Asymptotic Optimality of Speed-Aware JSQ for Heterogeneous Service Systems

Sanidhay Bhambay and Arpan Mukhopadhyay
University of Warwick

Abstract

The Join-the-Shortest-Queue (JSQ) load-balancing scheme is known to minimise the average delay of jobs in homogeneous systems consisting of identical servers. However, it performs poorly in heterogeneous systems where servers have different processing rates. Finding a delay optimal scheme remains an open problem for heterogeneous systems. In this paper, we consider a speed-aware version of the JSQ scheme for heterogeneous systems and show that it achieves delay optimality in the fluid limit. One of the key issues in establishing this optimality result for heterogeneous systems is to show that the sequence of steady-state distributions indexed by the system size is tight in an appropriately defined space. The usual technique for showing tightness by coupling with a suitably defined dominant system does not work for heterogeneous systems. To prove tightness, we devise a new technique that uses the drift of exponential Lyapunov functions. Using the non-negativity of the drift, we show that the stationary queue length distribution has an exponentially decaying tail - a fact we use to prove tightness. Another technical difficulty arises due to the complexity of the underlying state-space and the separation of two time-scales in the fluid limit. Due to these factors, the fluid-limit turns out to be a function of the invariant distribution of a multi-dimensional Markov chain which is hard to characterise. By using some properties of this invariant distribution and using the monotonicity of the system, we show that the fluid limit is has a unique and globally attractive fixed point.

1 Introduction

The average response time of user requests is a key performance measure in modern large-scale service systems such as web server farms and cloud data centers. How incoming user requests or *jobs* are assigned to *servers* can significantly impact the performance of such systems. The canonical model for studying the effect of job assignment schemes on the mean response time of jobs consists of N servers each having its own queue, a stream of incoming jobs arriving at rate $N\lambda$, and a job dispatcher. The job dispatcher assigns every incoming job to a server in the system according to a specific job assignment scheme. For homogeneous systems consisting of identical servers, a natural job assignment scheme to consider is the Join-the-Shortest-Queue (JSQ) scheme in which each incoming job is assigned to the server with the minimum queue length. For finite N, this scheme is known to achieve the minimum average response time of jobs in a variety of settings [1–4]. Furthermore, in the large system limit ($N \to \infty$) it has been shown that the average queuing delay of jobs (average time a job spends in the queue before its processing starts) under JSQ approaches to zero.

Variants of the JSQ scheme such as SQ(d) scheme and the Join-the-Idle-Queue (JIQ) scheme have also been extensively studied in the literature [5–7]. In the SQ(d) scheme, each job is assigned to the shortest of d randomly sampled queues; in the JIQ scheme, each job is assigned to an idle queue if one is available and is otherwise sent to a randomly sampled queue. These schemes require less messaging overhead than JSQ but achieve the same asymptotic performance as JSQ in the large system limit.

The (asymptotic) optimality of JSQ and its variants crucially relies on the assumption of homogeneity of server speeds. However, this assumption is not accurate in practice since data centres typically contain multiple-generations of physical devices with different processing capabilities [8]. Processing speeds of servers can vary also due to the presence of various types of acceleration devices such as GPUs, FPGAs and ASICs [9, 10]. Finding the delay optimal scheme for such systems remains an open problem to this date. For such systems, speed-unaware schemes such as JSQ and SQ(d), designed primarily for homogeneous systems, may perform very poorly [11–13]. Speed-aware schemes which assign jobs based on both queue lengths and server speeds can significantly outperform speed-unaware schemes. To see this, consider a speed-aware version of JSQ, henceforth referred to as the *Speed-Aware JSQ (SA-JSQ)* scheme, in which each arrival is assigned to a server with the highest speed among the ones with the minimum queue length. In Figure 1, we compare the JSQ scheme with the SA-JSQ scheme for a system with two types of servers. Inter-arrival and service times of jobs are assumed to be exponentially distributed with rates $N\lambda$ and 1, respectively. We observe that the average response time of jobs is almost 60% lower under the SA-JSQ scheme than under the JSQ scheme for loads (λ) less than 0.5. This example clearly highlights the need to design speed-aware schemes which can lead to much improved performance compared to speed-unaware schemes for heterogeneous systems.

There are various ways in which server speeds can be incorporated into job dispatching decisions. A natural scheme is to send each job to the queue with the shortest expected delay (SED), i.e., the queue where the ratio of the queue length to the server-speed is minimum. As we shall see later, for finite N this scheme performs slightly better than the SA-JSQ scheme discussed above but this difference vanishes quickly as the number of servers increases. Thus, the key questions in the context of heterogeneous systems are the following: *what is the minimum achievable mean response time of jobs in heterogeneous systems?* and *is it possible achieve this minimum with some simple speed-aware variants of the JSQ scheme in the limit as the system size becomes large?*

1.1 Contributions

To answer the questions above we need to find a lower bound on the mean-response time of jobs in heterogeneous systems and analytically characterise the performance of speed-aware dispatching schemes in the limit as $N \to \infty$. These are the primary aims of this

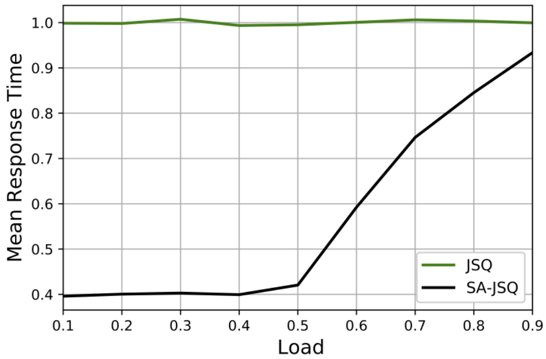


Figure 1: Comparison of mean response time of jobs under JSQ and SA-JSQ schemes. Both schemes are applied to a heterogeneous system consisting of two types of servers: a fraction γ_i of servers have rate μ_i. We choose $\gamma_1 = 1 - \gamma_2 = 1/5$, and $\mu_1 = 4\mu_2 = 20/8$.

paper. Our key finding is that in the fluid limit the SA-JSQ scheme achieves the lowest possible mean response time of jobs for heterogeneous systems. Note that to prove such optimality results, it is essential to develop the necessary tools to analyse job assignment schemes for heterogeneous systems. One would expect the fluid limit techniques applicable to homogeneous systems would easily generalise to heterogeneous systems. However, this turns out to be not the case. As explained later, the stochastic coupling technique, used to obtain uniform bounds on the stationary queue lengths in homogeneous systems, is not generalizable to heterogeneous systems. Consequently, we develop a new method based on exponential Lyapunov functions to obtain similar uniform bounds for heterogeneous systems. This method is sufficiently general to be applicable to other systems where a direct coupling is difficult to construct. Furthermore, the increased complexity of the underlying state space for heterogeneous systems results in a fluid limit which depends on the invariant distribution of a multi-dimensional birth-death process. Deriving exact analytical expressions for this invariant distribution is difficult. We find properties that are sufficient to show the existence, uniqueness and global stability of the fluid limit. From this point onward, we use the term 'optimality' to refer to optimality in the fluid limit unless otherwise mentioned. Our contributions are listed below:

1. *Stochastic comparison and lower bound*: Our first contribution is to compare the heterogeneous system, which consists of separate queues, with a similar heterogeneous system where all the servers serve a single central queue. We show that the system with separate queues stochastically dominates the system with the central queue. From this stochastic comparison result and the fluid limit of the system with central queue, we obtain a lower bound on the mean response time of jobs that holds for any dispatching scheme applied to the system with separate queues. It is important to note that studying fluid limit of the system with a single central queue and heterogeneous

servers can be of independent interest.

2. *Stability and tightness*: We next show that the heterogeneous system is stable under the SA-JSQ policy and obtain uniform bounds (not depending on the system size) on the tails of the stationary queue length distribution. The latter result is required to show the tightness of the sequence of stationary distributions indexed by the system size. The usual approach for proving tightness via coupling does not work in the heterogeneous setting as it is difficult to construct a coupling that maintains the desired stochastic dominance. Thus, we take a new approach and establish this result by analysing the drift of an exponential Lyapunov function. This technique is sufficiently general to be applicable to other systems where direct stochastic comparison is difficult.

3. *Fluid limit analysis*: Our final contribution is the fluid limit analysis of the SA-JSQ scheme in the heterogeneous system. This analysis turns out to be considerably more challenging than conventional fluid analysis due to the increased complexity of the underlying state space for heterogeneous systems and the separation of two time-scales in the fluid limit. The main idea in establishing the fluid limit of SA-JSQ is to first use the martingale functional central limit theorem (FCLT) and then characterise the limit of any convergent subsequence using Lemma 2 and Theorem 3 of [14]. The fluid limit in the heterogeneous setting turns out to be a function of the invariant distribution of a multi-dimensional Markov process whose exact analysis seems intractable. However, we are able to prove properties of the invariant distribution sufficient to characterise the fluid limit and its unique fixed point. Using these properties and the monotonicity of the system we also show that the fluid limit is globally stable. This final result establishes the asymptotic optimality of the SA-JSQ policy for heterogeneous systems.

1.2 Related works

There exists a vast literature on load-balancing policies for multi-server systems. Here, we only discuss the works that are most relevant to our paper. For a comprehensive review of existing works, we refer the reader to [15].

The JSQ policy was first shown to minimise the average delay of jobs for finite systems consisting of identical servers in [1] under the assumption of Poisson arrivals and exponential service times. This optimality result was later extended to general stochastic arrival processes and service-time distributions with non-decreasing hazard rates in [2], to queues with state-dependent service rates in [3], and to systems with finite buffer-buffers and general batch arrivals in [4]. Recent works [7, 16] have considered the fluid and diffusion limits of the JSQ scheme. In the fluid limit, it has been shown in [7] that the fraction of servers with two or more jobs converges to zero under the JSQ scheme. This implies that in the fluid limit all jobs find an idle server to join. In the Halfin-Whitt regime, where the normalized arrival rate λ varies with the system size N as $\lambda = 1 - \beta/\sqrt{N}$ for some $\beta > 0$, it has been shown in [16] that the diffusion-scaled process approaches to a two-dimensional reflected

Ornstein-Ulhenbeck (OU) process as $N \to \infty$. The stationary distribution of this OU process has been studied in [17] which establishes that the steady-state fraction of servers with exactly two jobs scales as $O(1/\sqrt{N})$ and the fraction of servers with more than three jobs scales as $O(1/N)$.

Another line of works explores load balancing schemes which require less communication between the servers and the job dispatcher. The SQ(d) scheme in which each arrival is assigned to the shortest among d randomly sampled queues was first analyzed independently in [18] and [19]. Using mean-field analysis, it was shown that for $d \geq 2$ the stationary queue length distribution has a super-exponentially decaying tail for large system sizes. Thus, by querying only $d \geq 2$ servers at every arrival instant a significant reduction in the average delay can be obtained in comparison to $d = 1$. The Join-the-Idle-Queue (JIQ) further reduces communication overhead by keeping track of only the idle servers in the system. At each arrival instant it sends the incoming job to an idle server if one is available; otherwise, the job is sent to a randomly sampled server. This scheme was first proposed and analyzed by Lu *et al.* in [20] and it was shown that in the fluid limit the JIQ scheme achieves the same performance as the JSQ scheme. Thus, the JIQ scheme is asymptotically optimal for homogeneous systems in the fluid limit.

Relatively few works consider load balancing in heterogeneous systems. The SQ(d) scheme for heterogeneous systems has been analyzed in [12, 21, 22]. While [22] and [21] analyze the performance of the SQ(d) policy in light and heavy traffic regimes for finite system sizes, respectively, [12] considers its performance in the mean-field regime. It has been shown that the SQ(d) scheme suffers from a reduced stability region in heterogeneous systems due to infrequent sampling of faster servers. Subsequent works [11, 23] have studied variations of the SQ(d) scheme, aimed at improving its performance in heterogeneous systems while retaining the maximal stability region. Heterogeneity-aware load balancing (HALO) schemes such Random, Round-Robin (RR), and SQ(d) have been analyzed in [24] for heterogeneous processor sharing systems. In particular, the optimal load splitting has been obtained for the Random scheme and then used for the other schemes. It has been shown that these schemes result in good performance for all system sizes. The JIQ scheme has been analyzed in the heterogeneous setting by Stolyar [25]; it has been shown that the average waiting time of jobs under the JIQ scheme approaches to zero in the fluid limit. From this result, it follows intuitively that a scheme which further distinguishes between idle servers by their speeds can only result in a smaller processing time of jobs. However, to establish this formally, it is necessary to construct a coupling that maintains the dominance of speed-aware schemes over speed-unaware schemes for sufficiently large N. Constructing such coupling in turn requires establishing the fluid limit for speed-aware schemes, which brings us back to the problem considered in this paper.

Some recent works, e.g., [26, 27], study load balancing schemes for systems where jobs are constrained to be served only by specific subsets of servers. In these works, the focus is on finding conditions on the compatibility constraints such that the performance of classical load balancing algorithms such as JSQ, JIQ and SQ(d) remain asymptotically the same as in systems without compatibility constraints. In the work by Weng *et al.* [27], a scheme similar to the SA-JSQ scheme has been considered for constrained heterogeneous systems with finite buffer sizes. They prove the asymptotic optimality of this scheme using Lyapunov drifts methods. Although the scheme is technically the same as the SA-JSQ scheme for constrained systems, their analysis crucially relies on the assumption that the queues have finite buffer sizes. Hence, in their setting tightness and stability results follow immediately. In contrast, in the setting considered in this paper, the queues have infinite buffer sizes and proving optimality in this setting requires showing the tightness of the stationary distributions. As discussed before, this is one of the key difficulties in our model and we employ a new technique to establish tightness. Furthermore, the drift technique, applicable to finite-buffer systems, is difficult to generalise to our setting. Instead, to prove asymptotic optimality, we use martingale methods outlined in [28].

In this paper, the drift of the fluid-scaled Markov chain depends on the un-scaled state of the system through an indicator function. This dependence ultimately gives rise to a time-scale separation in the fluid limit. A similar drift structure involving indicator functions was studied in [29]. Here, the challenge is to tackle the discontinuity of the drift at the boundaries of the state-space. It has been shown in [29] that in such cases the finite system can be viewed as a stochastic approximation of a differential inclusion. In [30], the author study optimal scheduling schemes for users competing for a common resource. They show that a tie-breaking rule similar to JSQ, where users with highest departure probability are preferred, leads to asymptotic optimality.

1.3 General Notations

We use the following notations throughout the paper. The sets $\mathbb{R}, \mathbb{N}, \mathbb{Z}_+$ denote the set of real numbers, natural numbers, and non-negative integers, respectively. We also use the set $\bar{\mathbb{Z}}_+$ to denote the extended set $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+ \bigcup \{\infty\}$. For $x, y \in \mathbb{R}$, we use $x \wedge y$, $x \vee y$, and $(x)_+$ to denote $\max(x, y)$, $\min(x, y)$, and $\max(x, 0)$, respectively. For any $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For any complete separable metric space E, we denote $D_E[0, \infty)$ to be the set of all *cadlag* functions from $[0, \infty)$ to E endowed with the Skorohod topology. Moreover, the notation $\mathscr{B}(E)$ is used to denote the Borel sigma algebra generated by the set E. The notation $\Rightarrow$ is used for weak convergence. We use $\mathbb{1}(A)$ to denote indicator function for set A.

1.4 Organisation

The rest of the paper is organized as follows. In Section 2, we introduce the system model and define the SA-JSQ policy. In Section 3, we obtain a lower bound on the mean response time of jobs that holds for any scheme in the heterogeneous system by comparing the system with a similar system having a central queue. In Section 4, we state our main results for the SA-JSQ policy. We prove the stability of the SA-JSQ scheme in Section 5 and the uniform bounds on the tails of the stationary queue lengths distribution in Section 6. The monotonicity property for SA-JSQ for a finite N system is shown in Section 7. In Section 8, we prove the process convergence result for SA-JSQ and characterise its fixed point in

Section 9. The proof of resource pooled optimality is given in Section 10. Numerical studies comparing different schemes under heterogeneous setting is given in Section 11. Finally, we conclude the paper in Section 12.

2 System Model

We consider a system $\mathcal{M}_N$ consisting of N parallel servers, each with its own queue of infinite buffer size. The servers are assumed to be *heterogeneous*, i.e., they have different service rates. Specifically, we assume that there are M different server *types* or *pools*. Each type $j \in [M]$ server has a service rate of μ_j. The proportion of type $j \in [M]$ servers in the system is assumed to be fixed at $\gamma_j \in [0,1]$ with $\sum_{j\in[M]} \gamma_j = 1$. We further assume without loss of generality that $\mu_1 > \mu_2 > \ldots > \mu_M$ and $\sum_{j\in[M]} \gamma_j \mu_j = 1$ (normalised system capacity is unity). For simplicity of exposition we also assume that there exists $N \in \mathbb{N}$ such that $N\gamma_j \in \mathbb{N}$ for all $j \in [M]$.[1]

Jobs are assumed to arrive at the system according to a Poisson process with a rate $N\lambda$. Each job requires a random amount of work, independent and exponentially distributed with mean 1. The inter-arrival and service times are assumed to be independent of each other. Upon arrival, each job is assigned to a server where it either immediately receives service (if the server is idle at that instant) or joins the corresponding queue to be served later. The queues are served according to the First-Come-First-Server (FCFS) scheduling discipline. For this system, the term *queue length* will refer to the total number of jobs in a queue including the current job in service.

Our main interest is to find a job assignment policy that minimises the steady state mean response time of jobs in the system. To this end, we shall analyse a modified version of the classical Join-the-Shortest-Queue (JSQ) policy which is known to be optimal for homogeneous systems [2]. We shall refer to the modified policy as the *Speed-Aware JSQ* or the *SA-JSQ* policy. It is defined as follows

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

Note that unlike the classical JSQ policy, which breaks ties uniformly at random, the SA-JSQ policy breaks ties among servers of different types by choosing the server type with the maximum service rate. Furthermore, unlike the classical JSQ policy, the SA-JSQ policy is not optimal for the heterogeneous system for finite N: numerical simulations shown in Figure 2 show that a scheme which assigns jobs to servers based on the shortest expected delay (SED) (i.e., queue length divided by the service rate) performs marginally better than the SA-JSQ policy for small system sizes. However, as we show later in the paper, the gap

[1]Our asymptotic results do not depend on this assumption. However, the results for finite systems need to be modified slightly if this is not the case.

vanishes as $N \to \infty$. More specifically, we show that in the limit as $N \to \infty$ there is no better scheme than SA-JSQ for heterogeneous systems.

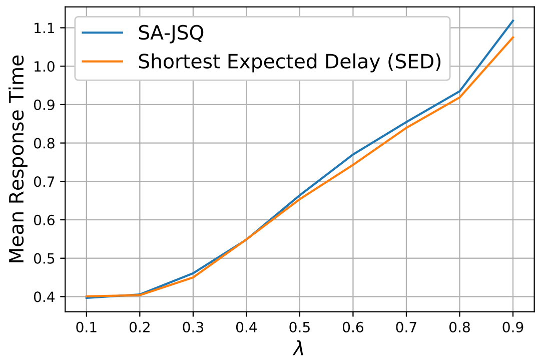


Figure 2: Comparison of mean response time of jobs between SED and SA-JSQ schemes under heterogeneous system with $N = 50$. We choose $\gamma_1 = 1 - \gamma_2 = 1/5$, and $\mu_1 = 4\mu_2 = 20/8$.

3 Lower Bound on the Mean Response Time

A key ingredient in establishing the optimality of the SA-JSQ policy is finding a uniform lower bound on the steady state mean response time of jobs in system $\mathcal{M}_N$ for all N under any stationary job assignment policy Π. We later show that this lower bound is achieved by the SA-JSQ policy as $N \to \infty$.

To obtain this lower bound, we show that the total number of jobs in system $\mathcal{M}_N$ under any stationary policy Π stochastically dominates the total number of jobs in a similar heterogeneous system $\mathcal{M}'_N$ working under a specific policy, referred to as the *Join-the-Fastest-Free-Server (JFFS)* policy.

Description of system $\mathcal{M}'_N$: In system $\mathcal{M}'_N$, the servers remain the same as in system $\mathcal{M}_N$, i.e., there are N servers in total with $N\gamma_j$ of them having service rate μ_j for $j \in [M]$. However, unlike system $\mathcal{M}_N$, we have a central queue for the whole system. Upon arrival, a job joins the central queue if all the servers are busy. Otherwise, the job is assigned to an idle server. Here, the central queue consists of only those jobs that are waiting for service and it excludes the jobs already in service.

JFFS policy: If multiple servers are idle and the queue is non-empty, a decision needs to be made on which server the next job should be assigned to. To this end, we consider the following job assignment policy referred to as the *Join-the-Fastest-Free-Server (JFFS)* policy.

