(n)}(0) + \hat{M}_A^{(n)}(t) - \sum_{j \in [M]} \hat{M}_{D,j,1}^{(n)}(t) - \sqrt{n}(1 - \lambda^{(n)})t - \sum_{j=1}^{M-1} \int_0^t \mu_j \hat{Y}_{j,1}^{(n)}(s) ds - \mu_M \int_0^t (\hat{Y}_{M,1}^{(n)}(s) - \hat{Y}_{[1,M],1}^{(n)}(s)) ds \\ &\quad - \int_0^t (\mu_M \hat{Y}_{[1,M],1}^{(n)}(s) - \mu_1 \hat{Y}_{1,2}^{(n)}(s)) ds + \sum_{j=2}^M \int_0^t \mu_j \hat{Y}_{j,2}^{(n)}(s) ds - \hat{V}_1^{(n)}(t), \end{aligned} \quad (5.17)$$

where

$$\hat{V}_1^{(n)}(t) = \frac{\hat{U}_{1,1}^{(n)}(t)}{\sqrt{n}} - \frac{\sum_{j=2}^M \hat{U}_{j,1}^{(n)}(t)}{\sqrt{n}}.$$

Similarly, using (5.10)–(5.12), we obtain

$$\hat{Y}_{1,2}^{(n)}(t) = \hat{Y}_{1,2}^{(n)}(0) - \hat{M}_{D,1,2}^{(n)}(t) - \int_0^t \mu_1 (\hat{Y}_{1,2}^{(n)}(s) - \hat{Y}_{1,3}^{(n)}(s)) ds + \hat{V}_1^{(n)}(t) - \hat{V}_2^{(n)}(t), \quad (5.18)$$

$$\hat{Y}_{j,2}^{(n)}(t) = \hat{Y}_{j,2}^{(n)}(0) - \hat{M}_{D,j,2}^{(n)}(t) - \int_0^t \mu_j (\hat{Y}_{j,2}^{(n)}(s) - \hat{Y}_{j,3}^{(n)}(s)) ds, \quad j \in \{2, \dots, M\}, \quad (5.19)$$

$$\hat{Y}_{j,i}^{(n)}(t) = \hat{Y}_{j,i}^{(n)}(0) - \hat{M}_{D,j,i}^{(n)}(t) - \int_0^t \mu_j (\hat{Y}_{j,i}^{(n)}(s) - \hat{Y}_{j,i+1}^{(n)}(s)) ds, \quad i \geq 3, j \in [M], \quad (5.20)$$

where

$$\hat{V}_2^{(n)}(t) = \hat{V}_1^{(n)}(t) - \frac{\hat{U}_{1,2}^{(n)}(t)}{\sqrt{n}}.$$

From (5.13) and (5.14), we see that $\hat{V}_1^{(n)}(t)$ is a nondecreasing function increasing only when an arrival occurs and $\hat{Q}_{j,1}^{(n)} = N_j^{(n)}$ for all $j \in [M]$; that is, for all $t \geq 0$,

$$\int_0^t \mathbb{1}(\hat{Y}_{[1,M],1}^{(n)}(s) < 0) d\hat{V}_1^{(n)}(s) = 0. \quad (5.21)$$

Also, by definition, $\hat{Y}_{1,2}^{(n)}(t) \leq N_1^{(n)}/\sqrt{n}$ and $\hat{V}_2^{(n)}(t)$ is a nondecreasing function and only increases when there is an arrival, $\hat{Q}_{j,1}^{(n)} = N_j^{(n)}$, for all $j \in [M]$ and $\hat{Q}_{1,2}^{(n)} = N_1^{(n)}$. Hence, for all $t \geq 0$,

$$\int_0^t \mathbb{1} \left(\hat{Y}_{1,2}^{(n)}(s) < \frac{N_1^{(n)}}{\sqrt{n}} \right) d\hat{V}_2^{(n)}(s) = 0. \quad (5.22)$$

Equations (5.17)–(5.22) have the same form as (5.1)–(5.7) with $B = B^{(n)}$, $z_1 = \hat{Z}_1^{(n)}$ and $z_{j,i} = \hat{Z}_{j,i}^{(n)}$, where

$$\begin{aligned} B^{(n)} &= N_1^{(n)} / \sqrt{n}, \\ \hat{Z}_1^{(n)}(t) &= \hat{M}_A^{(n)}(t) - \sum_{j \in [M]} \hat{M}_{D,j,1}^{(n)}(t) - \sqrt{n}(1 - \lambda^{(n)})t + \sum_{j=1}^{M-1} \int_0^t \mu_j \hat{Y}_{j,1}^{(n)}(s) ds - \mu_M \int_0^t (\hat{Y}_{M,1}^{(n)}(s) - \hat{Y}_{[1:M],1}^{(n)}(s)) ds, \\ \hat{Z}_{j,i}^{(n)}(t) &= -\hat{M}_{D,j,i}^{(n)}(t), \quad i \geq 2, j \in [M]. \end{aligned}$$

Once we show that these terms converge to the appropriate limits, we can use the continuous mapping theorem to prove Theorem 1. The convergence of the third term of $\hat{Z}_1^{(n)}(t)$ is implied by (3.16). Moreover, Proposition 1 implies that the fourth and the fifth terms converge to zero as $n \rightarrow \infty$. Hence, we focus on the limits of the martingales $\hat{M}_A^{(n)}(t)$ and $\hat{M}_{D,j,1}^{(n)}(t)$ in Section 5.1.

5.1. Martingale Convergence

The martingales $\hat{M}_A^{(n)}$ and $\hat{M}_{D,j,i}^{(n)}$ for all $i \in \mathbb{Z}_+$ and $j \in [M]$ are obtained by centering the time-changed unit Poisson processes with the corresponding random time changes. Hence, we can write the predictable quadratic variation process for the martingale $\hat{M}_A^{(n)}$ as

$$\langle \hat{M}_A^{(n)} \rangle(t) = \lambda^{(n)} t. \quad (5.23)$$

Assumption (A1) clearly implies that $\langle \hat{M}_A^{(n)} \rangle(t) \rightarrow t$ as $n \rightarrow \infty$. The following lemma characterizes the limit of $\langle \hat{M}_{D,j,i}^{(n)} \rangle(t)$.

Lemma 2. For any $j \in [M]$ as $n \rightarrow \infty$, we have

$$\frac{\hat{Q}_{j,1}^{(n)}}{n} \Rightarrow \gamma_j e, \quad (5.24)$$

where $e(t) = 1$ for all $t \geq 0$. Moreover, for $i \geq 2$ and for $j \in [M]$ as $n \rightarrow \infty$, we have

$$\frac{\hat{Q}_{j,i}^{(n)}}{n} \Rightarrow 0. \quad (5.25)$$

To prove Lemma 2, we need the following version of lemma 9 in Eschenfeldt and Gamarnik (2018) modified by incorporating the necessary coefficients. We provide its proof in Appendix B for completeness.

Lemma 3. For each $n \geq 1$, the following equations:

$$\begin{aligned} \hat{Y}_{[1,M],1}^{(n)}(t) &= \hat{Y}_{[1,M],1}^{(n)}(0) + \hat{Z}_1^{(n)}(t) - \int_0^t (\mu_M \hat{Y}_{[1,M],1}^{(n)}(s) - \mu_1 \hat{Y}_{1,2}^{(n)}(s)) ds - \hat{V}_1^{(n)}(t), \\ \hat{Y}_{1,2}^{(n)}(t) &= \hat{Y}_{1,2}^{(n)}(0) + \hat{Z}_{1,2}^{(n)}(t) - \int_0^t \mu_1 \hat{Y}_{1,2}^{(n)}(s) ds + \hat{V}_1^{(n)}(t) - \hat{V}_2^{(n)}(t), \\ \hat{Y}_{[1,M],1}^{(n)}(t) &\leq 0, \quad 0 \leq \hat{Y}_{1,2}^{(n)}(t) \leq B_n, \quad t \geq 0, \\ \int_0^t \mathbb{1}(\hat{Y}_{[1,M],1}^{(n)}(s) < 0) d\hat{V}_1^{(n)}(s) &= 0, \\ \int_0^t \mathbb{1}(\hat{Y}_{1,2}^{(n)}(s) < B_n) d\hat{V}_2^{(n)}(s) &= 0, \end{aligned}$$

have a unique solution $(\hat{Y}_{[1,M],1}^{(n)}, \hat{Y}_{1,2}^{(n)}, \hat{V}_1^{(n)}, \hat{V}_2^{(n)})$. Moreover, if the sequences $\{\hat{Y}_{[1,M],1}^{(n)}(0)\}_{n \geq 1}$, $\{\hat{Y}_{1,2}^{(n)}(0)\}_{n \geq 1}$, $\{\hat{Z}_1^{(n)}\}_{n \geq 1}$, and $\{\hat{Z}_{1,2}^{(n)}\}_{n \geq 1}$ are stochastically bounded, then the sequences $\{\hat{Y}_{[1,M],1}^{(n)}\}_{n \geq 1}$ and $\{\hat{Y}_{1,2}^{(n)}\}_{n \geq 1}$ are also stochastically bounded.

Proof of Lemma 2. Suppose $\{\hat{Y}_{[1,M],1}^{(n)}\}_n$ and $\{\hat{Y}_{j,i}^{(n)}\}_n$ for all $i \geq 2$ and $j \in [M]$ are stochastically bounded. Then by lemma 5.9 of Pang et al. (2007), we have

$$\frac{\hat{Y}_{[1,M],1}^{(n)}}{\sqrt{n}} = \frac{\sum_{j \in [M]} (\hat{Q}_{j,1}^{(n)} - N_j^{(n)})}{n} \Rightarrow 0 \text{ and } \frac{\hat{Y}_{j,i}^{(n)}}{\sqrt{n}} = \frac{\hat{Q}_{j,i}^{(n)}}{n} \Rightarrow 0.$$

Applying (A2) to the above proves the lemma. Hence, it remains to establish that $\{\hat{Y}_{[1,M],1}^{(n)}\}_n$ and $\{\hat{Y}_{j,i}^{(n)}\}_n$ for all $i \geq 2$ and $j \in [M]$ are stochastically bounded.

Because of truncation, we know that

$$\hat{Y}_{j,i}^{(n)}(t) = \hat{Q}_{j,i}^{(n)}(t)/\sqrt{n} \leq \hat{Q}_{j,i}^{(n)}(0)/\sqrt{n}.$$

This is for all $j \in [M]$ and $i \geq 3$ and $j \in \{2, \dots, M\}$ and $i \geq 2$. Hence, using Lemmas 1 and 3, we only need to prove that $\hat{Z}_1^{(n)}$ and $\hat{Z}_{1,2}^{(n)}$ are stochastically bounded, which follows if the respective predictive quadratic variations are stochastically bounded. From (5.23) it follows immediately that the sequence $\{\langle \hat{M}_A^{(n)} \rangle(t)\}_{n \geq 1}$ is bounded for each $t \geq 0$, and we have

$$\langle \hat{M}_{D,j,i}^{(n)} \rangle(t) \leq \frac{1}{n} \int_0^t \mu_j \hat{Q}_{j,i}^{(n)}(s) ds + \frac{1}{n} \int_0^t \mu_j \hat{Q}_{j,i+1}^{(n)}(s) ds \leq \frac{2\mu_j}{n} N_j^{(n)} t.$$

Hence, the stochastic boundedness of the sequence $\{\langle \hat{M}_{D,j,i}^{(n)} \rangle(t)\}_{n \geq 1}$ for each $t \geq 0$ follows from (A2) and Proposition 1. Hence, the lemma follows. $\square$

The following lemma characterizes the limits of the martingales $\hat{M}_A^{(n)}$ and $\hat{M}_{D,j,i}^{(n)}$ as $n \rightarrow \infty$.

Lemma 4. *The sequence of scaled martingales converges weakly to following limits as $n \rightarrow \infty$*

$$(\hat{M}_A^{(n)}, \hat{M}_{D,1,1}^{(n)}, \dots, \hat{M}_{D,M,1}^{(n)}, \dots, \hat{M}_{D,1,i}^{(n)}, \dots, \hat{M}_{D,M,i}^{(n)}, \dots) \Rightarrow (W_A, \mu_1 \gamma_1 W_{D,1,1}, \dots, \mu_M \gamma_M W_{D,M,1}, \dots, 0, \dots, 0, \dots),$$

where W_A and $W_{D,j,1}$ for $j \in [M]$ are independent standard Brownian motions.

To prove this lemma, we utilize the continuous mapping theorem and functional central limit theorem (FCLT) for Poisson processes, which we restate here for completeness.

Lemma 5 (Theorem 4.2 of Pang et al. 2007). *Let A , and $D_{j,i}$ (for $i \geq 1$ and $j \in [M]$) be independent unit-rate Poisson processes. For each such process C (i.e., $C = A$ or $C = D_{j,i}$), define*

$$\hat{M}_{C,n} = \frac{C(nt) - nt}{\sqrt{n}}.$$

Then, as $n \rightarrow \infty$, we have

$$(\hat{M}_{A,n}, \hat{M}_{D,1,1,n}, \dots, \hat{M}_{D,M,1,n}, \dots, \hat{M}_{D,1,i,n}, \dots, \hat{M}_{D,M,i,n}, \dots) \Rightarrow (W_A, W_{D,1,1}, \dots, W_{D,M,1}, \dots, W_{D,1,i}, \dots, W_{D,M,i}, \dots),$$

where W_A and $W_{D,j,i}$ are independent standard Brownian motions.

Proof of Lemma 4. Using the predictable quadratic variation processes, we can express the martingales $\hat{M}_A^{(n)}$ and $\hat{M}_{D,j,i}^{(n)}$ for $i \geq 1$ and $j \in [M]$ as the following compositions:

$$\hat{M}_A^{(n)} = \hat{M}_{A,n} \circ \langle \hat{M}_A^{(n)} \rangle, \hat{M}_{D,j,i}^{(n)} = \hat{M}_{D,j,i,n} \circ \langle \hat{M}_{D,j,i}^{(n)} \rangle, i \geq 1, j \in [M].$$

Now, applying Continuous Mapping Theorem (CMT) together with Lemma 5 and Lemma 2, we get

$$\begin{aligned} &(\hat{M}_A^{(n)}, \hat{M}_{D,1,1}^{(n)}, \dots, \hat{M}_{D,M,1}^{(n)}, \hat{M}_{D,1,2}^{(n)}, \dots, \hat{M}_{D,M,2}^{(n)}, \dots) \\ &\Rightarrow (W_A \circ e, W_{D,1,1} \circ \mu_1 \gamma_1 e, \dots, W_{D,M,1} \circ \mu_M \gamma_M e, W_{D,1,2} \circ 0, \dots, W_{D,M,2} \circ 0, \dots), \end{aligned}$$

which proves the lemma. $\square$

Now, we have all the necessary results to prove Theorem 1.

Proof of Theorem 1. Note that $B_n \rightarrow \infty$ as $n \rightarrow \infty$. Therefore, we have $\hat{V}_2^{(n)}$ converges to zero in the limit. Furthermore, Lemma 4, Proposition 1, and (A1) imply

$$\begin{aligned} &\hat{M}_A^{(n)}(t) - \sum_{j \in [M]} \hat{M}_{D,j,1}^{(n)}(t) - \sqrt{n}(1 - \lambda^{(n)})t + \sum_{j=1}^{M-1} \int_0^t \mu_j \hat{Y}_{j,1}^{(n)}(s) ds - \mu_M \int_0^t (\hat{Y}_{M,1}^{(n)}(s) - \hat{Y}_{[1,M],1}^{(n)}(s)) ds \\ &\Rightarrow W_A(t) - \sum_{j \in [M]} \mu_j \gamma_j W_{D,j,1}(t) - \beta t \\ &\stackrel{(d)}{=} \sqrt{2}W(t) - \beta t, \end{aligned}$$

where $\stackrel{(d)}{=}$ indicates equivalence in distribution, and W is the standard Wiener process. Moreover, Lemma 4 also implies that martingales terms in (5.19)–(5.20) converges weakly to zero as $n \rightarrow \infty$. Now using CMT, as $n \rightarrow \infty$, we have $\hat{Y}_{[1,M],1}^{(n)} \Rightarrow Y_{M,1}$, $\hat{Y}_{1,2}^{(n)} \Rightarrow Y_{1,2}$, $\hat{Y}_{j,2}^{(n)} \Rightarrow Y_{j,2}$, for $j \in \{2, \dots, M\}$ and $\hat{Y}_{j,i}^{(n)} \Rightarrow Y_{j,i}$ for $i \geq 3, j \in [M]$. Hence, for the truncated system, we have $\hat{\mathbf{Y}}^{(n)} \Rightarrow \mathbf{Y}$.

As the final step, we prove that both the truncated and the original system have the same limiting dynamics. To show this, we define

$$T_n = \inf\{t \geq 0 : Q_{1,2}^{(n)}(t) = N_1^{(n)}\}.$$

It is easy to see that starting from the same initial state at $t = 0$, the original system and that of the truncated system remain identical for all $t \in [0, T_n]$ because, in both systems, no arrival is sent to a busy server in the slowest $M - 1$ pools until all servers in the fastest pool have two jobs. Therefore, it suffices to prove that $\mathbb{P}(T_n \leq t) \rightarrow 0$, as $n \rightarrow \infty$ for all $t \geq 0$.

Because the truncated and the original systems are identical up to T_n , we have

$$\mathbb{P}(T_n \leq t) = \mathbb{P}\left(\sup_{0 \leq s \leq t} \hat{Q}_{1,2}^{(n)}(s) \geq N_1^{(n)}\right) = \mathbb{P}\left(\sup_{0 \leq s \leq t} \hat{Y}_{1,2}^{(n)}(s) \geq N_1^{(n)}/\sqrt{n}\right).$$

Hence, we obtain

$$\limsup_{n \rightarrow \infty} \mathbb{P}(T_n \leq t) \leq \limsup_n \mathbb{P}\left(\sup_{0 \leq s \leq t} \hat{Y}_{1,2}^{(n)}(s) \geq N_1^{(n)}/\sqrt{n}\right) = 0,$$

where the last equality follows because the sequence $\{\hat{Y}_{1,2}^{(n)}(s)\}_{n \in \mathbb{Z}_+}$ converges weakly and hence is stochastically bounded, and $\frac{N_1^{(n)}}{\sqrt{n}}$ is of the order of $O(\sqrt{n})$. This completes the proof. $\square$

6. Stationary Analysis: Proof of Theorem 2

The process level convergence in Theorem 1 only states convergence on compact time intervals and does not imply the convergence of stationary measures. To prove the interchangeability of many-servers limit and the limit as time goes to infinity and obtain the convergence of stationary distributions in Corollary 1, we need to prove the tightness of the stationary distributions for the scaled system sizes under consideration as stated in Theorem 2.

To prove (3.11) of Theorem 2, we establish through an appropriate Lyapunov function that $\mathbb{E}\left[\sum_{j=1}^{M-1} |Y_{j,1}^{(n)}(\infty)|\right] = O(1/\sqrt{n})$. Then, leveraging rate conservation and (3.11), we establish $\mathbb{E}[|Y_{M,1}^{(n)}(\infty)|] = O(1)$, thereby confirming (3.12). Additionally, employing Stein's method, we prove that $\mathbb{E}[Y_{1,2}^{(n)}(\infty)] = O(1)$, which establishes (3.13). Finally using a bootstrapping argument as in Braverman (2020), we prove (3.14).

Throughout the proof, we shall be using the infinitesimal generator $\mathcal{G}$ of the process $\mathbf{Q}^{(n)}$. To express the generator compactly, we write $(j, i) \prec (k, l)$ if $i < l$, or $i = l$ and $j < k$. We also use $(k, l) \succ (j, i)$ if $(j, i) \prec (k, l)$. The infinitesimal generator $\mathcal{G}$ applied to a function $f : S^{(n)} \rightarrow \mathbb{R}$ and evaluated at a state $\mathbf{q} \in S^{(n)}$ can be written as

$$\begin{aligned} \mathcal{G}f(\mathbf{q}) &= n\lambda^{(n)}\mathbb{1}_{(q_{1,1} < N_1^{(n)})}(f(\mathbf{q} + \mathbf{e}_{1,1}) - f(\mathbf{q})) \\ &\quad + \sum_{(k,i) \succ (1,1)} n\lambda^{(n)}\mathbb{1}_{(q_{l,j} = N_l^{(n)}) \text{ for } (l,j) \prec (k,i), q_{k,i} < N_k^{(n)}}(f(\mathbf{q} + \mathbf{e}_{k,i}) - f(\mathbf{q})) \\ &\quad + \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} \mu_k(q_{k,i} - q_{k,i+1})(f(\mathbf{q} - \mathbf{e}_{k,i}) - f(\mathbf{q})), \end{aligned} \quad (6.1)$$

where $\mathbf{e}_{i,j}$ denotes the infinite dimensional unit vector with one in the (i, j) th position. We will be using the following rate conservation result (Braverman 2020, lemma 1) that states that for any $f : S^{(n)} \rightarrow \mathbb{R}$ with $\mathbb{E}[|f(\mathbf{Q}^{(n)}(\infty))|] < \infty$, we have

$$\mathbb{E}[\mathcal{G}f(\mathbf{Q}^{(n)}(\infty))] = 0. \quad (6.2)$$

The above equation plays a key role in the proof using Stein's approach that chooses an appropriate f to plug into (6.2).

To use (6.2) within Stein's approach, for each $t \in [0, \infty]$ we define $X_1^{(n)}(t) = \sum_{k=1}^M (Q_{k,1}^{(n)}(t) - N_k^{(n)})/n$ and $X_2^{(n)}(t) = Q_{1,2}^{(n)}(t)/n$. Furthermore, for any function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, we define its lifted version $Af : S^{(n)} \rightarrow \mathbb{R}$ as $Af(\mathbf{q}) = f(\mathbf{x})$ where $\mathbf{q} \in S^{(n)}$ and $\mathbf{x}$ is the value of $\mathbf{X}^{(n)}(t)$ when $\mathbf{Q}^{(n)}(t) = \mathbf{q}$. Then, the generator acting on the lifted function can

be written as

$$\begin{aligned} \mathcal{G}Af(\mathbf{q}) &= n\lambda^{(n)}\mathbb{1}(x_1 < 0)(f(\mathbf{x} + \mathbf{e}_1/n) - f(\mathbf{x})) + n\lambda^{(n)}\mathbb{1}(x_1 = 0)(f(\mathbf{x} + \mathbf{e}_2/n) - f(\mathbf{x})) \\ &\quad - n\lambda^{(n)}\mathbb{1}(x_1 = 0, x_2 = N_1^{(n)}/n)(f(\mathbf{x} + \mathbf{e}_2/n) - f(\mathbf{x})) \\ &\quad + \sum_{k=1}^M \mu_k(q_{k,1} - q_{k,2})(f(\mathbf{x} - \mathbf{e}_1/n) - f(\mathbf{x})) + \mu_1(q_{1,2} - q_{1,3})(f(\mathbf{x} - \mathbf{e}_2/n) - f(\mathbf{x})), \end{aligned} \quad (6.3)$$

where $\mathbf{e}_i$ is the two-dimensional unit vector with one in the i th position. Furthermore, because $\mathbf{X}^{(n)}(\infty)$ can only take a finite many values for each n , we have $\mathbb{E}[|f(\mathbf{X}^{(n)}(\infty))|] < \infty$ for any $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, which by (6.2) implies that

$$\mathbb{E}[\mathcal{G}Af(\mathbf{X}^{(n)}(\infty))] = 0. \quad (6.4)$$

To find the appropriate function f to plug into the above equation, we observe that for any twice-differentiable $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, the Taylor expansion yields

$$\begin{aligned} f(\mathbf{x} + \mathbf{e}_1/n) - f(\mathbf{x}) &= \frac{1}{n} f_1(\mathbf{x}) + \int_{x_1}^{x_1+1/n} (x_1 + 1/n - u) f_{11}(u, x_2) du, \\ f(\mathbf{x} - \mathbf{e}_1/n) - f(\mathbf{x}) &= -\frac{1}{n} f_1(\mathbf{x}) + \int_{x_1-1/n}^{x_1} (u - x_1 + 1/n) f_{11}(u, x_2) du, \end{aligned}$$

where f_i denotes the first-order partial derivative $\partial f / \partial x_i$, and f_{ij} denotes the second-order partial derivative $\partial^2 f / \partial x_i \partial x_j$. Using the above expansion in (6.3), we have that for any twice-differentiable $f : \mathbb{R}^2 \rightarrow \mathbb{R}$:

$$\mathcal{G}Af(\mathbf{q}) = \mathcal{L}f(\mathbf{x}) - \lambda^{(n)}\mathbb{1}(x_1 = 0)(f_1(\mathbf{x}) - f_2(\mathbf{x})) + \varepsilon_f^n(\mathbf{q}), \quad (6.5)$$

where

$$\mathcal{L}f(\mathbf{x}) = \left(-\frac{\beta}{\sqrt{n}} - \mu_M x_1 + \mu_1 x_2 \right) f_1(\mathbf{x}) - \mu_1 x_2 f_2(\mathbf{x}) \quad (6.6)$$

and

$$\varepsilon_f^n(\mathbf{q}) = n\lambda^{(n)}\mathbb{1}(x_1 < 0) \int_{x_1}^{x_1+1/n} (x_1 + 1/n - u) f_{11}(u, x_2) du, \quad (6.7)$$

$$+ n\lambda^{(n)}\mathbb{1}(x_1 = 0) \int_{x_2}^{x_2+1/n} (x_2 + 1/n - u) f_{22}(x_1, u) du, \quad (6.8)$$

$$- n\lambda^{(n)}\mathbb{1}(x_1 = 0, x_2 = N_1^{(n)}/n) \int_{N_1^{(n)}/n}^{(N_1^{(n)}+1)/n} f_2(0, u) du, \quad (6.9)$$

$$+ \sum_{k=1}^{M-1} (\mu_k - \mu_M)(N_k^n - q_{k,1}) \int_{x_1-1/n}^{x_1} f_1(u, x_2) du, \quad (6.10)$$

$$+ n\mu_M x_1 \int_{x_1+1/n}^{x_1} (u - x_1 + 1/n) f_{11}(u, x_2) du, \quad (6.11)$$

$$+ \sum_{k=1}^M N_k^n \mu_k \int_{x_1-1/n}^{x_1} (u - x_1 + 1/n) f_{11}(u, x_2) du, \quad (6.12)$$

$$+ \sum_{k=2}^M \mu_k q_{k,2} \int_{x_1-1/n}^{x_1} f_1(u, x_2) du, \quad (6.13)$$

$$- n\mu_1 x_2 \int_{x_1-1/n}^{x_1} (u - x_1 + 1/n) f_{11}(u, x_2) du, \quad (6.14)$$

$$+ n\mu_1 x_2 \int_{x_2-1/n}^{x_2} (u - x_2 + 1/n) f_{22}(x_1, u) du, \quad (6.15)$$

$$+ \mu_1 q_{1,3} \int_{x_2-1/n}^{x_2} f_2(x_1, u) du, \quad (6.16)$$

$$+ \left(\lambda^{(n)} - \frac{\beta}{\sqrt{n}} - \frac{\sum_{k=1}^M \mu_k N_k^{(n)}}{n} \right) f_1(\mathbf{x}). \quad (6.17)$$

Thus, Expansion (6.5) of the generator $\mathcal{G}$ shows that for any twice-differentiable $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ the operator $\mathcal{G}$ acting on the lifted function Af can be approximated by the linear operator $\mathcal{L}$ (defined in (6.6)) acting on f as long as the error terms contained in $\varepsilon_f^n(\mathbf{q})$ remain small and first-order partial derivatives f_1 and f_2 coincide along the boundary $x_1 = 0$ (which makes the second term appearing the generator expansion (6.5) zero).

The key idea of our proof using Stein's approach is to restrict f to a class of functions for which $f_1(\mathbf{x})$ indeed matches with $f_2(\mathbf{x})$ along the reflection boundary $x_1 = 0$, and $\varepsilon_f^n(\mathbf{q})$ remains small for all sufficiently large n . For f belonging to such a class of functions, we have from (6.4)

$$\mathbb{E}[-\mathcal{L}f(\mathbf{X}^{(n)}(\infty))] = \mathbb{E}[\varepsilon_f^n(\mathbf{Q}^{(n)}(\infty))] \approx 0,$$

for all sufficiently large n . Thus, to show that a chosen metric h of the steady-state $\mathbf{X}^{(n)}(\infty)$ remains close to zero in expectation for all sufficiently large n , that is, to establish

$$\mathbb{E}[h(\mathbf{X}^{(n)}(\infty))] = \mathbb{E}[\varepsilon_f^n(\mathbf{Q}^{(n)}(\infty))] \approx 0,$$

it is enough to find f satisfying PDE $\mathcal{L}f(\mathbf{x}) = -h(\mathbf{x})$. This is the key idea we employ in our proof.

To prove (3.11) and (3.14), we will also make use the following lemma from Wang et al. (2022) that provides steady-state bounds on expected value of a Lyapunov function when its drift is negative outside a compact set.

Lemma 6 (Lemma 10 of Wang et al. 2022). *Let $Z = (Z(t), t \geq 0)$ be a continuous time Markov process with countable state space $\mathcal{Z}$ and infinitesimal generator $\mathcal{G}_Z$. Furthermore, assume that the chain Z is irreducible, nonexplosive, and ergodic and $Z(t) \Rightarrow Z(\infty)$ as $t \rightarrow \infty$. Let $r_{i,j}$ denote the transition rate of Z from state $i \in \mathcal{Z}$ to state $j \in \mathcal{Z} \setminus \{i\}$ and $r_{i,i} = -\sum_{j \in \mathcal{Z} \setminus \{i\}} r_{i,j}$. Suppose that $r_{\max} = \sup_{i \in \mathcal{Z}} |r_{i,i}| < \infty$ and a Lyapunov function $V: \mathcal{Z} \rightarrow \mathbb{R}_+$ exists such that $\mathcal{G}_Z V(i) \leq -\gamma$ for some $\gamma > 0$ whenever $i \in \mathcal{Z}$ is such that $V(i) > B$ for some $B > 0$. Then, for any integer $m \in \mathbb{Z}_+$ we have*

$$\mathbb{E}[V^m(Z(\infty))] \leq (2B)^m + (4v_{\max})^m \left(\frac{r_{\max} V_{\max} + \gamma}{\gamma} \right)^m m!,$$

where $v_{\max} = \sup_{i,j \in \mathcal{Z}: r_{i,j} > 0} |V(j) - V(i)|$.

Proof of Theorem 2. Define function

$$V(\mathbf{q}) = \sum_{j=1}^{M-1} (N_j^{(n)} - q_{j,1}). \quad (6.18)$$

To prove (3.11), it is enough to show that $\mathbb{E}[V(\mathbf{Q}^{(n)}(\infty))] = O(1)$. Plugging $V(\mathbf{q})$ into the generator in (6.1), we have

$$\mathcal{G}V(\mathbf{q}) = -n\lambda^{(n)} \mathbb{1}(V(\mathbf{q}) > 0) + \sum_{j=1}^{M-1} \mu_j (q_{j,1} - q_{j,2}).$$

Fix $\epsilon > 0$, and (A1) and (A2) imply that for large enough n and any $\mathbf{q}$ with $V(\mathbf{q}) \geq 1$, we have

$$\begin{aligned} \mathcal{G}V(\mathbf{q}) &\leq -n\lambda^{(n)} + \sum_{j=1}^{M-1} N_j^{(n)} \mu_j \\ &\leq -n\gamma_M \mu_M + (\beta + \epsilon)\sqrt{n} < 0. \end{aligned}$$

Hence, using Lemma 6 for $m \in \mathbb{Z}_+$, we have

$$\mathbb{E}[V(\mathbf{q})^m] \leq (2)^m + (4)^m \left(\frac{\lambda^{(n)} + \sum_{k=1}^M \gamma_k \mu_k - (\beta + \epsilon)n^{-1/2}}{\gamma_M \mu_M - (\beta + \epsilon)n^{-1/2}} \right)^m m!.$$

Hence, for any $n > \left(\frac{2(\beta + \epsilon)}{\gamma_M \mu_M} \right)^2$ and $m \in \mathbb{Z}_+$:

$$\mathbb{E}[V(\mathbf{q})^m] \leq (2)^m + \left(\frac{16}{\gamma_M \mu_M} \right)^m m!. \quad (6.19)$$

In particular, for $m = 1$, (6.19) becomes

$$\mathbb{E} \left[\sqrt{n} \sum_{j=1}^{M-1} |Y_{j,1}^{(n)}(\infty)| \right] \leq 2 + \frac{16}{\gamma_M \mu_M},$$

which proves (3.11).

Now we turn to the proofs of (3.12) and (3.13) for which we need the following lemma proved in Appendix C.

Lemma 7. For any $n \geq 1$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^M \mu_k Q_{k,1}^{(n)}(\infty)\right] &= n\lambda^{(n)}, \\ \mathbb{E}\left[\sum_{k=2}^M \mu_k Q_{k,2}^{(n)}(\infty) + \mu_1 Q_{1,3}^{(n)}(\infty)\right] &= n\lambda^{(n)}\mathbb{P}\left(\sum_{k=1}^M Q_{k,1}^{(n)}(\infty) = n, Q_{1,2}^{(n)}(\infty) = N_1^{(n)}\right).\end{aligned}$$

Now, we write

$$\begin{aligned}\mathbb{E}[\sqrt{n}|Y_{M,1}^{(n)}(\infty)|] &\leq \mu_M^{-1}\left(\sum_{k=1}^M N_k^{(n)}\mu_k - \mathbb{E}\left[\sum_{k=1}^M \mu_k Q_{k,1}^{(n)}(\infty)\right]\right) \\ &= \mu_M^{-1}\left(\sum_{k=1}^M N_k^{(n)}\mu_k - n\lambda^{(n)}\right) \\ &= \sqrt{n}\mu_M^{-1}\left(\sqrt{n}\sum_{k=1}^M (N_k^{(n)}/n - \gamma_k)\mu_k + \sqrt{n}(1 - \lambda^{(n)})\right),\end{aligned}$$

where the second line follows from part 1 of Lemma 7. The proof of (3.12) now follows by applying (A1) and (A2) to the above inequality.

To establish (3.13), it is enough to show that

$$\mathbb{E}\left[\left(X_2^{(n)}(\infty) - \frac{\kappa}{\sqrt{n}}\right)_+\right] = O\left(\frac{1}{\sqrt{n}}\right) \quad (6.20)$$

for any constant $\kappa > \beta/\mu_1$. To establish (6.20), we first take expectations of both sides of (6.5) and apply lemma 1 from Braverman (2020) to obtain

$$\mathbb{E}[GAf(\mathbf{Q}^n(\infty))] = \mathbb{E}[\mathcal{L}f(\mathbf{X}^n(\infty))] - \lambda^{(n)}\mathbb{E}[\mathbb{1}(X_1^n(\infty) = 0)(f_1(\mathbf{X}^n(\infty)) - f_2(\mathbf{X}^n(\infty)))] + \mathbb{E}[\epsilon_f^n(\mathbf{Q}^n(\infty))] = 0. \quad (6.21)$$

Lemma 8 provides the appropriate Lyapunov function $f^*(\mathbf{x})$ to plug in to (6.21) and prove (6.20). We construct the function $f^*(\mathbf{x})$ in Section 6.1 and prove that this function satisfies the conditions stated in Lemma 8 in Appendix F.

Lemma 8. Fix $\kappa > \beta/\mu_1$ and define $\Omega = (-\infty, 0] \times [0, \infty)$. Then, there exists a function $f^*(\mathbf{x})$ that solves the PDE

$$\mathcal{L}f^*(\mathbf{x}) = -(x_2 - \kappa/\sqrt{n})_+, \text{ for all } \mathbf{x} \in \Omega, \quad (6.22)$$

$$f_1^*(0, x_2) = f_2^*(0, x_2) = \frac{\sqrt{n}}{\beta} \left(x_2 - \frac{\kappa}{\sqrt{n}}\right)_+, \text{ for all } x_2 \geq 0, \quad (6.23)$$

with absolutely continuous first-order derivatives $f_1^*(\cdot, x_2)$ and $f_2^*(x_1, \cdot)$. Moreover, we have

$$f_1^*(\mathbf{x}) \geq 0, f_{12}^*(\mathbf{x}) \geq 0, f_{11}^*(\mathbf{x}) \geq 0, f_{22}^*(\mathbf{x}) \geq 0, \quad \mathbf{x} \in \Omega, \quad (6.24)$$

$$f_{11}(\mathbf{x}) = f_{22}(\mathbf{x}) = 0, \quad x_2 \in [0, \kappa/\sqrt{n}], \quad (6.25)$$

$$f_{11}^*(\mathbf{x}) \leq C_5(\kappa)\sqrt{n}, f_{22}^*(\mathbf{x}) \leq C_6(\kappa)\sqrt{n}, \quad \mathbf{x} \in \Omega, \quad (6.26)$$

where $C_5(\kappa) = \frac{1}{\beta} \frac{\mu_1 + \mu_M}{\mu_1} \frac{\kappa}{\kappa - \beta/\mu_1}$ and $C_6(\kappa) = \frac{1}{\mu_1\kappa} + \frac{1}{\beta} \left(\frac{\mu_1}{\mu_1 - \mu_M}\right)^2 \frac{\kappa}{\kappa - \beta/\mu_1} + \frac{1}{\beta} \frac{\mu_1(1 + \mu_2)}{\mu_1 - \mu_M}$.

Now, for f^* as given in Lemma 8, we have from (6.21)

$$\mathbb{E}\left[\left(X_2^{(n)}(\infty) - \frac{\kappa}{\sqrt{n}}\right)_+\right] = \mathbb{E}[\epsilon_{f^*}^n(\mathbf{Q}^{(n)}(\infty))]. \quad (6.27)$$

Suppose that we can show that the expected value of every individual term appearing in Lines (6.7)–(6.16) is $O(1/\sqrt{n})$. Then using (6.27), (6.7)–(6.17), and the fact that $f_1^*(\mathbf{x}) \geq 0$, we have

$$\begin{aligned}\mathbb{E}\left[\left(X_2^{(n)}(\infty) - \frac{\kappa}{\sqrt{n}}\right)_+\right] &\leq O\left(\frac{1}{\sqrt{n}}\right) + \left(\left|\lambda^{(n)} - 1 - \frac{\beta}{\sqrt{n}}\right| + \left|1 - \frac{\sum_{k \in [M]} \mu_k N_k^{(n)}}{n}\right|\right) \mathbb{E}[f_1^*(\mathbf{X}^{(n)}(\infty))], \\ &\leq O\left(\frac{1}{\sqrt{n}}\right) + o\left(\frac{1}{\sqrt{n}}\right) \mathbb{E}[f_1^*(0, X_2^{(n)}(\infty))], \\ &\leq O\left(\frac{1}{\sqrt{n}}\right) + o(1) \mathbb{E}\left[\left(X_2^{(n)}(\infty) - \frac{\kappa}{\sqrt{n}}\right)_+\right],\end{aligned} \quad (6.28)$$

where the second inequality follows due to Assumptions (A1) and (A2) and the fact that $f_{11}^*(\mathbf{x}) \geq 0$, and the last inequality follows because $f_1^*(0, x_2) = (\sqrt{n}/\beta)(x_2 - \kappa/\sqrt{n})_+$. From (6.28), it follows that

$$\mathbb{E} \left[(1 - o(1)) \left(X_2^{(n)}(\infty) - \frac{\kappa}{\sqrt{n}} \right)_+ \right] = O \left(\frac{1}{\sqrt{n}} \right),$$

which implies (6.20) and hence (3.13). Therefore, it suffices to show that the expected value of every individual term appearing in Lines (6.7)–(6.16) is $O(1/\sqrt{n})$.

Using the bounds on the second derivatives from Lemma 8, we have

$$n\lambda^{(n)} \mathbb{1}(x_1 < 0) \int_{x_1}^{x_1+1/n} (x_1 + 1/n - u) f_{11}^*(u, x_2) du \leq C_5(\kappa) \lambda^{(n)} / \sqrt{n}.$$

All other terms involving second derivatives can be bounded in a similar fashion.

The second derivatives of $f^*(\mathbf{x})$ being nonnegative implies that the first derivatives are nondecreasing functions in x_1 and x_2 , and hence,

$$f_1^*(\mathbf{x}) \leq f_1^* \left(0, \frac{N_1^{(n)}}{n} \right) \text{ and } f_2^*(\mathbf{x}) \leq f_2^* \left(0, \frac{N_1^{(n)}}{n} \right) = f_1^* \left(0, \frac{N_1^{(n)}}{n} \right) \quad (6.29)$$

for all $x_1 \leq 0$ and $0 \leq x_2 \leq \frac{N_1^{(n)}}{n}$. Hence, for the term appearing in (6.10), we have

$$\begin{aligned} \sum_{k=1}^{M-1} (\mu_k - \mu_M) \mathbb{E}[(N_k^n - Q_{k,1}^{(n)}(\infty)) \int_{X_1^{(n)}(\infty)-1/n}^{X_1^{(n)}(\infty)} f_1^*(u, X_2^{(n)}(\infty)) du] \\ \leq \frac{\mu_1}{\sqrt{n}} \mathbb{E} \left[\left| \sum_{k=1}^{M-1} Y_{k,1}^{(n)}(\infty) \right| \right] f_1^* \left(0, \frac{N_1^{(n)}}{n} \right) \\ \leq \frac{\mu_1 C_1(M-1)}{\beta \sqrt{n}} \left(\frac{N_1^{(n)}}{n} - \frac{\kappa}{\sqrt{n}} \right), \end{aligned}$$

where the first inequality follows from (6.29) and the last inequality follows from (3.11) and (6.23). From the above inequality, it is clear that the right-hand side converges to zero as $n \rightarrow \infty$.

Using (6.23), we can bound the remaining terms appearing in (6.9), (6.13), and (6.16) as

$$\begin{aligned} -n\lambda^{(n)} \mathbb{1} \left(x_1 = 0, x_2 = \frac{N_1^{(n)}}{n} \right) \int_{\frac{\kappa}{n}}^{\frac{N_1^{(n)}+1}{n}} f_2^*(0, u) du &\leq -\lambda^{(n)} \mathbb{1} \left(x_1 = 0, x_2 = \frac{N_1^{(n)}}{n} \right) f_2^* \left(0, \frac{N_1^{(n)}}{n} \right) \\ \sum_{k=2}^M \mu_k q_{k,2} \int_{x_1 - \frac{1}{n}}^{x_1} f_1^*(u, x_2) du &\leq \frac{1}{n} \sum_{k=2}^M \mu_k q_{k,2} f_2^* \left(0, \frac{N_1^{(n)}}{n} \right) \\ \mu_1 q_{1,3} \int_{x_2 - \frac{1}{n}}^{x_2} f_2^*(x_1, u) du &\leq \frac{1}{n} \mu_1 q_{1,3} f_2^* \left(0, \frac{N_1^{(n)}}{n} \right). \end{aligned}$$

Now, replacing $\mathbf{q}$ with $\mathbf{Q}^{(n)}(\infty)$, summing the above inequalities and taking expectations both sides with respect to π^n , we see that the sum of the right-hand sides is equal to zero due to part 2 of Lemma 7. Hence, the contribution of these terms to $\mathbb{E}[\epsilon_{f^*}^n(\mathbf{Q}^{(n)}(\infty))]$ is nonpositive and can be safely ignored to obtain an upper bound. This completes the proof of (3.13).

To establish (3.14), it is sufficient to prove that for any $1/2 < a < 1$, there exists a $C_8(a) > 0$ such that for all sufficiently large n

$$\mathbb{E} \left[\sum_{k=2}^M \mu_k Q_{k,2}^{(n)}(\infty) + \mu_1 Q_{1,3}^{(n)}(\infty) \right] \leq n^{1-a} C_8(a).$$

Using part 2 of Lemma 7, the above inequality follows if we can show that for any $1/2 < a < 1$, there exists a $C_8(a)$ such that for all sufficiently large n the following holds:

$$n\lambda^{(n)} \mathbb{P} \left(\sum_{k=1}^M Q_{k,1}^{(n)}(\infty) = n, Q_{1,2}^{(n)}(\infty) = N_1^{(n)} \right) \leq n^{1-a} C_8(a). \quad (6.30)$$

By Assumption (A2) for any fixed $\epsilon > 0$, we have $|N_1^{(n)}/n - \gamma_1| < \epsilon$ and $\max(\beta/(\mu_1\sqrt{n}), 1/n) < \gamma_1 - 2\epsilon$ for all sufficiently large n . Choose $\tilde{\kappa}$ in the interval $(\gamma_1 - 2\epsilon, \gamma_1 - \epsilon)$. Hence, for all sufficiently large n , we have $\tilde{\kappa} \in (\max(\beta/(\mu_1\sqrt{n}), 1/n), N_1^{(n)}/n)$. We have

$$\mathbb{P}(Q_{1,2}^{(n)}(\infty) = N_1^{(n)}) = \mathbb{P}\left(X_2^{(n)}(\infty) = \frac{N_1^{(n)}}{n}\right), \quad (6.31)$$

$$= \frac{n}{N_1^{(n)} - \tilde{\kappa}n} \mathbb{E}\left[\left(X_2^{(n)}(\infty) - \tilde{\kappa}\right) \mathbb{1}\left(X_2^{(n)}(\infty) = \frac{N_1^{(n)}}{n}\right)\right], \quad (6.32)$$

$$\leq \frac{n}{N_1^{(n)} - \tilde{\kappa}n} \mathbb{E}[(X_2^{(n)}(\infty) - \tilde{\kappa})_+]. \quad (6.33)$$

Consider the $\tilde{f}^*(\mathbf{x})$ and $\varepsilon_{\tilde{f}^*}^n(\mathbf{q})$ as in Lemma 8 defined by replacing $\kappa = \sqrt{n}\tilde{\kappa} > \beta/\mu_1$. Using (6.27) with $\varepsilon_{\tilde{f}^*}^n(\mathbf{q})$, the result will follow if we can show that for any $1/2 < a < 1$, there exists a $C_8(a)$ such that $\mathbb{E}[\varepsilon_{\tilde{f}^*}^n(\mathbf{Q}^{(n)}(\infty))] \leq C_8(a)/n^a$ for all large enough n .

The terms in (6.7), (6.8), (6.11), (6.12), and (6.15) of $\varepsilon_{\tilde{f}^*}^n(\mathbf{q}^{(n)}(\infty))$ are equal to zero when $x_2 < \tilde{\kappa} - 1/n$. On the other hand, when $x_2 \geq \tilde{\kappa} - 1/n$, we can find a $C_9 > 0$ such that for all large enough n

$$\begin{aligned} n\lambda^{(n)}\mathbb{1}(x_1 < 0) \int_{x_1}^{x_1+1/n} (x_1 + 1/n - u) \tilde{f}_{11}^*(u, x_2) du &\leq \frac{C_5(\sqrt{n}\tilde{\kappa})}{\sqrt{n}} < \frac{C_9}{\sqrt{n}}, \\ n\lambda^{(n)}\mathbb{1}(x_1 = 0) \int_{x_2}^{x_2+1/n} (x_2 + 1/n - u) \tilde{f}_{22}^*(x_1, u) du &\leq \frac{C_6(\sqrt{n}\tilde{\kappa})}{\sqrt{n}} < \frac{C_9}{\sqrt{n}}, \\ n\mu_M x_1 \int_{x_1+1/n}^{x_1} (u - x_1 + 1/n) \tilde{f}_{11}^*(u, x_2) du &\leq \frac{C_5(\sqrt{n}\tilde{\kappa})\mu_1}{\sqrt{n}} < \frac{C_9}{\sqrt{n}}, \\ \sum_{k=1}^M N_k^n \mu_k \int_{x_1-1/n}^{x_1} (u - x_1 + 1/n) \tilde{f}_{11}^*(u, x_2) du &\leq \frac{C_5(\sqrt{n}\tilde{\kappa})}{\sqrt{n}} \left(1 + o\left(\frac{1}{\sqrt{n}}\right)\right) < \frac{C_9}{\sqrt{n}}, \\ n\mu_1 x_2 \int_{x_2-1/n}^{x_2} (u - x_2 + 1/n) \tilde{f}_{22}^*(x_1, u) du &\leq \frac{C_6(\sqrt{n}\tilde{\kappa})\mu_1}{\sqrt{n}} < \frac{C_9}{\sqrt{n}}. \end{aligned}$$

Next, we prove a bound on the expected value of the third term of $\varepsilon_{\tilde{f}^*}^n(\mathbf{q})$, (6.10). Taking the expectation of (6.10) with respect to π^n , we can write

$$\begin{aligned} &\sum_{k=1}^{M-1} (\mu_k - \mu_M) \mathbb{E}\left[(N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) \int_{X_1^{(n)}(\infty)-1/n}^{X_1^{(n)}(\infty)} \tilde{f}_1^*(u, X_2^{(n)}(\infty)) du\right] \\ &\leq \mu_1 \mathbb{E}\left[\sum_{k=1}^{M-1} (N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) \int_{X_1^{(n)}(\infty)-1/n}^{X_1^{(n)}(\infty)} \tilde{f}_1^*(u, X_2^{(n)}(\infty)) du\right] \\ &\leq \frac{\mu_1}{n} \mathbb{E}\left[\sum_{k=1}^{M-1} (N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) \tilde{f}_1^*(0, X_2^{(n)}(\infty))\right] \\ &= \frac{\mu_1}{\beta\sqrt{n}} \mathbb{E}\left[\sum_{k=1}^{M-1} (N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) (X_2^{(n)}(\infty) - \tilde{\kappa})_+\right] \\ &= \frac{\mu_1}{\beta\sqrt{n}} \mathbb{E}\left[\sum_{k=1}^{M-1} (N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) (X_2^{(n)}(\infty) - \tilde{\kappa}) \mathbb{1}(X_2^{(n)}(\infty) > \tilde{\kappa})\right] \\ &\leq \frac{\mu_1}{\beta\sqrt{n}} \mathbb{E}\left[\sum_{k=1}^{M-1} (N_k^{(n)} - Q_{k,1}^{(n)}(\infty)) \mathbf{n}(X_2^{(n)}(\infty) > \tilde{\kappa})\right] \\ &\leq \frac{\mu_1}{\beta\sqrt{n}} (\mathbb{E}[V(\mathbf{Q}^{(n)}(\infty))^p])^{1/p} (\mathbb{P}(X_2^{(n)}(\infty) > \tilde{\kappa}))^{1/q}, \end{aligned} \quad (6.34)$$

where the first equality follows from (6.23), the second inequality follows from (6.29), the second to last inequality follows as $X_2^{(n)}(\infty) < 1$, and in the last inequality, we used (6.18), and the Holder's inequality with $p, q > 1$ satisfying $1/p + 1/q = 1$. Let $a = 1/2 + 1/2q$ and define $p = \frac{1}{2-2a}$, then using (6.19) for $m = \lceil p \rceil$, we have

$$\mathbb{E}[V(\mathbf{Q}^{(n)}(\infty))^p] \leq \mathbb{E}[V(\mathbf{Q}^{(n)}(\infty))^{\lceil p \rceil}]^{\frac{p}{\lceil p \rceil}} \leq \left((2)^{\lceil p \rceil} + \left(\frac{16}{\gamma_M \mu_M} \right)^{\lceil p \rceil} (\lceil p \rceil)! \right)^{\frac{p}{\lceil p \rceil}} = C_{10}(a).$$

Furthermore, using the Markov inequality and (3.13), for all sufficiently large n

$$\mathbb{P}(X_2^{(n)}(\infty) \geq \tilde{\kappa}) \leq \frac{\mathbb{E}[X_2^{(n)}(\infty)]}{\tilde{\kappa}} \leq \frac{C_3}{\tilde{\kappa} \sqrt{n}}.$$

Hence, the right-hand side of (6.34) can be bounded as

$$\sum_{k=1}^{M-1} (\mu_k - \mu_M) \mathbb{E} \left[(N_k^n - Q_{k,1}^{(n)}(\infty)) \int_{X_1^{(n)}(\infty) - 1/n}^{X_1^{(n)}(\infty)} f_1^*(u, X_2^{(n)}(\infty)) du \right] \leq \frac{\mu_1 C_3 C_{10}(a)}{\beta \tilde{\kappa}^{1/q} n^a}.$$

This is for any $a \in (1/2, 1)$ and all sufficiently large n .