Definition 2. Under the JFFS policy, if idle servers are available and there are jobs waiting in the central queue, then the head-of-the-line (HOL) job is assigned to the idle server with the highest service rate. Ties within the same server type are broken uniformly at random.

Hence, unlike system $\mathcal{M}_N$ where each server only serves its own queue, in system $\mathcal{M}'_N$, the central queue is served by all the servers. This system is, therefore, referred to as the *resource pooled system* [31].

The evolution of the resource pooled system $\mathcal{M}'_N$ under the JFFS policy can be described by the Markov chain $Z^{(N)} = (Z^{(N)}(t), t \geq 0)$, where $Z^{(N)}(t) \in \mathbb{Z}_+$ denotes the total number of jobs in system $\mathcal{M}'_N$ at time t under the JFFS policy. Clearly, under the JFFS policy, the number of busy type j servers at time $t \geq 0$ is given by $(Z^{(N)}(t) - \sum_{i=1}^{j-1} N\gamma_i)_+ \wedge N\gamma_j$. Hence, the transition rate $q^{Z^{(N)}}(k,l)$ of $Z^{(N)}$ from state k to state l is given by

$$q^{Z^{(N)}}(k,l) = \begin{cases} N\lambda, & \text{if } l = k+1, \\ \sum_{j\in[M]} \mu_j((k - \sum_{i=1}^{j-1} N\gamma_i)_+ \wedge N\gamma_j), & \text{if } l = k-1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The following two propositions characterise the stationary behaviour of the chain $Z^{(N)}$ for finite N and for $N \to \infty$, respectively.

Proposition 3. *For $\lambda < 1$, the Markov chain $Z^{(N)}$ is positive recurrent. Furthermore, if $Z^{(N)}(\infty)$ denotes the stationary number of jobs in the system, then*

$$\sup_N \frac{\mathbb{E}\left[Z^{(N)}(\infty)\right]}{N} \leq \lambda + \frac{\lambda}{1-\lambda}. \tag{2}$$

Proposition 4. *If $\lambda < 1$, then for system $\mathcal{M}'_N$ operating under the JFFS policy we have as $N \to \infty$*

$$\frac{Z^{(N)}(\infty)}{N} \Rightarrow z^* \triangleq \max_{j\in[M]} \left(\sum_{i=1}^{j-1} \gamma_i + \frac{\lambda - \sum_{i=1}^{j-1} \mu_i\gamma_i}{\mu_j} \right) \tag{3}$$

Proposition 3 implies that the resource pooled system is stable for $\lambda < 1$. It also provides a uniform upper bound on the steady-state expected (scaled) number of jobs in the system for all N. These results are utilised to prove Proposition 4 which shows that the steady-state (scaled) number of jobs in the resource pooled system converges weakly to z^* as $N \to \infty$, where z^* is as defined in (3).

In (3), the expression within the brackets on the RHS is maximised for $j = j^* \in [M]$ iff $\lambda \in \left[\sum_{i=1}^{j^*-1} \gamma_i\mu_i, \sum_{i=1}^{j^*} \gamma_i\mu_i\right)$. Note that the existence of such a j^* is guaranteed because $\lambda < 1 = \sum_{j\in[M]} \gamma_j\mu_j$. For j^* as defined above, it is easy to see that the fraction of busy type j servers is one for all $j \in [j^*-1]$ and zero for all $j \geq j^*+1$. Furthermore, in pool j^*, the fraction of busy servers is given by $\frac{\left(\lambda - \sum_{i=1}^{j^*-1} \gamma_i\mu_i\right)_+}{\gamma_{j^*}\mu_{j^*}}$. Since $z^* < 1$, the central queue in the limiting system is empty in the steady-state.

To show that the system $\mathcal{M}_N$ operating under any stationary policy Π stochastically dominates the system $\mathcal{M}'_N$ operating under the JFFS policy, we describe the state of the system $\mathcal{M}_N$ under a stationary policy Π by the Markov chain $\mathbf{X}^{(N,\Pi)} = (X_{i,j}^{(N,\Pi)}(t), t \geq 0)$, where $X_{i,j}^{(N,\Pi)}(t)$ denotes the number of type j servers with at least i jobs. Let $R^{(N,\Pi)}(t) = \sum_{i,j} X_{i,j}^{(N,\Pi)}(t)$ denote the total number of jobs in system $\mathcal{M}_N$ at time $t \geq 0$ under policy Π. We have the following stochastic comparison result.

Theorem 5. *For any stationary policy* Π *if* $Z^{(N)}(0) \leq R^{(N,\Pi)}(0)$*, then the processes* $Z^{(N)}$ *and* $\mathbf{X}^{(N,\Pi)}$ *can be constructed on the same probability space such that* $Z^{(N)}(t) \leq R^{(N,\Pi)}(t)$ *for all* $t \geq 0$*, i.e.,* $Z^{(N)}(t) \leq_{st} R^{(N,\Pi)}(t)$*, where* $\leq_{st}$ *implies stochastic dominance.*

The above theorem implies that if both $Z^{(N)}$ and $R^{(N,\Pi)}$ are ergodic chains, then $Z^{(N)}(\infty) \leq_{st} R^{(N,\Pi)}(\infty)$, where $Z^{(N)}(\infty) = \lim_{t\to\infty} Z^{(N)}(t)$ and $R^{(N,\Pi)}(\infty) = \lim_{t\to\infty} R^{(N,\Pi)}(t)$ denote the stationary number of jobs in $\mathcal{M}'_N$ and $\mathcal{M}_N$, respectively. This further implies that $\mathbb{E}\left[Z^{(N)}(\infty)\right] \leq \mathbb{E}\left[R^{(N,\Pi)}(\infty)\right]$ which by Little's law gives the desired lower bound on the stationary mean response time of jobs in $\mathcal{M}_N$ under any policy Π.

By Little's law, the steady-state mean response time, $\bar{T}'_N$, of jobs in $\mathcal{M}'_N$ under the JFFS policy is given by $\bar{T}'_N = \mathbb{E}\left[Z^{(N)}(\infty)\right]/N\lambda$. Hence, Proposition 4 implies $\lim_{N\to\infty} \bar{T}'_N = \frac{z^*}{\lambda}$. Combining the above result with the stochastic comparison result in Theorem 5, we can conclude the following.

Corollary 6. *The steady-state mean response time,* $\bar{T}_{N,\Pi}$*, of jobs in system* $\mathcal{M}_N$ *under any stationary policy* Π *satisfies*

$$\liminf_{N\to\infty} \bar{T}_{N,\Pi} \geq \frac{z^*}{\lambda} \tag{4}$$

In subsequent sections, we establish that the above lower bound is achieved with equality when SA-JSQ is employed as the job assignment policy in $\mathcal{M}_N$. This will imply the asymptotic optimality of the SA-JSQ policy for $\mathcal{M}_N$.

4 Analysis setup and main results

In this section, we introduce the key ingredients and notations for our analysis of the SA-JSQ policy. In particular, we discuss the different state-descriptors used in the analysis and discuss some key properties of the space they belong to. We also state our main results for the SA-JSQ policy and discuss their implications.

The state of the system $\mathcal{M}_N$ at any time $t \geq 0$ under the SA-JSQ policy can be described in two different ways. These are as defined below:

1. *Queue-length descriptor*: We define the queue-length vector at time $t \geq 0$ as

$$\mathbf{Q}^{(N)}(t) = (Q_{k,j}^{(N)}(t), k \in [N\gamma_j], j \in [M]),$$

where $Q_{k,j}^{(N)}(t)$ denotes the queue length of the k^{th} server of type j at time t.

2. *Empirical measure descriptor*: We define the empirical tail measure on the queue lengths at time t as
$$\mathbf{x}^{(N)}(t) = (x_{i,j}^{(N)}(t), i \geq 1, j \in [M]),$$
where $x_{i,j}^{(N)}(t)$ denotes the fraction of type j servers with at least i jobs at time t. For completeness, we set $x_{0,j}^{(N)}(t) = 1$ for all $j \in [M]$ and all $t \geq 0$.

It follows from the Poisson arrival and exponential job size assumptions that both processes $\mathbf{Q}^{(N)} = (\mathbf{Q}^{(N)}(t), t \geq 0)$ and $\mathbf{x}^{(N)} = (\mathbf{x}^{(N)}(t), t \geq 0)$ are Markov. It is possible to switch from first descriptor to the second by noting the following

$$x_{i,j}^{(N)}(t) = \frac{1}{N\gamma_j} \sum_{k \in [N\gamma_j]} \mathbb{1}\Big(Q_{k,j}^{(N)}(t) \geq i\Big), \ \forall i \geq 1, \ j \in [M]. \tag{5}$$

We use both descriptors above to state and prove our results.

One important observation to make is that the SA-JSQ policy does not distinguish between servers of the same type. Hence, if the distribution ν of the initial state of the system is *exchangeable* within each server type, i.e., if for each j we have $\mathbb{P}_\nu(Q_{k_l,j}(0) = r_l, l \in [r]) = \mathbb{P}_\nu(Q_{\sigma(k_l),j}(0) = r_l, l \in [r])$ for all permutations σ of the numbers $k_1, k_2, \ldots, k_r \in [N\gamma_j]$, then for any time $t \geq 0$, we must have $\mathbb{P}_\nu(Q_{k_l,j}(t) = r_l, l \in [r]) = \mathbb{P}_\nu(Q_{\sigma(k_l),j}(t) = r_l, l \in [r])$, where $\mathbb{P}_\nu$ is the probability measure conditional on the initial state being distributed according to ν. This property is crucial in connecting the distribution of $\mathbf{Q}^{(N)}$ to that of $\mathbf{x}^{(N)}$. In particular, by taking expectation on both sides of (5) we have

$$\mathbb{E}_\nu[x_{i,j}^{(N)}(t)] = \frac{1}{N\gamma_j} \sum_{k \in [N\gamma_j]} \mathbb{P}_\nu(Q_{k,j}(t) \geq i) = \mathbb{P}_\nu(Q_{k,j}(t) \geq i), \tag{6}$$

where in the second equality we employ the exhangeability property discussed above. Clearly, the process $\mathbf{Q}^{(N)}$ takes values in $\mathbb{Z}_+^N$ and the process $\mathbf{x}^{(N)}$ takes values in the space $S^{(N)}$ defined as

$$S^{(N)} \triangleq \{\mathbf{s} = (s_{i,j}) : N\gamma_j s_{i,j} \in \mathbb{Z}_+, 1 \geq s_{i,j} \geq s_{i+1,j} \geq 0 \ \forall i \geq 1, j \in [M]\}. \tag{7}$$

Note that for finite N, the space $S^{(N)}$ is countable since each $x_{i,j}^{(N)}$ can only take finitely many values. We further define the space S as follows

$$S \triangleq \{\mathbf{s} = (s_{i,j}) : 1 \geq s_{i,j} \geq s_{i+1,j} \geq 0, \forall i \geq 1, j \in [M], \|\mathbf{s}\|_1 < \infty\}, \tag{8}$$

where the ℓ_1-norm, denoted by $\|\cdot\|_1$, is defined as $\|\mathbf{s}\|_1 \triangleq \max_{j \in [M]} \sum_{i \geq 1} |s_{i,j}|$ for any $\mathbf{s} \in S$. It is easy to verify that the space S is complete and separable under the ℓ_1-norm. Furthermore, if the system's state $\mathbf{x}^{(N)}$ belongs to $S^{(N)} \cap S$, then there are finitely many jobs in the system. Starting with a state in $S^{(N)} \cap S$ we can ensure that chain $\mathbf{x}^{(N)}$ remains in $S \cap S^{(N)}$ for all $t \geq 0$ only if the process $\mathbf{x}^{(N)}$, or equivalently the process $\mathbf{Q}^{(N)}$, is positive recurrent. Our first main result states that this is indeed the case when $\lambda < 1$.

Theorem 7. *The process $\mathbf{Q}^{(N)}$ or equivalently $\mathbf{x}^{(N)}$ is positive recurrent for each $\lambda < 1$ and each N.*

The theorem above implies that for $\lambda < 1$ the stationary distributions of $\mathbf{Q}^{(N)}$ and $\mathbf{x}^{(N)}$ exist and they are unique. Let $\mathbf{Q}^{(N)}(\infty) = \lim_{t\to\infty} \mathbf{Q}^{(N)}(t)$ (resp. $\mathbf{x}^{(N)}(\infty) = \lim_{t\to\infty} \mathbf{x}^{(N)}(t)$) denote the state of the system distributed according to the stationary distribution of $\mathbf{Q}^{(N)}$ (resp. $\mathbf{x}^{(N)}$). The stability of the system guarantees that the steady-state expected number of jobs in the system is finite, i.e., $\mathbf{x}^{(N)}(\infty)$ belongs to S almost surely. However, to establish the asymptotic optimality of the SA-JSQ policy, we require the stronger result that the sequence $(\mathbf{x}^{(N)}(\infty))_N$ of stationary states indexed by the system size is tight in S. Here, it is important to observe that S is not a compact space. The relatively compact subsets of S and the tightness criterion for sequences in S are stated in E. The tightness criterion essentially requires the expected tail sums $\mathbb{E}[\sum_{j\in[M]}\sum_{i\geq l} x_{i,j}^{(N)}(\infty)]$ approach to zero as $l \to \infty$ uniformly in N. To show that $(\mathbf{x}^{(N)}(\infty))_N$ satisfies this criterion, we need the following important result which states that the stationary tail probabilities of the queue lengths decay exponentially and the decay rate is uniform in N.

Theorem 8. *If $\lambda < 1$, then under the SA-JSQ scheme, for each $j \in [M]$, each $k \in [N\gamma_j]$, and each $l \geq 1$ the following bound holds for all $\theta \in [0, -\log\lambda)$*

$$\sup_N \mathbb{P}(Q_{k,j}^{(N)}(\infty) \geq l) \leq C_j(\lambda,\theta)e^{-l\theta}, \tag{9}$$

where $C_j(\lambda,\theta) = (1-\lambda)/(\mu_j\gamma_j(1-\lambda e^{\theta})) > 0$.

Our next main result establishes a key monotonicity property of the system for finite N. For two states $\mathbf{q}$ and $\tilde{\mathbf{q}}$ in $\mathbb{Z}_+^N$ we say $\mathbf{q} \leq \tilde{\mathbf{q}}$ if $q_{k,j} \leq \tilde{q}_{k,j}$ for each $k \in [N\gamma_j], j \in [M]$. The inequality $\mathbf{s} \leq \tilde{\mathbf{s}}$ is similarly defined for $\mathbf{s}, \tilde{\mathbf{s}} \in S$. We have the following result.

Theorem 9. *Consider two systems with initial states $\mathbf{Q}^{(N)}(0)$ and $\tilde{\mathbf{Q}}^{(N)}(0)$ (resp. $\mathbf{x}^N(0)$ and $\tilde{\mathbf{x}}^N(0)$) satisfying $\mathbf{Q}^{(N)}(0) \leq \tilde{\mathbf{Q}}^{(N)}(0)$ (resp. $\mathbf{x}^N(0) \leq \tilde{\mathbf{x}}^N(0)$). Let $\mathbf{Q}^{(N)}$ and $\tilde{\mathbf{Q}}^{(N)}$ (resp. $\mathbf{x}^N$ and $\tilde{\mathbf{x}}^N$) denote the corresponding processes describing the two systems under the SA-JSQ policy. Then, the processes $\mathbf{Q}^{(N)}$ and $\tilde{\mathbf{Q}}^{(N)}$ (resp. $\mathbf{x}^N$ and $\tilde{\mathbf{x}}^N$) can be constructed on the same probability space such that $\mathbf{Q}^{(N)}(t) \leq \tilde{\mathbf{Q}}^{(N)}(t)$ (resp. $\mathbf{x}^N(t) \leq \tilde{\mathbf{x}}^N(t)$) for all $t \geq 0$.*

The monotonicity property stated above implies that if two systems, both working under the SA-JSQ policy, start at two different initial states such that the queue lengths in the first system are dominated by the corresponding queue lengths in the second system, then this dominance is maintained for all $t \geq 0$. This property turns out to be essential in establishing the global asymptotic stability of the fluid limit process.

Our next set of results characterise the asymptotic properties of the process $\mathbf{x}^{(N)}$ as $N \to \infty$. The first result states that the sequence $(\mathbf{x}^{(N)})_N$ of processes indexed by N converges weakly to a deterministic process $\mathbf{x} = (\mathbf{x}(t), t \geq 0)$ defined on S. To describe the limiting process, we define $l_j(\mathbf{s}) = \min\{i : s_{i+1,j} < 1\}$ for any state $\mathbf{s} \in S$ to be the minimum queue length in pool $j \in [M]$ in state $\mathbf{s}$.

Theorem 10. *(Process Convergence): Assume* $\mathbf{x}^{(N)}(0) \in S \cap S^{(N)}$ *for each* N *and* $\mathbf{x}^{(N)}(0) \Rightarrow \mathbf{x}(0) \in S$ *as* $N \to \infty$*. Then, the sequence* $(\mathbf{x}^{(N)})_{N \geq 1}$ *is relatively compact in* $D_S[0, \infty)$ *and any limit* $\mathbf{x} = (\mathbf{x}(t) = (x_{i,j}(t), i \geq 1, j \in [M]), t \geq 0)$ *of a convergent sub-sequence of* $(\mathbf{x}^{(N)})_{N \geq 1}$ *satisfies the following set of equations for all* $t \geq 0$*,* $i \geq 1$ *and* $j \in [M]$

$$x_{i,j}(t) = x_{i,j}(0) + \frac{\lambda}{\gamma_j} \int_0^t p_{i-1,j}(\mathbf{x}(u)) du - \int_0^t \mu_j (x_{i,j}(u) - x_{i+1,j}(u)) du, \tag{10}$$

where $p_{i-1,j}(\mathbf{s}) \in [0, 1]$ *is uniquely determined for each state* $\mathbf{s} \in S$*. Furthermore,* $\mathbf{p}(\mathbf{s}) = (p_{i-1,j}(\mathbf{s}), i \geq 1, j \in [M])$ *for each* $\mathbf{s} \in S$ *satisfies the following properties*

P1. $\sum_{j \in [M]} \sum_{i \geq 1} p_{i-1,j}(\mathbf{s}) = 1$ *for all* $\mathbf{s} \in S$*,*

P2. $p_{i-1,j}(\mathbf{s}) = 0$ *for all* $i \geq l_j(\mathbf{s}) + 2$ *and for all* $j \in [M]$*,*

P3. *If* $l_j(\mathbf{s}) > 0$ *for some* $j \in [M]$*, then* $p_{i-1,j}(\mathbf{s}) = 0$ *for all* $1 \leq i \leq l_j(\mathbf{s}) - 1$*.*

P4. *If* $l_1(\mathbf{s}) = 0$*, then* $p_{0,1}(\mathbf{s}) = 1$*.*

P5. *For some* $j \in \{2, \ldots, M\}$*. If* $l_k(\mathbf{s}) \geq 1$ *for all* $k \in [j-1]$ *and* $l_j(\mathbf{s}) = 0$*, then* $p_{0,j}(\mathbf{s}) = 1$*.*

In the theorem above, $p_{i-1,j}(\mathbf{s})$ can be interpreted as the limiting probability of an incoming arrival being assigned to a type j server with queue length $i-1$ when the system is in state $\mathbf{s} \in S$. With this interpretation, the properties of $p_{i,j}(\mathbf{s})$ listed above follow intuitively from the assignment rule under the SA-JSQ policy. Indeed, properties P4 and P5 state that if in state $\mathbf{s} \in S$ pool $j \in [M]$ is the pool with the highest speed containing idle servers, then with probability 1, incoming jobs are assigned to idle servers in pool j in state $\mathbf{s}$. Similarly, properties P2 and P3 imply that for any state $\mathbf{s} \in S$ and any pool $j \in [M]$ with minimum queue-length $l_j(\mathbf{s})$, jobs can only be assigned to queues with lengths $l_j(\mathbf{s}) - 1$ and $l_j(\mathbf{s})$ in the limiting system.

The exact expressions for $p_{i,j}(\mathbf{s})$ are complicated as they depend on the stationary probabilities of a multi-dimensional Markov chain (we explain this more in Section 8). In Theorem 10, we only list the properties essential to characterise the fixed point $\mathbf{x}^* = (x^*_{i,j})$ of the fluid limit $\mathbf{x}$. In our final result stated below, we show that the fixed point $\mathbf{x}^*$ is unique and globally attractive.