In proving (3.13), we have already shown that the expected sum of the terms appearing in (6.9), (6.13), and (6.16) is at most zero for any κ , and therefore, we can safely ignore the contributions of these terms to obtain an upper bound on $\mathbb{E}[\varepsilon_{f^*}^n(\mathbf{Q}^{(n)}(\infty))]$. Hence, combining all the bounds obtained, we have

$$\mathbb{E}[\varepsilon_{f^*}^n(\mathbf{Q}^{(n)}(\infty))] \leq \frac{4C_9}{\sqrt{n}} \mathbb{P} \left(X_2^{(n)}(\infty) > \tilde{\kappa} - \frac{1}{n} \right) + \frac{\mu_1 C_3 C_{10}(a)}{\beta \tilde{\kappa}^{1/q} n^a}. \quad (6.35)$$

But using the Markov inequality and (3.13) again, we have

$$\mathbb{P} \left(X_2^{(n)}(\infty) > \tilde{\kappa} - \frac{1}{n} \right) = \mathbb{P} \left(Y_2^{(n)}(\infty) > \sqrt{n} \tilde{\kappa} - \frac{1}{\sqrt{n}} \right) \leq \frac{\mathbb{E}[Y_2^{(n)}(\infty)]}{\sqrt{n} \tilde{\kappa} - \frac{1}{\sqrt{n}}} \leq \frac{C_3}{\tilde{\kappa} \sqrt{n}}, \quad (6.36)$$

for all sufficiently large n . Hence,

$$\mathbb{E}[\varepsilon_{f^*}^n(\mathbf{Q}(\infty))] \leq \frac{4C_9 C_3}{n \tilde{\kappa}} + \frac{\mu_1 C_3 C_{10}(a)}{\beta \tilde{\kappa} n^a},$$

and the result follows. $\square$

6.1. Construction of the Lyapunov Function for Lemma 8

In this section, we construct a candidate Lyapunov function $f^*(\mathbf{x})$ that solves the PDE (6.22)–(6.23) using the drift-based fluid limit (DFL) approach of Stolyar (2015b). This approach relates to the method of characteristics for solving the PDE by reducing the problem to solving a set of Ordinary Differential Equations (ODEs) (Renardy and Rogers 2004, chapter 2). The key idea behind this approach is that a solution to the PDE:

$$\mathcal{L} f^{(h)}(\mathbf{x}) = -h(\mathbf{x}), \quad (6.37)$$

where the operator $\mathcal{L}$ satisfies $\mathcal{L} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{F}(\mathbf{x})$ for some continuous vector field $\mathbf{F}$ is given by

$$f^{(h)}(\mathbf{x}) = \int_0^\infty h(\mathbf{v}^{\mathbf{x}}(s)) ds, \quad (6.38)$$

provided that the right-hand side of (6.38) is finite, and the function $\mathbf{v}^{\mathbf{x}}(t)$ is the solution of the ODE $\dot{\mathbf{v}}(t) = \mathbf{F}(\mathbf{v}(t))$ with initial condition $\mathbf{v}(0) = \mathbf{x}$.

To apply the DFL approach to our problem, observe that a function f^* solves the PDE (6.22)–(6.23) if and only if it is also the solution of the following PDE:

$$\bar{\mathcal{L}} f^*(\mathbf{x}) = - \left(x_2 - \frac{\kappa}{\sqrt{n}} \right)_+, \quad (6.39)$$

$$f_1^*(0, x_2) = f_2^*(0, x_2) = \frac{\sqrt{n}}{\beta} \left(x_2 - \frac{\kappa}{\sqrt{n}} \right)_+, \quad (6.40)$$

for all $\mathbf{x} \in \mathbb{R}^2$ with $x_2 \geq 0$, where the operator $\tilde{\mathcal{L}}$ is defined as

$$\tilde{\mathcal{L}}f(\mathbf{x}) = \left(-\frac{\beta}{\sqrt{n}} - \mu_M x_1 + \mu_1 x_2 \right) \mathbb{1}(\mathbf{x} \in \Omega \setminus \bar{\Omega}) f_1(\mathbf{x}) - \mu_1 x_2 \mathbb{1}(\mathbf{x} \in \Omega \setminus \bar{\Omega}) f_2(\mathbf{x}) - \frac{\beta}{\sqrt{n}} \mathbb{1}(\mathbf{x} \in \bar{\Omega}) f_2(\mathbf{x}), \quad (6.41)$$

and $\bar{\Omega} = \left\{ \mathbf{x} \in \Omega : x_1 = 0, x_2 > \frac{\beta}{\mu_1 \sqrt{n}} \right\}$. Clearly, the operator $\tilde{\mathcal{L}}$ satisfies $\tilde{\mathcal{L}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{F}(\mathbf{x})$ with $\mathbf{F}$ defined as

$$\mathbf{F}(\mathbf{x}) = \begin{cases} \left(0, -\frac{\beta}{\sqrt{n}} \right), & \mathbf{x} \in \bar{\Omega} \\ \left(-\frac{\beta}{\sqrt{n}} - \mu_M x_1 + \mu_1 x_2, -\mu_1 x_2 \right), & \mathbf{x} \in \Omega \setminus \bar{\Omega}. \end{cases} \quad (6.42)$$

Let $\mathbf{v}^{\mathbf{x}}(t)$ denote the solution to the ODE:

$$\dot{\mathbf{v}}(t) = \mathbf{F}(\mathbf{v}(t)), \quad (6.43)$$

$$\mathbf{v}(0) = \mathbf{x} \in \Omega. \quad (6.44)$$

Thus, taking $h(\mathbf{x}) = (x_2 - \kappa/\sqrt{n})_+$, the DFL approach implies that

$$f^*(\mathbf{x}) = \int_0^\infty (v_2^{\mathbf{x}}(s) - \kappa/\sqrt{n})_+ ds, \quad (6.45)$$

is a candidate for the solution to the PDE (6.39)–(6.40). In the remainder of this section, we derive the expression of f^* given in (6.45) by solving (6.43)–(6.44) and computing the integral in (6.45) explicitly. However, it is to be noted that f^* constructed in this way is not guaranteed to solve the PDE (6.39)–(6.40) because the vector field $\mathbf{F}$ in (6.42) is *not continuous*. Hence, we need to additionally verify that the constructed f^* indeed solves (6.39)–(6.40).

Remark 2. We note that if the two-dimensional process $\mathbf{v}^{\mathbf{x}}(t) = (v_1^{\mathbf{x}}(t), v_2^{\mathbf{x}}(t))$ solves (6.43)–(6.44) for each $\mathbf{x} = (x_1, x_2) \in \Omega$, then it must also solve the integral equations:

$$v_1^{\mathbf{x}}(t) = x_1 - \frac{\beta}{\sqrt{n}}t - \int_0^t (\mu_M v_1^{\mathbf{x}}(s) - \mu_1 v_2^{\mathbf{x}}(s)) ds - U_1(t), \quad (6.46)$$

$$v_2^{\mathbf{x}}(t) = x_2 - \int_0^t \mu_1 v_2^{\mathbf{x}}(s) ds + U_1(t), \quad (6.47)$$

where the function $U_1(t)$ is the unique nondecreasing function satisfying

$$\int_0^\infty v_1^{\mathbf{x}}(s) dU_1(s) = 0, U_1(t) \geq 0, U_1(0) = 0. \quad (6.48)$$

Proposition 2 ensures that the process $\mathbf{v}^{\mathbf{x}}(t) = (v_1^{\mathbf{x}}(t), v_2^{\mathbf{x}}(t)) \in \Omega$ is well defined. Note that Equations (6.46)–(6.48) closely resemble those used to define the fluid model in Braverman (2020), except for the coefficients of v_1 and v_2 . Although this may appear to be a subtle difference, it significantly alters the solution, requiring a full rederivation of the solution.

The drift of the process $\mathbf{v}^{\mathbf{x}}$, given by the function $\mathbf{F}$, clearly has a switching behavior depending on whether the process $\mathbf{v}^{\mathbf{x}}$ enters the set $\bar{\Omega}$ or not, which in turn is determined by the starting point $\mathbf{x}$. More specifically, if we define $\tau(\mathbf{x}) = \inf\{t \geq 0 : v_1^{\mathbf{x}}(t) = 0\}$ and set $\tau(\mathbf{x}) = \infty$ if $v_1^{\mathbf{x}}(t) < 0$ for all $t \geq 0$, then $\tau(\mathbf{x}) = \infty$ implies that the process never enters the set $\bar{\Omega}$. Furthermore, for all $t \in [0, \tau(\mathbf{x})]$, we have

$$\frac{dv_1^{\mathbf{x}}(t)}{dt} = -\frac{\beta}{\sqrt{n}} - \mu_M v_1^{\mathbf{x}}(t) + \mu_1 v_2^{\mathbf{x}}(t), \quad (6.49)$$

$$\frac{dv_2^{\mathbf{x}}(t)}{dt} = -\mu_1 v_2^{\mathbf{x}}(t). \quad (6.50)$$

Solving the second equation, we obtain

$$v_2^{\mathbf{x}}(t) = x_2 e^{-\mu_1 t}. \quad (6.51)$$

Plugging this into the first equation, we obtain

$$v_1^x(t) = -\frac{\beta}{\mu_M \sqrt{n}} - \frac{\mu_1}{\mu_1 - \mu_M} x_2 e^{-\mu_1 t} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} \right) e^{-\mu_M t}, \quad (6.52)$$

$$= -\frac{\beta}{\mu_M \sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}} \right) e^{-\mu_M t} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} (e^{-\mu_M t} - e^{-\mu_1 t}). \quad (6.53)$$

In the lemmas below, we determine the conditions needed on the initial point $\mathbf{x}$ for $\tau(\mathbf{x})$ to be finite and show that, if $\tau(\mathbf{x})$ is finite, then $v_2^x(\tau(\mathbf{x})) \geq \beta/\mu_1 \sqrt{n}$, implying that the trajectory $\mathbf{v}^x$ indeed enters the set $\bar{\Omega}$.

Lemma 9. Consider the nonlinear system of equations:

$$-\frac{\beta}{\mu_M \sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}} \right) e^{-\mu_M \tau} - \frac{\mu_1 x_2}{\mu_1 - \mu_M} (e^{-\mu_1 \tau} - e^{-\mu_M \tau}) = 0, \quad (6.54)$$

$$x_2 e^{-\mu_1 \tau} = \kappa / \sqrt{n}. \quad (6.55)$$

The above system of equations has a unique solution in $(x_2, \tau) \in [\kappa/\sqrt{n}, \infty) \times [0, \infty)$ denoted by $(x_2^*(x_1, \kappa), \tau^*(x_1, \kappa))$ for any $x_1 \leq 0$ and any $\kappa \geq \beta/\mu_1$. In particular, we have $(x_2^*(0, \kappa), \tau^*(0, \kappa)) = (\kappa/\sqrt{n}, 0)$.

Lemma 9, proved in Appendix D, allows us to define the curve

$$\Gamma^\kappa = \{\mathbf{x} \in \Omega : x_2 = x_2^*(x_1, \kappa)\} \quad (6.56)$$

for each $\kappa \geq \beta/\mu_1$. Clearly, we have $(0, \kappa/\sqrt{n}) \in \Gamma^\kappa$. We write $\mathbf{x} \geq \Gamma^\kappa$ when $x_2 \geq x_2^*(x_1, \kappa)$ and define $\mathbf{x} > \Gamma^\kappa$, $\mathbf{x} \leq \Gamma^\kappa$ similarly. The following lemma relates the curves Γ^κ , $\kappa \geq \beta/\sqrt{n}$, to $\tau(\mathbf{x})$.

Lemma 10. The following statements hold:

1. If $\mathbf{x} \geq \Gamma^\kappa$ for some $\kappa \geq \beta/\mu_1$, then $\tau(\mathbf{x}) < \infty$ and $x_2 e^{-\mu_1 \tau(\mathbf{x})} \geq \kappa/\sqrt{n}$.
2. If $\mathbf{x} \in \Gamma^\kappa$ for some $\kappa \geq \beta/\mu_1$, then $\tau(\mathbf{x}) = \tau^*(x_1, \kappa)$, where $\tau^*(\cdot)$ is as defined in Lemma 9. Hence, $x_2 e^{-\mu_1 \tau(\mathbf{x})} = \kappa/\sqrt{n}$.
3. We have that $\tau(\mathbf{x})$ is differentiable if $\mathbf{x} \geq \Gamma^\kappa$ for some $\kappa \geq \beta/\mu_1$. Furthermore, the derivatives $\tau_1(\mathbf{x}) = \partial \tau(\mathbf{x})/\partial x_1$ and $\tau_2(\mathbf{x}) = \partial \tau(\mathbf{x})/\partial x_2$ satisfy

$$\tau_1(\mathbf{x}) = -e^{-\mu_M \tau(\mathbf{x})} \left(\mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}} \right)^{-1} \leq 0, \quad (6.57)$$

$$\tau_2(\mathbf{x}) = \frac{\mu_1}{\mu_1 - \mu_M} \left(e^{-\mu_1 \tau(\mathbf{x})} - e^{-\mu_M \tau(\mathbf{x})} \right) \left(\mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}} \right)^{-1} \leq 0, \quad (6.58)$$

where $\tau_1(\mathbf{x})$ is understood to be the left-derivative when $x_1 = 0$ and $\mathbf{x} \geq \Gamma^\kappa$ for some $\kappa \geq \beta/\mu_1$.

4. If $\kappa_1 > \kappa_2 \geq \beta/\mu_1$, then $\Gamma^{\kappa_1} > \Gamma^{\kappa_2}$, that is, $\mathbf{x} \geq \Gamma^{\kappa_1}$ implies $\mathbf{x} > \Gamma^{\kappa_2}$.
5. If $\mathbf{x} < \Gamma^{\beta/\mu_1}$, then $\tau(\mathbf{x}) = \infty$.

Thus, from Lemma 10, it follows that $\tau(\mathbf{x}) < \infty$ if $\mathbf{x} \in \Gamma^\kappa$, for some $\kappa \geq \beta/\mu_1$, or equivalently, if $\mathbf{x} \geq \Gamma^{\beta/\mu_1}$, and in this case, $v_2^x(\tau(\mathbf{x})) = x_2 e^{-\mu_1 \tau(\mathbf{x})} \geq \beta/\mu_1 \sqrt{n}$, that is, at time $\tau(\mathbf{x})$ the trajectory $\mathbf{v}^x$ enters the set $\bar{\Omega}$. On the other hand, if $\mathbf{x} < \Gamma^{\beta/\mu_1}$, then the trajectory $\mathbf{v}^x$ never enters the set $\bar{\Omega}$. This behavior is shown in Figure 3. Once the trajectory $\mathbf{v}^x$ enters $\bar{\Omega}$, we have from (6.42)–(6.44) that

$$dv_1^x(t) = 0, \quad (6.59)$$

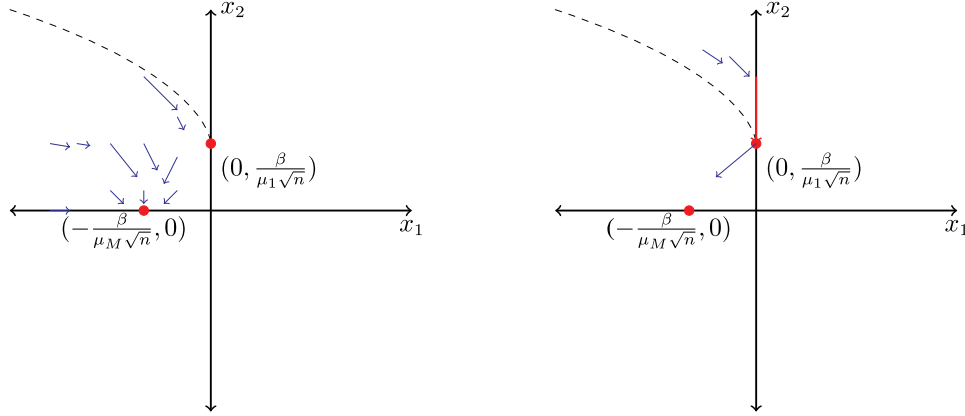
$$dv_2^x(t) = -\frac{\beta}{\sqrt{n}} dt. \quad (6.60)$$

The above equations hold as long as $\mathbf{v}^x$ remains within $\bar{\Omega}$. Hence, for any $\kappa \geq \beta/\mu_1$, if we define $\tau^\kappa(\mathbf{x}) = \inf\{t \geq 0 : v_2^x(t) = \kappa/\sqrt{n}\}$, then $\tau^{\beta/\mu_1}(\mathbf{x}) \geq \tau(\mathbf{x})$ if $\mathbf{x} \geq \Gamma^{\beta/\mu_1}$. Furthermore, for $t \in [\tau(\mathbf{x}), \tau^{\beta/\mu_1}(\mathbf{x})]$, we can characterize $v_2^x(t)$ using (6.60) as

$$v_2^x(t) = -\frac{\beta}{\sqrt{n}}(t - \tau(\mathbf{x})) + x_2 e^{-\mu_1 \tau(\mathbf{x})}. \quad (6.61)$$

After leaving the set $\bar{\Omega}$ at time $\tau^{\beta/\mu_1}(\mathbf{x})$, the trajectory of $\mathbf{v}^x$ never re-enters the set $\bar{\Omega}$ because $v_2^x(\cdot)$ is nonincreasing in time. From Lemma 10, it also follows that if $\mathbf{x} < \Gamma^{\beta/\mu_1}$, then $\tau(\mathbf{x}) = \infty$, and therefore, $v_1^x(t)$ and $v_2^x(t)$ are given by (6.53) and (6.51), respectively, for all $t \geq 0$.

Figure 3. Any Trajectory That Begins Below the Dashed Curve Representing the Curve Γ^{β/μ_1} Will Not Intersect the Vertical Axis, While Trajectories Starting Above the Curve Will Eventually Reach the Axis and Move Downward Until Arriving at the Point $(0, \beta = \mu_1 \sqrt{n})$



Taking κ as in Lemma 8, Equation (6.55) implies that the set $\{\mathbf{x} \in \Omega : x_2 \leq \kappa/\sqrt{n}, \mathbf{x} > \Gamma^\kappa\}$ is empty. Hence, the set Ω can be partitioned as $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$, where

$$\Omega_1 = \{x_2 \leq \kappa/\sqrt{n}\}, \Omega_2 = \{x_2 > \kappa/\sqrt{n} \text{ and } \mathbf{x} \leq \Gamma^\kappa\} \text{ and } \Omega_3 = \{\mathbf{x} > \Gamma^\kappa\}.$$

We are now ready to evaluate (6.45) for each $\mathbf{x} \in \Omega_i$ for $i \in [3]$. Because v_2^* is nonincreasing in time, we have $v_2^*(s) \leq v_2^*(0) = x_2 \leq \kappa/\sqrt{n}$ for all $s \geq 0$ and for all $\mathbf{x} \in \Omega_1$. Hence, from (6.45), we have $f^*(\mathbf{x}) = 0$ for all $\mathbf{x} \in \Omega_1$.

For $\mathbf{x} \in \Omega_2$, we have $\tau^\kappa(\mathbf{x}) \leq \tau(\mathbf{x})$ because otherwise we would have $x_2 e^{-\mu_1 \tau(\mathbf{x})} > \kappa/\sqrt{n}$, which contradicts the fact that $\mathbf{x} \leq \Gamma^\kappa$ according to Lemma 10. Because v_2^* is given by (6.51) for all $t \in [0, \tau(\mathbf{x})]$, we have $\tau^\kappa(\mathbf{x}) = \mu_1^{-1} \log(\sqrt{n}x_2/\kappa)$ and therefore

$$f^*(\mathbf{x}) = \int_0^{\frac{1}{\mu_1} \log(\frac{\sqrt{n}x_2}{\kappa})} \left(x_2 e^{-\mu_1 t} - \frac{\kappa}{\sqrt{n}} \right) dt = -\frac{x_2}{\mu_1} e^{-\mu_1 t} - \frac{\kappa}{\sqrt{n}} t \Big|_0^{\frac{1}{\mu_1} \log(\frac{\sqrt{n}x_2}{\kappa})} = \frac{x_2}{\mu_1} - \frac{\kappa}{\mu_1 \sqrt{n}} - \frac{\kappa}{\mu_1 \sqrt{n}} \log\left(\frac{\sqrt{n}x_2}{\kappa}\right).$$

Finally, for $\mathbf{x} \in \Omega_3$, we have $\tau^\kappa(\mathbf{x}) \geq \tau(\mathbf{x})$ and therefore

$$\begin{aligned} f^*(\mathbf{x}) &= \int_0^{\tau(\mathbf{x})} \left(x_2 e^{-\mu_1 t} - \frac{\kappa}{\sqrt{n}} \right) dt + \int_0^{-\frac{\sqrt{n}}{\beta} \left(\frac{\kappa}{\sqrt{n}} + x_2 e^{-\mu_1 \tau(\mathbf{x})} \right)} \left(-\frac{\beta}{\sqrt{n}} t + x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right) dt \\ &= -\frac{x_2}{\mu_1} e^{-\mu_1 t} - \frac{\kappa}{\sqrt{n}} t \Big|_0^{\tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}} t^2 + \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right) t \Big|_0^{-\frac{\sqrt{n}}{\beta} \left(\frac{\kappa}{\sqrt{n}} + x_2 e^{-\mu_1 \tau(\mathbf{x})} \right)} \\ &= \frac{x_2}{\mu_1} - \frac{x_2}{\mu_1} e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \tau(\mathbf{x}) + \frac{\beta}{2\sqrt{n}} \left(\frac{\sqrt{n}}{\beta} x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\beta} \right)^2 \\ &= \frac{x_2}{\mu_1} - \frac{x_2}{\mu_1} e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \tau(\mathbf{x}) + \frac{\sqrt{n}}{2\beta} \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right)^2. \end{aligned}$$

Aggregating the above expressions, we obtain the candidate solution $f^*(\mathbf{x})$ to the PDE (6.22)–(6.23) as

$$f^*(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \Omega_1, \\ \frac{x_2}{\mu_1} - \frac{\kappa}{\mu_1 \sqrt{n}} - \frac{\kappa}{\mu_1 \sqrt{n}} \log\left(\frac{\sqrt{n}x_2}{\kappa}\right), & \mathbf{x} \in \Omega_2, \\ \frac{x_2}{\mu_1} - \frac{x_2}{\mu_1} e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \tau(\mathbf{x}) + \frac{\sqrt{n}}{2\beta} \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right)^2, & \mathbf{x} \in \Omega_3. \end{cases} \quad (6.62)$$

In Appendix F, we complete the proof of Lemma 8 by verifying that the above function indeed satisfies (6.22)–(6.23).

7. Asymptotic Optimality of the SA-JSQ Scheme: Proof of Theorem 4

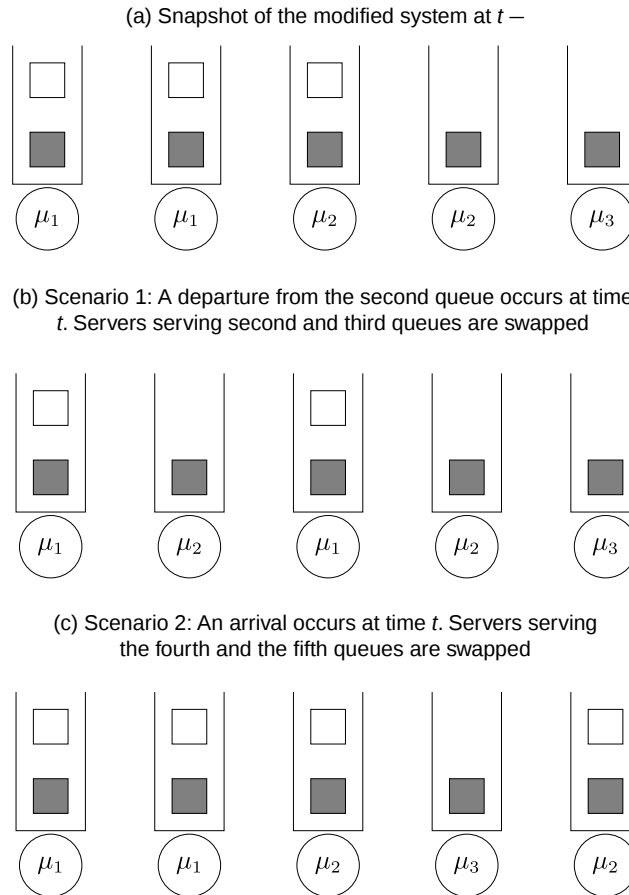
In this section, we establish the asymptotic optimality of the SA-JSQ scheme in the Halfin-Whitt regime. To demonstrate this optimality, we first derive a lower bound on the performance achievable by any admissible load balancing scheme for the original system. This lower bound is obtained by coupling the original system with a modified system in which the servers are not attached to the queues but can freely move between queues ensuring that faster servers always handle the longer queues. We then show that this modified system shares the same diffusion limit as the original system when operating under the SA-JSQ scheme.

7.1. Modified System with Free Servers and Blocking

To identify a lower bound on the achievable performance of admissible load balancing policies, we consider a *modified system* where, similar to the original system, in each pool $j \in [M]$, there are $N_j^{(n)}$ servers with rate μ_j and $\sum_{j=1}^M N_j^{(n)} = n$ queues. However, contrary to the original system, the queues are not attached to the servers. Instead, at any time $t \geq 0$, the servers rearrange themselves in a preemptive fashion so that longer queues are always served by faster servers. We further assume that each queue has a buffer capacity of two including the job in service at that queue and arrivals occurring when all queues have two jobs are *rejected/blocked*. Each arriving job is irrevocably assigned to a queue according to the conventional JSQ policy that picks the shortest queue (breaking ties uniformly at random) for each incoming arrival.

To better understand the dynamics of the modified system, consider a system where there are two type 1 servers with rate μ_1 , two type 2 servers with rate μ_2 , and one type 3 server with rate μ_3 . Suppose that right before time t (i.e., at time $t-$), there are three queues with two jobs and two queues with one jobs as illustrated in Figure 4(a), where at each queue the first job at the head-of-the-line (shaded job) is being served. As the servers always need to organize themselves so that the faster servers serve the longer queues, we need to have two type 1 servers and one type two server serving queues with two jobs and one type 2 server and one type 3 server serving queues with one job at $t-$. Now, consider the following two scenarios. For the first scenario, suppose that a departure occurs at t

Figure 4. Modified System with Two Type 1 Servers, Two Type 2 Servers, and One Type 3 Server



from the second queue with two jobs being served by a type 1 server. Then, to ensure that the longer queues are served by faster servers, the type 1 server serving this queue will change places with the type 2 server serving the queue with two jobs as illustrated in Figure 4(b). As another scenario, suppose that an arrival occurs at time t , and following the JSQ policy, it is assigned to the fifth queue that was being served by a type 3 server. Then, at time t , the type 2 server serving the queue with one job should change places with the type 3 server as illustrated in Figure 4(c).

Similar to the original system, we define the following quantities for the modified system:

- $\tilde{Q}_i^{(n)}(t) :=$ The number of queues with i or more jobs at time t
- $\tilde{A}^{(n)}(t) :=$ The cumulative number of external arrivals by time t
- $\tilde{A}_1^{(n)}(t) :=$ The number of arrivals routed to queues with exactly 1 job by time t
- $\tilde{A}_R^{(n)}(t) :=$ The number of rejected jobs by time t
- $\tilde{D}_i^{(n)}(t) :=$ The total number of departures from servers with i or more jobs by time t .

As the arrivals are routed in a JSQ fashion where the ties are broken randomly, we have

$$\begin{aligned}\tilde{A}_1^{(n)}(t) &:= \int_0^t \mathbb{1} \left(\tilde{Q}_1^{(n)}(s-) = \sum_{j=1}^M N_j^{(n)} \text{ and } \tilde{Q}_2^{(n)}(s-) < \sum_{j=1}^M N_j^{(n)} \right) d\tilde{A}^{(n)}(s) \\ \tilde{A}_R^{(n)}(t) &:= \int_0^t \mathbb{1} \left(\tilde{Q}_2^{(n)}(s-) = \sum_{j=1}^M N_j^{(n)} \right) d\tilde{A}^{(n)}(s).\end{aligned}$$

Then, the number of arrivals routed to idle servers is given by $\tilde{A}^{(n)}(t) - \tilde{A}_R^{(n)}(t) - \tilde{A}_1^{(n)}(t)$ for any time $t \geq 0$. Finally, because we always have the faster servers serving the longer queues, the departure rate for queues with i or more jobs at time t is given by

$$\tilde{\mu}^{(n)}(\tilde{Q}_i^{(n)}(t-)) := \sum_{j=1}^M \mu_j \left[\left(\tilde{Q}_i^{(n)}(t-) - \sum_{k=1}^{j-1} N_k^{(n)} \right)_+ \wedge N_j^{(n)} \right]. \quad (7.1)$$

From lemma 16 of Bhambay and Mukhopadhyay (2022), it follows that for any $0 \leq Q_{j,i}^{(n)}(t-) \leq N_j^{(n)}$, we have

$$\tilde{\mu}^{(n)} \left(\sum_{j=1}^M Q_{j,i}^{(n)}(t-) \right) \geq \sum_{j=1}^M \mu_j Q_{j,i}^{(n)}(t-). \quad (7.2)$$

Now, we have the following balance equations:

$$\tilde{Q}_1^{(n)}(t) = \tilde{Q}_1^{(n)}(0) + \tilde{A}^{(n)}(t) - \tilde{A}_R^{(n)}(t) - \tilde{A}_1^{(n)}(t) - \tilde{D}_1^{(n)}(t) + \tilde{D}_2^{(n)}(t), \quad (7.3)$$

$$\tilde{Q}_2^{(n)}(t) = \tilde{Q}_2^{(n)}(0) + \tilde{A}_1^{(n)}(t) - \tilde{D}_2^{(n)}(t). \quad (7.4)$$

To present key results for the modified system, we define the following quantity for any $t \geq 0$:

$$Q_{+2}^{(n,\Pi)}(t) = \sum_{j=1}^M \sum_{i=2}^{\infty} Q_{j,i}^{(n)}(t),$$

which represents the number of servers with at least two jobs in the original system under any admissible load balancing policy Π . With this definition in place, we now introduce the following coupling proposition.

Proposition 3. Consider the original system under any admissible load balancing policy Π and the modified system with initial states

$$Q_1^{(n,\Pi)}(0) = \tilde{Q}_1^{(n)}(0) \text{ and } Q_{+2}^{(n,\Pi)}(0) = \tilde{Q}_2^{(n)}(0) \leq \sum_{j=1}^M N_j^{(n)}.$$

Then, there exists a coupling between the two systems such that for any $t \geq 0$, we have $\tilde{Q}_1^{(n)}(t) + \tilde{Q}_2^{(n)}(t) \leq Q_1^{(n,\Pi)}(t) + Q_{+2}^{(n,\Pi)}(t)$ and $\tilde{Q}_2^{(n)}(t) \leq Q_{+2}^{(n,\Pi)}(t)$.

The proposition above implies that if Π is any stable admissible policy then $\mathbb{P}(\tilde{Q}_1^{(n)}(\infty) + \tilde{Q}_2^{(n)}(\infty) > y) \leq \mathbb{P}(Q_1^{(n,\Pi)}(\infty) + Q_{+2}^{(n,\Pi)}(\infty) > y)$ for all $y \in \mathbb{R}$.

We now proceed in the next proposition to prove that the diffusion limit of the modified system is identical to that of the original system operating under the SA-JSQ scheme. To state the proposition, we need the following additional notations. We define the scaled process $\tilde{\mathbf{Y}}^{(n)} = (\tilde{\mathbf{Y}}(t) : t \geq 0)$, where $\tilde{\mathbf{Y}}(t) = (\tilde{Y}_1^{(n)}(t), \tilde{Y}_2^{(n)}(t))$ with

$$\tilde{Y}_1^{(n)}(t) = \frac{\tilde{Q}_1^{(n)}(t) - n}{\sqrt{n}}, \quad \tilde{Y}_2^{(n)}(t) = \frac{\tilde{Q}_2^{(n)}(t)}{\sqrt{n}}. \quad (7.5)$$

Proposition 4. Suppose that (A1) and (A2) hold, and there exists a random vector $\tilde{\mathbf{Y}}(0) = (\tilde{Y}_1(0), \tilde{Y}_2(0))$ with $\tilde{Y}_1(0) = 0$ and $\tilde{Y}_2(0) \in \mathbb{R}_+$ such that

$$\tilde{\mathbf{Y}}^{(n)}(0) = \tilde{\mathbf{Y}}(0), \quad \forall n \geq 1. \quad (7.6)$$

Then, we have

$$\tilde{\mathbf{Y}}^{(n)} \Rightarrow \tilde{\mathbf{Y}}, \text{ as } n \rightarrow \infty,$$

where $\tilde{\mathbf{Y}} = (\tilde{Y}_i, i \in [2])$ is the unique solution in $\mathbb{D}_{\mathbb{R}^2}[0, \infty)$ of the following stochastic integral equations:

$$\tilde{Y}_1(t) = \tilde{Y}_1(0) + \sqrt{2}W(t) - \beta t - \int_0^t (\mu_M \tilde{Y}_1(s) - \mu_1 \tilde{Y}_2(s))ds - \tilde{U}_1(t), \quad (7.7)$$

$$\tilde{Y}_2(t) = \tilde{Y}_2(0) - \mu_1 \int_0^t \tilde{Y}_2(s)ds + \tilde{U}_1(t). \quad (7.8)$$

In the above, W is a standard Brownian motion, and $\tilde{U}_1$ is the unique nondecreasing and nonnegative process satisfying

$$\int_0^\infty \mathbb{1}(\tilde{Y}_1(t) < 0) d\tilde{U}_1(t) = 0. \quad (7.9)$$

Note that the limiting diffusion process $\tilde{\mathbf{Y}}$ given in the above theorem is the same as the two-dimensional process defined by (3.9)–(3.10). Hence, we can now prove the asymptotic optimality of the SA-JSQ scheme as follows.

Proof of Theorem 4. The diffusion-scaled total number of jobs $\tilde{Y}_{+1}^{(n)}$ and the diffusion-scaled number of waiting jobs $\tilde{Y}_{+2}^{(n)}$ in the modified system are given by $\tilde{Y}_{+1}^{(n)} = (\tilde{Q}_1^{(n)}(t) + \tilde{Q}_2^{(n)}(t) - n)/\sqrt{n}$ and $\tilde{Y}_{+2}^{(n)} = \tilde{Q}_2^{(n)}(t)/\sqrt{n}$, respectively. The same quantities for the original system operating under any general policy Π are given by $\tilde{Y}_{+1}^{(n,\Pi)} = (\tilde{Q}_1^{(n,\Pi)}(t) + \tilde{Q}_{+2}^{(n,\Pi)}(t) - n)/\sqrt{n}$ and $\tilde{Y}_{+2}^{(n,\Pi)} = \tilde{Q}_{+2}^{(n,\Pi)}(t)/\sqrt{n}$, respectively. Now, by Proposition 3, we have $\tilde{Q}_1^{(n)}(t) + \tilde{Q}_2^{(n)}(t) \leq_{\text{st}} Q_1^{(n,\Pi)}(t) + Q_{+2}^{(n,\Pi)}(t)$ for any $t \geq 0$. Therefore, assuming that the steady-state distribution exists under policy Π , we obtain after centering and scaling

$$\mathbb{P}(\tilde{Y}_{+1}^{(n)}(\infty) > y) \leq \mathbb{P}(Y_{+1}^{(n,\Pi)}(\infty) > y), \quad \forall y \in \mathbb{R}.$$

Hence, we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(\tilde{Y}_{+1}^{(n)}(\infty) > y) = \lim_{n \rightarrow \infty} \mathbb{P}(Y_{+1}^{(n)}(\infty) > y) \leq \liminf_{n \rightarrow \infty} \mathbb{P}(Y_{+1}^{(n,\Pi)}(\infty) > y), \quad \forall y \in \mathbb{R},$$

where the equality follows from Proposition 4. This proves (3.16). The proof of (3.17) follows similarly using stochastic ordering of $\tilde{Q}_2^{(n)}(t) \leq Q_{+2}^{(n,\Pi)}(t)$ for any $t \geq 0$. $\square$

7.2. Coupling Construction: Proof of Proposition 3

In this section, we first construct a coupling that enables us to relate the modified system presented above to the original system operating under any load balancing policy. To construct the original system under any

admissible load balancing policy Π , we define

$A_{j,i}^{(n,\Pi)}(t) :=$ The number of arrivals assigned to queues with i or more jobs in pool $j \in [M]$ by time t

$A_i^{(n,\Pi)}(t) :=$ The total number of arrivals assigned to queues with i or more jobs by time t

$D_{j,i}^{(n,\Pi)}(t) :=$ The number of departures from queues with i or more jobs in pool $j \in [M]$ by time t

$D_i^{(n,\Pi)}(t) :=$ The total number of departures from queues with i or more jobs by time t

$$Q_1^{(n,\Pi)}(t) := \sum_{j=1}^M Q_{j,1}^{(n,\Pi)}(t)$$

$$Q_{+2}^{(n,\Pi)}(t) := \sum_{j=1}^M \sum_{i=2}^{\infty} Q_{j,i}^{(n,\Pi)}(t).$$

We note that the processes $A_{j,i}^{(n,\Pi)}(t)$ is determined by the load balancing policy Π and is not used explicitly in the remainder of this section. We only need that for all $t \geq 0$ the following equations hold:

$$A_i^{(n,\Pi)}(t) = \sum_{j=1}^M A_{j,i}^{(n,\Pi)}(t) \text{ and } D_i^{(n,\Pi)}(t) = \sum_{j=1}^M D_{j,i}^{(n,\Pi)}(t).$$

Furthermore, we have the following balance equations:

$$Q_1^{(n,\Pi)}(t) = Q_1^{(n,\Pi)}(0) + A_0^{(n,\Pi)}(t) - A_1^{(n,\Pi)}(t) - D_1^{(n,\Pi)}(t) + D_2^{(n,\Pi)}(t), \quad (7.10)$$

$$Q_{+2}^{(n,\Pi)}(t) = Q_{+2}^{(n,\Pi)}(0) + A_1^{(n,\Pi)}(t) - D_2^{(n,\Pi)}(t). \quad (7.11)$$

Now, to couple the original system and the modified system, we take $A(t)$ and $D(t)$ to be two unit rate Poisson processes and assume that, for both original and modified systems, we have

$$A_0^{(n,\Pi)}(t) = \tilde{A}^{(n)}(t) = A(n\lambda^{(n)}t), \quad (7.12)$$

that is, both original and the modified systems have the same arrival streams. For both systems, we define the “potential” departure process as

$$D_p^{(n)}(t) := D \left(\sum_{j=1}^M \mu_j N_j^{(n)} \right), \text{ for all } t \geq 0,$$

and we let $\theta_l^{(n)}$ to be the l th event epoch for $D_p^{(n)}(t)$ and take $\{U_l^{(n)}\}_{l \in \mathbb{Z}_+}$ to be a sequence of independent uniform(0,1) random variables.

To convert the potential departure process to the actual departure process in the original system, we first order the pairs of positive integers (j, i) such that $(j, i) \geq (j', i')$ if $i > i'$ or $i = i'$ and $j < j'$. Then, the potential departure at epoch $\theta_l^{(n)}$ is assumed to contribute an actual departure for $D_{j,i}^{(n,\Pi)}(t)$, that is, $D_{j,i}^{(n,\Pi)}(\theta_l^{(n)}) - D_{j,i}^{(n,\Pi)}(\theta_l^{(n)} -) = 1$, if and only if (j, i) is the supremum pair for which the following holds:

$$U_l^{(n)} \leq \frac{\sum_{k=1}^j \mu_k Q_{k,i}^{(n,\Pi)}(\theta_l^{(n)} -)}{\sum_{j=1}^M \mu_j N_j^{(n)}}, \quad (7.13)$$

where the supremum should be understood in terms of the ordering above. If no pair (j, i) satisfying the above inequality exists, then the potential departure assumed to be lost; that is, it does not correspond to an actual departure.

In the modified system, the potential departure occurring at time $\theta_l^{(n)}$ is an actual departure from a queue with two customers if

$$U_l^{(n)} \leq \frac{\tilde{\mu}^{(n)}(\tilde{Q}_2^{(n)}(t-))}{\sum_{j=1}^M \mu_j N_j^{(n)}}, \quad (7.14)$$

or an actual departure from a queue with one customer if

$$\frac{\tilde{\mu}^{(n)}(\tilde{Q}_2^{(n)}(t-))}{\sum_{j=1}^M \mu_j N_j^{(n)}} < U_l^{(n)} \leq \frac{\tilde{\mu}^{(n)}(\tilde{Q}_1^{(n)}(t-))}{\sum_{j=1}^M \mu_j N_j^{(n)}}. \quad (7.15)$$

The potential departure is assumed to be lost (i.e., does not correspond to an actual departure) if none of the above two conditions is satisfied.

Proof of Proposition 3. We prove the result with a contradiction. First, define $\tau_1^{(n)} := \inf\{t \geq 0 : \tilde{Q}_2^{(n)}(t) > Q_{+2}^{(n,\Pi)}(t)\}$. If $\tau_1^{(n)} < \infty$, we need to have $\tilde{Q}_2^{(n)}(\tau_1^{(n)}-) = Q_{+2}^{(n,\Pi)}(\tau_1^{(n)}-)$, which together with the assumption on initial conditions, (7.4) and (7.11), yields

$$A_1^{(n,\Pi)}(\tau_1^{(n)}-) - D_2^{(n,\Pi)}(\tau_1^{(n)}-) = \tilde{A}_1^{(n)}(\tau_1^{(n)}-) - \tilde{D}_2^{(n)}(\tau_1^{(n)}-), \text{ and} \quad (7.16)$$

$$A_1^{(n,\Pi)}(t) - D_2^{(n,\Pi)}(t) \geq \tilde{A}_1^{(n)}(t) - \tilde{D}_2^{(n)}(t) \text{ for all } t < \tau_1^{(n)}. \quad (7.17)$$

Equations (7.2), (7.13), and (7.14) imply that the event occurring at time $\tau_1^{(n)}$ cannot be a departure. Hence, we need to have an arrival at $\tau_1^{(n)}$, this arrival should be routed to a queue with one job in the modified system, and it should be routed to an idle queue in the original system, which can only happen if

$$Q_1^{(n,\Pi)}(\tau_1^{(n)}-) < \tilde{Q}_1^{(n)}(\tau_1^{(n)}-) = \sum_{j=1}^M N_j^{(n)}.$$

Combining (7.12), (7.16), (7.3), and (7.10), we need to have

$$\tilde{A}_R^{(n)}(\tau_1^{(n)}-) + \tilde{D}_1^{(n)}(\tau_1^{(n)}-) < D_1^{(n,\Pi)}(\tau_1^{(n)}-).$$

Hence, there should exist a minimal $\tau_2^{(n)} < \tau_1^{(n)}$ where the equality $\tilde{A}_R^{(n)}(s) + \tilde{D}_1^{(n)}(s) = D_1^{(n,\Pi)}(s)$ is violated (note that this equality trivially holds at $s=0$). We now show that this is not possible even when $\tau_1^{(n)} = \infty$. Using a similar logic as above, if such a $\tau_2^{(n)}$ exists, we need to have

$$\tilde{A}_R^{(n)}(\tau_2^{(n)}-) + \tilde{D}_1^{(n)}(\tau_2^{(n)}-) = D_1^{(n,\Pi)}(\tau_2^{(n)}-).$$

This can only happen if there is a departure only in the original system at time $\tau_2^{(n)}$, that is,

$$\tilde{\mu}^{(n)}(\tilde{Q}_1^{(n)}(\tau_2^{(n)}-)) < \sum_{j=1}^M \mu_j Q_{1,j}^{(n,\Pi)}(\tau_2^{(n)}-).$$

However, using the inequality in (7.16) and (7.17) in combination with (7.3) and (7.10), we see that $\tilde{Q}_1^{(n)}(\tau_2^{(n)}-) \geq Q_1^{(n)}(\tau_2^{(n)}-)$. The monotonicity of $\tilde{\mu}(\cdot)$ and (7.2) implies that a departure only in the original system at $\tau_2^{(n)}$ is not possible. Hence, we can conclude first that $\tau_1^{(n)} = \infty$ and $\tilde{Q}_2^{(n)}(t) \leq Q_{+2}^{(n,\Pi)}(t)$. The nonexistence of $\tau_2^{(n)} < \tau_1^{(n)} = \infty$ implies that $\tilde{A}_R^{(n)}(t) + \tilde{D}_1^{(n)}(t) \geq D_1^{(n,\Pi)}(t)$ for all $t \geq 0$, and hence summing (7.3), (7.4), (7.10), and (7.11), we see that $\tilde{Q}_1^{(n)}(t) + \tilde{Q}_2^{(n)}(t) \leq Q_1^{(n,\Pi)}(t) + Q_{+2}^{(n,\Pi)}(t)$. $\square$

7.3. Diffusion Limit of the Modified System: Proof of Proposition 4

To prove Proposition 4, we follow similar steps as in the proof of Theorem 1. In particular, (i) we first define a stopping time $\tilde{T}_n$ to be the first time the system reaches a state with at least $N_M^{(n)}/2$ idle servers or at least $N_1^{(n)}$ queues with two jobs, (ii) we then derive the diffusion limit of the system for all $t \in [0, \tilde{T}_n]$, (iii) and finally show that, for any finite t , the probability of $T_n \leq t$ decreases to zero as $n \rightarrow \infty$ to obtain process-level convergence. Finally, we present a Lyapunov-type argument to argue tightness of the stationary distributions to conclude interchange of limits.