Theorem 11. *(i) For* $\lambda < 1$*, there exists a unique fixed point* $\mathbf{x}^* = (x^*_{i,j}, i \geq 1, j \in [M]) \in S$ *of the fluid limit* $\mathbf{x}$ *described by* (10)*, i.e., if* $\mathbf{x}(0) = \mathbf{x}^*$ *then* $\mathbf{x}(t) = \mathbf{x}^*$ *for all* $t \geq 0$*. Furthermore, the fixed point* $\mathbf{x}^*$ *is given by*

$$x^*_{1,j} = \left(1 \wedge \frac{(\lambda - \sum_{k=1}^{j-1} \mu_k \gamma_k)_+}{\mu_j \gamma_j}\right) \forall j \in [M], \quad x^*_{i,j} = 0 \quad \forall\ i \geq 2, j \in [M]. \tag{11}$$

(ii) (Global Stability): If $\lambda < 1$*, then for any solution* $\mathbf{x}$ *of* (10) *with* $\mathbf{x}(0) \in S$ *converges to* $\mathbf{x}^*$ *component-wise, i.e.,* $x_{i,j}(t) \to x^*_{i,j}$ *as* $t \to \infty$ *for all* $i \geq 1$ *and for all* $j \in [M]$*.*

(iii) (Interchange of Limits): Let $\lambda < 1$. Then, the sequence $(\mathbf{x}^{(N)}(\infty))_N$ converges weakly to $\mathbf{x}^*$, i.e., $\mathbf{x}^{(N)}(\infty) \Rightarrow \mathbf{x}^*$ as $N \to \infty$.

Note that the last statement of the theorem implies that the sequence of stationary distributions indexed by the system size concentrates on the point $\mathbf{x}^*$. Furthermore, the first statement of the theorem implies that in state $\mathbf{x}^* \in S$ the fraction of servers with two or more jobs is zero. This is similar to the classical JSQ result except that in this case if $\lambda \in [\sum_{k=1}^{j-1} \mu_k \gamma_k, \sum_{k=1}^{j} \mu_k \gamma_k)$, then all servers in pools $k \in [j-1]$ have exactly one job and in pool j a fraction $(\lambda - \sum_{k=1}^{j-1} \mu_k \gamma_k)/\mu_j \gamma_j$ of servers have exactly one job; all remaining servers are idle. Thus, the total (scaled) number of jobs in state x^* is equal to z^* as defined by (3). Thus, by Little's law, the mean response time of jobs $\bar{T}_N$ under the SA-JSQ policy converges to z^*/λ as $N \to \infty$, which by Corollary 6 implies the asymptotic optimality of the SA-JSQ policy.

5 Stability

To show that the process $\mathbf{Q}^{(N)}$ is positive recurrent for all $\lambda < 1$, we use an appropriately defined Lyapunov function and show that its drift along any trajectory of $\mathbf{Q}^{(N)}$ is negative outside a compact subset of the state space. For any function $f : \mathbb{Z}_+^N \to \mathbb{R}$, defined on the state space of the process $\mathbf{Q}^{(N)}$, the drift evaluated at a state $\mathbf{Q} \in \mathbb{Z}_+^N$ is given by

$$G_{\mathbf{Q}^{(N)}} f(\mathbf{Q}) = \sum_{j \in [M]} \sum_{k \in [N\gamma_j]} [r_{k,j}^{+,N}(\mathbf{Q})(f(\mathbf{Q} + \mathbf{e}_{k,j}^{(N)}) - f(\mathbf{Q})) + r_{k,j}^{-,N}(\mathbf{Q})(f(\mathbf{Q} - \mathbf{e}_{k,j}^{(N)}) - f(\mathbf{Q}))], \tag{12}$$

where $G_{\mathbf{Q}^{(N)}}$ is the generator of $\mathbf{Q}^{(N)}$; $\mathbf{e}_{k,j}^{(N)}$ denotes the N-dimensional unit vector with one in the $(k,j)^{\text{th}}$ position; $r_{k,j}^{\pm,N}(\mathbf{Q})$ are the transition rates from the state $\mathbf{Q}$ to the states $\mathbf{Q} \pm \mathbf{e}_{k,j}^{(N)}$. Intuitively, the drift, defined above, represents the expected infinitesimal rate at which $f(\mathbf{Q}^{(N)}(t))$ changes when $\mathbf{Q}^{(N)}(t) = \mathbf{Q}$. Under the SA-JSQ policy, for each state $\mathbf{Q} \in \mathbb{Z}_+^N$ and each $k \in [N\gamma_j], j \in [M]$ we have

$$r_{k,j}^{+,N}(\mathbf{Q}) = \begin{cases} \frac{N\lambda}{|I_{\min,j}(\mathbf{Q})|}, & \text{if } j = j^\dagger(\mathbf{Q}) \text{ and } k \in I_{\min,j}(\mathbf{Q}) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

$$r_{k,j}^{-,N}(\mathbf{Q}) = \mu_j \mathbb{1}(Q_{k,j} > 0), \tag{14}$$

where $I_{\min,j}(\mathbf{Q})$ denotes the set of servers with the minimum queue length in pool j in state $\mathbf{Q}$ and $j^\dagger(\mathbf{Q})$ denotes the fastest pool that contains a server with the minimum queue length in state $\mathbf{Q}$. The upward transition rate $r_{k,j}^{+,N}(\mathbf{Q})$ is obtained by multiplying the total arrival rate with the probability that the k^{th} server in the j^{th} pool receives an arrival. Similarly, the downward transition rate $r_{k,j}^{-,N}(\mathbf{Q})$ is simply the service rate of the k^{th} server in the j^{th} pool if the server is busy.

Proof of Theorem 7: To prove Theorem 7, we compute the drift of the Lyapunov function $\Phi : \mathbb{Z}_+^N \to [0,\infty)$ defined as follows

$$\Phi(\mathbf{Q}) = \sum_{j\in[M]} \sum_{k\in[N\gamma_j]} Q_{k,j}^2. \tag{15}$$

From (12) we have that for any $\mathbf{Q} \in \mathbb{Z}_+^N$

$$\begin{aligned} G_{\mathbf{Q}^{(N)}}\Phi(\mathbf{Q}) &= \sum_{j\in[M]} \sum_{k\in[N\gamma_j]} r_{k,j}^{+,N}(\mathbf{Q})(2Q_{k,j}+1) + r_{k,j}^{-,N}(\mathbf{Q})(-2Q_{k,j}+1) \\ &= 2\sum_{j\in[M]} \sum_{k\in[N\gamma_j]} r_{k,j}^{+,N}(\mathbf{Q})Q_{k,j} + N\lambda - 2\sum_{j\in[M]} \sum_{k\in[N\gamma_j]} r_{k,j}^{-,N}(\mathbf{Q})Q_{k,j} \\ &\quad + \sum_{j\in[M]} \mu_j B_j(\mathbf{Q}), \end{aligned} \tag{16}$$

where $B_j(\mathbf{Q})$ denotes the number of busy servers in pool j in state $\mathbf{Q}$. In the second equality, we have used the facts that $\sum_{j,k} r_{k,j}^{+,N}(\mathbf{Q}) = N\lambda$ and $\sum_k r_{k,j}^{-,N}(\mathbf{Q}) = \mu_j B_j(\mathbf{Q})$ which follow easily from (13) and (14), respectively. We further note from (13) and (14) that $\sum_{k,j} r_{k,j}^{+,N}(\mathbf{Q})Q_{k,j} = N\lambda Q_{\min}$ and $\sum_k r_k^{-,N}(\mathbf{Q})Q_{k,j} = \mu_j \sum_k Q_{k,j}$, where $Q_{\min}$ denotes the minimum queue length in state $\mathbf{Q}$. Hence, from (16) we have

$$G_{\mathbf{Q}^{(N)}}\Phi(\mathbf{Q}) = 2N\lambda Q_{\min} - 2\sum_j \mu_j \sum_k Q_{k,j} + N\lambda + \sum_j \mu_j B_j(\mathbf{Q}) \tag{17}$$

$$\leq -2(1-\lambda)\sum_j \mu_j \sum_k Q_{k,j} + N\lambda + N. \tag{18}$$

In the second line, we have used the facts that $\sum_j \mu_j \sum_k Q_{k,j} \geq \sum_j \mu_j \sum_k Q_{\min} = NQ_{\min}$ and $B_j(\mathbf{Q}) \leq N\gamma_j$. Hence, it follows from the above that if $\lambda < 1$, then the drift is strictly negative whenever $\sum_j \mu_j \sum_k Q_{k,j} > \frac{N(1+\lambda)}{1-\lambda}$, and is bounded by $N(1+\lambda)$ otherwise. Thus, using the Foster-Lyapunov criterion for positive recurrence (see, e.g., Proposition D.3 of [32]) we conclude that $\mathbf{Q}^{(N)}$ is positive recurrent. ∎

6 Uniform bounds and tightness

In this section, we first prove Theorem 8 which shows that the stationary queue length distribution has a uniformly decaying tail for all system sizes. We then use this uniform bound to establish the tightness of the sequence $(\mathbf{x}^{(N)}(\infty))_N$ in S.

For homogeneous systems working under the classical JSQ policy, uniform bounds, similar to (9) can be obtained by coupling the system with another similar system working under a 'worse performing' policy such as the *random policy* in which the destination server for each job is chosen uniformly at random from the set of all servers. A coupling similar to the one described in [33] ensures that the total number of jobs in the first system is always smaller than that in the second system. However, in the heterogeneous setting, it is difficult to construct a similar coupling since the arrival of a job to a given pool in one system does not guarantee that the job will join the same pool in the other system.

To overcome this difficulty, we use a completely different approach based on analysing the drift of an exponential Lyapunov function. In particular, we analyse the drift of the Lyapunov function $\Psi_\theta : \mathbb{Z}_+^N \to \mathbb{R}_+$ defined as

$$\Psi_\theta(\mathbf{Q}) \triangleq \sum_{j \in [M]} \sum_{k \in [N\gamma_j]} \exp(\theta Q_{k,j}), \tag{19}$$

for some $\theta > 0$. They key idea we employ to prove Theorem 8 is the following: we show that for some positive values of θ the steady-state expected drift $\mathbb{E}[G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}^{(N)}(\infty))]$ of Ψ_θ is non-negative. From this, we obtain bounds on the weighted sum of moment generating functions of the stationary queue lengths of different pools, i.e., on $\mathbb{E}[\sum_{j\in[M]} \mu_j \gamma_j \exp(\theta Q_{k,j}^{(N)}(\infty))]$ for some positive θ. Using Chernoff bounds, we then obtain the bounds on the tail probabilities as in Theorem 8.

Note that although we would normally expect the steady-state expected drift to satisfy

$$\mathbb{E}[G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}^{(N)}(\infty))] = 0,$$

proving this requires showing $\mathbb{E}[\Psi_\theta(\mathbf{Q}^{(N)}(\infty))] < \infty$ which is only true for some $\theta > 0$ if all moments of the stationary queue lengths exist. However, the stability condition only guarantees the existence of the first moment, but it does not guarantee the existence of higher moments of the queue lengths. Thus, to prove Theorem 8, we use a weaker condition given by Proposition 1 of [34] which states that for any non-negative function $f : \mathbb{Z}_+^N \to \mathbb{R}_+$, if $\sup_{\mathbf{Q}\in\mathbb{Z}_+^N} G_{\mathbf{Q}^{(N)}} f(\mathbf{Q}) < \infty$ then $\mathbb{E}[G_{\mathbf{Q}^{(N)}} f(\mathbf{Q}^{(N)}(\infty))] \geq 0$. However, to use this result we first need to show that the function Ψ_θ satisfies the above condition for some positive θ. This is shown in the following lemma.

Lemma 12. *For Ψ_θ as defined in (19) and $\theta \geq 0$, we have*

$$G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}) \leq (1 - e^{-\theta}) \left[(\lambda e^\theta - 1) \sum_{j\in[M]} \sum_{k\in[N\gamma_j]} \mu_j \exp(\theta Q_{k,j}) + \sum_{j\in[M]} \mu_j I_j(\mathbf{Q}) \right], \tag{20}$$

where $I_j(\mathbf{Q})$ denotes the number of idle servers in pool $j \in [M]$ in state $\mathbf{Q}$. In particular, for all $\theta \in [0, -\log \lambda)$ we have

$$\sup_{\mathbf{Q}\in\mathbb{Z}_+^N} G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}) < \infty, \tag{21}$$

and the steady-state drift of Ψ_θ satisfies

$$\mathbb{E}[G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}^{(N)}(\infty))] \geq 0. \tag{22}$$

Proof. From (12) and (19), we have that for any $\mathbf{Q} \in \mathbb{Z}_+^N$

$$G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q}) = \sum_{j\in[M]}\sum_{k\in[N\gamma_j]}\Big[r_{k,j}^{+,N}(\mathbf{Q})\Big(\exp(\theta Q_{k,j})(e^\theta - 1)\Big) + r_{k,j}^{-,N}(\mathbf{Q})\Big(\exp(\theta Q_{k,j})(e^{-\theta} - 1)\Big)\Big]. \tag{23}$$

First, note that from (13), we can write

$$\begin{aligned}
&\sum_{j\in[M]}\sum_{k\in[N\gamma_{j^\dagger(\mathbf{Q})}]} r_{k,j}^{+,N}(\mathbf{Q})\exp(\theta Q_{k,j}) \\
&= \sum_{k\in[N\gamma_{j^\dagger(\mathbf{Q})}]}\frac{N\lambda\mathbb{1}\Big(k\in I_{\min,j^\dagger(\mathbf{Q})}(\mathbf{Q})\Big)\exp(\theta Q_{k,j^\dagger(\mathbf{Q})})}{|I_{\min,j^\dagger(\mathbf{Q})}(\mathbf{Q})|}, \\
&= N\lambda\sum_{k\in[N\gamma_{j^\dagger(\mathbf{Q})}]}\frac{\mathbb{1}\Big(k\in I_{\min,j^\dagger(\mathbf{Q})}(\mathbf{Q})\Big)\exp(\theta Q_{\min})}{|I_{\min,j^\dagger(\mathbf{Q})}(\mathbf{Q})|}, \\
&= N\lambda\exp(\theta Q_{\min}),
\end{aligned} \tag{24}$$

where $Q_{\min}$ denotes the minimum queue length in state $\mathbf{Q}$ and the equality on the second line follows from the fact that for $k \in I_{\min,j^\dagger(\mathbf{Q})}(\mathbf{Q})$ the queue length at the k^{th} server of type $j^\dagger(\mathbf{Q})$ is $Q_{\min}$. Furthermore, using (14), we can write

$$\begin{aligned}
\sum_{j\in[M]}\sum_{k\in[N\gamma_j]} r_{k,j}^{-,N}(\mathbf{Q})\exp(\theta Q_{k,j}) &= \sum_{j\in[M]}\sum_{k\in[N\gamma_j]}\mu_j\mathbb{1}(Q_{k,j}>0)\exp(\theta Q_{k,j}), \\
&= \sum_{j\in[M]}\mu_j\sum_{k\in[N\gamma_j]}\Big(1-\mathbb{1}(Q_{k,j}=0)\Big)\exp(\theta Q_{k,j}), \\
&= \sum_{j\in[M]}\mu_j\sum_{k\in[N\gamma_j]}\exp(\theta Q_{k,j}) - \sum_j \mu_j I_j(\mathbf{Q}).
\end{aligned} \tag{25}$$

Hence, using (24) and (25), we can write

$$G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q}) = (e^\theta - 1)\Big[N\lambda\exp(\theta Q_{\min}) - \frac{1}{e^\theta}\sum_{j\in[M]}\mu_j\sum_{k\in[N\gamma_j]}\exp(\theta Q_{k,j}) + \frac{1}{e^\theta}\sum_j\mu_j I_j(\mathbf{Q})\Big]. \tag{26}$$

We further note that for any $\theta > 0$ we have

$$\begin{aligned}
\sum_j\mu_j\sum_k\exp(\theta Q_{k,j}) &\geq \sum_j\mu_j\sum_k\exp(\theta Q_{\min}) \\
&= N\exp(\theta Q_{\min})\sum_j\mu_j\gamma_j \\
&= N\exp(\theta Q_{\min}).
\end{aligned}$$

Using this fact in (26), we have that for any $\theta \geq 0$

$$G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}) \leq (1 - e^{-\theta}) \left[\left(\lambda e^\theta - 1 \right) \sum_{j \in [M]} \mu_j \sum_{k \in [N\gamma_j]} \exp(\theta Q_{k,j}) + \sum_{j \in [M]} \mu_j I_j(\mathbf{Q}) \right].$$

This proves the first statement of the lemma. In order to prove the next statement, note that for all $\theta \in [0, -\log \lambda)$, we have $\lambda e^\theta - 1 < 0$. Therefore, from (20) it follows that for all $\theta \in [0, -\log \lambda)$ we have

$$G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}) \leq (1 - e^{-\theta}) \sum_{j \in [M]} \mu_j I_j(\mathbf{Q}) \leq (1 - e^{-\theta}) N, \tag{27}$$

where in the second inequality we have used the fact that $I_j(\mathbf{Q}) \leq N\gamma_j$. Hence, for all $\theta \in [0, -\log \lambda)$ we have $\sup_{\mathbf{Q} \in \mathbb{Z}_+^N} G_{\mathbf{Q}^{(N)}} \Psi_\theta(\mathbf{Q}) < \infty$ The last statement of the lemma now follows from the application of Proposition 1 of [34]. ■

Proof of Theorem 8: We are now equipped to prove Theorem 8 using the result of the lemma above. Taking expectation of (20) with respect to the stationary distribution and applying (22) we obtain

$$\begin{aligned}\left(1 - \lambda e^\theta\right) \mathbb{E}\left[\sum_j \mu_j \sum_k \exp(\theta Q_{k,j}^{(N)}(\infty))\right] &\leq \mathbb{E}\left[\sum_j \mu_j (N\gamma_j - B_j(\mathbf{Q}^{(N)}(\infty)))\right] \\ &= N(1 - \lambda),\end{aligned} \tag{28}$$

where $B_j(\mathbf{Q}) = N\gamma_j - I_j(\mathbf{Q})$ denotes the number of busy servers in pool j in state $\mathbf{Q}$. In the last inequality we have used the fact that due to erogodicity of the process $\mathbf{Q}^{(N)}$, the steady state rate of departure from the system $\mathbb{E}[\sum_j \mu_j B_j(\mathbf{Q})]$ is equal to the arrival rate $N\lambda$. Hence, from (28) we have that for all $\theta \in [0, -\log \lambda)$

$$\begin{aligned}\frac{N(1-\lambda)}{1 - \lambda e^\theta} &\geq \mathbb{E}\left[\sum_{j \in [M]} \mu_j \sum_{k \in [N\gamma_j]} \exp(\theta Q_{k,j}^{(N)}(\infty))\right] \\ &= N\mathbb{E}\left[\sum_{j \in [M]} \mu_j \gamma_j \exp(\theta Q_{k,j}^{(N)}(\infty))\right] \\ &\geq N\mathbb{E}\left[\mu_j \gamma_j \exp(\theta Q_{k,j}^{(N)}(\infty))\right]\end{aligned}$$

where the second equality follows from the exchangeability of the stationary measure. Thus, for each $j \in [M]$ and $\theta \in [0, -\log \lambda)$ we have

$$\mathbb{E}\left[\exp(\theta Q_{k,j}^{(N)}(\infty))\right] \leq \frac{1}{\mu_j \gamma_j} \frac{1 - \lambda}{1 - \lambda e^\theta}. \tag{29}$$

Now, for each positive θ we have

$$\mathbb{P}(Q_{k,j}^{(N)}(\infty) \geq l) = \mathbb{P}(\exp(\theta Q_{k,j}^{(N)}(\infty)) \geq \exp(\theta l)) \leq \frac{\mathbb{E}\left[\exp\left(\theta Q_{k,j}^{(N)}(\infty)\right)\right]}{\exp(\theta l)}.$$

The statement of the theorem now follows by using (29) on the above inequality. ■

Using Theorem 8, we now show that the sequence of stationary states $(\mathbf{x}^{(N)}(\infty))_N$ is tight in the space S under the ℓ_1-norm. According to Prohorov's theorem [35] the tightness of this sequence will imply that the sequence has convergent subsequences with limits in S. We shall show later that all convergent subsequences of $(\mathbf{x}^{(N)}(\infty))_N$ have the same limit, thereby establishing the convergence of the original sequence $(\mathbf{x}^{(N)}(\infty))_N$ to the same limit.

Proposition 13. *The sequence $(\mathbf{x}^{(N)}(\infty))_N$ of stationary states is tight in S under the ℓ_1-norm.*

The necessary and sufficient criterion for tightness of the sequence $(\mathbf{x}^{(N)}(\infty))_N$ in S under the ℓ_1-norm is derived in Lemma 22 of E and is given by

$$\lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(\infty) > \epsilon\Big) = 0, \ \forall \epsilon > 0. \tag{30}$$

The proof of Proposition 13 consists of verifying this condition using the uniform bounds derived in Theorem 8. The formal proof is given in B.