Proof of Proposition 4. To derive the diffusion limit for the modified system, we first define the stopping time $\tilde{T}_n$ as follows:

$$\tilde{T}_n = \inf\{t \geq 0 : n - \tilde{Q}_1^{(n)}(t) \geq N_M^{(n)}/2 \text{ or } \tilde{Q}_2^{(n)}(t) = N_1^{(n)}\}. \quad (7.18)$$

This definition implies before time $\tilde{T}_n$ only the slowest servers (i.e., servers in pool M) can be idle and no job can be blocked. Hence, using the unit rate Poisson processes $\tilde{A}(t)$, $\tilde{D}_{u1}(t)$ and $\tilde{D}_{u2}(t)$, the evolution of the modified

system up to time $\tilde{T}_n$ can be described by the following equations:

$$\tilde{Q}_1^{(n)}(t) = \tilde{Q}_1^{(n)}(0) + \tilde{A}(n\lambda^{(n)}t) - \tilde{D}_{u1} \left(\sum_{j=1}^M N_j^{(n)} \mu_j t - \int_0^t (\mu_M(n - \tilde{Q}_1^{(n)}(s)) + \mu_1 \tilde{Q}_2^{(n)}(s)) ds \right) - \tilde{U}_1^{(n)}(t), \quad (7.19)$$

$$\tilde{Q}_2^{(n)}(t) = \tilde{Q}_2^{(n)}(0) - \tilde{D}_{u2} \left(\int_0^t \mu_1 \tilde{Q}_2^{(n)}(s) ds \right) + \tilde{U}_1^{(n)}(t), \quad (7.20)$$

where $\tilde{U}_1^{(n)}(t) = \tilde{A}_1^{(n)}(t) = \int_0^t \mathbb{1}(Q_1^{(n)}(s) = n) d\tilde{A}(n\lambda^{(n)}s)$ for all $0 \leq t \leq \tilde{T}_n$. Upon centering and scaling the processes as in Section 5, we get

$$\tilde{Y}_1^{(n)}(t) = \tilde{Y}_1^{(n)}(0) + \tilde{M}_A^{(n)}(t) - \tilde{M}_{D,1}^{(n)}(t) + \frac{n\lambda^{(n)} - \sum_{j=1}^M \mu_j N_j^{(n)}}{\sqrt{n}} t - \mu_M \int_0^t \tilde{Y}_1^{(n)}(s) ds + \mu_1 \int_0^t \tilde{Y}_2^{(n)}(s) ds - \tilde{V}_1^{(n)}(t), \quad (7.21)$$

$$\tilde{Y}_2^{(n)}(t) = \tilde{Y}_2^{(n)}(0) - \tilde{M}_{D,2}^{(n)}(t) - \int_0^t \mu_1 \tilde{Y}_2^{(n)}(s) ds + \tilde{V}_1^{(n)}(t), \quad (7.22)$$

where

$$\begin{aligned} \tilde{M}_A^{(n)}(t) &= \frac{1}{\sqrt{n}} (\tilde{A}(n\lambda^{(n)}t) - n\lambda^{(n)}t), \\ \tilde{M}_{D,1}^{(n)}(t) &= \frac{1}{\sqrt{n}} \tilde{D}_{u1} \left(\sum_{j=1}^M N_j^{(n)} \mu_j t - \int_0^t (\mu_M(n - \tilde{Q}_1^{(n)}(s)) + \mu_1 \tilde{Q}_2^{(n)}(s)) ds \right) \\ &\quad - \frac{1}{\sqrt{n}} \left(\sum_{j=1}^M N_j^{(n)} \mu_j t - \int_0^t (\mu_M(n - \tilde{Q}_1^{(n)}(s)) + \mu_1 \tilde{Q}_2^{(n)}(s)) ds \right), \\ \tilde{M}_{D,2}^{(n)}(t) &= \frac{1}{\sqrt{n}} \left(\tilde{D}_{u2} \left(\mu_1 \int_0^t \tilde{Q}_2^{(n)}(s) ds \right) - \int_0^t \mu_1 \tilde{Q}_2^{(n)}(s) ds \right), \\ \tilde{V}_1^{(n)}(t) &= \frac{\tilde{U}_1^{(n)}(t)}{\sqrt{n}}. \end{aligned}$$

Assumptions (A1) and (A2) imply that the following convergence holds uniformly on compact intervals

$$\frac{n\lambda^{(n)} - \sum_{j=1}^M \mu_j N_j^{(n)}}{\sqrt{n}} t \rightarrow -\beta e(t),$$

where $e(t) = t$ for all t . Hence, the continuous mapping theorem and Lemma A.1 in Appendix A imply that the solution of (7.21) and (7.22) converges to the solution of (7.7)–(7.9), once we prove that $\tilde{M}_A^{(n)}(t)$ and $\tilde{M}_{D,1}^{(n)}(t)$ each converges weakly to standard Brownian motions and $\tilde{M}_{D,2}^{(n)}(t)$ converges weakly to the zero function. As in the proof of Lemma 4, we have that $\tilde{M}_A^{(n)} \Rightarrow W_{\tilde{A}}$ because $\langle \tilde{M}_A^{(n)} \rangle(t) = \lambda^{(n)} t \rightarrow t$, where $W_{\tilde{A}}$ is a standard Brownian motion. Furthermore, the predictable quadratic variations $\langle \tilde{M}_{D,1}^{(n)} \rangle$ and $\langle \tilde{M}_{D,2}^{(n)} \rangle$ of the martingales $\tilde{M}_{D,1}^{(n)}$ and $\tilde{M}_{D,2}^{(n)}$ satisfy

$$\begin{aligned} \langle \tilde{M}_{D,1}^{(n)} \rangle(t) &= \frac{\sum_{j=1}^M N_j^{(n)} \mu_j t}{n} - \int_0^t \left(\frac{\mu_M(n - \tilde{Q}_1^{(n)}(s))}{n} + \frac{\mu_1 \tilde{Q}_2^{(n)}(s)}{n} ds \right), \\ &\leq \frac{\sum_{j=1}^M N_j^{(n)} \mu_j t}{n} \\ \langle \tilde{M}_{D,2}^{(n)} \rangle(t) &= \frac{\int_0^t \mu_1 \tilde{Q}_2^{(n)}(s) ds}{n} \leq \frac{\mu_1 N_1^{(n)} t}{n}, \end{aligned}$$

which by (A2) and Lemma 1 imply that $\{\tilde{M}_{D,1}^{(n)}\}_n$ and $\{\tilde{M}_{D,2}^{(n)}\}_n$ are stochastically bounded. Hence, Lemma 3 implies that $\{\tilde{Y}_1^{(n)}\}_n$ and $\{\tilde{Y}_2^{(n)}\}_n$ are stochastically bounded. Therefore, by lemma 5.9 of Pang et al. (2007), we

have $\tilde{Y}_1^{(n)}/\sqrt{n} = (\tilde{Q}_1^{(n)} - n)/n \Rightarrow 0$ and $\tilde{Y}_2^{(n)}/\sqrt{n} = \tilde{Q}_2^{(n)}/n \Rightarrow 0$. Therefore, by (A2) and continuous mapping theorem, we have $\langle \tilde{M}_{D,1}^{(n)} \rangle \Rightarrow e$ and $\langle \tilde{M}_{D,2}^{(n)} \rangle \Rightarrow 0$. This proves that the solution of (7.21) and (7.22) converges to the solution of (7.7) and (7.8).

To complete the proof, we need to show that $\mathbb{P}(\tilde{T}_n \leq t) \rightarrow 0$ for any finite $t \geq 0$. We have

$$\mathbb{P}(\tilde{T}_n \leq t) \leq \mathbb{P}\left(\sup_{0 \leq s \leq t} |\tilde{Y}_1^{(n)}(s)| \geq \frac{N_M^{(n)}}{2\sqrt{n}}\right) + \mathbb{P}\left(\sup_{0 \leq s \leq t} \tilde{Y}_2^{(n)}(s) \geq \frac{N_1^{(n)}}{\sqrt{n}}\right).$$

Assumption (A2) and the stochastic boundedness of $\tilde{Y}_1^{(n)}$ and $\tilde{Y}_2^{(n)}$ imply that the right-hand side (RHS) of the above inequality goes to zero as $n \rightarrow \infty$.

To prove the convergence of stationary distributions, we need to prove that the sequences $\{\tilde{Y}_1^{(n)}(\infty)\}_n$ and $\{\tilde{Y}_2^{(n)}(\infty)\}_n$ are tight. From Proposition 3, we have $0 \leq \tilde{Y}_2^{(n)}(\infty) \leq_{st} Y_{+2}^{(n)}(\infty)$. Furthermore, Theorem 3 implies that the sequence $\{Y_{+2}^{(n)}(\infty)\}_n$ weakly converges to $Y_{1,2}(\infty)$ and is therefore tight. Hence, $\{\tilde{Y}_2^{(n)}(\infty)\}_n$ is tight. To prove the tightness of $\{\tilde{Y}_1^{(n)}(\infty)\}_n$, we consider the Lyapunov function $\tilde{V}$ defined for any state $\tilde{q}$ of the modified system as

$$\tilde{V}(\tilde{q}) = -\tilde{y}_1 \geq 0.$$

By definition, we have

$$\tilde{\mu}^{(n)}(\tilde{q}_1) \leq \sum_{j=1}^M N_j^{(n)} \mu_j - \sqrt{n} \mu_M \tilde{V}(\tilde{q}).$$

Hence, applying the generator $\tilde{\mathcal{G}}$ of the process $\tilde{\mathbf{Q}}^{(n)}$ to $\tilde{V}$, we obtain

$$\tilde{\mathcal{G}}\tilde{V}(\tilde{q}) = -\sqrt{n}\lambda^{(n)}\mathbb{1}(\tilde{y}_1 < 0) + \frac{\tilde{\mu}^{(n)}(\tilde{q}_1)}{\sqrt{n}} \leq -\sqrt{n}\lambda^{(n)} + \frac{\sum_{j=1}^M N_j^{(n)} \mu_j}{\sqrt{n}} - \mu_M \tilde{V}(\tilde{q}) \leq \beta - \mu_M \tilde{V}(\tilde{q}) + o(1),$$

where the last inequality follows from (A1) and (A2). Hence, for $\tilde{q}$ satisfying $\tilde{V}(\tilde{q}) > 2\beta/\mu_M$, we have $\tilde{\mathcal{G}}\tilde{V}(\tilde{q}) \leq -(\beta - o(1))$. Then, applying lemma 10 of Wang et al. (2022), we have

$$\mathbb{E}[\tilde{V}(\tilde{\mathbf{Q}}^{(n)}(\infty))] = \mathbb{E}[-\tilde{Y}_1^{(n)}(\infty)] \leq \frac{4\beta}{\mu_M} + 4 \frac{(\lambda^{(n)} + \sum_{j \in [M]} \mu_j N_j^{(n)}/n) + \beta - o(1)}{\beta - o(1)} = O(1),$$

which shows that $\{\tilde{Y}_1^{(n)}(\infty)\}_n$ is tight. This completes the proof of interchange of limits. $\square$

8. Conclusion

In this paper, we established that the SA-JSQ policy is asymptotically optimal in the Halfin-Whitt traffic regime for a system with heterogeneous servers. In order to prove this result, we needed to do both transient and steady-state analysis of the SA-JSQ policy in addition to finding a suitable system that provides asymptotically tight performance lower bounds for any admissible load balancing policy. The key difficulty in our transient and steady-state analyzes arises from the heterogeneity of the servers that results in a more complex state space for the underlying Markov process. This makes generalizing existing analyzes hard as we need to establish additional state space collapse results and construct Lyapunov functions suited to heterogeneous systems. To the best of our knowledge, our proof of asymptotic optimality is the first in the literature for a heterogeneous system where each server has a dedicate queue. It uses a system that differs fundamentally from those used in earlier works where the number of queues did not scale with the number of servers.

The paper opens up several interesting avenues for further exploration. We stated our optimality result in terms of minimizing the steady-state distribution of the total diffusion-scaled number of jobs and the diffusion-scaled number of waiting jobs in the system. Another natural way to define optimality is through the diffusion-scaled mean response time of jobs and the diffusion-scaled mean waiting time of jobs in the steady state. Although one can relate the expected number of waiting jobs and expected total number of jobs in the system at steady-state to the mean waiting time and mean response times of jobs, respectively, one cannot conclude the optimality in terms of (diffusion-scaled) averages from our result as this requires showing uniform integrability of stationary queue lengths that we have not established. An interesting direction of future research would be to show asymptotic optimality in terms of averages by establishing uniform integrability results as in Braverman (2023).

In this paper, we focused on analyzing the SA-JSQ policy under the condition $\lambda^{(n)} = 1 - \beta/n^\alpha$, with $\alpha = 1/2$. In a related analysis in Bhambay and Mukhopadhyay (2022), the optimality of the SA-JSQ was established for $\alpha = 0$,

corresponding to the fluid limit. However, asymptotically optimal policies remain unknown for other values of α . Analyzing the SA-JSQ scheme for $\alpha = 1$, corresponding to the NDS regime, is of particular interest as it could reveal interesting dynamics.

Finding optimal load balancing policies for nonexponential service time distributions remains another challenging open problem.

Acknowledgments

The authors thank the anonymous referees whose comments have helped us improve the quality of the paper immensely. Their suggestions ultimately motivated us to prove Theorem 4.

Appendix A. Proof of Proposition 2

To prove Proposition 2, we first prove an important result for the first two dimensions involving reflection terms.

Lemma A.1. *For a given $B \in \bar{\mathbb{R}}_+$, $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$, and $\mathbf{z} = (z_1, z_2) \in \mathbb{D}_{\mathbb{R}^2}[0, \infty)$, consider the following integral equations*

$$y_1(t) = b_1 + z_1(t) - \int_0^t (\mu_M y_1(s) - \mu_1 y_2(s)) ds - u_1(t), \quad (\text{A.1})$$

$$y_{1,2}(t) = y_2(t) = b_2 + z_2(t) - \int_0^t \mu_1 y_2(s) ds + u_1(t) - u_2(t), \quad (\text{A.2})$$

$$y_1(t) \leq 0, 0 \leq y_2(t) \leq B, t \geq 0, \quad (\text{A.3})$$

where u_1 and u_2 are nondecreasing and nonnegative functions satisfying

$$\int_0^t \mathbb{1}(y_1(s) < 0) du_1(s) = 0, \quad (\text{A.4})$$

$$\int_0^t \mathbb{1}(y_2(s) < B) du_2(s) = 0. \quad (\text{A.5})$$

Then (A.1)–(A.3) have a unique solution in $(\mathbf{y}, \mathbf{u}) \in \mathbb{D}_{\mathbb{R}^2}[0, \infty) \times \mathbb{D}_{\mathbb{R}^2}[0, \infty)$. Moreover, there exists a well-defined function $(f, g) : \bar{\mathbb{R}}_+ \times \mathbb{R}^2 \times \mathbb{D}_{\mathbb{R}^2}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{R}^2}[0, \infty) \times \mathbb{D}_{\mathbb{R}^2}[0, \infty)$ that maps $(B, \mathbf{b}, \mathbf{z})$ to $\mathbf{y} = f(B, \mathbf{b}, \mathbf{z})$ and $\mathbf{u} = g(B, \mathbf{b}, \mathbf{z})$. Furthermore, the function (f, g) is continuous on $\bar{\mathbb{R}}_+ \times \mathbb{R}^2 \times \mathbb{D}_{\mathbb{R}^2}[0, \infty)$. Finally, $\mathbf{z}$ continuous implies that $\mathbf{y}$ and $\mathbf{u}$ are also continuous.

To prove Lemma A.1, we first define a one sided reflection map function with upper barrier κ as $(\phi_\kappa, \psi_\kappa) : \mathbb{D}_{\mathbb{R}}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{R}^2}[0, \infty)$, where for any $x \in \mathbb{D}_{\mathbb{R}}[0, \infty)$ and for all $t \geq 0$

$$\begin{aligned} \phi_\kappa(x) &= x - \psi_\kappa(x) \leq \kappa, \\ \int_0^\infty \mathbb{1}(\phi_\kappa(x) < \kappa) d\psi_\kappa(x) &= 0 \text{ and} \\ \psi_\kappa(x) &\text{ is nondecreasing and } \psi_\kappa(x(0)) = 0. \end{aligned}$$

It is well known in the literature that the reflection map $(\phi_\kappa, \psi_\kappa)$ is well defined (Whitt 2001, sections 5.2 and 13.5) and can be expressed as

$$\psi_\kappa(x)(t) = \sup_{0 \leq s \leq t} (x(s) - \kappa)^+, \quad (\text{A.6})$$

$$\phi_\kappa(x)(t) = x(t) - \psi_\kappa(x)(t). \quad (\text{A.7})$$

The reflection mapping can be trivially extended to $\kappa = \infty$ as $(\phi_\infty(x), \psi_\infty(x)) = (x, 0)$. The Lipschitz continuity of the reflection mapping, expressed as

$$\|\psi_\kappa(x) - \psi_{\kappa'}(x')\|_t \leq \|x - x'\|_t + |\kappa - \kappa'|, \quad (\text{A.8})$$

$$\|\phi_\kappa(x) - \phi_{\kappa'}(x')\|_t \leq 2\|x - x'\|_t + |\kappa - \kappa'|, \quad (\text{A.9})$$

is used in our proofs. To be able to use the continuous mapping theorem, we need the continuity to be extended to cover the order topology for extended reals.

Lemma A.2 (Eschenfeldt and Gamarnik 2018, Lemma 2). *The reflection mapping $(\phi, \psi) : \bar{\mathbb{R}}_+ \times \mathbb{D}_{\mathbb{R}}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{R}^2}[0, \infty)$ is continuous with respect to the product topology, where $\bar{\mathbb{R}}_+$ and $\mathbb{D}_{\mathbb{R}}[0, \infty)$ are equipped with the order topology and topology of uniform convergence on compact sets, respectively.*

We also need the following Gronwall-type inequality.

Lemma A.3 (Gronwall Inequality; Greene 1977, Das 1979). Let $K_1, K_2 \in \mathbb{R}_+$ and $h_i \in \mathbb{R}$ for $i = 1, \dots, 4$ and let f and g to be continuous nonnegative functions such that

$$f(t) \leq K_1 + h_1 \int_0^t f(s)ds + h_2 \int_0^t g(s)ds,$$

$$g(t) \leq K_2 + h_3 \int_0^t f(s)ds + h_4 \int_0^t g(s)ds$$

for all $t \geq 0$. Then,

$$f(t) \leq (K_1 + K_2)e^{ht} \text{ and } g(t) \leq (K_1 + K_2)e^{ht},$$

for all $t \geq 0$, where $h = \max\{h_1 + h_3, h_2 + h_4\}$.

The following lemma is a version of Lemma 3 that addresses (A.4) and (A.5) internally using the reflection mapping.

Lemma A.4. Consider the following equations

$$w_1(t) = b_1 + z_1(t) - \int_0^t (\mu_M \phi_0(w_1)(s) - \mu_1 \phi_B(w_2)(s))ds, \quad (\text{A.10})$$

$$w_2(t) = b_2 + z_2(t) + \psi_0(w_1)(t) - \int_0^t \mu_1 \phi_B(w_2)(s)ds. \quad (\text{A.11})$$

For given $B \in \bar{\mathbb{R}}_+$, $\mathbf{b} \in \mathbb{R}^2$ and $\mathbf{z} \in \mathbb{D}_{\mathbb{R}^2}[0, \infty)$, (A.10) and (A.11) have a unique solution w and define a function $\xi = (\xi_1, \xi_2) : \bar{\mathbb{R}}_+ \times \mathbb{D}_{\mathbb{R}}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{R}^2}[0, \infty)$ such that $\mathbf{w} = \xi(B, \mathbf{b}, \mathbf{z})$. The function ξ is continuous and the continuity of $\mathbf{z}$ implies the continuity of w .

Proof. In the first part of the proof, we show the existence of a solution using a fixed point argument. We consider the version of (A.10) and (A.11) by replacing $(\tilde{w}_1, \tilde{w}_2) = (w_1, w_2 - \psi_0(w_1))$ as

$$\tilde{w}_1(t) = b_1 + z_1(t) - \int_0^t (\mu_M \phi_0(\tilde{w}_1)(s) - \mu_1 \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s))ds, \quad (\text{A.12})$$

$$\tilde{w}_2(t) = b_2 + z_2(t) - \int_0^t \mu_1 \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))ds. \quad (\text{A.13})$$

For fixed $(B, \mathbf{b}, \mathbf{y})$, a solution to (A.12) and (A.13) is the fixed point of the operator $\mathcal{T} : \mathbb{D}_{\mathbb{R}^2}[0, \infty) \rightarrow \mathbb{D}_{\mathbb{R}^2}[0, \infty)$ defined by

$$\mathcal{T}(\tilde{\mathbf{w}})_1(t) = b_1 + z_1(t) - \int_0^t (\mu_M \phi_0(\tilde{w}_1)(s) - \mu_1 \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s))ds,$$

$$\mathcal{T}(\tilde{\mathbf{w}})_2(t) = b_2 + z_2(t) - \int_0^t \mu_1 \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))ds.$$

For any $\tilde{\mathbf{w}}, \tilde{\mathbf{v}} \in \mathbb{D}_{\mathbb{R}^2}[0, \infty)$,

$$\begin{aligned} \|\mathcal{T}(\tilde{\mathbf{w}})_1 - \mathcal{T}(\tilde{\mathbf{v}})_1\|_t &\leq \mu_M t (\|\phi_0(\tilde{w}_1) - \phi_0(\tilde{v}_1)\|_t + \|\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) - \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))\|_t) \\ &\leq \mu_1 t (2\|\tilde{w}_1 - \tilde{v}_1\|_t + 2\|\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)\|_t) \\ &\leq \mu_1 t (4\|\tilde{w}_1 - \tilde{v}_1\|_t + 2\|\tilde{w}_2 - \tilde{v}_2\|_t) \leq 6\mu_1 t \|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_t, \end{aligned}$$

and

$$\begin{aligned} \|\mathcal{T}(\tilde{\mathbf{w}})_2 - \mathcal{T}(\tilde{\mathbf{v}})_2\|_t &\leq \mu_M t (\|\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) - \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))\|_t) \\ &\leq \mu_1 t (2\|\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)\|_t) \\ &\leq \mu_1 t (2\|\tilde{w}_1 - \tilde{v}_1\|_t + 2\|\tilde{w}_2 - \tilde{v}_2\|_t) \leq 4\mu_1 t \|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_t. \end{aligned}$$

Therefore, we have

$$\|\mathcal{T}(\tilde{\mathbf{w}}) - \mathcal{T}(\tilde{\mathbf{v}})\|_t \leq 6\mu_1 t \|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_t.$$

This implies that the mapping $\mathcal{T}$ is a contraction on $\mathbb{D}_{\mathbb{R}^2}[0, t_0]$ for any $t_0 \leq (6\mu_1)^{-1}$. Now, repeating the same argument iteratively for intervals $[t_0, 2t_0], [2t_0, 3t_0], \dots$, and using the contraction mapping theorem (Rudin 1987, theorem 9.23), we conclude that a fixed point to the operator $\mathcal{T}$. Hence, a solution to (A.12) and (A.13) exists, and $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \psi_0(\tilde{w}_1))$

solves (A.10) and (A.11). Now, we prove uniqueness. If $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{w}}'$ both solve (A.10) and (A.11), then

$$\begin{aligned}\|w_1 - w'_1\|_t &\leq \mu_M \int_0^t \|\phi_0(w_1) - \phi_0(w'_1)\|_s ds + \mu_1 \int_0^t \|\phi_B(w_2) - \phi_B(w'_2)\|_s ds \\ &\leq 2\mu_M \int_0^t \|w_1 - w'_1\|_s ds + 2\mu_1 \int_0^t \|w_2 - w'_2\|_s ds \\ &\leq 4\mu_1 \int_0^t \|w_1 - w'_1\|_s ds + 2\mu_1 \int_0^t (\|w_2 - w'_2\|_s - \|w_1 - w'_1\|_s)^+ ds\end{aligned}$$

and

$$\begin{aligned}\|w_2 - w'_2\|_t &\leq \|\psi_0(w_1) - \psi_0(w'_1)\|_t + \mu_1 \int_0^t \|\phi_B(w_2) - \phi_B(w'_2)\|_s ds \\ \|w_2 - w'_2\|_t - \|w_1 - w'_1\|_t &\leq 2\mu_1 \int_0^t \|w_1 - w'_1\|_s ds + 2\mu_1 \int_0^t (\|w_2 - w'_2\|_s - \|w_1 - w'_1\|_s)^+ ds \\ (\|w_2 - w'_2\|_t - \|w_1 - w'_1\|_t)^+ &\leq 2\mu_1 \int_0^t \|w_1 - w'_1\|_s ds + 2\mu_1 \int_0^t (\|w_2 - w'_2\|_s - \|w_1 - w'_1\|_s)^+ ds,\end{aligned}$$

where the last inequality follows due to the nonnegativity of the right-hand side. Now, applying Lemma A.3, we have

$$\|w_1 - w'_1\|_t = \|w_2 - w'_2\|_t = 0,$$

which implies uniqueness of the solution.

To prove continuity, consider a sequence $(B^n, \mathbf{b}^n, \mathbf{z}^n) \rightarrow (B, \mathbf{b}, \mathbf{z})$. There exists solutions $\mathbf{w}^n$ and $\mathbf{w}$ corresponding to the solution of (A.10) and (A.11) for $(B^n, \mathbf{b}^n, \mathbf{z}^n)$ and $(B, \mathbf{b}, \mathbf{z})$, respectively. Lemma A.2 implies that for any $\delta > 0$, there exists an N_δ such that for $n > N_\delta$ implies

$$|\mathbf{b}^n - \mathbf{b}| + \|\mathbf{z}^n - \mathbf{z}\|_t + \|\phi_{B^n}(w_2) - \phi_B(w_2)\| < \delta.$$

For $n > N_\delta$,

$$\begin{aligned}\|w_1^n - w_1\|_t &\leq \delta + \mu_M \int_0^t \|\phi_0(w_1^n) - \phi_0(w_1)\|_s ds + \mu_1 \int_0^t \|\phi_{B^n}(w_2^n) - \phi_B(w_2)\|_s ds \\ &\leq \delta + \mu_M \int_0^t \|\phi_0(w_1^n) - \phi_0(w_1)\|_s ds + \mu_1 \int_0^t \|\phi_{B^n}(w_2^n) - \phi_{B^n}(w_2)\|_s ds + \mu_1 \int_0^t \|\phi_{B^n}(w_2) - \phi_B(w_2)\|_s ds \\ &\leq \delta(1 + \mu_1 t) + 2\mu_M \int_0^t \|w_1^n - w_1\|_s ds + 2\mu_1 \int_0^t \|w_2^n - w_2\|_s ds \\ &\leq \delta(1 + \mu_1 t) + 4\mu_M \int_0^t \|w_1^n - w_1\|_s ds + 2\mu_1 \int_0^t (\|w_2^n - w_2\|_s - \|w_1^n - w_1\|_s)^+ ds\end{aligned}$$

and similarly,

$$(\|w_2^n - w_2\|_t - \|w_1^n - w_1\|_t)^+ \leq \delta(1 + \mu_1 t) + 2\mu_1 \int_0^t \|w_1^n - w_1\|_s ds + 2\mu_1 \int_0^t (\|w_2^n - w_2\|_s - \|w_1^n - w_1\|_s)^+ ds.$$

Choosing $\delta = (4 + 4\mu_1 t)^{-1} e^{-6\mu_1 t} \epsilon$ and applying Lemma A.3,

$$\|w_1^n - w_1\|_t < \epsilon \text{ and } \|w_2^n - w_2\|_t < 2\epsilon,$$

which proves the continuity of the mapping ξ . Finally, to prove that continuity of $\mathbf{z}$ implies the continuity of $\mathbf{w}$, we first show that the function w is bounded using a similar approach as above. Choosing $x=w$ and x' as the zero function, for any $\kappa \geq 0$, (A.8) and (A.9) imply

$$\|\psi_\kappa(w)\|_t \leq \|w\|_t, \|\phi_\kappa(w)\|_t \leq 2\|w\|_t.$$

Now, reorganizing (A.10) and (A.11), we have

$$\begin{aligned}\|w_1\|_t &\leq |b_1| + \|z_1\|_t + (\mu_1 + \mu_M) \int_0^t \|w_1\|_s ds + \mu_1 \int_0^t (\|w_2\|_s - \|w_1\|_s) ds, \\ \|w_2\|_t - \|w_1\|_t &\leq |b_2| + \|z_2\|_t + \mu_1 \int_0^t \|w_1\|_s ds + \mu_1 \int_0^t (\|w_2\|_s - \|w_1\|_s) ds.\end{aligned}$$

Now, applying Lemma A.3 again, we get

$$\begin{aligned}\|w_1\|_t &\leq (|b_1| + \|z_1\|_t + |b_2| + \|z_2\|_t) e^{(2\mu_1 + \mu_M)t} \text{ and} \\ \|w_2\|_t - \|w_1\|_t &\leq (|b_1| + \|z_1\|_t + |b_2| + \|z_2\|_t) e^{(2\mu_1 + \mu_M)t},\end{aligned}$$

which implies that $\mathbf{w}$ is bounded. Then, again using (A.10) and (A.11), we get

$$\begin{aligned} |w_1(t+s) - w_1(t)| &\leq |z_1(t+s) - z_1(t)| + \mu_1 \int_t^{t+s} |\phi_0(w_1)(z)| + |\phi_B(w_2)(z)| dz, \\ |w_2(t+s) - w_2(t)| &\leq |z_1(t+s) - z_1(t)| + |w_1(t+s) - w_1(t)| + \mu_1 \int_t^{t+s} |\phi_B(w_2)(z)| dz, \end{aligned}$$

and along with the boundedness of $\mathbf{w}$, this implies the desired continuity. $\square$

Proof of Lemma A.1. The lemma follows using the same arguments in the proof of lemma 1 in Eschenfeldt and Gamarnik (2018) by taking

$$\begin{aligned} y_1 &= (\phi_0 \circ \xi_1)(\mathbf{b}, \mathbf{z}), & u_1 &= (\psi_0 \circ \xi_1)(\mathbf{b}, \mathbf{z}) \\ y_2 &= (\phi_B \circ \xi_2)(\mathbf{b}, \mathbf{z}), & u_2 &= (\psi_B \circ \xi_2)(\mathbf{b}, \mathbf{z}). \end{aligned} \quad \square$$

Proof of Proposition 2. The existence, uniqueness and continuity of $y_{j,i}$ solving (5.3) and (5.5) can be shown by fixing j and applying lemma 5 in Eschenfeldt and Gamarnik (2018). Defining

$$\hat{z}_{1,2}(t) = z_{1,2}(t) + \mu_1 \int_0^t y_{1,3}(s) ds,$$

and applying Lemma A.1 with $\hat{z}_{1,2}$, we can prove existence, uniqueness and continuity of the solution. To show the continuity of the mapping (f, g) , suppose that $(B^n, \mathbf{b}^n, \mathbf{z}^n) \rightarrow (B, \mathbf{b}, \mathbf{z})$ as $n \rightarrow \infty$ and $\mathbf{y}^n$ and $\mathbf{y}$ solve the respective (5.1)–(5.3). Lemma 5 in Eschenfeldt and Gamarnik (2018) implies that $y_{j,i}^n \rightarrow y_{j,i}$ for all $i \geq 3$ and $(j, 2)$ for $j \geq 2$, which implies

$$z_{1,2}^n(t) + \mu_1 \int_{0,t} y_{1,3}^n(s) ds \rightarrow z_{1,2}(t) + \mu_1 \int_{0,t} y_{1,3}(s) ds.$$

Combining this with the continuity in Lemma A.1, we obtain the desired continuity. $\square$

Appendix B. Proof of Lemma 3

Proof of Lemma 3. Let $\hat{\mathbf{Y}}^{(n)} = (\hat{Y}_{1,2}^{(n)}, \hat{Y}_{[1,M],1}^{(n)})$, $\hat{\mathbf{Z}}^{(n)} = (\hat{Z}_1^{(n)}, \hat{Z}_{1,2}^{(n)})$ and fix $t \geq 0$. To prove the result of proposition we establish the following bound

$$\|\hat{\mathbf{Y}}^{(n)}\|_t \leq 8e^{6\mu_1 t} (|\hat{\mathbf{Y}}^{(n)}(0)| + \|\hat{\mathbf{Z}}^{(n)}\|_t). \quad (\text{B.1})$$

To prove the above bound, we show that the similar bound holds for the unreflected process $\mathbf{w}^{(n)}$ defined in Lemma A.4. Then (B.1) follows from the Lipschitz continuity of the reflected maps ϕ_0 and ϕ_{B_n} . We write $\hat{Y}_{[1,M],1}^{(n)}(t) = \phi_0(w_1^{(n)}(t))$ and $\hat{Y}_{1,2}^{(n)}(t) = \phi_{B_n}(w_2^{(n)}(t))$, where $w_1^{(n)}(t)$ and $w_2^{(n)}(t)$ satisfies

$$\begin{aligned} w_1^{(n)}(t) &= \hat{Y}_{[1,M],1}^{(n)}(0) + \hat{Z}_1^{(n)}(t) - \int_0^t (\mu_M \phi_0(w_1^{(n)})(s) - \mu_1 \phi_B(w_2^{(n)})(s)) ds, \\ w_2^{(n)}(t) &= \hat{Y}_{1,2}^{(n)}(0) + \hat{Z}_{1,2}^{(n)}(t) + \psi_0(w_1^{(n)})(t) - \int_0^t \mu_1 \phi_B(w_2^{(n)})(s) ds. \end{aligned}$$

Now using Grownwall's inequality from Lemma A.3 and using Lipschitz property of ϕ_0 , ϕ_{B_n} , and ψ_0 , we have for any $t \geq 0$

$$\begin{aligned} \|w_1^{(n)}\|_t &\leq |\hat{Y}_{[1,M],1}^{(n)}(0)| + \|\hat{Z}_1^{(n)}\|_t + 2\mu_1 \int_0^t (\|w_2^{(n)}\|_s + \|w_1^{(n)}\|_s) ds, \\ \|w_2^{(n)}\|_t &\leq |\hat{Y}_{1,2}^{(n)}(0)| + \|\hat{Z}_{1,2}^{(n)}\|_t + \|\psi_0(w_1^{(n)})\|_t + \int_0^t \mu_1 \|w_2^{(n)}\|_s ds. \end{aligned}$$

Note that we have $\|\psi_0(w_1^{(n)})\|_t \leq \|w_1^{(n)}\|_t$. Moreover, we define $r_1(t) = \|w_1^{(n)}\|_t$ and $r_2(t) = (\|w_2^{(n)}\|_t - \|w_1^{(n)}\|_t)^+$ and note that $\|w_2^{(n)}\|_t \leq r_1(t) + r_2(t)$. Therefore, we can write above inequalities as

$$\begin{aligned} r_1(t) &\leq |\hat{Y}_{[1,M],1}^{(n)}(0)| + \|\hat{Z}_1^{(n)}\|_t + 4\mu_1 \int_0^t r_1(s) ds + 2\mu_1 \int_0^t r_2(s) ds, \\ r_2(t) &\leq |\hat{Y}_{1,2}^{(n)}(0)| + \|\hat{Z}_{1,2}^{(n)}\|_t + \int_0^t \mu_1 r_1(s) ds + \int_0^t \mu_1 r_2(s) ds. \end{aligned}$$

Let $|\hat{\mathbf{Y}}^{(n)}(0)| + \|\hat{\mathbf{Z}}^{(n)}\|_t = m$. Now from Lemma A.3, we have $r_1(t) \leq 2me^{6\mu_1 t}$ and $r_2(t) \leq 2me^{6\mu_1 t}$. Finally, using the definition of r_1 , r_2 and from the fact that ϕ_0 , ϕ_{B_n} are Lipschitz with a constant of two, we get

$$\|\hat{Y}_{[1,M],1}^{(n)}\|_t \leq 4me^{6\mu_1 t}, \quad \|\hat{Y}_{1,2}^{(n)}\|_t \leq 8me^{6\mu_1 t},$$

which proves (B.1). Hence, the proof is complete. $\square$

Appendix C. Proof of Lemma 7

To prove part 1 of the lemma, we define a Lyapunov function $g(\mathbf{q}) = \min\{T, \sum_{k=1}^M \sum_{i=1}^{\infty} q_{k,i}\}$. Then,

$$\mathcal{G}g(\mathbf{q}) = n\lambda^{(n)} \mathbb{1}\left(\sum_{k=1}^M \sum_{i=1}^{\infty} q_{k,i} < T\right) - \sum_{k=1}^M \mu_k q_{k,1} \mathbb{1}\left(\sum_{k=1}^M \sum_{i=1}^{\infty} q_{k,i} \leq T\right).$$

Now taking the expectation of both sides with respect to the appropriate stationary measure, applying (6.2) and the monotone convergence theorem, we have $\mathbb{E}[\sum_{k=1}^M \mu_k Q_{k,1}^{(n)}(\infty)] = n\lambda^{(n)}$. Hence,

$$0 \leq \mu_M \mathbb{E}\left[n^{-1/2} \left(\sum_{k=1}^M (N_k^n - Q_{k,1}^{(n)}(\infty))\right)\right] \leq \mathbb{E}\left[n^{-1/2} \left(\sum_{k=1}^M \mu_k (N_k^n - Q_{k,1}^{(n)}(\infty))\right)\right] = n^{1/2} \left((1 - \lambda^{(n)}) + \sum_{k=1}^M \mu_k \left(\frac{N_k^n}{n} - \gamma_k\right)\right),$$

which implies tightness using (A1) and (A2). Part 2 of the lemma can be proved in a similar fashion by using the Lyapunov function

$$g(\mathbf{q}) = \min\left\{T, \sum_{k=2}^M \sum_{i=2}^{\infty} q_{k,i} + \sum_{i=3}^{\infty} q_{1,i}\right\}. \quad \square$$

Appendix D. Proof of Lemma 9

For fixed $x_1 \leq 0$, plugging in $x_2 e^{-\mu_1 \tau} = \kappa/\sqrt{n}$, (6.54) becomes

$$-\frac{\beta}{\mu_M \sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}}\right) e^{-\mu_M \tau} - \frac{\mu_1 \kappa}{(\mu_1 - \mu_M) \sqrt{n}} (1 - e^{-(\mu_M - \mu_1) \tau}) = 0. \quad (\text{D.1})$$

When $\tau = 0$, the left-hand side is equal to x_1 , which is nonpositive, and as $\tau \rightarrow \infty$, the left-hand side converges to ∞ . Furthermore, the derivative of the left-hand side of (D.1) with respect to τ is given by

$$\left(-\mu_M x_1 - \frac{\beta}{\sqrt{n}} + \frac{\mu_1 \kappa e^{\mu_1 \tau}}{\sqrt{n}}\right) e^{-\mu_M \tau} = \left(-\mu_M x_1 + \frac{\mu_1}{\sqrt{n}} \left(\kappa e^{\mu_1 \tau} - \frac{\beta}{\mu_1}\right)\right) e^{-\mu_M \tau},$$

which is clearly nonnegative for any $\kappa \geq \beta/\mu_1$. Hence, we can conclude that there exists a unique $\tau^* \in [0, \infty)$ such that Equation (D.1) holds, which in turn implies the uniqueness of x_2^* . The second statement of the lemma follows directly by solving for (x_1, τ) as a function of x_2 . $\square$

Appendix E. Proof of Lemma 10

If $\mathbf{x} \geq \Gamma^\kappa$ for some $\kappa \geq \beta/\mu_1$, then by definition of Γ^κ and Lemma 9 there exist $\delta \geq 0$ and $\tau' = \tau^*(x_1, \kappa) \in [0, \infty)$ such that the following equations hold:

$$-\frac{\beta}{\mu_M \sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}}\right) e^{-\mu_M \tau'} - \frac{\mu_1 (x_2 - \delta)}{\mu_1 - \mu_M} (e^{-\mu_1 \tau'} - e^{-\mu_M \tau'}) = 0, \quad (\text{E.1})$$

$$(x_2 - \delta) e^{-\mu_1 \tau'} = \kappa/\sqrt{n}. \quad (\text{E.2})$$

To simplify the notation, we define

$$g_x(\tau) = -\frac{\beta}{\mu_M \sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M \sqrt{n}}\right) e^{-\mu_M \tau} - \frac{\mu_1 x_2}{\mu_1 - \mu_M} (e^{-\mu_1 \tau} - e^{-\mu_M \tau}). \quad (\text{E.3})$$

Clearly, $g_x(0) = x_1 \leq 0$, and using (E.1), we also have $g_x(\tau') = -\frac{\mu_1 \delta}{\mu_1 - \mu_M} (e^{-\mu_1 \tau'} - e^{-\mu_M \tau'}) \geq 0$. Hence, by the continuity of $g_x(\cdot)$, there exists $\tau \in [0, \tau']$, where $g_x(\tau) = 0$. Furthermore, for such a $\tau \leq \tau'$, we also have $x_2 e^{-\mu_1 \tau} \geq x_2 e^{-\mu_1 \tau'} \geq \kappa/\sqrt{n}$. This proves the first part of the lemma because $\tau(\mathbf{x})$ is defined as the smallest nonnegative root of $g_x(\cdot)$.

To prove the second part, observe that $\mathbf{x} \in \Gamma^\kappa$ implies $\delta = 0$ in Equations (E.1) and (E.2). Hence, we have $g_x(\tau') = 0$, where $\tau' = \tau^*(x_1, \kappa)$. Now, we shall prove that there exists no root of $g_x(\cdot)$ in $[0, \tau')$, which will establish that $\tau(\mathbf{x}) = \tau'$, thereby completing the proof of this part. Let us suppose that it is not true. Hence, there exists at least one point $0 \leq \tau_1 < \tau'$ such that $g_x(\tau_1) = 0$ and $g'_x(\tau_1) g'_x(\tau') \leq 0$, where $g'_x(\cdot)$ denotes the derivative of $g_x(\cdot)$ w.r.t τ . But from (E.3), it is easy to see that $g'_x(\tau) = -\mu_M g(\tau) + \mu_1 x_2 e^{-\mu_1 \tau} - \beta/\sqrt{n}$. Hence, $g'_x(\tau_1) = \mu_1 x_2 e^{-\mu_1 \tau_1} - \beta/\sqrt{n}$ and $g'_x(\tau') = \mu_1 x_2 e^{-\mu_1 \tau'} - \beta/\sqrt{n}$. Furthermore, $0 \leq \tau_1 < \tau'$ implies that $x_2 e^{-\mu_1 \tau_1} > x_2 e^{-\mu_1 \tau'} = \kappa/\sqrt{n} > \beta/\mu_1 \sqrt{n}$. Hence, $g'_x(\tau_1) g'_x(\tau') > 0$, which contradicts $g'_x(\tau_1) g'_x(\tau') \leq 0$.

To prove the third part of the lemma, we take the partial derivatives of the equation $g_x(\tau(\mathbf{x})) = 0$ with respect to x_1 and x_2 . This yields the expressions given by (6.57) and (6.58), respectively, after manipulation using $g_x(\tau(\mathbf{x})) = 0$. The inequalities $\tau_1(\mathbf{x}) \leq 0$ and $\tau_2(\mathbf{x}) \leq 0$ follow by observing that $x_2 e^{-\mu_1 \tau(\mathbf{x})} \geq \kappa/\sqrt{n} \geq \beta/\mu_1 \sqrt{n}$.

To prove the fourth part of the lemma, we fix $\mathbf{x}$ and assume that $\mathbf{x} \geq \Gamma^{k_1}$. From the first part, we know that $x_2 e^{-\mu_1 \tau(\mathbf{x})} \geq \kappa_1 / \sqrt{n} > \kappa_2 / \sqrt{n}$. Taking derivative of $x_2 e^{-\mu_1 \tau(\mathbf{x})}$ with respect to x_2 we get

$$\frac{d}{dx_2} x_2 e^{-\mu_1 \tau(\mathbf{x})} = e^{-\mu_1 \tau(\mathbf{x})} (1 - \mu_1 x_2 \tau_2(\mathbf{x})) > 0,$$

where the inequality follows from (6.58). Therefore, $x_2 e^{-\mu_1 \tau(\mathbf{x})}$ is nondecreasing in x_2 , and we can write for any $\delta \geq 0$

$$(x_2 + \delta) e^{-\mu_1 \tau(x_1, x_2 + \delta)} \geq \kappa_1 / \sqrt{n} > \kappa_2 / \sqrt{n}. \quad (\text{E.4})$$

The above inequality implies that $(x_1, x_2 + \delta) \notin \Gamma^{k_2}$ for all $\delta \geq 0$ because by part 2 of this lemma, $(x_1, x_2 + \delta) \in \Gamma^{k_2}$ will imply $\tau(x_1, x_2 + \delta) = \tau^*(x_1, \kappa_2)$, which violates (E.4). On the other hand, from Lemma 9, there must exist $x'_2 \geq 0$ such that $(x_1, x'_2) \in \Gamma^{k_2}$, implies that $x'_2 = x_2 - \delta'$ for some $\delta' > 0$. Hence, we have $\mathbf{x} > \Gamma^{k_2}$.

To prove the last part of the lemma, let us assume $\tau(\mathbf{x}) < \infty$ for some $\mathbf{x} < \Gamma^{\beta/\mu_1}$. Then we must have $\mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} < \beta / \sqrt{n}$ because otherwise we would have $\mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} = \kappa / \sqrt{n}$ for some $\kappa \geq \beta / \mu_1$, and combined with $g_{\mathbf{x}}(\tau(\mathbf{x})) = 0$, it would imply that $\mathbf{x} \in \Gamma^{\kappa}$ for some $\kappa \geq \beta / \mu_1$ that contradicts with $\mathbf{x} < \Gamma^{\beta/\mu_1}$ due to the fourth part of the lemma. Furthermore, because $\tau(\mathbf{x})$ is the smallest non-negative root of $g_{\mathbf{x}}(\cdot)$ and $g_{\mathbf{x}}(0) = x_1 \leq 0$, we must have $g'_{\mathbf{x}}(\tau(\mathbf{x})) \geq 0$. But from the proof of the previous part, we know that $g'_{\mathbf{x}}(\tau(\mathbf{x})) = -\mu_M g(\tau(\mathbf{x})) + \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} - \beta / \sqrt{n} = \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})} - \beta / \sqrt{n} < 0$, which contradicts with $g'_{\mathbf{x}}(\tau(\mathbf{x})) \geq 0$. $\square$

Appendix F. Proof of Lemma 8

In this section, we show that the Lyapunov function given by (6.62) solves the PDE (6.22)–(6.23) and satisfies the properties stated in Lemma 8.

For convenience, we repeat the expression of $f^*(\mathbf{x})$ below:

$$f^*(\mathbf{x}) = \begin{cases} 0, & \text{for } \mathbf{x} \in \Omega_1, \\ \frac{x_2}{\mu_1} - \frac{\kappa}{\mu_1 \sqrt{n}} - \frac{\kappa}{\mu_1 \sqrt{n}} \log\left(\frac{\sqrt{n} x_2}{\kappa}\right), & \text{for } \mathbf{x} \in \Omega_2, \\ \frac{x_2}{\mu_1} - \frac{x_2}{\mu_1} e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \tau(\mathbf{x}) + \frac{\sqrt{n}}{2\beta} \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right)^2, & \text{for } \mathbf{x} \in \Omega_3. \end{cases} \quad (\text{F.1})$$

Here, the sets Ω_i , $i \in [3]$ are defined as

$$\Omega_1 = \{x_2 \leq \kappa / \sqrt{n}\}, \Omega_2 = \{x_2 > \kappa / \sqrt{n} \text{ and } \mathbf{x} \leq \Gamma^{\kappa}\} \text{ and } \Omega_3 = \{\mathbf{x} > \Gamma^{\kappa}\},$$

and by Lemma 10, we have $(-\infty, 0] \times [0, \infty) = \Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$. Also recall that the operator $\mathcal{L}$ is defined as

$$\mathcal{L}f(\mathbf{x}) = \left(-\frac{\beta}{\sqrt{n}} - \mu_M x_1 + \mu_1 x_2 \right) f_1(\mathbf{x}) - \mu_1 x_2 f_2(\mathbf{x}), \quad (\text{F.2})$$

for any differentiable function f .

We first calculate the derivatives of f^* on the sets Ω_1 , Ω_2 , and Ω_3 . When $\mathbf{x} \in \Omega_1$, we clearly have $f_1^*(\mathbf{x}) = f_2^*(\mathbf{x}) = 0$. By (6.6), this implies that $\mathcal{L}f^*(\mathbf{x}) = 0 = -(x_2 - \kappa / \sqrt{n})_+$ when $\mathbf{x} \in \Omega_1$.

When $\mathbf{x} \in \Omega_2$, we have $f_1^*(\mathbf{x}) = 0$ and

$$f_2^*(\mathbf{x}) = \frac{1}{\mu_1} - \frac{\kappa}{\mu_1 \sqrt{n}} \frac{\kappa}{\sqrt{n} x_2} = \frac{1}{\mu_1} - \frac{\kappa}{\mu_1 \sqrt{n} x_2},$$

and therefore, we have

$$\mathcal{L}f^*(\mathbf{x}) = -\mu_1 x_2 f_2^*(\mathbf{x}) = -x_2 + \frac{\kappa}{\sqrt{n}} = -(x_2 - \kappa / \sqrt{n})_+.$$

When $\mathbf{x} \in \Omega_3$, we have

$$\begin{aligned} f_1^*(\mathbf{x}) &= \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} - \frac{\sqrt{n}}{\beta} \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right) x_2 \mu_1 e^{-\mu_1 \tau(\mathbf{x})} \right) \tau_1(\mathbf{x}), \\ &= -\frac{\sqrt{n}}{\beta} \tau_1(\mathbf{x}) \left(x_2 \mu_1 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}} \right) \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right), \\ &= \frac{\sqrt{n}}{\beta} e^{-\mu_M \tau(\mathbf{x})} \left(x_2 e^{-\mu_1 \tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}} \right), \end{aligned} \quad (\text{F.3})$$

where the last line follows by using the expression of $\tau_1(\mathbf{x})$ given in Lemma 10. Furthermore, in the same region, we have

$$\begin{aligned} f_2^*(\mathbf{x}) &= \frac{1}{\mu_1}(1 - e^{-\mu_1\tau(\mathbf{x})}) + \left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right)\tau_2(\mathbf{x}) + \frac{\sqrt{n}}{\beta}\left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right)(e^{-\mu_1\tau(\mathbf{x})} - \mu_1 x_2 e^{-\mu_1\tau(\mathbf{x})}\tau_2(\mathbf{x})) \\ &= \frac{1}{\mu_1}(1 - e^{-\mu_1\tau(\mathbf{x})}) + \left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right)\left(\left(1 - \frac{\mu_1\sqrt{n}}{\beta}x_2 e^{-\mu_1\tau(\mathbf{x})}\right)\tau_2(\mathbf{x}) + \frac{\sqrt{n}}{\beta}e^{-\mu_1\tau(\mathbf{x})}\right) \\ &= \frac{1}{\mu_1}(1 - e^{-\mu_1\tau(\mathbf{x})}) + \left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right)\frac{\sqrt{n}}{\beta}\left(\frac{\mu_1}{\mu_1 - \mu_M}(e^{-\mu_M\tau(\mathbf{x})} - e^{-\mu_1\tau(\mathbf{x})}) + e^{-\mu_1\tau(\mathbf{x})}\right). \end{aligned} \quad (\text{F.4})$$

Also, by definition of $\tau(\mathbf{x})$, we have

$$-\frac{\beta}{\mu_M\sqrt{n}} + \left(x_1 + \frac{\beta}{\mu_M\sqrt{n}}\right)e^{-\mu_M\tau(\mathbf{x})} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(e^{-\mu_1\tau(\mathbf{x})} - e^{-\mu_M\tau(\mathbf{x})}) = 0. \quad (\text{F.5})$$

After some straightforward algebraic manipulations using (F.3), (F.4), and (F.5), it follows that $\mathcal{L}f^*(\mathbf{x}) = -x_2 + \kappa/\sqrt{n}$ when $\mathbf{x} \in \Omega_3$. Hence, we have verified that (6.22) is satisfied for all $\mathbf{x} \in \Omega$.