7 Monotonicity

In this section, we prove the monotonicity property stated in Theorem 9. The key idea here is to couple the arrivals and the departures of the two systems such that if the inequality $\mathbf{Q}^{(N)}(t^-) \leq \tilde{\mathbf{Q}}^{(N)}(t^-)$ is satisfied just before the arrival or departure event at time t then it is also satisfied after the event has taken place.

Proof of Theorem 9: We refer to the systems corresponding to the processes $\mathbf{Q}^{(N)}$ and $\tilde{\mathbf{Q}}^{(N)}$ as the *smaller* and *larger* systems, respectively. Furthermore, in both systems, we call the k^{th} server in the j^{th} pool as the server with index (k, j).

To couple the departures, we first generate a sequence of potential departure instants at the points of a Poisson process with rate N. At each potential departure instant, a server index (k, j) is chosen as follows: First, a server type $j \in [M]$ is chosen with probability $\mu_j\gamma_j$ (recall that $\sum_{j\in[M]} \mu_j\gamma_j = 1$). Then, any server $k \in [N\gamma_j]$ within the chosen type j is selected uniformly at random. Once the server index (k, j) is chosen as described above, the server with the index (k, j) is selected for departure in both systems. In each system, an actual departure occurs from the selected server if the selected server is busy; otherwise, no departure occurs from the selected server (this is why the term *potential departure* is used to describe the event). Let (k, j) denote the index of the chosen server and D denote the potential departure instant. Assume that the inequality $Q_{k,j}^{(N)}(D^-) \leq \tilde{Q}_{k,j}^{(N)}(D^-)$ holds just before the departure. Then, we must have $Q_{k,j}^{(N)}(D) \leq \tilde{Q}_{k,j}^{(N)}(D)$. This is because $Q_{k,j}^{(N)}(D^-) > 0$ only if $\tilde{Q}_{k,j}^{(N)}(D^-) > 0$ in which case a departure occurs from both systems.

To couple the arrivals, we generate a common Poisson arrival stream with rate $N\lambda$ for both systems. At each arrival instant A, the job assignment decision is made following the SA-JSQ policy in each system independently of the other system, unless the pool containing the destination server is the same for both systems. If $j^\dagger(\mathbf{Q}^{(N)}(A^-)) = j^\dagger(\tilde{\mathbf{Q}}^{(N)}(A^-)) = j$, then we perform the following steps sequentially: (1) In the smaller system, the destination server is chosen uniformly at random from the servers having the minimum queue length in pool j. Let the index of this chosen server be (k, j). (2) We check if the server having the same index (k, j) in the larger system has the minimum queue length. If so, then this server is chosen as the destination server in the larger system. (3) Otherwise, the destination server in the larger system is chosen uniformly at random (and independently of the smaller system) from the servers having the minimum queue length in pool j in the larger system. Let (k_s, j_s) and (k_l, j_l) be the indices of the destination servers in the smaller and the larger systems, respectively, at an arrival instant A. Let $S = Q_{k_s,j_s}^{(N)}(A^-)$ and $L = \tilde{Q}_{k_l,j_l}^{(N)}(A^-)$. Assume that $\mathbf{Q}^{(N)}(t) \leq \tilde{\mathbf{Q}}^{(N)}(t)$ holds for all $t < A$. Hence, $S \leq L$. It is sufficient to show that the inequality

$$\mathbf{Q}^{(N)}(t) \leq \tilde{\mathbf{Q}}^{(N)}(t) \text{ holds at } t = A. \tag{31}$$

If $(k_s, j_s) = (k_l, j_l)$, then the inequality trivially holds at $t = A$. If $(k_s, j_s) \neq (k_l, j_l)$, then we have the following possibilities:

1. *If $j_s < j_l$*: In the larger system, the incoming arrival joins the server with index (k_l, j_l) and there is no arrival to the server with index (k_l, j_l) in smaller system. Hence, after the arrival we have

$$Q_{k_l,j_l}^{(N)}(A) = Q_{k_l,j_l}^{(N)}(A^-) < \tilde{Q}_{k_l,j_l}^{(N)}(A^-) + 1 = \tilde{Q}_{k_l,j_l}^{(N)}(A).$$

   Thus, for (31) to be violated we must have $Q_{k_s,j_s}^{(N)}(A^-) = \tilde{Q}_{k_s,j_s}^{(N)}(A^-) = S$. But since $S \leq L$, we must have $\tilde{Q}_{k_s,j_s}^{(N)}(A^-) \leq \tilde{Q}_{k_l,j_l}^{(N)}(A^-)$ which contradicts with the fact that the server with index (k_l, j_l) is the destination server in the larger system, i.e., it is the fastest server with the minimum queue length in the larger system, because a faster server of type $j_s (< j_l)$ has smaller queue length. Hence, (31) must hold in this case.

2. *If $j_s = j_l = j, k_s \neq k_l$*: Similarly as before, the inequality $Q_{k_l,j}^{(N)}(A) < \tilde{Q}_{k_l,j}^{(N)}(A)$ holds after the arrival. For violation of (31), we therefore must have $Q_{k_s,j}^{(N)}(A^-) = \tilde{Q}_{k_s,j}^{(N)}(A^-) = S$ before the arrival. Since the arrival is assigned to the server with index (k_l, j) in the larger system, we must have $L = \tilde{Q}_{k_l,j}^{(N)}(A^-) \leq \tilde{Q}_{k_s,j}^{(N)}(A^-) = S$. Hence, we must have $S = L$. Also, since $S = Q_{k_s,j}^{(N)}(A^-) \leq Q_{k_l,j}^{(N)}(A^-) \leq \tilde{Q}_{k_l,j}^{(N)}(A^-) = L$, we must have $Q_{k_l,j}^{(N)}(A^-) = S = L$. Thus, we have $S = Q_{k_s,j}^{(N)}(A^-) = Q_{k_l,j}^{(N)}(A^-) = \tilde{Q}_{k_s,j}^{(N)}(A^-) = \tilde{Q}_{k_l,j}^{(N)}(A^-) = L$. In this case, our coupling rule dictates that if we have chosen the

index (k_s, j) as the destination server for the smaller system, then we must also choose the same indexed server in the larger system as the destination server. This leads to a contradiction because $k_s \neq k_l$. Hence, (31) must hold in this case.

3. *If* $j_s > j_l$: Similar to the first two cases, the inequality $Q^{(N)}_{k_l,j_l}(A) < \tilde{Q}^{(N)}_{k_l,j_l}(A)$ holds. For violation of (31), we must have

$$Q^{(N)}_{k_s,j_s}(A^-) = \tilde{Q}^{(N)}_{k_s,j_s}(A^-) = S$$

just before the arrival. Since $L = \tilde{Q}^{(N)}_{k_l,j_l}(A^-) \leq \tilde{Q}^{(N)}_{k_s,j_s}(A^-) = S$, this implies $S = L$. Furthermore, since $S = Q^{(N)}_{k_s,j_s}(A^-) \leq Q^{(N)}_{k_l,j_l}(A^-) \leq \tilde{Q}^{(N)}_{k_l,j_l}(A^-) = L$, we must have

$$S = Q^{(N)}_{k_s,j_s}(A^-) = Q^{(N)}_{k_l,j_l}(A^-) = \tilde{Q}^{(N)}_{k_s,j_s}(A^-) = \tilde{Q}^{(N)}_{k_l,j_l}(A^-) = L.$$

Hence, there must be a tie between the servers of indices (k_s, j_s) and (k_l, j_l) in both systems. Therefore, according to the SA-JSQ policy the incoming arrival should be assigned to the server having index (i_l, j_l) in smaller system because $j_l < j_s$. This leads to the contradiction because job has been assigned to the server having index (i_s, j_s) in the smaller system. Hence, (31) must hold in this case.

This completes the proof of the theorem. ■

8 Process convergence

In this section, we outline the proof of Theorem 10 using the martingale approach of [28] and the time-scale separation technique of [14]. Here, we discuss the main steps of the proof and characterise the fluid limit process $\mathbf{x}$. The proof consists of the following three steps

Step 1: Martingale representation: The first step is to express the evolution of each component of the process $\mathbf{x}^{(N)}$ in terms of suitably defined martingales and a process $\mathbf{V}^{(N)}$ which evolves at a faster time scale than the components of $\mathbf{x}^{(N)}$. In particular, the process $\mathbf{V}^{(N)} = (\mathbf{V}^{(N)}(t) = (V^{(N)}_{i,j}(t), i \geq 1, j \in [M]), t \geq 0)$ is defined as $V^{(N)}_{i,j}(t) = N\gamma_j - N\gamma_j x^{(N)}_{i,j}(t)$. Thus, $V^{(N)}_{i,j}(t)$ counts the number of type j servers with at most $i-1$ jobs at time t. It is easy to see that $\mathbf{V}^{(N)}$ is a Markov process defined on $\mathbf{E} = (\bar{\mathbb{Z}}^{\infty}_{+})^M$ with transition rates

$$\mathbf{V}^{(N)} \to \begin{cases} \mathbf{V}^{(N)} + \mathbf{e}_{i,j}, & \text{at rate } N\gamma_j\mu_j(x^{(N)}_{i,j} - x^{(N)}_{i+1,j}) \\ \mathbf{V}^{(N)} - \mathbf{e}_{i,j}, & \text{at rate } N\lambda\mathbb{1}\Big(\mathbf{V}^{(N)} \in R_{i,j}\Big) \end{cases}, \; i \geq 1, \; j \in [M], \tag{32}$$

where $R_{i,j}$ for all $i \geq 1$ and $j \in [M]$ is defined as

$$\begin{aligned} R_{i,j} = \Big\{ \mathbf{v} = (v_{i,j}) \in \mathbf{E} : v_{i,k} = 0 \; \forall k \in [j-1], v_{i-1,l} = 0 \; \forall l \in \{j+1, \ldots, M\}, \\ 0 = v_{i-1,j} < v_{i,j}, \Big\}. \end{aligned} \tag{33}$$

The set $R_{i,j}$ represents the set of states where the minimum queue length is at least i for the pools in $\{1, \dots, j-1\}$, exactly $i-1$ for pool j, and at least $i-1$ for the pools in $\{j+1, \dots, M\}$. Thus, when $V^{(N)}(t) \in R_{i,j}$, an incoming job under the SA-JSQ scheme will be assigned to a type j server with queue length $i-1$. We can therefore express $x_{i,j}^{(N)}$ for each $i \geq 1$ and $j \in [M]$ as follows

$$x_{i,j}^{(N)}(t) = x_{i,j}^{(N)}(0) + \frac{\lambda}{\gamma_j}\int_0^t \mathbb{1}_{\{\mathbf{V}^{(N)}(s) \in R_{i,j}\}} ds - \mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s)) ds + \frac{1}{N\gamma_j}(M_{i,j}^{(A,N)}(t) - M_{i,j}^{(D,N)}(t)), \quad (34)$$

where $M_{i,j}^{(A,N)}$ and $M_{i,j}^{(D,N)}$ are martingales corresponding to the arrivals and departures at the component $x_{i,j}^{(N)}$. The precise definitions of these martingales in terms of the counting processes are given in are given in C. It is important to note the difference in the time-scale for the processes $\mathbf{x}^{(N)}$ and $\mathbf{V}^{(N)}$. In a small interval $[t, t+\delta]$, the process $\mathbf{V}^{(N)}$ experiences $O(N\delta)$ transitions whereas the $\mathbf{x}^{(N)}$ changes only by $O(\delta)$. Hence, for large N the process $\mathbf{V}^{(N)}$ reaches its steady-state while $\mathbf{x}^{(N)}$ remains almost constant in this interval. This separation of the two time-scales becomes crucial in characterising the limit of the indicator function $\mathbb{1}\big(\mathbf{V}^{(N)}(s) \in R_{i,j}\big)$ appearing in (34).

Since the time-scales of $\mathbf{V}^{(N)}$ and $\mathbf{x}^{(N)}$ are different, they have different limits as $N \to \infty$. To treat them as a single object and characterise its limit, we define the joint process $(\mathbf{x}^{(N)}, \beta^{(N)})$ where $\beta^{(N)}$ is a random measure defined on $[0, \infty) \times \mathbf{E}$ as

$$\beta^{(N)}(A_1 \times A_2) = \int_{A_1} \mathbb{1}\Big(V^{(N)}(s) \in A_2\Big) ds.$$

for any $A_1 \in \mathcal{B}([0, \infty))$ and $A_2 \in \mathcal{B}(\mathbf{E})$. Hence, (34) can be rewritten in terms of $\beta^{(N)}$ for $i \geq 1$ and $j \in [M]$ as follows

$$x_{i,j}^{(N)}(t) = x_{i,j}^{(N)}(0) + \frac{\lambda}{\gamma_j}\beta^{(N)}([0,t] \times R_{i,j}) - \mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s)) ds + \frac{1}{N\gamma_j}(M_{i,j}^{(A,N)}(t) - M_{i,j}^{(D,N)}(t)). \quad (35)$$

Step 2: Relative compactness: The second step in proving Theorem 10 consists of showing that the sequence of processes $((\mathbf{x}^{(N)}, \beta^{(N)}))_N$ is relatively compact in $D_S[0, \infty) \times \mathcal{L}_0$ where $\mathcal{L}_0$ is defined as the space of measures on $[0, \infty) \times \mathbf{E}$ satisfying $\beta([0,t] \times \mathbf{E}) = t$ for each $t \geq 0$ and each $\beta \in \mathcal{L}_0$. We equip $\mathcal{L}_0$ with the topology of weak convergence of measures restricted to $[0,t] \times \mathbf{E}$ for each t. In the next lemma, we show that $((\mathbf{x}^{(N)}, \beta^{(N)}))_N$ is a relatively compact sequence in $D_S[0, \infty) \times \mathcal{L}_0$ and characterise the limit of any convergent subsequence.

Lemma 14. *If* $\mathbf{x}^N(0) \Rightarrow \mathbf{x}(0) \in S$ *as* $N \to \infty$*, then the sequence* $((\mathbf{x}^{(N)}, \beta^{(N)}))_N$ *is relatively compact in* $D_S[0,\infty) \times \mathcal{L}_0$ *and the limit* $(\mathbf{x}, \beta)$ *of any convergent subsequence satisfies for all* $t \geq 0, i \geq 1, j \in [M]$

$$x_{i,j}(t) = x_{i,j}(0) + \frac{\lambda}{\gamma_j}\beta([0,t] \times R_{i,j}) - \mu_j \int_0^t (x_{i,j}(s) - x_{i+1,j}(s))ds. \tag{36}$$

The proof of Lemma 14 consists of verifying standard conditions of relative compactness in $D_S[0,\infty)$ given in Proposition 3.2.4 of [36]. This is achieved using an approach similar to [7]. The details are given in F.

Step 3: Characterisation of the limit: The final step in proving Theorem 10 is the characterisation the limit $\beta([0,t], R_{i,j})$ appearing in (36). To do so, we define for any $\mathbf{x} \in S$ a Markov process $\mathbf{V}_\mathbf{x}$ on $\mathbf{E}$ with transition rates

$$\mathbf{V}_\mathbf{x} \to \begin{cases} \mathbf{V}_\mathbf{x} + \mathbf{e}_{i,j}, & \text{at rate } \gamma_j \mu_j (x_{i,j} - x_{i+1,j}) \\ \mathbf{V}_\mathbf{x} - \mathbf{e}_{i,j}, & \text{at rate } \lambda \mathbb{1}(\mathbf{V}_\mathbf{x} \in R_{i,j}) \end{cases}, \ i \geq 1, \ j \in [M]. \tag{37}$$

From Lemma 2 and Theorem 3 of [14], it follows that the limit $\beta([0,t] \times R_{i,j})$ satisfies

$$\beta([0,t] \times R_{i,j}) = \int_0^t \pi_{\mathbf{x}(s)}(R_{i,j})ds, \ i \geq 1, j \in [M],$$

where $\pi_\mathbf{x}$ is a stationary measure of the process $\mathbf{V}_\mathbf{x}$ and $\pi_\mathbf{x}$ satisfies the following for all $j \in [M]$

$$\pi_\mathbf{x}(\{\mathbf{V} \in \mathbf{E} : V_{i,j} = \infty\}) = 1, \text{ if } x_{i,j} < 1. \tag{38}$$

Hence, we can write (36) in terms of $\pi_\mathbf{x}$ as

$$x_{i,j}(t) = x_{i,j}(0) + \frac{\lambda}{\gamma_j}\int_0^t \pi_{\mathbf{x}(s)}(R_{i,j})ds - \mu_j \int_0^t (x_{i,j}(s) - x_{i+1,j}(s))ds, \ i \geqslant 1, \ j \in [M]. \tag{39}$$

We set $p_{i-1,j}(\mathbf{x}) = \pi_\mathbf{x}(R_{i,j})$. Hence, to complete the proof of Theorem 10, it remains to show that $\mathbf{x}$ uniquely determines the stationary measure $\pi_\mathbf{x}$ and it satisfies the properties P1-P5 listed in Theorem 10.

First note from (33) that the sets $R_{i,j}$ for $i \geq 1, j \in [M]$ form a partition of $\mathbf{E}$. Since $\pi_\mathbf{x}$ is a probability measure on $\mathbf{E}$, it follows that $\sum_{i\geq 1, j \in [M]} p_{i-1,j}(\mathbf{x}) = \sum_{i \geq 1, j\in[M]} \pi_\mathbf{x}(R_{i,j}) = 1$. This proves P1. Since $\|\mathbf{x}\|_1 < \infty$ for each $\mathbf{x} \in S$, there exists $l_j(\mathbf{x}) = \min\{i : x_{i+1,j} < 1\}$ for all $j \in [M]$ and $\mathbf{x} \in S$. Observe that for each $j \in [M]$ we have $x_{i,j} = 1$ for all $0 \leq i \leq l_j(\mathbf{x})$, and $x_{i,j} < 1$ for all $i \geq l_j(\mathbf{x}) + 1$. Therefore, from (38), it follows that

$$\pi_\mathbf{x}\Big(\{V_{l_j(\mathbf{x})+1,j} = V_{l_j(\mathbf{x})+2,j} = \cdots = \infty : \forall j \in [M]\}\Big) = 1. \tag{40}$$

Hence, from the definition of the set $R_{i,j}$ in (33) it follows that $p_{i-1,j}(\mathbf{x}) = \pi_\mathbf{x}(R_{i,j}) = 0$ for all $i \geq l_j(\mathbf{x}) + 2$ and $j \in [M]$. This proves the property P2.

To prove P3, we note that if $x_{i,j} = 1$ for some $i \geq 1, j \in [M]$, then $dx_{i,j}/dt \leq 0$. Using this fact in (39) we conclude that $\lambda \pi_{\mathbf{x}}(R_{i,j}) \leq \mu_j \gamma_j (x_{i,j} - x_{i+1,j})$ for all (i,j) for which $x_{i,j} = 1$. The definition $l_j(\mathbf{x})$ implies if $l_j(\mathbf{x}) > 0$ for some $\mathbf{x}$ and some j, then that $x_{i,j} = 1$ for all $i \leq l_j(\mathbf{x})$. Hence, $p_{i-1,j}(\mathbf{x}) = \pi_{\mathbf{x}}(R_{i,j}) = 0$ for $i \in [l_j(\mathbf{x}) - 1]$. This shows P3. If $l_j(\mathbf{x}) > 0$ for some $j \in [M]$, then it can be verified using property P3 and the definition of $R_{i,j}$ that

$$\pi_{\mathbf{x}}\Big(\{V_{1,j} = V_{2,j} = \cdots = V_{l_j(\mathbf{x})-1,j} = 0\}\Big) = 1. \tag{41}$$

Now suppose $l_1(\mathbf{x}) = 0$. From (33), it follows that $p_{0,1}(\mathbf{x}) = \pi_{\mathbf{x}}(R_{1,1}) = \pi_{\mathbf{x}}(\{0 < V_{1,1}\}) = 1$ since $\pi_{\mathbf{x}}(\{V_{1,1} = \infty\}) = 1$ according to (38). This proves the P4. The proof of P5 follows similarly using (33), (38), and P2.