Now, we verify the boundary conditions for the derivatives. It is easy to verify from (F.5) that when $x_1 = 0$ and $\mathbf{x} \in \Omega_3$ we have $\tau(\mathbf{x}) = 0$. Hence, from (F.3) and (F.4), we have

$$f_1^*(0, x_2) = f_2^*(0, x_2) = \frac{\sqrt{n}}{\beta}\left(x_2 - \frac{\kappa}{\sqrt{n}}\right), \quad (\text{F.6})$$

which shows that (6.23) holds for all $\mathbf{x} \in \Omega$.

Now, we show that the second-order derivatives of $f^*(\mathbf{x})$ satisfy the bounds given in Lemma 8. To summarize our development above, we have derived the following expressions for the first derivatives of f^* :

$$f_1^*(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in \Omega_1 \cup \Omega_2 \\ \frac{\sqrt{n}}{\beta}e^{-\mu_M\tau(\mathbf{x})}\left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right) & \text{if } \mathbf{x} \in \Omega_3, \end{cases} \quad \text{and} \quad (\text{F.7})$$

$$f_2^*(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in \Omega_1 \\ \frac{1}{\mu_1} - \frac{\kappa}{\mu_1\sqrt{n}x_2} & \text{if } \mathbf{x} \in \Omega_2 \\ \frac{1}{\mu_1}(1 - e^{-\mu_1\tau(\mathbf{x})}) + \left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\right)\frac{\sqrt{n}}{\beta}\theta(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_3, \end{cases} \quad (\text{F.8})$$

where

$$\theta(\mathbf{x}) = \frac{\mu_1}{\mu_1 - \mu_M}(e^{-\mu_M\tau(\mathbf{x})} - e^{-\mu_1\tau(\mathbf{x})}) + e^{-\mu_1\tau(\mathbf{x})} \geq 0. \quad (\text{F.9})$$

Taking the derivative of (F.7) with respect to x_1 for $\mathbf{x} \in \Omega_3$:

$$f_{11}^*(\mathbf{x}) = \frac{\sqrt{n}(\mu_1 + \mu_M)}{\beta}\tau_1(\mathbf{x})e^{-\mu_M\tau(\mathbf{x})}\left(-x_2 e^{-\mu_1\tau(\mathbf{x})} + \frac{\kappa}{\sqrt{n}}\frac{\mu_M}{\mu_1 + \mu_M}\right).$$

Now, $\mathbf{x} > \Gamma^\kappa$ implies that $x_2 e^{-\mu_1\tau(\mathbf{x})} \geq \kappa/\sqrt{n}$, which in turn implies that the term inside the parenthesis in the above expression is negative. Furthermore, because by Lemma 10, we have $\tau_1(\mathbf{x}) \leq 0$, we conclude that $f_{11}^*(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. Plugging in the expression of $\tau_1(\mathbf{x})$ from Lemma 10 in the above equation, we obtain

$$\begin{aligned} f_{11}^*(\mathbf{x}) &= \frac{\sqrt{n}(\mu_1 + \mu_M)}{\beta} \frac{e^{-2\mu_M\tau(\mathbf{x})}}{\mu_1 x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}}} \left(x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\kappa}{\sqrt{n}}\frac{\mu_M}{\mu_1 + \mu_M}\right), \\ &\leq \frac{\sqrt{n}(\mu_1 + \mu_M)}{\beta} \frac{e^{-2\mu_M\tau(\mathbf{x})}x_2 e^{-\mu_1\tau(\mathbf{x})}}{\mu_1 x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\beta}{\sqrt{n}}}, \\ &\leq \frac{\sqrt{n}(\mu_1 + \mu_M)}{\mu_1\beta} \frac{x_2 e^{-\mu_1\tau(\mathbf{x})}}{x_2 e^{-\mu_1\tau(\mathbf{x})} - \frac{\beta}{\mu_1\sqrt{n}}}, \\ &\leq \frac{\sqrt{n}\mu_1 + \mu_M}{\beta} \frac{\kappa}{\mu_1 \kappa - \frac{\beta}{\mu_1}}, \end{aligned}$$

where the last inequality follows because $x_2 e^{-\mu_1 \tau(x)} \geq \kappa/\sqrt{n}$ for $\mathbf{x} > \Gamma^\kappa$ and $z \mapsto \frac{z}{z-c}$ is a nonincreasing function for any constant $c \geq 0$.

Now, taking derivative of $f_1^*(\mathbf{x})$ with respect to x_2 for $\mathbf{x} \in \Omega_3$, we obtain

$$f_{12}^*(\mathbf{x}) = -\frac{\sqrt{n}(\mu_1 + \mu_M)}{\beta} e^{-\mu_M \tau(x)} \tau_2(\mathbf{x}) \left(x_2 e^{-\mu_1 \tau(x)} - \frac{\kappa}{\sqrt{n}} \frac{\mu_M}{\mu_1 + \mu_M} \right).$$

As $\tau_2(\mathbf{x}) \leq 0$ (by Lemma 10) and $\mathbf{x} > \Gamma_\kappa$, it follows that $f_{12}^*(\mathbf{x}) \geq 0$.

Now, we concentrate on $f_{22}^*(\mathbf{x})$, which is trivially zero when $\mathbf{x} \in \Omega_1$. When $\mathbf{x} \in \Omega_2$ from (F.8), we have

$$f_{22}^*(\mathbf{x}) = \frac{\kappa}{\mu_1 \sqrt{n} x_2^2} \leq \frac{\sqrt{n}}{\mu_1 \kappa},$$

and hence $f_{22}^*(\mathbf{x}) \in \left[0, \frac{\sqrt{n}}{\mu_1 \kappa}\right]$. When $\mathbf{x} \in \Omega_3$ again from (F.8), we have

$$f_{22}^*(\mathbf{x}) = e^{-\mu_1 \tau(x)} \left(\tau_2(\mathbf{x}) \left(1 - \frac{\sqrt{n}}{\beta} \mu_1 x_2 \theta(\mathbf{x}) \right) + \frac{\sqrt{n}}{\beta} \theta(\mathbf{x}) \right) + \frac{\sqrt{n}}{\beta} \left(x_2 e^{-\mu_1 \tau(x)} - \frac{\kappa}{\sqrt{n}} \right) \theta_2(\mathbf{x}), \quad (\text{F.10})$$

where $\theta_2(\mathbf{x})$ is the derivative of $\theta(\mathbf{x})$ with respect to x_2 and is given by

$$\theta_2(\mathbf{x}) = -\frac{\mu_1 \mu_M}{\mu_1 - \mu_M} (e^{-\mu_M \tau(x)} - e^{-\mu_1 \tau(x)}) \tau_2(\mathbf{x}), \quad (\text{F.11})$$

for $\mathbf{x} \in \Omega_3$. Because $\tau_2(\mathbf{x}) \leq 0$, we have $\theta_2(\mathbf{x}) \geq 0$. Furthermore, by the definition of $\theta(\mathbf{x})$, we have $\theta(\mathbf{x}) \geq e^{-\mu_1 \tau(x)}$. Now, because $\mathbf{x} \in \Omega_3$, we have $x_2 \theta(\mathbf{x}) \geq x_2 e^{-\mu_1 \tau(x)} \geq \kappa/\sqrt{n} \geq \beta/\mu_1 \sqrt{n}$. In other words, we have $1 - \frac{\sqrt{n}}{\beta} \mu_1 x_2 \theta(\mathbf{x}) \leq 0$, which implies that $f_{22}^*(\mathbf{x}) \geq 0$ for $\mathbf{x} \in \Omega_3$. Hence, we have established that $f_{22}^*(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \Omega$.

Now we proceed to find an upper bound on $f_{22}^*(\mathbf{x})$ on Ω_3 . To do this, we first note that

$$\theta(\mathbf{x}) = \frac{\mu_1}{\mu_1 - \mu_M} e^{-\mu_M \tau(x)} - \frac{\mu_M}{\mu_1 - \mu_M} e^{-\mu_1 \tau(x)} \leq \frac{\mu_1}{\mu_1 - \mu_M}.$$

Now, plugging in the expression of $\tau_2(\mathbf{x})$ from Lemma 10 into (F.11), we obtain

$$\theta_2(\mathbf{x}) = \frac{\mu_1 \mu_M}{\mu_1 - \mu_M} \frac{(e^{-\mu_M \tau(x)} - e^{-\mu_1 \tau(x)})^2}{x_2 e^{-\mu_1 \tau(x)} - \frac{\beta}{\mu_1 \sqrt{n}}} \leq \frac{\mu_1 \mu_M}{\mu_1 - \mu_M} \frac{1}{x_2 e^{-\mu_1 \tau(x)} - \frac{\beta}{\mu_1 \sqrt{n}}}.$$

From the above and the fact that $\kappa > \beta/\mu_1$, the second term in (F.10) can be bounded above by

$$\frac{\sqrt{n}}{\beta} \frac{\mu_1 \mu_M}{\mu_1 - \mu_M}.$$

Using the expression of $\tau_2(\mathbf{x})$ from Lemma 10, again we have

$$\begin{aligned} e^{-\mu_1 \tau(x)} \tau_2(\mathbf{x}) \left(1 - \frac{\sqrt{n}}{\beta} \mu_1 x_2 \theta(\mathbf{x}) \right) &= \frac{e^{-\mu_1 \tau(x)}}{\mu_1 - \mu_M} \frac{(e^{-\mu_M \tau(x)} - e^{-\mu_1 \tau(x)})}{\left(x_2 e^{-\mu_1 \tau(x)} - \frac{\beta}{\mu_1 \sqrt{n}} \right)} \left(\frac{\sqrt{n}}{\beta} \mu_1 x_2 \theta(\mathbf{x}) - 1 \right) \\ &\leq \frac{\sqrt{n}}{\beta} \left(\frac{\mu_1}{\mu_1 - \mu_M} \right)^2 \frac{(e^{-\mu_M \tau(x)} - e^{-\mu_1 \tau(x)})}{\left(x_2 e^{-\mu_1 \tau(x)} - \frac{\beta}{\mu_1 \sqrt{n}} \right)} x_2 e^{-\mu_1 \tau(x)} \\ &\leq \frac{\sqrt{n}}{\beta} \left(\frac{\mu_1}{\mu_1 - \mu_M} \right)^2 \frac{x_2 e^{-\mu_1 \tau(x)}}{\left(x_2 e^{-\mu_1 \tau(x)} - \frac{\beta}{\mu_1 \sqrt{n}} \right)} \\ &\leq \frac{\sqrt{n}}{\beta} \left(\frac{\mu_1}{\mu_1 - \mu_M} \right)^2 \frac{\kappa}{\kappa - \frac{\beta}{\mu_1}}. \end{aligned} \quad (\text{F.12})$$

Hence, combining the bounds, we have

$$f_{22}^*(\mathbf{x}) \leq \frac{\sqrt{n}}{\beta} \left(\left(\frac{\mu_1}{\mu_1 - \mu_M} \right)^2 \frac{\kappa}{\kappa - \frac{\beta}{\mu_1}} + \frac{\mu_1(1 + \mu_M)}{\mu_1 - \mu_M} \right), \quad (\text{F.13})$$

for $\mathbf{x} \in \Omega_3$. The bound on $f_{22}^*(\mathbf{x})$ stated in Lemma 8 now follows by combining the bounds for the sets Ω_2 and Ω_3 .

The absolute continuity of $f_1^*(\cdot, x_2)$ and $f_2^*(x_1, \cdot)$ follows using the same argument as in Braverman (2020).

Appendix G. Positive Recurrence of The Limiting Diffusion: Proof of Theorem 3

The diffusion limit is expressed as

$$Y_{M,1}(t) = Y_{1,M}(0) + \sqrt{2}W(t) - \beta t + \int_0^t (\mu_1 Y_{2,1}(s) - \mu_M Y_{M,1}(s)) ds - V_1(t), \quad (G.1)$$

$$Y_{1,2}(t) = Y_{1,2}(0) - \mu_1 \int_0^t Y_{1,2}(s) ds + V_1(t), \quad (G.2)$$

and using Ito's lemma, for any $f(\mathbf{x}) \in C^2(\Omega)$,

$$\begin{aligned} & \mathbb{E}_x[f(Y_{M,1}(t), Y_{1,2}(t))] - \mathbb{E}_x[f(Y_{M,1}(0), Y_{1,2}(0))] \\ &= \mathbb{E}_x \left[\int_0^t (\mu_1 Y_{1,2}(s) - \mu_M Y_{M,1}(s) - \beta) f_1(Y_{M,1}(s), Y_{1,2}(s)) - \mu_1 Y_{1,2}(s) f_1(Y_{M,1}(s), Y_{1,2}(s)) ds \right] \\ &+ \mathbb{E}_x \left[\int_0^t f_{1,1}(Y_{M,1}(s), Y_{1,2}(s)) ds \right] + \mathbb{E}_x \left[\int_0^t (f_1(0, Y_{1,2}(s)) - f_2(0, Y_{1,2}(s))) dU(s) \right]. \end{aligned}$$

Therefore, for any function $f(\mathbf{x})$ with $f_1(0, x_2) = f_2(0, x_2)$, the extended generator becomes

$$\mathcal{G}_Y f(\mathbf{x}) = (-\beta - \mu_M x_1 + \mu_1 x_2) f_1(\mathbf{x}) - \mu_1 x_2 f_2(\mathbf{x}) + f_{11}(\mathbf{x}), \mathbf{x} \in \Omega.$$

The geometric ergodicity of the diffusion will follow if we can find a suitable Lyapunov function $V(\mathbf{x})$, constants $c, d > 0$, and a compact set K such that

$$\mathcal{G}_Y V(\mathbf{x}) \leq -cV(\mathbf{x}) + d\mathbb{1}(x \in K).$$

In the rest of this section, our goal is to prove the theorem by identifying the function $V(\mathbf{x})$, the compact set K , and the constants c and d following the same steps as Braverman (2020). The compact set K has the form

$$K = [-\kappa_2, 0] \times [0, \kappa_2],$$

for some $\kappa_2 > \beta$. Unfortunately, the indicator function $\mathbb{1}(x \notin K)$ is not smooth, and hence, we need to define a function $\phi(x) \geq \mathbb{1}(x \notin K)$ for all $x \in \Omega$ with regularity properties to approximate the indicator function. The function we need has the form

$$\phi(x) = \begin{cases} 0 & \text{for } x \leq \kappa_1 \\ (x - \kappa_1)^2 \left(\frac{-(x - \kappa_1)}{((\kappa_2 + \kappa_1)/2 - \kappa_1)^2 (\kappa_2 - \kappa_1)} + \frac{2}{((\kappa_2 + \kappa_1)/2 - \kappa_1)(\kappa_2 - \kappa_1)} \right) & \text{for } x \in [\kappa_1, (\kappa_2 + \kappa_1)/2] \\ 1 - (x - \kappa_2)^2 \left(\frac{-(x - \kappa_1)}{((\kappa_2 + \kappa_1)/2 - \kappa_1)^2 (\kappa_2 - \kappa_1)} - \frac{2}{((\kappa_2 + \kappa_1)/2 - \kappa_1)(\kappa_2 - \kappa_1)} \right) & \text{for } x \in [(\kappa_2 + \kappa_1)/2, \kappa_2] \\ 1 & \text{for } x \geq \kappa_2, \end{cases}$$

for some $\kappa_2 > \kappa_1 > \beta$. We need the first and second derivatives of this function, and as shown in Braverman (2020), we have

$$\begin{aligned} \phi'(\kappa_1) &= \phi'(\kappa_2) = 0, \\ |\phi'(x)| &\leq \frac{4}{\kappa_2 - \kappa_1} \text{ and } |\phi''(x)| \leq \frac{12}{(\kappa_2 - \kappa_1)^2}. \end{aligned}$$

We choose the Lyapunov function $V(\mathbf{x})$ of the form

$$V(\mathbf{x}) = \exp(\alpha(f^{(1)}(\mathbf{x}) + f^{(2)}(\mathbf{x}))), \quad (G.3)$$

where $\alpha > 0$ and

$$\begin{aligned} \mathcal{L}_1 f^{(1)}(\mathbf{x}) &:= (-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(1)}(x) - \mu_1 x_2 f_2^{(1)}(x) = -\phi(-x_1), \mathbf{x} \in \Omega, \\ f_1^{(1)}(0, x_2) &= f_2^{(1)}(0, x_2), x_2 \leq 0, \end{aligned} \quad (G.4)$$

$$\begin{aligned} \mathcal{L}_1 f^{(2)}(\mathbf{x}) &= -\phi(x_2), \mathbf{x} \in \Omega, \\ f_1^{(2)}(0, x_2) &= f_2^{(2)}(0, x_2), x_2 \leq 0. \end{aligned} \quad (G.5)$$

It is easy to check $V_1(0, x_2) = V_2(0, x_2)$ for all $x_2 \geq 0$. For convenience, we use the notation

$$f^{(\Sigma)}(\mathbf{x}) = f^{(1)}(\mathbf{x}) + f^{(2)}(\mathbf{x}).$$

We need the following derivatives of $V(\mathbf{x})$:

$$\begin{aligned} V_1(\mathbf{x}) &= \alpha f_1^{(\Sigma)}(\mathbf{x}) V(\mathbf{x}), \\ V_2(\mathbf{x}) &= \alpha f_2^{(\Sigma)}(\mathbf{x}) V(\mathbf{x}), \\ V_{11}(\mathbf{x}) &= \alpha (f_{11}^{(\Sigma)}(\mathbf{x})) V(\mathbf{x}) + \alpha^2 (f_1^{(\Sigma)}(\mathbf{x}))^2 V(\mathbf{x}). \end{aligned}$$

Now, we have

$$\begin{aligned} \mathcal{G}V(\mathbf{x}) &= (\mathcal{L}f^{(\Sigma)}(\mathbf{x}))\alpha V(\mathbf{x}) + \frac{\alpha}{n} (f_{11}^{(\Sigma)}(\mathbf{x})) V(\mathbf{x}) + \frac{\alpha^2}{n} (f_1^{(\Sigma)}(\mathbf{x}))^2 V(\mathbf{x}) \\ &= (-\phi(-x_1) - \phi(x_2))\alpha V(\mathbf{x}) + (\alpha(f_{11}^{(\Sigma)}(\mathbf{x})) + \alpha^2(f_1^{(\Sigma)}(\mathbf{x}))^2) V(\mathbf{x}) \\ &\leq -\alpha V(\mathbf{x}) 1(\mathbf{x} \notin K) + (\alpha(f_{11}^{(\Sigma)}(\mathbf{x})) + \alpha^2(f_1^{(\Sigma)}(\mathbf{x}))^2) V(\mathbf{x}) \\ &= \alpha(-1 + f_{11}^{(\Sigma)}(\mathbf{x}) + \alpha(f_1^{(\Sigma)}(\mathbf{x}))^2) V(\mathbf{x}) + (\alpha(f_{11}^{(\Sigma)}(\mathbf{x})) + \alpha^2(f_1^{(\Sigma)}(\mathbf{x}))^2) V(\mathbf{x}) 1(\mathbf{x} \in K). \end{aligned}$$

Our goal now is to find $f^{(1)}(\mathbf{x})$ and $f^{(2)}(\mathbf{x})$ that solve the two PDEs in (G.4) and (G.5) and then show that κ_1 and κ_2 can be chosen so that

$$\alpha(-1 + f_{11}^{(\Sigma)}(\mathbf{x}) + \alpha(f_1^{(\Sigma)}(\mathbf{x}))^2) \leq -c, \text{ for all } \mathbf{x} \in \Omega, \quad (\text{G.6})$$

$$(\alpha(f_{11}^{(\Sigma)}(\mathbf{x})) + \alpha^2(f_1^{(\Sigma)}(\mathbf{x}))^2) V(\mathbf{x}) \leq d \text{ for all } \mathbf{x} \in K. \quad (\text{G.7})$$

Lemma G.1. For any $\kappa_2 > \kappa_1 > \beta$, there exist solutions $f^{(1)}(\mathbf{x}), f^{(2)}(\mathbf{x}) \in C^2(\Omega)$ that solves (G.4) and (G.5). Moreover, κ_1 and κ_2 can be chosen so that

$$|f^{(1)}(\mathbf{x})| \leq \frac{1}{\mu_M} \log 2 \quad \mathbf{x} \in [-\kappa_2, 0] \times [0, \kappa_2], \quad (\text{G.8})$$

$$|f^{(2)}(\mathbf{x})| \leq \frac{\log 2}{\mu_1} + \frac{\epsilon}{\beta} \quad \mathbf{x} \in [-\kappa_2, 0] \times [0, \kappa_2], \quad (\text{G.9})$$

$$|f_1^{(1)}(\mathbf{x})| \leq \frac{4 \log 2}{\epsilon}, \quad |f_{11}^{(1)}(\mathbf{x})| \leq \frac{12 \log 2}{\mu_M \epsilon^2}, \quad (\text{G.10})$$

$$|f_1^{(2)}(\mathbf{x})| \leq \frac{1}{\beta}, \quad |f_{11}^{(2)}(\mathbf{x})| \leq \frac{\mu_M + 4}{\beta \epsilon}. \quad (\text{G.11})$$

Proof. For $\kappa_2 > \kappa_1 > \beta$, we first identify functions $f^{(1)}(\mathbf{x})$ and $f^{(2)}(\mathbf{x})$ that solves (G.4) and (G.5) as

$$\begin{aligned} f^{(1)}(\mathbf{x}) &= \int_0^\infty \phi(-v_1^x(t)) dt \\ f^{(2)}(\mathbf{x}) &= \int_0^\infty \phi(v_2^x(t)) dt, \end{aligned}$$

where $v_1^x(t)$ and $v_2^x(t)$ are as in Section 6.1 with $n=1$. To evaluate the first integral, we need $t \geq 0$ that solves

$$\frac{\beta}{\mu_M} - \kappa = \left(x_1 + \frac{\beta}{\mu_M} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)t}) \right) e^{-\mu_M t}.$$

There exists a solution to this equation when $x_1 < -\kappa$, as the right-hand side is x_1 for $t=0$ and zero as $t \rightarrow \infty$. However, the solution may not be unique. Hence, we define

$$\tau^\kappa(\mathbf{x}) = \min \left\{ t \geq 0 : \frac{\beta}{\mu_M} - \kappa \leq \left(x_1 + \frac{\beta}{\mu_M} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)t}) \right) e^{-\mu_M t} \right\}.$$

The continuity of the right-hand side implies that, for any $x_1 < -\kappa_2 < -\kappa_1 < 0$, we have $\tau^{\kappa_2}(\mathbf{x}) < \tau^{\kappa_1}(\mathbf{x})$. We also have

$$\begin{aligned} \frac{\mu_M \kappa_2 - \beta}{\mu_M \kappa_1 - \beta} &= \frac{\left(x_1 + \frac{\beta}{\mu_M} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)\tau^{\kappa_2}(\mathbf{x})}) \right) e^{-\mu_M \tau^{\kappa_2}(\mathbf{x})}}{\left(x_1 + \frac{\beta}{\mu_M} + \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)\tau^{\kappa_1}(\mathbf{x})}) \right) e^{-\mu_M \tau^{\kappa_1}(\mathbf{x})}} \\ &\geq e^{\mu_M(\tau^{\kappa_1}(\mathbf{x}) - \tau^{\kappa_2}(\mathbf{x}))} \\ \frac{1}{\mu_M} \log \left(\frac{\mu_M \kappa_2 - \beta}{\mu_M \kappa_1 - \beta} \right) &\geq \tau^{\kappa_1}(\mathbf{x}) - \tau^{\kappa_2}(\mathbf{x}). \end{aligned} \quad (\text{G.12})$$

Now, we can write

$$f^{(1)}(\mathbf{x}) = \begin{cases} \tau^{\kappa_2}(\mathbf{x}) + \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} \phi\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt & \text{for } x_1 < -\kappa_2 \\ \int_0^{\tau^{\kappa_1}} \phi\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt & \text{for } x_1 \in [-\kappa_2, -\kappa_1] \\ 0 & \text{for } x_1 \in [-\kappa_1, 0]. \end{cases}$$

We have $f_1^{(1)}(0, x_2) = f_2^{(1)}(0, x_2) = 0$, and hence, $f^{(1)}(\mathbf{x})$ satisfies the boundary condition. To check whether the suggested $f^{(1)}(\mathbf{x})$ satisfies the PDE, we now calculate its derivatives. First, consider the derivative with respect to x_1 when $x_1 < -\kappa_2$:

$$\begin{aligned} f_1^{(1)}(\mathbf{x}) &= \tau_1^{\kappa_2}(\mathbf{x}) + \phi\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)\tau^{\kappa_1}}(x))\right)e^{-\mu_M \tau^{\kappa_1}(\mathbf{x})}\right) \tau_1^{\kappa_1}(\mathbf{x}) \\ &\quad - \phi\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)\tau^{\kappa_2}}(x))\right)e^{-\mu_M \tau^{\kappa_2}(\mathbf{x})}\right) \tau_1^{\kappa_2}(\mathbf{x}) \\ &\quad + \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} e^{-\mu_M t} \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt \\ &= \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} e^{-\mu_M t} \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt. \end{aligned}$$

Note that the boundary terms reduce to $\phi(\kappa_1) = 0$ and $\phi(\kappa_2) = 1$. Now, considering the case $x_1 \in [-\kappa_2, -\kappa_1]$, and using the same trick for the boundary term, we get

$$f_1^{(1)}(\mathbf{x}) = \int_0^{\tau^{\kappa_1}} e^{-\mu_M t} \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt.$$

Combining, the results we have

$$f_1^{(1)}(\mathbf{x}) = \begin{cases} \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} e^{-\mu_M t} \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt, & \text{for } x_1 < -\kappa_2 \\ \int_0^{\tau^{\kappa_1}} e^{-\mu_M t} \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt, & \text{for } x_1 \in [-\kappa_2, -\kappa_1]. \end{cases}$$

Using a similar argument, we get

$$f_2^{(1)}(\mathbf{x}) = \begin{cases} \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} \frac{\mu_1}{\mu_1 - \mu_M} (e^{-\mu_M t} - e^{-\mu_1 t}) \times \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt, & \text{for } x_1 < -\kappa_2 \\ \int_0^{\tau^{\kappa_1}} \frac{\mu_1 x_2}{\mu_1 - \mu_M} (e^{-\mu_M t} - e^{-\mu_1 t}) \times \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt, & \text{for } x_1 \in [-\kappa_2, -\kappa_1] \\ 0, & \text{for } x_1 \in [-\kappa_1, 0]. \end{cases}$$

Now, we can show that for $x_1 < -\kappa_2$

$$\begin{aligned} (-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(1)}(\mathbf{x}) + \mu_1 x_2 f_2^{(1)}(\mathbf{x}) &= \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} \left((-\beta - \mu_M x_1 + \mu_1 x_2) e^{-\mu_M t} - \mu_1 x_2 \frac{\mu_1}{\mu_1 - \mu_M} (e^{-\mu_M t} - e^{-\mu_1 t}) \right) \\ &\quad \times \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt \\ &= \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} \left((-\beta - \mu_M x_1) e^{-\mu_M t} - \frac{\mu_1 \mu_M x_2}{\mu_1 - \mu_M} e^{-\mu_M t} + \frac{\mu_1^2 x_2}{\mu_1 - \mu_M} e^{-\mu_1 t} \right) \\ &\quad \times \phi'\left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1}{\mu_1 - \mu_M}(1 - e^{-(\mu_1 - \mu_M)t})\right)e^{-\mu_M t}\right) dt. \end{aligned}$$

Now, making the variable change

$$u = \frac{\beta}{\mu_M} - x_1 e^{-\mu_M t} + \frac{\beta}{\mu_M} e^{-\mu_M t} - \frac{\mu_1}{\mu_1 - \mu_M} e^{-\mu_M t} - \frac{\mu_1}{\mu_1 - \mu_M} e^{-\mu_1 t},$$

$$du = (\beta + \mu_M x_1) e^{-\mu_M t} + \frac{\mu_1 \mu_M x_2}{\mu_1 - \mu_M} e^{-\mu_M t} - \frac{\mu_1^2 x_2}{\mu_1 - \mu_M} e^{-\mu_1 t} dt,$$

we have

$$(-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(1)}(\mathbf{x}) + \mu_1 x_2 f_2^{(1)}(\mathbf{x}) = \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} -\phi'(u) du = -1 = -\phi(-x_1).$$

A similar result holds for all other cases; hence, $f^{(1)}(\mathbf{x})$ satisfy the PDE along with the boundary conditions. To prove a similar result for $f^{(2)}(\mathbf{x})$, we will again follow the analysis in Braverman closely. First, we define four subdomains

$$\begin{aligned} S_0 &= \{\mathbf{x} \in \Omega : x_2 \leq \kappa_1\} \\ S_1 &= \{\mathbf{x} \in \Omega : x_2 \geq \kappa_1, \mathbf{x} \leq \Gamma_{\kappa_1}\} \\ S_2 &= \{\mathbf{x} \in \Omega : x_2 \geq \kappa_1, \Gamma_{\kappa_1} \geq \mathbf{x} \geq \Gamma_{\kappa_2}\} \\ S_3 &= \{\mathbf{x} \in \Omega : x_2 \geq \kappa_1, \mathbf{x} \geq \Gamma_{\kappa_2}\}. \end{aligned}$$

Then, we can define the function as

$$f^{(2)}(\mathbf{x}) = \begin{cases} 0, & \text{for } \mathbf{x} \in S_0 \\ \int_0^{\frac{\log(x_2/\kappa_1)}{\mu_1}} \phi(x_2 e^{-\mu_1 t}) dt, & \text{for } x_2 \leq \kappa_2, \mathbf{x} \in S_1 \\ \frac{\log(x_2/\kappa_2)}{\mu_1} + \int_0^{\frac{\log(x_2/\kappa_1)}{\mu_1}} \phi(\kappa_2 e^{-\mu_1 t}) dt, & \text{for } x_2 \geq \kappa_2, \mathbf{x} \in S_1 \\ \int_0^{\tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 t}) dt + \frac{\sqrt{n}}{\beta} \int_{\kappa_1}^{x_2 e^{-\mu_1 \tau(\mathbf{x})}} \phi(t) dt, & \text{for } x_2 \leq \kappa_2, \mathbf{x} \in S_2 \\ \frac{\log(x_2/\kappa_2)}{\mu_1} + \int_{\frac{\log(x_2/\kappa_2)}{\mu_1}}^{\tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 t}) dt + \frac{1}{\beta} \int_{\kappa_1}^{x_2 e^{-\mu_1 \tau(\mathbf{x})}} \phi(t) dt, & \text{for } x_2 \geq \kappa_2, \mathbf{x} \in S_2 \\ \tau(\mathbf{x}) + \frac{x_2 e^{-\mu_1 \tau(\mathbf{x})} - \kappa_2}{\beta} + \frac{1}{\beta} \int_{\kappa_1/\sqrt{n}}^{\kappa_2} \phi(t) dt, & \text{for } \mathbf{x} \in S_3. \end{cases}$$

Now, we calculate the first derivatives with respect to x_1 and x_2 :

$$f_1^{(2)}(\mathbf{x}) = \begin{cases} 0, & \text{for } \mathbf{x} \in S_0 \cup S_1 \\ \frac{1}{\beta} e^{-\mu_M \tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}), & \text{for } \mathbf{x} \in S_2 \\ \frac{1}{\beta} e^{-\mu_M \tau(\mathbf{x})}, & \text{for } \mathbf{x} \in S_3 \end{cases}$$

$$f_2^{(2)}(\mathbf{x}) = \begin{cases} 0, & \text{for } \mathbf{x} \in S_0 \\ \frac{1}{\mu_1 x_2} \phi(x_2), & \text{for } \mathbf{x} \in S_1 \\ \frac{1}{\mu_1 x_2} (\phi(x_2) - \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})})) + \frac{1}{\beta} \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \left(\frac{\mu_1 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} - \frac{\mu_M e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right), & \text{for } \mathbf{x} \in S_2 \\ \frac{1}{\beta} \frac{\mu_1 e^{-\mu_M \tau(\mathbf{x})} - \mu_M e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M}, & \text{for } \mathbf{x} \in S_3. \end{cases}$$

Now, we prove that the suggested $f^{(2)}(\mathbf{x})$ satisfies the PDE.

1. When $\mathbf{x} \in S_0$, $\mathcal{L}f(\mathbf{x}) = 0 = -\phi(0)$
2. When $\mathbf{x} \in S_1$,

$$(-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(2)}(\mathbf{x}) + \mu_1 x_2 f_2^{(2)}(\mathbf{x}) = -\mu_1 x_2 f_2^{(2)}(\mathbf{x}) = -\phi(x_2)$$

3. When $\mathbf{x} \in S_2$,

$$\begin{aligned}
 & (-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(2)}(\mathbf{x}) + \mu_1 x_2 f_2^{(2)}(\mathbf{x}) \\
 &= -e^{-\mu_M \tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) - \frac{1}{\beta} \mu_M x_1 e^{-\mu_M \tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) + \frac{1}{\beta} \mu_1 x_2 e^{-\mu_M \tau(\mathbf{x})} \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) - \phi(\mathbf{x}) + \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \\
 &\quad - \frac{1}{\beta} \mu_1 x_2 \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \left(\frac{\mu_1 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} - \frac{\mu_M e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) \\
 &= -\phi(\mathbf{x}) + \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \left(1 - e^{-\mu_M \tau(\mathbf{x})} - \frac{1}{\beta} \mu_M x_1 e^{-\mu_M \tau(\mathbf{x})} + \frac{1}{\beta} \frac{\mu_1^2 x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} - \frac{1}{\beta} \frac{\mu_1 \mu_M x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} - \frac{1}{\beta} \frac{\mu_1^2 x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} + \frac{1}{\beta} \frac{\mu_M \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) \\
 &= -\phi(\mathbf{x}) + \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \left(1 - e^{-\mu_M \tau(\mathbf{x})} - \frac{1}{\beta} \mu_M x_1 e^{-\mu_M \tau(\mathbf{x})} - \frac{1}{\beta} \frac{\mu_1 \mu_M x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} + \frac{1}{\beta} \frac{\mu_M \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) \\
 &= -\phi(\mathbf{x}) + \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) \left(1 - e^{-\mu_M \tau(\mathbf{x})} - \left(1 - e^{-\mu_M \tau(\mathbf{x})} - \frac{\mu_M}{\beta} \frac{\mu_1 x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} - \frac{\mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) - \frac{1}{\beta} \frac{\mu_1 \mu_M x_2 e^{-\mu_M \tau(\mathbf{x})}}{\mu_1 - \mu_M} + \frac{1}{\beta} \frac{\mu_M \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) \\
 &= -\phi(\mathbf{x}).
 \end{aligned}$$

4. When $\mathbf{x} \in S_3$,

$$\begin{aligned}
 & (-\beta - \mu_M x_1 + \mu_1 x_2) f_1^{(2)}(\mathbf{x}) + \mu_1 x_2 f_2^{(2)}(\mathbf{x}) = (-\beta - \mu_M x_1 + \mu_1 x_2) \frac{1}{\beta} e^{-\mu_M \tau(\mathbf{x})} + \frac{1}{\beta} \frac{\mu_1^2 x_2 e^{-\mu_M \tau(\mathbf{x})} - \mu_1 \mu_M x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \\
 &= -e^{-\mu_M \tau(\mathbf{x})} - \frac{1}{\beta} \mu_M \left(\frac{\beta}{\mu_M} - \frac{\beta}{\mu_M} e^{-\mu_M \tau(\mathbf{x})} - \frac{\mu_1 x_2 e^{-\mu_M \tau(\mathbf{x})} - \mu_1 x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \right) \\
 &\quad + \frac{1}{\beta} \frac{\mu_1^2 x_2 e^{-\mu_1 \tau(\mathbf{x})} - \mu_1 \mu_M x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} + \frac{1}{\beta} \frac{\mu_1^2 x_2 e^{-\mu_M \tau(\mathbf{x})} - \mu_1 \mu_M x_2 e^{-\mu_1 \tau(\mathbf{x})}}{\mu_1 - \mu_M} \\
 &= -1 = -\phi(\mathbf{x}).
 \end{aligned}$$

Now, in a similar fashion we check the boundary conditions.

1. When $\mathbf{x} \in S_0$, the equality is trivial as both derivatives are equal to zero.
2. When $\mathbf{x} \in S_1$, $f_1^{(2)}(0, x_2) = 0$. Note that the only point in S_1 with $x_1 = 0$ is $x_2 = \kappa_1$ and $f_2^{(2)}(0, \kappa_1) = 0$. Hence, the boundary condition is satisfied.
3. When $\mathbf{x} \in S_2$ and $x_1 = 0$, $\tau(\mathbf{x}) = 0$, which implies $f_1^{(2)}(0, x_2) = f_2^{(2)}(0, x_2) = \frac{1}{\beta} \phi(x_2)$.
4. Similarly, when $\mathbf{x} \in S_3$, $f_1^{(2)}(0, x_2) = f_2^{(2)}(0, x_2) = \frac{1}{\beta}$.

Hence, the boundary condition is also satisfied for all values of x_2 .

Now, we prove that the suggested $f^{(1)}(\mathbf{x})$ satisfy the bounds in (G.8)–(G.11) as

$$f^{(1)}(\mathbf{x}) \leq \tau^{\kappa_1}(\mathbf{x}) \leq \tau^{\kappa_1}(\kappa_2) - \tau^{\kappa_2}(\kappa_2) \leq \frac{1}{\mu_M} \log \left(\frac{\mu_M \kappa_2 - \beta}{\mu_M \kappa_1 - \beta} \right) \text{ for all } x_1 \leq \kappa_2. \quad (\text{G.13})$$

Similarly, for $f^{(2)}(\mathbf{x})$, we have

$$f^{(2)}(\mathbf{x}) \leq \frac{\log(\kappa_2/\kappa_1)}{\mu_1}, \text{ for } x_2 \leq \kappa_2, \mathbf{x} \in S_1, \quad (\text{G.14})$$

$$f^{(2)}(\mathbf{x}) \leq \tau(\mathbf{x}) + \frac{x_2 e^{-\mu_1 \tau(\mathbf{x})} - \kappa_1}{\beta} \leq \frac{\log(\kappa_2/\kappa_1)}{\mu_1} + \frac{\kappa_2 - \kappa_1}{\beta}, \text{ for } x_2 \leq \kappa_2, \mathbf{x} \in S_2, \quad (\text{G.15})$$

where the second inequality in (G.15) follows as

$$x_2 e^{-\mu_1 \tau(\mathbf{x})} \geq \kappa_1 \text{ and } x_2 \leq \kappa_2, \text{ when } \mathbf{x} \in S_2.$$

Now, we prove the bounds for $f^{(1)}(\mathbf{x})$ and $f^{(2)}(\mathbf{x})$. As shown in Braverman, we have

$$|\phi'(\mathbf{x})| \leq 4/(\kappa_2 - \kappa_1),$$

and hence,

$$f_1^{(1)}(\mathbf{x}) \leq \frac{4\tau^{\kappa_1}(\mathbf{x})}{\kappa_2 - \kappa_1} \leq \frac{4 \log \left(\frac{\mu_M \kappa_2 - \beta}{\mu_M \kappa_1 - \beta} \right)}{\kappa_2 - \kappa_1} \text{ for all } x_1 \leq \kappa_2, \quad (\text{G.16})$$

$$|f_1^{(2)}(\mathbf{x})| \leq \frac{1}{\beta} \text{ for all } \mathbf{x} \in \Omega. \quad (\text{G.17})$$

Now, we will focus on $f_{11}^{(1)}(\mathbf{x})$ and $f_{11}^{(2)}(\mathbf{x})$:

$$f_{11}^{(1)}(\mathbf{x}) = \begin{cases} \int_{\tau^{\kappa_2}}^{\tau^{\kappa_1}} e^{-2\mu_M t} \phi'' \left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)t}) \right) e^{-\mu_M t} \right) dt, & \text{for } x_1 < -\kappa_2 \\ \int_0^{\tau^{\kappa_1}} e^{-2\mu_M t} \phi'' \left(\frac{\beta}{\mu_M} - \left(x_1 + \frac{\beta}{\mu_M} - \frac{\mu_1 x_2}{\mu_1 - \mu_M} (1 - e^{-(\mu_1 - \mu_M)t}) \right) e^{-\mu_M t} \right) dt & \text{for } x_1 \in [-\kappa_2, -\kappa_1] \\ 0 & \text{for } x_1 \in [-\kappa_1, 0]. \end{cases}$$

$$f_{11}^{(2)}(\mathbf{x}) = \begin{cases} 0 & \text{for } \mathbf{x} \in S_0 \cup S_1 \\ -\frac{1}{\beta} \mu_M \tau_1(\mathbf{x}) \phi(x_2 e^{-\mu_1 \tau(\mathbf{x})}) - \frac{1}{\beta} e^{-(\mu_1 + \mu_M) \tau(\mathbf{x})} \phi'(x_2 e^{-\mu_1 \tau(\mathbf{x})}) x_2 \mu_1 \tau_1(\mathbf{x}) & \text{for } \mathbf{x} \in S_2 \\ -\frac{1}{\beta} \mu_M \tau_1(\mathbf{x}) & \text{for } \mathbf{x} \in S_3. \end{cases}$$

Again, as shown in Braverman, $|\phi''(\mathbf{x})| \leq 12/(\kappa_2 - \kappa_1)$, and hence,

$$|f_{11}^{(1)}(\mathbf{x})| \leq \frac{12 \log \left(\frac{\mu_M \kappa_2 - \beta}{\mu_M \kappa_1 - \beta} \right)}{\mu_M (\kappa_2 - \kappa_1)^2}. \quad (\text{G.18})$$

We also have

$$|f_{11}^{(2)}(\mathbf{x})| \leq \frac{\mu_M}{\beta(\mu_1 x_2 - \beta)} + \frac{4}{\beta(\kappa_2 - \kappa_1)} \text{ for all } x_2 \geq \kappa_1. \quad (\text{G.19})$$

Choosing $\kappa_2 = \frac{\beta}{\mu_M} + 2\epsilon$ and $\kappa_1 = \frac{\beta}{\mu_M} + \epsilon$ and realizing $\mu_M \leq 1 \leq \mu_1$, (G.13)–(G.19) yields (G.8)–(G.11). $\square$

Proof of Theorem 3. Lemma G.1 implies that for any $\epsilon > 0$, we can choose $f^{(1)}(\mathbf{x})$ and $f^{(2)}(\mathbf{x})$ solving (G.4) and (G.5), respectively, such that

$$\alpha(-1 + f_{11}^{(\Sigma)}(\mathbf{x}) + \alpha(f_{11}^{(\Sigma)}(\mathbf{x}))^2) \leq \alpha \left(-1 + \frac{12 \log 2}{\mu_M \epsilon^2} + \frac{\mu_M + 4}{\beta \epsilon} + \alpha \left(\frac{4 \log 2}{\epsilon} + \frac{1}{\beta} \right)^2 \right),$$

and

$$(\alpha(f_{11}^{(\Sigma)}(\mathbf{x})) + \alpha^2(f_{11}^{(\Sigma)}(\mathbf{x}))^2) \leq \left(\alpha \left(\frac{12 \log 2}{\mu_M \epsilon^2} + \frac{\mu_M + 4}{\beta \epsilon} \right) + \alpha^2 \left(\frac{4 \log 2}{\epsilon} + \frac{1}{\beta} \right)^2 \right).$$

Choosing α sufficiently small and ϵ sufficiently large, we can find constants $c, d > 0$ to satisfy (G.6) and (G.7), which proves the first part of the theorem. Then, the positive recurrence of the diffusion follows from theorem 5.2 in Down et al. (1995). $\square$

References

- Armony M (2005) Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51:287–329.
- Atar R (2012) A diffusion regime with nondegenerate slowdown. *Oper. Res.* 60(2):490–500.
- Bhambay S, Mukhopadhyay A (2022) Asymptotic optimality of speed-aware JSQ for heterogeneous service systems. *Performance Evaluation* 157:102320.
- Bramson M, Lu Y, Prabhakar B (2012) Asymptotic independence of queues under randomized load balancing. *Queueing Systems* 71(3):247–292.
- Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin-Whitt regime. *Math. Oper. Res.* 45(3):1069–1103.
- Braverman A (2023) The join-the-shortest-queue system in the Halfin-Whitt regime: Rates of convergence to the diffusion limit. *Stochastic Systems* 13(1):1–39.
- Büke B, Qin W (2023) Many-server queues with random service rates: A unified framework based on measure-valued processes. *Math. Oper. Res.* 48(2):748–783.
- Das K (1979) A note on an inequality due to Greene. *Proc. Amer. Math. Soc.* 77(3):424–425.
- Down D, Meyn SP, Tweedie RL (1995) Exponential and uniform ergodicity of Markov processes. *Ann. Probability* 23(4):1671–1691.
- Duato J, Peña AJ, Silla F, Mayo R, Quintana-Ortí ES (2010) rCUDA: Reducing the number of gpu-based accelerators in high performance clusters. *Proc. Internat. Conf. High Performance Comput. Simulation* (IEEE, Piscataway, NJ), 224–231.
- Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Oper. Res.* 43(3):867–886.
- Gardner K, Jaleel JA, Wickeham A, Doroudi S (2021) Scalable load balancing in the presence of heterogeneous servers. *Performance Evaluation* 145:102151.
- Greene D (1977) An inequality for a class of integral systems. *Proc. Amer. Math. Soc.* 62(1):101–104.
- Gupta V, Walton N (2019) Load balancing in the nondegenerate slowdown regime. *Oper. Res.* 67(1):281–294.
- Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

- Hordijk A, Koole G (1990) On the optimality of the generalized shortest queue policy. *Probability Engrg. Inform. Sci.* 4(4):477–487.
- Huang M, Wu D, Yu CH, Fang Z, Interlandi M, Condie T, Cong J (2016) Programming and runtime support to blaze FPGA accelerator deployment at datacenter scale. *Proc. 7th ACM Sympos. Cloud Comput.* (Association for Computing Machinery, New York), 456–469.
- Hyttia E, Richter R, Samúelsson SG (2017) Beyond shortest queue routing with heterogeneous servers and general cost functions. *Proc. 11th EAI Internat. Conf. Performance Evaluation Methodologies Tools* (ACM, New York), 206–213.
- Izagirre A, Makowski AM (2014) Light traffic performance under the power of two load balancing strategy: The case of server heterogeneity. *SIGMETRICS Performance Evaluation Rev.* 42(2):18–20.
- Johri PK (1989) Optimality of the shortest line discipline with state-dependent service rates. *Eur. J. Oper. Res.* 41(2):157–161.
- Krishnan K (1987) Joining the right queue: A Markov decision-rule. *Proc. 26th IEEE Conf. Decision Control*, vol. 26 (IEEE, Piscataway, NJ), 1863–1868.
- Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.
- Mitzenmacher M (1996) The power of two choices in randomized load balancing. PhD thesis, University of California, Berkeley.
- Mukherjee D, Borst SC, Van Leeuwen JS, Whiting PA (2016) Universality of load balancing schemes on the diffusion scale. *J. Appl. Probability* 53(4):1111–1124.
- Mukhopadhyay A, Mazumdar RR (2015) Analysis of randomized join-the-shortest-queue (JSQ) schemes in large heterogeneous processor-sharing systems. *IEEE Trans. Control Network Systems* 3(2):116–126.
- Mukhopadhyay A, Karthik A, Mazumdar RR (2016) Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay. *Stochastic Systems* 6(1):90–131.
- Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probab. Surveys* 4:193–267.
- Renardy M, Rogers R (2004) *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics (Springer, New York).
- Rudin W (1987) *Real and Complex Analysis*, 3rd ed. (McGraw-Hill, New York).
- Rutten D, Mukherjee D (2023) Load balancing under strict compatibility constraints. *Math. Oper. Res.* 48(1):227–256.
- Stolyar AL (2015a) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80:341–361.
- Stolyar AL (2015b) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stochastic Systems* 5(2):239–267.
- Tezcan T (2008) Optimal control of distributed parallel server systems under the halfin and whitt regime. *Math. Oper. Res.* 33(1):51–90.
- Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Inform.* 32(1):20–34.
- Wang W, Maguluri ST, Srikant R, Ying L (2022) Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies. *Math. Oper. Res.* 47(4):2691–2720.
- Weber RR (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Probability* 15(2):406–413.
- Weng W, Zhou X, Srikant R (2021) Optimal load balancing with locality constraints. *Proc. 2021 ACM SIGMETRICS/Internat. Conf. Measurement Modeling Comput. Systems* (ACM, New York), 49–50.
- Whitt W (2001) *Stochastic-Process Limits* (Springer-Verlag, New York).
- Winston W (1977) Optimality of the shortest line discipline. *J. Appl. Probability* 14(1):181–189.
- Zhou X, Wu F, Tan J, Sun Y, Shroff N (2017) Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms. *Proc. ACM Measurement Anal. Comput. Systems* 1(2):1–30.