Hence, the only part left to prove Theorem 10 is to show that $\pi_{\mathbf{x}}$ is uniquely determined by $\mathbf{x}$ for all $\mathbf{x} \in S$. To show this, it is sufficient to prove that the stationary distribution of $(V_{l_j(\mathbf{x}),j} : j \in [M])$ is uniquely determined by $\mathbf{x}$ because the stationary distribution of all other components of $\mathbf{V}_{\mathbf{x}}$ has already been uniquely characterised by (40) and (41). The transition rates of the individual components of the chain $(V_{l_1(\mathbf{x}),1}, V_{l_2(\mathbf{x}),2}, \ldots, V_{l_M(\mathbf{x}),M})$ are given by

$$V_{l_k(\mathbf{x}),k} \to \begin{cases} V_{l_k(\mathbf{x}),k} + 1, & \text{at rate } \gamma_k \mu_k (x_{l_k(\mathbf{x}),k} - x_{l_k(\mathbf{x})+1,k}) \\ V_{l_k(\mathbf{x}),k} - 1, & \text{at rate } \lambda \mathbb{1}\big(\mathbf{V} \in R_{l_k(\mathbf{x}),k}\big) \end{cases}, \forall k \in [M]. \tag{42}$$

Note that the Markov chain given by (42) is defined on $\bar{\mathbb{Z}}_+^M$ and has 2^M communicating classes since each component of the chain can be either finite or infinite. To show the uniqueness $\pi_{\mathbf{x}}$ we need to show that $\pi_{\mathbf{x}}$ is concentrated only on a single communicating class among these 2^M classes. This is equivalent to showing $\pi_{\mathbf{x}}(V_{l_j(\mathbf{x}),j} = \infty) = 0$ or 1 for each $j \in [M]$. To show the above, we use the result of the next lemma which characterises the stationary distribution of a finite dimensional Markov chain whose transition rates have a form similar to (42).

Lemma 15. *For $K \in \mathbb{N}$, let $\mathbf{U} = (\mathbf{U}(t) = (U_j(t), j \in [K]) \in \mathbb{Z}_+^K : t \geq 0)$ be a Markov chain with transition rates*

$$\mathbf{U} \to \begin{cases} \mathbf{U} + \mathbf{e}_i, & \textit{at rate } \nu_i \\ \mathbf{U} - \mathbf{e}_i, & \textit{at rate } \lambda \mathbb{1}(0 = U_1 = \cdots = U_{i-1} < U_i) \end{cases}, \forall i \in [K], \tag{43}$$

where $\mathbf{e}_i$ denotes the K-dimensional unit vector with entry 1 at i^{th} position. The Markov chain $\mathbf{U}$ is positive recurrent if and only if $\sum_{i \in [K]} \frac{\nu_i}{\lambda} < 1$. Furthermore, if π denotes the stationary distribution of the chain $\mathbf{U}$, then we have

$$\pi\left\{0 = U_1 = \cdots = U_{i-1} < U_i\right\} = \frac{\nu_i}{\lambda}, \ \forall i \in [K]. \tag{44}$$

Using Lemma 15, we show that $\pi_{\mathbf{x}}(V_{l_j(\mathbf{x}),j} = \infty) = 0$ or 1 for each $j \in [M]$ in I.

9 Fixed Point Characterisation

In this section, we prove Theorem 11 which characterises the fixed point $\mathbf{x}^*$ of the fluid limit $\mathbf{x}$ and shows that the fixed point is globally attractive. For the proof of uniqueness of the fixed point, we explicitly use properties P1-P5 listed in Theorem 10. Moreover, to prove global stability we use the monotonicity of the process $\mathbf{x}^{(N)}$ proved in Theorem 9.

Proof of Theorem 11.(i): From (10) it follows that for $\mathbf{x}^* \in S$ to be a fixed point of the fluid limit $\mathbf{x}$, we must have

$$\frac{\lambda}{\gamma_j} p_{i-1,j}(\mathbf{x}^*) = \mu_j(x^*_{i,j} - x^*_{i+1,j}),\ i \geqslant 1,\ j \in [M]. \tag{45}$$

Summing (45) over all $i \geq 1$ and for all $j \in [M]$, we get $\lambda \sum_{i\geq 1}\sum_{j\in[M]} p_{i-1,j}(\mathbf{x}^*) = \sum_{j\in[M]} \mu_j\gamma_j x^*_{1,j}$. Using P1, this implies that

$$\lambda = \sum_{j\in[M]} \mu_j\gamma_j x^*_{1,j}. \tag{46}$$

Thus, $x^*_{1,j} = 1$ for all $j \in [M]$ is not possible because the stability condition requires $\lambda < 1$. Hence, we must have $x^*_{1,j} < 1$ for at least one $j \in [M]$. In the following, we consider different cases based on the interval in which λ belongs

If $0 < \lambda < \mu_1\gamma_1$: For $\lambda \in (0, \mu_1\gamma_1)$, we show that $x^*_{1,1} = \lambda/\mu_1\gamma_1$ and $x^*_{i,j} = 0$ for all $(i,j) \neq (1,1)$. Suppose $x^*_{1,1} < 1$, this means that $l_1(\mathbf{x}^*) = 0$. Therefore, from property P4, we have $p_{0,1}(\mathbf{x}^*) = 1$. Hence, summing (45) over all $i \geq 1$ and for $j = 1$, we get $x^*_{1,1} = \lambda/\mu_1\gamma_1$. Similarly, summing (45) for all $i \geq m$ and for $j = 1$, we get $x^*_{m,1} = 0$ for any $m \geq 2$. By similar line of arguments as above, we can easily verify that $x^*_{i,j} = 0$ for all $i \geq 1$ and for all $j \in \{2, \dots, M\}$. Now, suppose $x^*_{1,1} = 1$. Then from (46), with $x^*_{1,1} = 1$ implies that $\sum_{j=2}^{M} \mu_j\gamma_j x^*_{1,j} = \lambda - \mu_1\gamma_1 < 0$, which leads to a contradiction as $\mathbf{x}^* \in S$.

If $\sum_{i=1}^{j-1} \mu_i\gamma_i \leq \lambda < \sum_{i=1}^{j} \mu_i\gamma_i$, *for* $j \in \{2, \dots, M\}$: For this case we show that $x^*_{1,k} = 1$ for all $k \in [j-1]$, $x^*_{1,j} = (\lambda - \sum_{i=1}^{j-1}\mu_i\gamma_i)/\mu_j\gamma_j$, $x_{1,k} = 0$ for all $k \geq j+1$, and $x^*_{l,k} = 0$ for all $k \in [M]$ and for all $l \geq 2$. First, we use induction to prove that $x^*_{1,k} = 1$ for all $k \in [j-1]$. Suppose $x^*_{1,1} < 1$. This means that $l_1(\mathbf{x}^*) = 0$. Therefore, using P4 and summing (45) for all $i \geq 1$ and $j = 1$, we get $x^*_{1,1} = \lambda/\mu_1\gamma_1 \geq 1$, which contradicts the assumption that $x^*_{1,1} < 1$. Therefore, $x^*_{1,1} = 1$ and $l_1(\mathbf{x}) \geq 1$. This proves base case for the induction. Now assume $x^*_{1,k} = 1$ for all $k \in [j-2]$, which implies that $l_k(\mathbf{x}^*) \geq 1$ for all $k \in [j-2]$. Using the assumption that $x^*_{1,k} = 1$ for all $k \in [j-2]$, we show that $x^*_{1,j-1} = 1$. Suppose $x^*_{1,j-1} < 1$, which implies that $l_{j-1}(\mathbf{x}^*) = 0$. Hence, using property P5 and using (45) we get $\frac{\lambda}{\gamma_{j-1}\mu_{j-1}} \sum_{i\geq 1} p_{i-1,j-1}(\mathbf{x}^*) = x_{1,j-1} = \frac{\lambda}{\gamma_{j-1}\mu_{j-1}} \geq 1$, which contradicts the assumption that $x^*_{1,j-1} < 1$. Therefore, we must have $x^*_{1,j-1} = 1$ and $l_{j-1}(\mathbf{x}^*) \geq 1$. Next, we prove that $x^*_{1,j} = (\lambda - \sum_{i=1}^{j-1}\mu_i\gamma_i)/\mu_j\gamma_j$. Suppose $l_j(\mathbf{x}^*) \geq 1$. This implies that $x^*_{1,j} = 1$. Therefore, using (46), we have $\sum_{i=j+1}^{M} \gamma_i\mu_i x^*_{1,i} = \lambda - \sum_{i=1}^{j}\gamma_i\mu_i < 0$, which is not possible as $\mathbf{x}^* \in S$. Hence, we have $l_j(\mathbf{x}^*) = 0$. So far we have proved that $l_k(\mathbf{x}^*) \geq 1$ for all $k \in [j-1]$, $l_j(\mathbf{x}^*) = 0$

and $l_k(\mathbf{x}^*) \geq 0$ for all $k \geq j+1$. Therefore, using property P5 and equation (45), we can easily get $x^*_{1,k} = 0$ for all $k \geq j+1$. Now using (46), we obtain $x^*_{1,j} = (\lambda - \sum_{i=1}^{j-1} \mu_i \gamma_i)\mu_j \gamma_j$. Similarly, using property P5 and (45), we can easily verify that $x^*_{l,k} = 0$ for all $k \in [M]$ and for all $l \geq 2$. ■

The proofs of global stability and limit interchange are given in H.

10 Optimality of the Resource Pooled System

In this section, we prove Theorem 5. We use the following lemma which establishes that when both $\mathcal{M}'_N$ and $\mathcal{M}_N$ have the same total number of jobs i, the rate of departure $q^{Z^{(N)}}(i, i-1)$ in $\mathcal{M}'_N$ is higher than the rate of departure in $\mathcal{M}_N$. Note that the departure rate in $\mathcal{M}_N$ at any time $t \geq 0$ under a policy Π is $\sum_{j \in [M]} \mu_j X^{(N,\Pi)}_{1,j}$.

Lemma 16. *For any stationary policy* Π*, if* $Z^{(N)}(t) = R^{(N,\Pi)}(t) = i$ *for some* $t \geq 0$*, then*

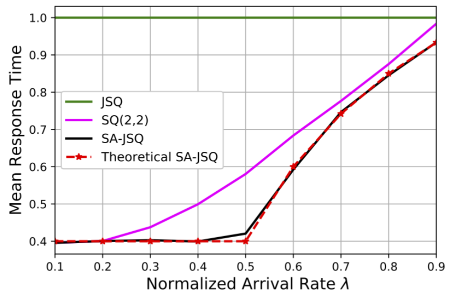
$$\sum_{j \in [M]} \mu_j X^{(N,\Pi)}_{1,j}(t) \leq q^{Z^{(N)}}(i, i-1) = \sum_{j=1}^{M} \mu_j \left(\left(i - \sum_{i=1}^{j-1} N\gamma_i \right)_+ \wedge N\gamma_j \right). \tag{47}$$

Proof of Theorem 5: We construct a coupling between the processes $Z^{(N)}$ and $\mathbf{X}^{(N,\Pi)}$ such that if $Z^{(N)}(0) \leq R^{(N,\Pi)}(0)$ then $Z^{(N)}(t) \leq R^{(N,\Pi)}(t)$ for all $t \geq 0$. Let the current instant be t and assume that $Z^{(N)}(t) \leq R^{(N,\Pi)}(t)$. We shall describe a way of generating the next event and the time for the next event s such that $Z^{(N)}(s) \leq R^{(N,\Pi)}(s)$ is maintained right after the event has taken place.

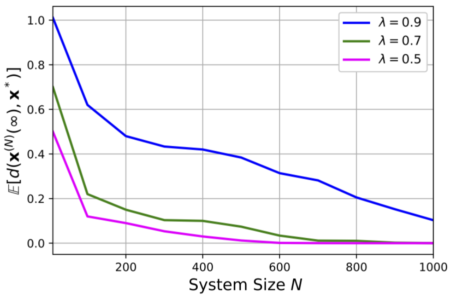
We generate the time until the next potential arrival for both systems as an exponentially distributed random variable with mean $N\lambda$. Hence, for both systems, arrivals occur at the same instants.

If $Z^{(N)}(t) < R^{(N,\Pi)}(t)$ then the time until the next potential departure is generated independently for each system as exponential random variables with means $q^{Z^{(N)}}(Z^{(N)}(t), Z^{(N)}(t) - 1)$ and $\sum_j \mu_j X^{(N,\Pi)}_{1,j}(t)$ for systems $\mathcal{M}'_N$ and $\mathcal{M}_N$, respectively. If $Z^{(N)}(t) = R^{(N,\Pi)}(t) = i$, then we generate the time D until the next potential departure for $\mathcal{M}_N$ as an exponential random variable with mean $\sum_j \mu_j X^{(N,\Pi)}_{1,j}(t)$. We also generate another independent exponential random variable C with mean $q^{Z^{(N)}}(i, i-1) - \sum_j \mu_j X^{(N,\Pi)}_{1,j}(t) \geq 0$. Note that we can do so by Lemma 16. Now we generate the time until the next potential departure for $\mathcal{M}'_N$ as $D' = \min(D, C)$. Therefore, $D' \leq D$ and D' is exponentially distributed with mean $q^{Z^{(N)}}(i, i-1)$.

Once all the potential arrival and departures are generated as described above, the next *true event* is taken to be the potential event which occurs the earliest; all other potential events are discarded (which is possible because of the memoryless property of the exponential distribution). Due to the construction above, it is clear that $Z^{(N)}(s) \leq R^{(N,\Pi)}(s)$ is maintained right after the next true event. This completes the proof. ■



(a) Mean response time as a function of normalized arrival rate λ with $N = 1000$ servers. We set $\mu_1 = 4\mu_2 = 20/8$, $\gamma_1 = 1 - \gamma_2 = 1/5$, and $d_1 = d_2 = 2$.



(b) $d(\mathbf{x}^{(N)}(\infty), \mathbf{x}^*) = \sum_{i,j} |x_{i,j}^{(N)}(\infty) - x_{i,j}^*|$ as a function of system size N. We set $\mu_1 = 2\mu_2 = 4/3$, $\gamma_1 = 1 - \gamma_2 = 1/2$.

Figure 3: Simulation plots

11 Numerical Studies

In this section, we present simulation results for different load balancing schemes. For all simulations, we have assumed $M = 2$ and taken the number of arrivals to be 3×10^5. In Figure 3(a), we have plotted the mean response time of jobs for different schemes as a function of the normalised arrival rate λ. For performance comparison we also simulated a scheme proposed in [37] and referred to as the SQ(d_1, d_2) scheme. For the SQ(d_1, d_2) scheme, upon job arrival d_j servers of type j are sampled uniformly at random from the set of $N\gamma_j$ servers for $j \in [2]$. The job is then sent to the server with the minimum queue length among the sampled servers. Ties within different types servers are broken by selecting the type with the maximum rate. We see that with SA-JSQ we obtain upto 60% reduction in average response time of jobs compared to classical JSQ. As expected, the performance of SQ$(2, 2)$ lies in between classical JSQ and SA-JSQ. To investigate the convergence rate to the fixed point of the fluid limit, in Figure 3(b), we have plotted the distance $d(\mathbf{x}^{(N)}(\infty), \mathbf{x}^*) = \sum_{i,j} |x_{i,j}^{(N)}(\infty) - x_{i,j}^*|$ as a function of N for $\lambda \in \{0.5, 0.7, 0.9\}$. We note that for large values of λ the distance is higher than that for smaller values of λ.

12 Conclusion and Future Works

In this paper, we have investigated speed-aware JSQ-type load balancing schemes for heterogeneous systems. We obtained a lower bound on the mean response time of jobs under any load balancing scheme by comparing the system with an appropriate resource pooled system. We showed that the lower bound is achieved by the SA-JSQ scheme in the fluid limit, thereby establishing the asymptotic optimality of SA-JSQ. Moreover, in establishing the fluid limit of SA-JSQ, we have proved uniform bounds on the stationary measures of queue lengths which are required to prove tightness. Using coupling, we have also shown that any stationary load balancing scheme in a system of parallel queues stochastically dominates the JFFS scheme in the resource pooled system. There are many interesting avenues for future work. Characterising the performance of the SA-JSQ scheme in the Halfin-Whitt regime remains as an open problem. It is also interesting to analytically characterise the distance between $\mathbf{x}^{(N)}(\infty)$ and the fixed point $\mathbf{x}^*$ as a function of the system size N.

References

[1] W. Winston, "Optimality of the shortest line discipline," *Journal of applied probability*, vol. 14, no. 1, pp. 181–189, 1977.

[2] R. R. Weber, "On the optimal assignment of customers to parallel servers," *Journal of Applied Probability*, vol. 15, no. 2, pp. 406–413, 1978.

[3] P. K. Johri, "Optimality of the shortest line discipline with state-dependent service rates," *European Journal of Operational Research*, vol. 41, no. 2, pp. 157–161, 1989.

[4] A. Hordijk and G. Koole, "On the optimality of the generalized shortest queue policy," *Probability in the Engineering and Informational Sciences*, vol. 4, no. 4, pp. 477–487, 1990.

[5] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Performance Evaluation*, vol. 68, no. 11, pp. 1056–1071, 2011.

[6] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia, "Delay, memory, and messaging tradeoffs in distributed service systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 1–12, 2016.

[7] D. Mukherjee, S. C. Borst, J. S. Van Leeuwaarden, and P. A. Whiting, "Universality of power-of-d load balancing in many-server systems," *Stochastic Systems*, vol. 8, no. 4, pp. 265–292, 2018.

[8] R. Govindan, I. Minei, M. Kallahalla, B. Koley, and A. Vahdat, "Evolve or die: High-availability design principles drawn from googles network infrastructure," in *Proceedings of the 2016 ACM SIGCOMM Conference*, ser. SIGCOMM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 58–72. [Online]. Available: https://doi.org/10.1145/2934872.2934891

[9] J. Duato, A. J. Peña, F. Silla, R. Mayo, and E. S. Quintana-Ortí, "rCUDA: Reducing the number of gpu-based accelerators in high performance clusters," in *2010 International Conference on High Performance Computing Simulation*, 2010, pp. 224–231.

[10] M. Huang, D. Wu, C. H. Yu, Z. Fang, M. Interlandi, T. Condie, and J. Cong, "Programming and runtime support to blaze fpga accelerator deployment at datacenter

scale," in *Proceedings of the Seventh ACM Symposium on Cloud Computing*, ser. SoCC '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 456–469. [Online]. Available: https://doi.org/10.1145/2987550.2987569

[11] K. Gardner, J. A. Jaleel, A. Wickeham, and S. Doroudi, "Scalable load balancing in the presence of heterogeneous servers," *Performance Evaluation*, vol. 145, p. 102151, 2021.

[12] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized join-the-shortest-queue (jsq) schemes in large heterogeneous processor-sharing systems," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 2, pp. 116–126, 2015.

[13] M. Bramson, Y. Lu, and B. Prabhakar, "Asymptotic independence of queues under randomized load balancing," *Queueing Systems*, vol. 71, no. 3, pp. 247–292, 2012.

[14] P. Hunt and T. Kurtz, "Large loss networks," *Stochastic Processes and their Applications*, vol. 53, no. 2, pp. 363–378, 1994.

[15] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, and D. Mukherjee, "Scalable load balancing in networked systems: A survey of recent advances," *arXiv preprint arXiv:1806.05444*, 2018.

[16] P. Eschenfeldt and D. Gamarnik, "Join the shortest queue with many servers. the heavy-traffic asymptotics," *Mathematics of Operations Research*, vol. 43, no. 3, pp. 867–886, 2018.

[17] A. Braverman, "Steady-state analysis of the join-the-shortest-queue model in the halfin–whitt regime," *Mathematics of Operations Research*, vol. 45, no. 3, pp. 1069–1103, 2020.

[18] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: An asymptotic approach," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 20–34, 1996.

[19] M. Mitzenmacher, "The power of two choices in randomized load balancing," *PhD thesis, University of California at Berkeley*, 1996.

[20] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Perform. Eval.*, vol. 68, no. 11, p. 1056–1071, nov 2011. [Online]. Available: https://doi.org/10.1016/j.peva.2011.07.015

[21] X. Zhou, F. Wu, J. Tan, Y. Sun, and N. Shroff, "Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, dec 2017. [Online]. Available: https://doi.org/10.1145/3154498

[22] A. Izagirre and A. M. Makowski, "Light traffic performance under the power of two load balancing strategy: The case of server heterogeneity," *SIGMETRICS Perform. Eval. Rev.*, vol. 42, no. 2, p. 18–20, sep 2014. [Online]. Available: https://doi.org/10.1145/2667522.2667527

[23] A. Mukhopadhyay, A. Karthik, and R. R. Mazumdar, "Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay," *Stochastic Systems*, vol. 6, no. 1, pp. 90–131, 2016. [Online]. Available: https://doi.org/10.1287/15-SSY179

[24] A. Gandhi, X. Zhang, and N. Mittal, "Halo: heterogeneity-aware load balancing," in *2015 IEEE 23rd International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*. IEEE, 2015, pp. 242–251.

[25] A. L. Stolyar, "Pull-based load distribution in large-scale heterogeneous service systems," *Queueing Systems*, vol. 80, no. 4, pp. 341–361, 2015.

[26] D. Rutten and D. Mukherjee, "Load balancing under strict compatibility constraints," ser. SIGMETRICS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 51–52. [Online]. Available: https://doi.org/10.1145/3410220.3456275

[27] W. Weng, X. Zhou, and R. Srikant, "Optimal load balancing with locality constraints," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 3, nov 2020. [Online]. Available: https://doi.org/10.1145/3428330

[28] W. Whitt *et al.*, "Proofs of the martingale FCLT," *Probability Surveys*, vol. 4, pp. 268–302, 2007.

[29] N. Gast and B. Gaujal, "Markov chains with discontinuous drifts have differential inclusion limits," *Performance Evaluation*, vol. 69, no. 12, pp. 623–642, 2012.

[30] U. Ayesta, M. Erausquin, M. Jonckheere, and I. M. Verloop, "Scheduling in a random environment: Stability and asymptotic optimality," *IEEE/ACM Transactions on networking*, vol. 21, no. 1, pp. 258–271, 2012.

[31] H. Song, A. L. Tucker, and K. L. Murrell, "The diseconomies of queue pooling: An empirical investigation of emergency department length of stay," *Management Science*, vol. 61, no. 12, pp. 3032–3053, 2015.

[32] F. Kelly and E. Yudovina, *Stochastic Networks*. Cambridge University Press, 2014.

[33] S. R. Turner, "The effect of increasing routing choice on resource pooling," *Probability in the Engineering and Informational Sciences*, vol. 12, no. 1, pp. 109–124, 1998.

[34] P. W. Glynn and A. Zeevi, "Bounding stationary expectations of markov processes," in *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*. Institute of Mathematical Statistics, 2008, pp. 195–214.

[35] P. Billingsley, *Convergence of probability measures.* John Wiley & Sons, 2013.

[36] S. N. Ethier and T. G. Kurtz, *Markov processes: characterization and convergence.* John Wiley & Sons, 2009, vol. 282.

[37] A. Mukhopadhyay, A. Karthik, and R. R. Mazumdar, "Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay," *Stochastic Systems*, vol. 6, no. 1, pp. 90–131, 2016.

[38] G. Pang, R. Talreja, W. Whitt *et al.*, "Martingale proofs of many-server heavy-traffic limits for markovian queues," *Probability Surveys*, vol. 4, pp. 193–267, 2007.

A Resource Pooled System under JFFS: Proofs of Proposition 3 and Proposition 4

In this section, we analyse the resource pooled system $\mathcal{M}'_N$ as described in Section 3 under the JFFS policy and prove the results stated in Propositions 3 and 4.

A.1 Proof of Proposition 3

To prove the proposition, we construct a coupling between the resource pooled system $\mathcal{M}'_N$ and an M/M/1/N system with Poisson arrival rate $N\lambda$ and N identical serves, each with rate 1. Let $Y^{(N)}(t)$ denote the number of jobs in the M/M/1/N system at time $t \geq 0$. The transition rates $q^{Y^{(N)}}(k,l)$, for $k,l \in \mathbb{Z}_+$, of the chain $Y^{(N)} = (Y^{(N)}(t), t \geq 0)$ are given by

$$q^{Y^{(N)}}(k,l) = \begin{cases} N\lambda, & \text{if } l = k+1, \\ k \wedge N, & \text{if } l = k-1, \\ 0, & \text{otherwise.} \end{cases} \tag{48}$$

From the standard results on M/M/1/N queues, it follows that the process $Y^{(N)} = (Y^{(N)}(t), t \geq 0)$ is positive recurrent for $\lambda < 1$. Furthermore, if $Y^{(N)}(\infty)$ denotes the stationary number of jobs in the system, then from the standard results on M/M/1/N queues we have

$$\frac{\mathbb{E}\left[Y^{(N)}(\infty)\right]}{N} = \lambda + \frac{\lambda}{1-\lambda}\frac{\mathbb{P}\left[Y^{(N)}(\infty) \geq N\right]}{N} \leq \lambda + \frac{\lambda}{1-\lambda}. \tag{49}$$

Hence, to prove the proposition it suffices to construct a coupling between $Z^{(N)}$ and $Y^{(N)}$ such that $Z^{(N)}(0) \leq Y^{(N)}(0)$ implies $Z^{(N)}(t) \leq Y^{(N)}(t)$ for all $t \geq 0$. First, note that for each Markov chain, the transition rate out of any state is bounded above by the constant $B = N(\lambda+1)$. Hence, we can generate both $Z^{(N)}$ and $Y^{(N)}$ processes by constructing the corresponding uniformized discrete-time Markov chains $\tilde{Z}^{(N)} = (\tilde{Z}^{(N)}(m), m \in \mathbb{Z}_+)$ and

$\tilde{Y}^{(N)} = (\tilde{Y}^{(N)}(m), m \in \mathbb{Z}_+)$. The one-step transition probabilities of these two chains from state k to state l are respectively given by

$$p^{\tilde{Z}^{(N)}}(k,l) = \begin{cases} \frac{q^{Z^{(N)}}(k,l)}{B}, & \text{for } l \neq k \\ 1 - \sum_{l' \neq k} p^{\tilde{Z}^{(N)}}(k,l'), & \text{for } l = k, \end{cases}$$

$$p^{\tilde{Y}^{(N)}}(k,l) = \begin{cases} \frac{q^{Y^{(N)}}(k,l)}{B}, & \text{for } j \neq i \\ 1 - \sum_{l' \neq k} p^{\tilde{Y}^{(N)}}(k,l'), & \text{for } l = k. \end{cases}$$

where $q^{Z^{(N)}}$ and $q^{Y^{(N)}}$ denote the transition rates for $Z^{(N)}$ and $Y^{(N)}$, respectively. To construct the continuous-time sample paths of the original chains $Z^{(N)}$ and $Y^{(N)}$ on the same probability space, we generate a common Poisson process with rate B and embed the time-steps of both $\tilde{Z}^{(N)}$ and $\tilde{Y}^{(N)}$ into the points of the Poisson process.

It is easy to see from the transition rates that $p^{\tilde{Z}^{(N)}}(k,k-1) \geq p^{\tilde{Y}^{(N)}}(k,k-1)$ for all $k \in \mathbb{Z}_+$. We let the chains $\tilde{Z}^{(N)}$ and $\tilde{Y}^{(N)}$ evolve independently of each other except at instants when they become equal. If $\tilde{Z}^{(N)}(m) = \tilde{Y}^{(N)}(m) = k \in \mathbb{Z}_+$ for some time step $m \in \mathbb{Z}_+$, we first construct $\tilde{Y}^{(N)}(m+1)$ according to the transition probabilities $p^{\tilde{Y}^{(N)}}$. Then we generate $\tilde{Z}^{(N)}(m+1)$ as follows:

$$\tilde{Z}^{(N)}(m+1) = \begin{cases} \tilde{Y}^{(N)}(m+1), & \text{if } \tilde{Y}^{(N)}(m+1) \in \{i+1, i-1\} \\ \tilde{Y}^{(N)}(m) - \theta, & \text{otherwise }, \end{cases}$$

where $\theta \in \{0,1\}$ is a Bernoulli random variable with

$$\mathbb{P}[\theta = 1 | \tilde{Z}^{(N)}(m) = \tilde{Y}^{(N)}(m) = k] = \frac{p^{\tilde{Z}^{(N)}}(k,k-1) - p^{\tilde{Y}^{(N)}}(k,k-1)}{p^{\tilde{Y}^{(N)}}(k,k)} \geq 0.$$

Clearly, under the coupling described above $\tilde{Z}^{(N)}(m) \leq \tilde{Y}^{(N)}(m)$ for all $m \in \mathbb{Z}_+$ if $Z^{(N)}(0) = \tilde{Z}^{(N)}(0) \leq \tilde{Y}^{(N)}(0) = Y^{(N)}(0)$. Hence, we have $Z^{(N)}(t) \leq Y^{(N)}(t)$ for all $t \geq 0$. ■

A.2 Proof of Proposition 4

To prove Proposition 4, we study the limit of the process $z^{(N)} = \big(z^{(N)}(t), t \geq 0\big)$ defined as

$$z^{(N)}(t) = \frac{Z^{(N)}(t)}{N}, \quad t \geq 0.$$

Thus, $z^{(N)}(t)$ denotes the scaled number of jobs in $\mathcal{M}'_N$ under the JFFS scheme at time t. We first characterise the limit of the sequence of processes $(z^{(N)})_N$ in the lemma below.

Lemma 17. *If $z^{(N)}(0) \Rightarrow z(0) \in \mathbb{R}$ as $N \to \infty$, then $z^{(N)} \Rightarrow z$ as $N \to \infty$, where $z = (z(t), t \geq 0)$ is the unique deterministic process satisfying the following integral equation*

$$z(t) = z(0) + \lambda t - \int_0^t \sum_{j=1}^{M} \mu_j \left(\left(z(s) - \sum_{i=1}^{j-1} \gamma_i \right)_+ \wedge \gamma_j \right) ds. \tag{50}$$

Furthermore, the process z satisfying (50) *has a unique fixed point z^* given by*

$$z^* = \max_{j\in[M]} \left(\sum_{i=1}^{j-1} \gamma_i + \frac{\lambda - \sum_{i=1}^{j-1} \mu_i\gamma_i}{\mu_j} \right). \tag{51}$$

Proof. Let $(A(t) : t \geq 0)$ and $(D(t) : t \geq 0)$ be independent unit-rate Poisson processes. We can express the evolution of the total number $Z^{(N)}(t)$ of jobs in the system $\mathcal{M}'_N$ as follows:

$$Z^{(N)}(t) = Z^{(N)}(0) + A(N\lambda t) - D(\int_0^t T(Z^{(N)}(s))\, ds), \quad t \geq 0, \tag{52}$$

where $Z^{(N)}(0) = Nz^{(N)}(0)$ and $T : \mathbb{R}_+ \to \mathbb{R}_+$ is defined as

$$T(Z) = \sum_{j=1}^{M} \mu_j \left(\left(Z - \sum_{i=1}^{j-1} N\gamma_i \right)_+ \wedge N\gamma_j \right).$$

Hence, for $Z \in \mathbb{Z}_+$, $T(Z)$ represents the total rate at which jobs depart the system in state Z. Dividing (52) by N, we have

$$z^{(N)}(t) = z^{(N)}(0) + M_A^{(N)}(t) - M_D^{(N)}(t) + \lambda t - \int_0^t T'(z^{(N)}(s))\, ds, \ \forall t \geq 0, \tag{53}$$

where

$$\begin{aligned} M_A^{(N)}(t) &= \frac{A(N\lambda t) - N\lambda t}{N}, \\ M_D^{(N)}(t) &= \frac{D(\int_0^t T(Z^{(N)}(s))\, ds) - \int_0^t T(Z^{(N)}(s))\, ds}{N}, \\ T'(z) &= \frac{T(Z)}{N} = \sum_{j=1}^{M} \mu_j((z - \sum_{i=1}^{j-1} \gamma_i)_+ \wedge \gamma_j). \end{aligned}$$

Using Lemma 3.2 from [38], it can be easily verified that $M_A^{(N)}$ and $M_D^{(N)}$ are square-integrable martingales with respect to the filtration $\mathbf{F}^{(N)} = (\mathcal{F}_{N,t}, t \geq 0)$, where

$$\mathcal{F}_{N,t} = \sigma\Big[Z^{(N)}(0), A(N\lambda s), D\Big(\int_0^s T(Z^{(N)}(u))du\Big), 0 \leq s \leq t\Big].$$

Moreover, quadratic variation process for $M_A^{(N)}$ and $M_D^{(N)}$ are given by

$$\begin{aligned} [M_A^{(N)}](t) &= \frac{A(N\lambda t)}{N^2}, \ t \geq 0 \\ [M_D^{(N)}](t) &= \frac{1}{N^2} D(\int_0^t T(Z^{(N)}(s))\, ds), \ t \geq 0. \end{aligned}$$

Using the martingale functional central limit theorem [28], we have

$$M_A^{(N)} \Rightarrow 0, \quad M_D^{(N)} \Rightarrow 0 \text{ as } N \to \infty.$$

From (53) and (50), we see that both $z^{(N)}$ and z can be expressed as $z^{(N)} = f(z^{(N)}(0), M_A^{(N)} - M_D^{(N)})$ and $z = f(z(0), 0)$, where $f : \mathbb{R} \ \times \mathcal{D}[0,\infty) \ \to \mathcal{D}[0,\infty)$ is defined as the mapping that takes (b, y) to x determined by the following integral equation

$$x(t) = b + y(t) + \lambda t - \int_0^t T'(x(s))ds.$$

Hence, if f is well-defined and continuous, then the continuous mapping theorem proves the first statement of the lemma. From Theorem 4.1 of [38]), it follows that to show that f is well-defined and continuous, it is sufficient to show that T' is Lipschitz continuous. Since, $\mu_j((z - \sum_{i=1}^{j-1} \gamma_i)_+ \wedge \gamma_j)$ is Lipschitz with constant μ_j for all $j \in [M]$. Therefore, the sum of Lipschitz functions is again Lipschitz with constant $(\mu_1 \vee \mu_2 \vee \ldots \vee \mu_M)$. This establishes the first statement of the lemma.

To prove the second statement of the lemma, note that we can can express (50) in its differential form as follows

$$\frac{d}{dt} z(t) = \lambda - T'(z(t)).$$

Hence, any fixed point z^* of the process z must satisfy the equation $\lambda - T'(z^*) = 0$. Since T' is a piecewise-linear map, it is easy to solve the above equation in closed form and find the unique solution to be (51). ∎

Lemma 18. *Let $z(u,t)$ denote the solution to* (50) *for $z(0) = u$. Then for any $u \in \mathbb{R}$, $z(u,t) \to z^*$ as $t \to \infty$. Furthermore, the sequence $(z^{(N)}(\infty))_N$ of stationary states converges weakly to z^* as $N \to \infty$.*

Proof. We prove the first statement of the lemma by considering the following two cases (i) $u \geq z^*$ and (ii) $u < z^*$. We only provide the proof for the first case as the proof of the second case is similar.

If $u \geq z^*$ then $z(u,t) \geq z^*$ for all $t \geq 0$. To see this, assume on the contrary that $z(u,t) < z^*$ for some $t > 0$. Since $z(0,u) = u \geq z^*$ and $z(u,\cdot)$ is continuous, there must exist a $t_1 \in (0,t)$ such that $z(u,t_1) = z^*$. But since z^* is a fixed point, this implies that $z(t) = z^*$ for all $t \geq t_1$ which leads to a contradiction.

Now consider the Lypunov function $V : \mathbb{R}_+ \to \mathbb{R}_+$ defined as $V(z(u,t)) = z(u,t) - z^* \geq 0$. Using (50) we have

$$\frac{d}{dt} V(z(u,t)) = \lambda - T'(z(u,t)), \tag{54}$$

where $T'(z) = \sum_{j=1}^{M} \mu_j((z - \sum_{i=1}^{j-1} \gamma_i)_+ \wedge \gamma_j)$. Since the fixed point z^* solves $\lambda - T'(z^*) = 0$, we have

$$\frac{d}{dt} V(z(u,t)) = -(T'(z(u,t)) - T'(z^*)).$$

We further note that T' is a strictly increasing function of its argument. Hence, $\dot{V}(z(u,t)) < 0$ when $z(u,t) > z^*$. This implies that $V(z(u,t)) \to 0$ as $t \to \infty$, thereby proving the first part of the lemma.

To prove the second part of the lemma, we first show that the sequence $\left(z^{(N)}(\infty)\right)_N$ is tight. By the application of Markov inequality and the bound in (2), we have

$$\sup_{N\geq 1} \mathbb{P}(z^{(N)}(\infty) > a) \leq \sup_{N\geq 1} \frac{\mathbb{E}[z^{(N)}(\infty)]}{a} \leq \frac{1}{a}\Big(\lambda + \frac{\lambda}{1-\lambda}\Big),$$

This shows that to make $\sup_{N\geq 1} \mathbb{P}(z^{(N)}(\infty) > a) < \epsilon$ for any $\epsilon > 0$, there exists appropriate choice $a(\epsilon)$ not dependent on N. This shows that the sequence $\left(z^{(N)}(\infty)\right)_N$ is tight. The rest of the lemma now follows from the same line of arguments as in the proof of the second statement of Theorem 11. ■

The proof of Proposiion 4 follows directly from Lemma 18 by noting that $z^{(N)}(\infty) = Z^{(N)}(\infty)/N$. ■

B Proof of Proposition 13

Fix any $\epsilon > 0$ and $l \geq 1$. Using Markov inequality, we obtain

$$\mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(\infty) > \epsilon\Big) \leq \frac{\mathbb{E}\left[\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(\infty)\right]}{\epsilon} \leq \frac{1}{\epsilon}\mathbb{E}\left[\sum_{j\in[M]}\sum_{i\geq l} x_{i,j}^{(N)}(\infty)\right].$$

Since $\left(x_{i,j}^{(N)}(\infty)\right)_i$ is a sequence non-negative random variables for each $j \in [M]$, using monotone convergence theorem we can interchange the sum and the expectation on the RHS. Hence, we have

$$\mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(\infty) > \epsilon\Big) \leq \frac{1}{\epsilon}\sum_{j\in[M]}\sum_{i\geq l}\mathbb{E}\left[x_{i,j}^{(N)}(\infty)\right] = \frac{1}{\epsilon}\sum_{j\in[M]}\sum_{i\geq l}\mathbb{P}\left[Q_{k,j}^{(N)}(\infty) \geq i\right], \quad (55)$$

where the last equality follows from (5). Now, from Theorem 8 we know that for any $\theta \in [0, -\log\lambda)$ we have

$$\sum_{j\in[M]}\sum_{i\geq l}\mathbb{P}\left[Q_{k,j}^{(N)}(\infty) \geq i\right] \leq \sum_{j\in[M]}\sum_{i\geq l} C_j(\lambda,\theta)e^{-i\theta} = C(\theta)e^{-l\theta},$$

where $C(\theta) = \frac{(1-\lambda)}{(1-\lambda e^{\theta})(1-e^{-\theta})}\sum_{j\in[M]}\frac{1}{\mu_j\gamma_j}$. Since the RHS of the above inequality is not dependent on N, using (55) we have

$$\limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(\infty) > \epsilon\Big) \leq C(\theta)e^{-l\theta} \quad (56)$$

for all $\theta \in [0, -\log\lambda)$ and all $l \geq 1$. Hence, the condition of tightness given by (30) is verified by fixing some $\theta \in (0, -\log\lambda)$ and letting $l \to \infty$. ■

C Martingale Representation and Its Convergence

We first write the evolution of $x_{i,j}^{(N)}(t)$ for all $t \geq 0$ in terms of number of arrivals and departures from system till time t, for $i \geq 1$ and $j \in [M]$ as

$$x_{i,j}^{(N)}(t) = x_{i,j}^{(N)}(0) + \frac{1}{N\gamma_j} A_{i,j}\Big(N\lambda \int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds\Big) - \frac{1}{N\gamma_j} D_{i,j}\Big(N\gamma_j\mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds\Big), \tag{57}$$

where $A_{i,j}$ and $D_{i,j}$ are mutually independent unit-rate Poisson processes and $p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s)) = \mathbb{1}_{\{\mathbf{V}^{(N)}(s) \in R_{i,j}\}}$. Define for all $i \geq 1$ and for all $j \in [M]$

$$\begin{aligned}
M_{i,j}^{(A,N)}(t) &= A_{i,j}\Big(N\lambda \int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds\Big) - N\lambda \int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds,\\
M_{i,j}^{(D,N)}(t) &= D_{i,j}\Big(N\gamma_j\mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds\Big) - N\gamma_j\mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds.
\end{aligned}$$

We next show that the processes $M_{i,j}^{(A,N)} = (M_{i,j}^{(A,N)}(t) : t \geq 0)$ and $M_{i,j}^{(D,N)} = (M_{i,j}^{(D,N)}(t) : t \geq 0)$ are martingales with respect to the filtration $\mathbf{F}^{(N)} = \Big\{F_t^{(N)} : t \geq 0\Big\}$ augmented with all null sets, where $F_t^N = \bigcup_{j\in[M]} G_{t,j}^{(N)}$ with

$$\begin{aligned}
G_{t,j}^{(N)} = \bigcup_{i\geq 1} \sigma\Big(& x_{i,j}^{(N)}(0), A_{i,j}\Big(N\lambda \int_0^s p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(u))du\Big),\\
& D_{i,j}\Big(N\gamma_j\mu_j \int_0^s (x_{i,j}^{(N)}(u) - x_{i+1,j}^{(N)}(u))du\Big), 0 \leqslant s \leqslant t\Big).
\end{aligned}$$

Lemma 19. *The processes $M_{i,j}^{(A,N)}$ and $M_{i,j}^{(D,N)}$, are square integrable $\mathbf{F}^{(N)}$-martingales for all $i \geqslant 1$ and for all $j \in [M]$. Moreover, the predictable quadratic variation processes are given by*

$$\begin{aligned}
\langle M_{i,j}^{(A,N)} \rangle(t) &= N\lambda \int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds,\ i \geq 1,\ j \in [M],\ t \geq 0,\\
\langle M_{i,j}^{(D,N)} \rangle(t) &= N\gamma_j\mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds,\ i \geq 1,\ j \in [M],\ t \geq 0.
\end{aligned}$$

Proof. Let $I_{i,j}(t) = N\mu_j\gamma_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds$ and $L_{i,j}(t) = N\lambda \int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds$. Next, we prove that

$$\begin{aligned}
&\mathbb{E}(I_{i,j}(t)) < \infty, \mathbb{E}\Big(D_{i,j}(I_{i,j}(t))\Big) < \infty,\ i \geq 1,\ j \in [M],\\
&\mathbb{E}(L_{i,j}(t)) < \infty, \mathbb{E}\Big(A_{i,j}(L_{i,j}(t))\Big) < \infty,\ i \geq 1,\ j \in [M].
\end{aligned}$$

The proof of square-integarble martingales and its corresponding predictable quadratic variation processes then follows immediately using Lemma 3.2 from [38]. Note that using crude inequality we can write

$$\begin{aligned}\mathbb{E}(I_{i,j}(t)) &\leq \mu_j t\Big(N\gamma_j \mathbb{E}(x_{i,j}^{(N)}(0)) + \mathbb{E}\Big(A_{i,j}(L_{i,j}(t))\Big)\Big) \\ &\leq \mu_j\gamma_j N t + \mu_j N\lambda t^2 < \infty,\ t\geq 0,\ i\geq 1,\ j\in[M],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\Big[D_{i,j}(I_{i,j}(t))\Big] &\leq \mathbb{E}\Big[D_{i,j}\Big(\mu_j t\Big(N\gamma_j x_{i,j}^{(N)}(0) + A_{i,j}(L_{i,j}(t))\Big)\Big)\Big] \\ &= \mathbb{E}\left\{\mathbb{E}\Big[D_{i,j}\Big(\mu_j t\Big(N\gamma_j x_{i,j}^{(N)}(0) + A_{i,j}(L_{i,j}(t))\Big)\Big)|A_{i,j}(L_{i,j}(t))\Big]\right\} \\ &\leq \mu_j t(N\gamma_j + N\lambda t) < \infty, t\geq 0,\ i\geq 1,\ j\in[M].\end{aligned}$$

Similarly, we can show $\mathbb{E}(L_{i,j}(t)) < \infty$ and $\mathbb{E}\Big(A_{i,j}(L_{i,j}(t))\Big) < \infty$. ■

Now we can write the martingale representation of equation (57) for all $t\geq 0$, $i\geq 1$, and $j\in[M]$ as

$$\begin{aligned}x_{i,j}^{(N)}(t) = x_{i,j}^{(N)}(0) + \frac{\lambda}{\gamma_j}\int_0^t \mathbb{1}_{\{\mathbf{V}^{(N)}(s)\in R_{i,j}\}}ds - \mu_j\int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds \\ + \frac{1}{N\gamma_j}(M_{i,j}^{(A,N)}(t) - M_{i,j}^{(D,N)}(t)).\end{aligned} \quad (58)$$

In below lemma, we prove that the martingale part in (58) converges to 0 with respect to ℓ_1.

Lemma 20. *Following convergence holds as $N\to\infty$*

$$\left\{\max_{j\in[M]}\frac{1}{N\gamma_j}\sum_{i\geq 1}(|M_{i,j}^{(A,N)}(t)|+|M_{i,j}^{(D,N)}(t)|)\right\}_{t\geq 0} \Rightarrow 0.$$

Proof. The proof is similar to the proof of Proposition 4.3 in [7] using Doob's inequality. ■

D Proof of Lemma 16

To prove (47), we consider the following linear optimisation problem

$$\begin{aligned}\max\quad & \sum_{j\in[M]}\mu_j X_{1,j}^{(N,\Pi)}(t), \\ \text{s.t.}\quad & \sum_{j\in[M]} X_{1,j}^{(N,\Pi)}(t)\leq i, \\ & 0\leq X_{1,j}^{(N,\Pi)}(t)\leq N\gamma_j\ \forall j\in[M].\end{aligned} \quad (59)$$

The first constraint in (59) is true as the total number of busy servers $\sum_{j\in[M]} X_{1,j}^{(N,\Pi)}(t)$ at time t in $\mathcal{M}_N$ is always less than equal to the total number of customers in system that is i. The above optimisation problem has a unique maximum which is obtained in following way. First, note that the objective function is given as the weighted sum of $X_{1,j}^{(N,\Pi)}(t)$ with weight μ_j for $j \in [M]$. We know that the maximum weight is μ_1. Therefore, the maximum value that $X_{1,1}^{(N,\Pi)}(t)$ takes is $(i \wedge N\gamma_1)$. Moreover, the second maximum weight is μ_2, hence the maximum value that $X_{1,2}^{(N,\Pi)}(t)$ takes is $((i - N\gamma_1)_+ \wedge N\gamma_2)$. Proceeding in this way the maximum value that $X_{1,j}^{(N,\Pi)}(t)$ takes is $\left(\left(i - \sum_{i=1}^{j-1} N\gamma_i\right)_+ \wedge N\gamma_j\right)$ for $j \in [M]$, which completes the proof. ■

E Characterisation of Compact Sets and Tightness Criteria

In the lemma below, we characterise compact sets in the space S.

Lemma 21. *A set $\mathbf{B} \subseteq S$ is relative compact in S if and only if*

$$\lim_{l\to\infty} \sup_{\mathbf{y}\in\mathbf{B}} \max_{j\in[M]} \sum_{i\geq l} y_{i,j} = 0. \tag{60}$$

Proof. Suppose any $\mathbf{B} \subseteq S$ satisfying (60). To show $\mathbf{B}$ is relatively compact in S, we need to show that any sequence $(\mathbf{y}^{(N)})_{N\geq 1}$ in $\mathbf{B}$ has a Cauchy subsequence. Since S is complete under ℓ_1, therefore the sequence $(\mathbf{y}^{(N)})_{N\geq 1}$ has a convergent subsequence whose limit lies in $\bar{\mathbf{B}}$ which will complete the proof.

Next we show that the sequence $(\mathbf{y}^{(N)})_{N\geq 1}$ has a Cauchy subsequence. Fix any $\epsilon > 0$ and choose $l \geq 1$ such that

$$\max_{j\in[M]} \sum_{i\geq l} |y_{i,j}^{(N)}| < \frac{\epsilon}{4},\ \forall N \geq 1. \tag{61}$$

Now consider the sequence of first coordinates $(y_{1,j}^{(N)})_{N\geq 1}$ for each $j \in [M]$. The sequence $(y_{1,j}^{(N)})_{N\geq 1}$ lies in $[0, 1]$. Therefore, by Bolzano-Wiestrass theorem it has a convergent subsequence $(y_{1,j}^{(N_k)})_{k\geq 1}$. Moreover, along the indices $(N_k)_{k\geq 1}$, the sequence $(y_{2,j}^{(N_k)})_{k\geq 1}$ has a further convergent subsequence. Proceeding this way, we get a sequence of indices $(N_m)_{m\geq 1}$ along which all first $l-1$ coordinates converges. This implies that there exists a $N^\epsilon \in \mathbb{N}$ such that

$$\max_{j\in[M]} \sum_{i<l} |y_{i,j}^{(N)} - y_{i,j}^{(R)}| < \frac{\epsilon}{2},\ \forall R, N \geq N^\epsilon. \tag{62}$$

Using (61), and (62), for all $R, N \geq N^\epsilon$ and for $\mathbf{y}^{(N)}, \mathbf{y}^{(R)} \in \mathbf{B}$ we have

$$\begin{aligned}\left\|\mathbf{y}^{(N)} - \mathbf{y}^{(R)}\right\|_1 &= \max_{j\in[M]} \sum_{i\geq 1} |y_{i,j}^{(N)} - y_{i,j}^{(R)}| \\ &\leq \max_{j\in[M]} \sum_{i<l} |y_{i,j}^{(N)} - y_{i,j}^{(R)}| + \max_{j\in[M]} \sum_{i\geq l} |y_{i,j}^{(N)}| + \max_{j\in[M]} \sum_{i\geq l} |y_{i,j}^{(R)}| \\ &< \epsilon,\end{aligned}$$

along the sequence of indices $(N_m)_{m\geq 1}$. This shows the existence of a Cauchy subsequence. Moreover, the limit point lies in S follows from the completeness of ℓ_1 space and the fact that S is a closed subset of ℓ_1.

Now for the only if part, let $\mathbf{B}$ be a relatively compact set in S. Assume that there exists a $\epsilon > 0$ such that

$$\lim_{l\to\infty} \sup_{\mathbf{y}\in\mathbf{B}} \max_{j\in[M]} \sum_{i\geq l} y_{i,j} > \epsilon. \tag{63}$$

This implies that for each $k \geq 1$, there exists a $\mathbf{y}^{(k)} \in \mathbf{B}$, such that $\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{(k)} \geq \frac{\epsilon}{2}$. Therefore, if $\mathbf{y}^*$ be the limit of the sequence $(\mathbf{y}^{(k)})_{k\geq 1}$, then (63) implies that $\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^* \geq \frac{\epsilon}{2}$ for all $l \geq 1$, this leads to the contradiction that $\mathbf{y}^* \in \ell_1$.

■

Next, we prove the criteria for a sequence to be tight in S.

Lemma 22. *A sequence $(\mathbf{y}^{(N)})_{N\geq 1}$ of random elements in S is tight iff for all $\epsilon > 0$ we have*

$$\lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{(N)} > \epsilon\Big) = 0. \tag{64}$$

Proof. For if part, we construct a relatively compact set $\mathbf{B}^\epsilon$ for any $\epsilon > 0$ such that

$$\mathbb{P}\Big(\mathbf{y}^{(N)} \notin \mathbf{B}^\epsilon\Big) < \epsilon, \ \forall N \in \mathbb{N}.$$

As $(\mathbf{y}^{(N)})_{N\geq 1}$ satisfies (64), therefore there exists a $l(\epsilon) \geq 1$ for all $\epsilon > 0$ such that

$$\limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l(\epsilon)} y_{i,j}^{(N)} > \epsilon\Big) < \epsilon,$$

and there exists a $N^\epsilon \geq 1$ such that

$$\mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l(\epsilon)} y_{i,j}^{(N)} > \epsilon\Big) < \epsilon, \ \forall N > N^\epsilon.$$

Moreover, since $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N^\epsilon)}$ are random elements of S, there exists $k(\epsilon) = \max\{l_1(\epsilon), \ldots, l_{N^\epsilon}(\epsilon)\}$ such that

$$\mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq k(\epsilon)} y_{i,j}^{(N)} > \epsilon\Big) < \epsilon, \ \forall N \in \mathbb{N}.$$

This implies that there exists an increasing sequence $(l(n))_{n\geq 1}$ such that

$$\mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l(n)} y_{i,j}^{(N)} > \frac{\epsilon}{2^n}\Big) < \frac{\epsilon}{2^n}, \ \forall N \in \mathbb{N}.$$

Define

$$\mathbf{B}^\epsilon = \left\{\mathbf{y} \in S : \max_{j\in[M]} \sum_{i\geq l(n)} y_{i,j} \leq \frac{\epsilon}{2^n}, \forall n \geq 1\right\}.$$

From Lemma 21, the set $\mathbf{B}^\epsilon$ is relatively compact in S. Therefore, we have

$$\mathbb{P}\Big(\mathbf{y}^{(N)} \notin \mathbf{B}^\epsilon\Big) = \mathbb{P}\Big(\bigcup_{n\geq 1}\left\{\max_{j\in[M]} \sum_{i\geq l(n)} y_{i,j}^{(N)} > \frac{\epsilon}{2^n}\right\}\Big) \leq \sum_{n\geq 1} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l(n)} y_{i,j}^{(N)} > \frac{\epsilon}{2^n}\Big) < \epsilon,$$

where the first inequality follows from union bound. For only if part, assume that there exists a $\epsilon > 0$ such that

$$\lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{(N)} > \epsilon\Big) > \epsilon. \tag{65}$$

Since $(\mathbf{y}^{(N)})_{N\geq 1}$ is tight in S, therefore there exists a convergent subsequence $(\mathbf{y}^{(N_k)})_{k\geq 1}$ with limit $\mathbf{y}^*$. From (65), we can write

$$\begin{aligned}\epsilon < \lim_{l\to\infty} \limsup_{k\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{(N_k)} > \epsilon\Big) &\leq \lim_{l\to\infty} \limsup_{k\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{(N_k)} \geq \epsilon\Big) \\ &\leq \lim_{l\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} y_{i,j}^{*} \geq \epsilon\Big),\end{aligned}$$

where the last inequality follows from Portmanteau's theorem for closed set. This leads to the contradiction that $\mathbf{y}^* \in \ell_1$.

■

F Proof of Lemma 14

To prove relative compactness of the sequence $((\mathbf{x}^{(N)}, \beta^{(N)}))_{N\geq 1}$, we start with proving that for all finite time t the system occupancy state lies in some compact set.

Lemma 23. *Assume* $\mathbf{x}^{(N)}(0) \Rightarrow \mathbf{x}(0) \in S$, *as* $N \to \infty$. *Then for any* $T \geq 0$, *there exists a* $L(T, \mathbf{x}(0)) > 2$ *such that under the SA-JSQ policy, the probability that an arriving job joins a server with at-least* $L(T, \mathbf{x}(0)) - 1$ *active jobs upto time* T *tends to* 0 *as* $N \to \infty$.

Proof. Let $A^{(N)}(t)$ denote the total number of arrivals up to time t. Arrivals are happening with rate $N\lambda$. Then for any $\epsilon > 0$ we have

$$\mathbb{P}\Big(A^{(N)}(t) \geq (\lambda t + \epsilon)N\Big) \to 0, \ \textit{as} \ N \to \infty.$$

Let $m^{(N)}(t) = \min_{j \in [M]} m_j^{(N)}(t)$, where $m_j^{(N)}(t)$ is the minimum queue length in the j^{th} pool at time t for N^{th} system. The probability at which an arrival joins a server with at-least $L(T, \mathbf{x}(0)) - 1$ jobs during the time interval $[0, T]$ is given by

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{m^{(N)}(t) \geq L(T, \mathbf{x}(0)) - 1, \ t \in [0, T]\Big\} \cap \{\textit{an arrival occur at } t\}\Big) \\
&\leq \mathbb{P}\Big(m^{(N)}(t) \geq L(T, \mathbf{x}(0)) - 1, \ t \in [0, T]\Big) \\
&= \mathbb{P}\Big(X^{(N)}(t) \geq N(L(T, \mathbf{x}(0)) - 1), \ t \in [0, T]\Big),
\end{aligned} \tag{66}$$

where $X^{(N)}(t)$the is the total number of jobs in system at time t. From conservation of flow, we can write $X^{(N)}(t) = X^{(N)}(0) + A^{(N)}(t) - D^{(N)}(t)$, where $D^{(N)}(t)$ is the total number of departures from system till time t. Therefore, from (66), we can write

$$\begin{aligned}
&\mathbb{P}\Big(X^{(N)}(t) \geq N(L(T, \mathbf{x}(0)) - 1), \ t \in [0, T]\Big) \\
&= \mathbb{P}\Big(X^{(N)}(0) + A^{(N)}(t) - D^{(N)}(t) \geq N(L(T, \mathbf{x}(0)) - 1), \ t \in [0, T]\Big) \\
&\leq \mathbb{P}\Big(X^{(N)}(0) + A^{(N)}(t) \geq N(L(T, \mathbf{x}(0)) - 1), \ t \in [0, T]\Big) \\
&= \mathbb{P}\Big(M^{(N)}(t) \geq (L(T, \mathbf{x}(0)) - 1) - x^{(N)}(0) - \lambda t, \ t \in [0, T]\Big),
\end{aligned} \tag{67}$$

where $M^{(N)}(t) = \frac{A^{(N)}(t) - N\lambda t}{N}$, and $x^{(N)}(0) = \frac{X^{(N)}(0)}{N}$. Now observer that for all $\epsilon > 0$ we have

$$\mathbb{P}\Big(\sup_{t \in [0,T]} |M^{(N)}(t)| > \epsilon\Big) \to 0, \ \textit{as} \ N \to \infty.$$

Therefore, we have

$$\mathbb{P}\Big(|M^{(N)}(t)| > 1, \ t \in [0, T]\Big) \to 0, \ \textit{as} \ N \to \infty. \tag{68}$$

From (66), (67), (68) and choosing $L(T, \mathbf{x}(0)) > 2 + x(0) + \lambda T$, where $x(0) = \sum_{j \in [M]} \sum_{i \geq 1} \gamma_j x_{i,j}(0)$, we have

$$\mathbb{P}\Big(\Big\{m^{(N)}(t) \geq L(T, \mathbf{x}(0)) - 1, \ t \in [0, T]\Big\} \cap \{\textit{an arrival occur at } t\}\Big) \to 0, \ \textit{as} \ N \to \infty.$$

■

We now prove the relative compactness of the sequence $((\mathbf{x}^{(N)}, \beta^{(N)}))_{N\geq 1}$ by showing the relative compactness of individual component. Note that the space $\mathbf{E}$ is compact. Therefore, relative compactness of the sequence $(\beta^{(N)})_{N\geq 1}$ follows from Prohorov's theorem [35]. To prove relative compactness of the sequence $(\mathbf{x}^{(N)})_{N\geq 1}$, we need to verify following conditions.

1. For every $\eta > 0$ and rational $t \geq 0$, there exists a compact set $\mathbf{B}_{\eta,t} \subset S$ such that

$$\liminf_{N\to\infty} \mathbb{P}(\mathbf{x}^{(N)}(t) \in \mathbf{B}_{\eta,t}) \geq 1 - \eta \tag{69}$$

2. For every $\eta > 0$ and for $T > 0$ there exists a $\delta > 0$ and a finite partition $\{t_1, t_2, \ldots, t_n\}$ of $[0, T]$ with $\min_{l\in[n]}|t_l - t_{l-1}| > \delta$ such that

$$\limsup_{N\to\infty} \mathbb{P}\Big(\max_{l\in[n]} \sup_{s,t\in[t_{l-1},t_l)} \Big\|\mathbf{x}^{(N)}(s) - \mathbf{x}^{(N)}(t)\Big\|_1 \geq \eta\Big) < \eta. \tag{70}$$

We first prove condition (69). From Lemma 23, for any fix $t \geq 0$ we have

$$\lim_{N\to\infty} \mathbb{P}\Big(x_{i,j}^{(N)}(t) \leq x_{i,j}^{(N)}(0),\ \forall i \geq L(t, \mathbf{x}(0)), j \in [M]\Big) = 1.$$

Now it can be easily verified that the sequence $(\mathbf{x}^{(N)}(0))_{N\geq 1}$ is tight in S. Therefore, using previous condition we can write for any $\epsilon > 0$

$$\lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(t) > \epsilon\Big) \leq \lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big(\max_{j\in[M]} \sum_{i\geq l} x_{i,j}^{(N)}(0) > \epsilon\Big) = 0. \tag{71}$$

Hence, from Lemma 22 the sequence $(\mathbf{x}^{(N)}(t))_{N\geq 1}$ is tight in S. This implies that the condition (69) is satisfied. Next for any $t_1 < t_2$, we consider

$$|x_{i,j}^{(N)}(t_1) - x_{i,j}^{(N)}(t_2)| \leq \frac{\lambda}{\gamma_j} \beta^{(N)}([t_1, t_2] \times R_{i,j}) + \mu_j \int_{t_1}^{t_2} (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds$$
$$+ \frac{1}{N\gamma_j} |M_{i,j}^{(A,N)}(t_1) - M_{i,j}^{(D,N)}(t_1) - M_{i,j}^{(A,N)}(t_2) + M_{i,j}^{(D,N)}(t_2)| + o(1),\ i \geq 1, j \in [M].$$

Using the above equation we can write ℓ_1 distance between $\mathbf{x}^{(N)}(t_1)$ and $\mathbf{x}^{(N)}(t_2)$ as

$$\begin{aligned}
&\left\|\mathbf{x}^{(N)}(t_1) - \mathbf{x}^{(N)}(t_2)\right\|_1 \\
&\leq \max_{j\in[M]} \sum_{i\geq 1} \frac{\lambda}{\gamma_j} \beta^{(N)}([t_1, t_2] \times R_{i,j}) + \max_{j\in[M]} \sum_{i\geq 1} \int_{t_1}^{t_2} \mu_j (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s)) ds \\
&+ \max_{j\in[M]} \frac{1}{N\gamma_j} \sum_{i\geq 1} |M_{i,j}^{(A,N)}(t_1) - M_{i,j}^{(D,N)}(t_1) - M_{i,j}^{(A,N)}(t_2) + M_{i,j}^{(D,N)}(t_2)| + o(1) \\
&\leq \frac{\lambda}{\gamma_{min}}(t_2 - t_1) + \max_{j\in[M]} \int_{t_1}^{t_2} \mu_j x_{1,j}^{(N)}(s) ds + \max_{j\in[M]} \frac{1}{N\gamma_j} \sum_{i\geq 1} |M_{i,j}^{(A,N)}(t_1) - M_{i,j}^{(D,N)}(t_1)(s) ds \\
&- M_{i,j}^{(A,N)}(t_2) + M_{i,j}^{(D,N)}(t_2)| + o(1) \\
&\leq (\frac{\lambda}{\gamma_{min}} + \mu_1)(t_2 - t_1) + \max_{j\in[M]} \frac{1}{N\gamma_j} \sum_{i\geq 1} |M_{i,j}^{(A,N)}(t_1) - M_{i,j}^{(D,N)}(t_1) - M_{i,j}^{(A,N)}(t_2) + M_{i,j}^{(D,N)}(t_2)| + o(1),
\end{aligned} \tag{72}$$

where $\gamma_{min} = \min_{j\in[M]} \gamma_j$. From Lemma 20, the martingale part in (72) converges to 0 as $N \to \infty$. Moreover, from (72), it implies that for any finite partition $\{t_1, t_2, \ldots, t_n\}$ of $[0, T]$ with $\min_{l\in[n]}|t_l - t_{l-1}| > \delta$, we have

$$\max_{l\in[n]} \sup_{s,t\in[t_{l-1},t_l)} \left\|\mathbf{x}^{(N)}(s) - \mathbf{x}^{(N)}(t)\right\|_1 \leq (\frac{\lambda}{\gamma_{min}} + \mu_1) \max_{l\in[n]}(t_l - t_{l-1}) + \zeta^{(N)},$$

where $\mathbb{P}(\zeta^{(N)} > \frac{\eta}{2}) < \eta$ for all sufficiently large N. Take $\delta = \eta/(4(\frac{\lambda}{\gamma_{min}} + \mu_1))$ and any partition with $\max_{l\in[n]}(t_l - t_{l-1}) < \eta/(2(\frac{\lambda}{\gamma_{min}} + \mu_1))$ and $\min_{l\in[n]}|t_l - t_{l-1}| > \delta$, we have

$$\max_{l\in[n]} \sup_{s,t\in[t_{l-1},t_l)} \left\|\mathbf{x}^{(N)}(s) - \mathbf{x}^{(N)}(t)\right\|_1 \leq \eta,$$

on the event $\{\zeta^{(N)} \leq \frac{\eta}{2}\}$. Therefore, for sufficiently large N we obtain

$$\mathbb{P}\Big(\max_{l\in[n]} \sup_{s,t\in[t_{l-1},t_l)} \left\|\mathbf{x}^{(N)}(s) - \mathbf{x}^{(N)}(t)\right\|_1 \geq \eta\Big) \leq \mathbb{P}(\zeta^{(N)} > \frac{\eta}{2}) < \eta.$$

Hence the condition (70) is satisfied. Next, we show that the limit $(\mathbf{x}, \beta)$ of any convergent subsequence of the sequence $((\mathbf{x}^{(N)}, \beta^{(N)}))_{N\geq 1}$ satisfies (36). We first show that the right side of (35) is a continuous map and then the result follows from an application of continuous mapping theorem. Consider,

$$\begin{aligned}
W_{i,j}\Big(\mathbf{x}^{(N)}(t), \beta^{(N)}, \mathbf{x}^{(N)}(0), \mathbf{m}^{(N)}\Big)(t) &= x_{i,j}^{(N)}(0) + \frac{\lambda}{\gamma_j} \beta^{(N)}([0, t] \times R_{i,j}) \\
&- \mu_j \int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s)) ds + m_{i,j}^{(N)}(t) \ i \geq 1, \ j \in [M],
\end{aligned}$$

where $m_{i,j}^{(N)}(t) = \frac{1}{N\gamma_j}(M_{i,j}^{(A,N)}(t) - M_{i,j}^{(D,N)}(t))$. Now we next prove that the map $\mathbf{W} = (W_{i,j})_{i,j}$ is continuous. First, assume that the sequence $((\mathbf{x}^{(N)}, \mathbf{m}^N))_{N\geq 1}$ converges to $(\mathbf{x}, \mathbf{m})$, then there exists a $N_1 \in \mathbb{N}$ such that $\sup_{t\in[0,T]} \left\|\mathbf{x}^{(N)}(t) - \mathbf{x}(t)\right\|_1 < \epsilon/(4T)$ for all $N \geq N_1$. Therefore, we can write

$$\sup_{t\in[0,T]} \int_0^t |x_{1,j}^{(N)}(t) - x_{1,j}(t)| ds \leq T \sup_{t\in[0,T]} \left\|\mathbf{x}^{(N)}(t) - \mathbf{x}(t)\right\|_1 < \epsilon/4.$$

Also, there exists a $N_2 \in \mathbb{N}$ such that $\sup_{t\in[0,T]} \left\|\mathbf{m}^{(N)}(t) - \mathbf{m}(t)\right\|_1 < \epsilon/4$. Second, assume that the sequence $(\mathbf{x}^{(N)}(0))$ converges to $\mathbf{x}(0)$ with respect to ℓ_1. Therefore, there exists a $N_3 \in \mathbb{N}$ such that $\left\|\mathbf{x}^{(N)}(0) - \mathbf{x}(0)\right\|_1 < \epsilon/4$. Now we claim that there exists a $N_4 \in \mathbb{N}$ such that

$$\frac{\lambda}{\gamma_j} \max_{j\in[M]} \sum_{i\geq 1} |\beta^{(N)}([0,T] \times R_{i,j}) - \beta([0,T] \times R_{i,j})| \leq \epsilon/4. \tag{73}$$

The equation (73) implies that the convergence of the sequence $((\beta^{(N)}([0,t] \times R_{i,j}))_{i,j})_{N\geq 1}$ for any $t \geq 0$ is ℓ_1 convergence. However, we know only the weak convergence of the sequence of measures $(\beta^{(N)})_{N\geq 1}$, which does not directly implies (73). Therefore, we show using weak convergence of the sequence $(\beta^{(N)})_{N\geq 1}$ that (73) is indeed true in our case. Since $\mathbf{x}(0) \in S$, there exists a $m'_j(\mathbf{x}(0))$ for all $j \in [M]$ such that $x_{i,j}(0) < 1$ for all $i \geq m'_j(\mathbf{x}(0))$. Furthermore, from Lemma 23 we can write

$$\lim_{N\to\infty} \mathbb{P}\Big(\sup_{t\in[0,T]} x_{i,j}^{(N)}(t) \leq x_{i,j}^{(N)}(0),\ \forall i \geq L(T, \mathbf{x}(0)), j \in [M]\Big) = 1.$$

Therefore, for $N' = \max\left\{m'_j(\mathbf{x}(0)), L(T, \mathbf{x}(0))\right\}$ we have

$$\lim_{N\to\infty} \mathbb{P}\Big(\sup_{t\in[0,T]} x_{i,j}^{(N)}(t) \leq 1,\ \forall i \geq N', j \in [M]\Big) = 1. \tag{74}$$

Using (74) and (33) we get

$$\lim_{N\to\infty} \max_{j\in[M]} \sum_{i\geq N'} \beta^{(N)}([0,T] \times R_{i,j}) = \max_{j\in[M]} \sum_{i\geq N'} \beta([0,T] \times R_{i,j}) = 0. \tag{75}$$

Also, weak convergence of $(\beta^{(N)})_N$ implies that

$$\lim_{N\to\infty} \max_{j\in[M]} \sum_{i<N'} \beta^{(N)}([0,T] \times R_{i,j}) = \max_{j\in[M]} \sum_{i<N'} \beta([0,T] \times R_{i,j}). \tag{76}$$

Hence, using (75) and (76) we get the desired result that is (73). Let $\bar{N} = \max\{N_1, N_2, N_3, N_4\}$, then we have

$$\sup_{t\in[0,T]} \left\|\mathbf{W}\Big(\mathbf{x}^{(N)}(t), \beta^{(N)}, \mathbf{x}^{(N)}(0), \mathbf{m}^{(N)}\Big) - \mathbf{W}\Big(\mathbf{x}(t), \beta, \mathbf{x}(0), \mathbf{m}\Big)\right\|_1 (t) < \epsilon.$$

This shows that the map $\mathbf{W}$ is continuous, which completes the proof.

G Proof of Lemma 15

To prove positive recurrent of the Markov chain $\mathbf{U}$, we consider the Lyapunov function $f : \mathbb{Z}_+^K \to [0, \infty)$ as

$$f(\mathbf{R}) = \sum_{j \in [K]} R_j, \ \mathbf{R} \in \mathbb{Z}_+^K.$$

Moreover, from (43) we can write the generator $G_{\mathbf{U}}$ acting on the function f for a state $\mathbf{R} \in \mathbb{Z}_+^K$ as

$$G_{\mathbf{U}} f(\mathbf{R}) = \sum_{j \in [K]} \nu_j - \lambda \sum_{j \in [K]} \mathbb{1}(0 = R_1 = \cdots = R_{j-1} < R_j).$$

Define the set $B = \left\{\mathbf{S} \in \mathbb{Z}_+^K : S_i = 0 \forall i \in [K]\right\}$. Observe that if $\sum_{j \in [K]} \nu_j < \lambda$, then $G_{\mathbf{U}} f(\mathbf{R}) < 0$ for all $\mathbf{R} \in B^c$, where B^c is the compliment of the set B and is given by

$$B^c = \left\{\mathbf{S} \in \mathbb{Z}_+^K : \mathbf{S} \in \bigcup_{i=1}^K \{0 = S_1 = \cdots = S_{i-1} < S_i\}\right\}.$$

Otherwise $G_{\mathbf{U}} f(\mathbf{R}) < \sum_{j \in [K]} \nu_j$ for $\mathbf{R} \in B$. Hence, using the Foster-Lyapunov criterion for positive recurrence from [32], we conclude that the chain $\mathbf{U}$ is positive recurrent if $\sum_{j \in [K]} \nu_j < \lambda$.

Next, we proceed to prove that the unique stationary distribution π of the chain $\mathbf{U}$ satisfies (44). We prove this using induction on K. First observer that $\pi(U_1 > 0) = \nu_1/\lambda$ if the birth-death process corresponding to the component U_1 is stable. This proves the base case, i.e., $K = 1$. Now for the induction hypothesis assume that (44) is true for all $j \in [K-1]$. Note that the positive recurrence of the chain $\mathbf{U}$ implies that $\mathbb{E} f(\mathbf{U}) < \infty$, where the expectation is with respect to the stationary distribution π. Therefore, we can set the steady-state expected drift to 0 which gives

$$\mathbb{E}[G_{\mathbf{U}} f(\mathbf{U})] = \sum_{j \in [K]} \nu_j - \lambda \sum_{j \in [K]} \pi(0 = U_1 = \cdots = U_{j-1} < U_j) = 0. \tag{77}$$

Hence, using the induction hypothesis for all $j \in [K-1]$ in the above expression we get

$$\pi(0 = U_1 = \cdots = U_{K-1} < U_K) = \frac{\nu_K}{\lambda},$$

which completes the proof for (44).

Now we show that the condition $\sum_{j \in [K]} \nu_j < \lambda$ is necessary for the chain $\mathbf{U}$ to be positive recurrent. Suppose the chain $\mathbf{U}$ is positive recurrent and, therefore, it has a unique stationary measure π. From (77) it is clear that π must satisfy (44). Therefore, we can write

$$\pi\left\{U_i = 0 \ \forall i \in [K]\right\} = 1 - \sum_{i=1}^K \pi\left\{0 = U_1 = \cdots = U_{i-1} < U_i\right\} = 1 - \sum_{i=1}^K \frac{\nu_i}{\lambda} > 0,$$

where the last inequality follows since the chain $\mathbf{U}$ is positive recurrent. ■

H Proof of Theorem 11

Proof of Theorem 11.(ii): To prove this, we need the following lemma which extends the monotonicity of the process $\mathbf{x}^{(N)}$ for finite N to the monotonicity of the limiting process $\mathbf{x}$.

Lemma 24. *Let* $\mathbf{x}(\cdot, \mathbf{u}) = (\mathbf{x}(t, \mathbf{u}), t \geq 0)$ *denote a solution to* (10) *with* $\mathbf{x}(0) = \mathbf{u} \in S$. *Then, for any* $\mathbf{u}, \mathbf{v} \in S$ *satisfying* $\mathbf{u} \leq \mathbf{v}$ *we have* $\mathbf{x}(t, \mathbf{u}) \leq \mathbf{x}(t, \mathbf{v})$ *for all* $t \geq 0$.

Proof. First, note that for any $\mathbf{u} \in S$, there exists a sequence $\left(\mathbf{u}^{(N)}\right)_{N \geq 1}$ with $\mathbf{u}^{(N)} = (u_{i,j}^{(N)}, i \geq 1, j \in [M]) \in S \cap S^{(N)}$ such that $\left\|\mathbf{u}^{(N)} - \mathbf{u}\right\|_1 \to 0$ as $N \to \infty$. We can simply construct such a sequence by setting $u_{i,j}^{(N)} = \frac{\lfloor u_{i,j} N \gamma_j \rfloor}{N \gamma_j}$ for each $i \geq 1$ and $j \in [M]$. This construction also satisfies the property that if $\mathbf{u}, \mathbf{v} \in S$ are such that $\mathbf{u} \leq \mathbf{v}$ and if the sequences $(\mathbf{u}^{(N)})_N$ and $(\mathbf{v}^{(N)})_N$ are constructed from their corresponding limits $\mathbf{u}$ and $\mathbf{v}$ as described above, then $\mathbf{u}^{(N)} \leq \mathbf{v}^{(N)}$ for all N.

Let $\mathbf{x}^{(N)}(\cdot, \mathbf{u}^{(N)}) = (\mathbf{x}^{(N)}(t, \mathbf{u}^{(N)}), t \geq 0)$ denote the process $\mathbf{x}^{(N)}$ started at $\mathbf{x}^{(N)}(0) = \mathbf{u}^{(N)}$. Then by Theorem 9 we have that

$$\mathbf{x}^{(N)}(t, \mathbf{u}^{(N)}) \leq \mathbf{x}^{(N)}(t, \mathbf{v}^{(N)}), \forall t \geq 0. \tag{78}$$

Now letting $N \to \infty$ and applying Theorem 10 gives the desired result. ∎

For $\mathbf{u} \in S$, we define $v_{n,j}(t, \mathbf{u}) = \sum_{i \geq n} x_{i,j}(t, \mathbf{u})$ and $v_{n,j}(\mathbf{u}) = \sum_{i \geq n} u_{i,j}$ for each $n \geq 1$ and $j \in [M]$. Furthermore, let $v_n(t, \mathbf{u}) = \sum_{j \in [M]} \gamma_j v_{n,j}(t, \mathbf{u})$ and $v_n(\mathbf{u}) = \sum_{j \in [M]} \gamma_j v_{n,j}(\mathbf{u})$ for each $n \geq 1$ and $\mathbf{u} \in S$.

Lemma 25. *For* $\mathbf{u} \in S$ *let* $\mathbf{x}(\mathbf{u}, \cdot)$ *denote a solution of* (10) *in* S. *Then for all* $t \geq 0$ *we have*

$$\frac{dv_n(t, \mathbf{u})}{dt} = \lambda \sum_{j \in [M]} \sum_{i \geq n} p_{i-1,j}(\mathbf{x}(t, \mathbf{u})) - \sum_{j \in [M]} \mu_j \gamma_j x_{n,j}(t, \mathbf{u}), \ \forall n \geq 1. \tag{79}$$

In particular, we have

$$\frac{dv_1(t, \mathbf{u})}{dt} = \lambda - \sum_{j \in [M]} \mu_j \gamma_j x_{1,j}(t, \mathbf{u}). \tag{80}$$

Proof. We can write the differential form of (10) as

$$\frac{dx_{i,j}(t, \mathbf{u})}{dt} = \frac{\lambda}{\gamma_j} p_{i-1,j}(\mathbf{x}(t, \mathbf{u})) - \mu_j (x_{i,j}(t, \mathbf{u}) - x_{i+1,j}(t, \mathbf{u})). \tag{81}$$

Multiplying the above by γ_j and summing first over $i \geq n$ and then over $j \in [M]$ we obtain (79). Equation (80) follows from (79) by using P1. ∎

From Lemma 24 it follows that for any $\mathbf{x}(0) \in S$ and any $t \geq 0$ we have

$$\mathbf{x}(t, \min(\mathbf{x}(0), \mathbf{x}^*)) \leq \mathbf{x}(t, \mathbf{x}(0)) \leq \mathbf{x}(t, \max(\mathbf{x}(0), \mathbf{x}^*)), \tag{82}$$

where $\min(\mathbf{u}, \mathbf{v})$ with $\mathbf{u}, \mathbf{v} \in S$ is defined by taking the component-wise minimum. Hence, to prove global stability it is sufficient to show that the (component-wise) convergence $\mathbf{x}(t, \mathbf{x}(0)) \to \mathbf{x}^*$ holds for initial states satisfying either of the following two conditions: (i) $\mathbf{x}(0) \geq \mathbf{x}^*$ and (ii) $\mathbf{x}(0) \leq \mathbf{x}^*$.

To prove the above, we first show that for any solution $\mathbf{x}(\cdot, \mathbf{x}(0)) \in S$, $v_n(t, \mathbf{x}(0))$ is uniformly bounded in t for all $n \geq 1$. Consider the case when $\mathbf{x}(0) \geq \mathbf{x}^*$. From Lemma 24 it follows that for $\mathbf{x}(0) \geq \mathbf{x}^*$, we have $\mathbf{x}(t, \mathbf{x}(0)) \geq \mathbf{x}^*$ for all $t \geq 0$. Therefore, we can write

$$\sum_{j \in [M]} \gamma_j \mu_j x_{1,j}(t, \mathbf{x}(0)) \geq \sum_{j \in [M]} \gamma_j \mu_j x^*_{1,j} = \lambda,$$

where the last equality follows from (46). Hence, from (80) we have $\frac{dv_1(t,\mathbf{x}(0))}{dt} \leq 0$ from which it follows that $0 \leq v_1(t, \mathbf{x}(0)) \leq v_1(\mathbf{x}(0))$ for all $t \geq 0$. Since the sequence $(v_n(t, \mathbf{x}(0)))_{n \geq 1}$ is non-increasing, we have $0 \leq v_n(t, \mathbf{x}(0)) \leq v_1(\mathbf{x}(0))$ for all $n \geq 1$ and for all $t \geq 0$. This proves that $v_n(t, \mathbf{x}(0))$ is uniformly bounded in t for each $n \geq 1$ if $\mathbf{x}(0) \geq \mathbf{x}^*$. Now consider the case $\mathbf{x}(0) \leq \mathbf{x}^*$. From Lemma 24 it follows that for $\mathbf{x}(0) \leq \mathbf{x}^*$, we have $\mathbf{x}(t, \mathbf{x}(0)) \leq \mathbf{x}^*$ for all $t \geq 0$. Therefore, we have $v_1(t, \mathbf{x}(0)) \leq v_1(\mathbf{x}^*)$ for all $t \geq 0$. This shows that the component $v_n(t, \mathbf{x}(0))$ is uniformly bounded in t for each $n \geq 1$ for $\mathbf{x}(0) \leq \mathbf{x}^*$.

Since $v_n(t, \mathbf{x}(0))$ is uniformly bounded in t, the convergence $x_{i,j}(t, \mathbf{x}(0)) \to x^*_{i,j}$ for all $i \geq 1$ and for all $j \in [M]$ will follow from

$$\int_0^\infty (x_{i,j}(t, \mathbf{x}(0)) - x^*_{i,j}) dt < \infty, \ \forall j \in [M], \ \forall i \geq 1, \tag{83}$$

for the case $\mathbf{x}(0) \geq \mathbf{x}^*$ and from

$$\int_0^\infty (x^*_{i,j} - x_{i,j}(t, \mathbf{x}(0))) dt < \infty, \ j \in [M], \ i \geq 1, \tag{84}$$

for the case $\mathbf{x}(0) \leq \mathbf{x}^*$. We now prove (83) to show convergence for the case $\mathbf{x}(0) \geq \mathbf{x}^*$; the proof of other case follows similarly. To show (83) it is sufficient to prove that

$$\int_0^\infty \sum_{j \in [M]} \mu_j \gamma_j (x_{i,j}(t, \mathbf{x}(0)) - x^*_{i,j}) dt < \infty, \ \forall i \geq 1. \tag{85}$$

For $i = 1$, we can write (85) as

$$\begin{aligned}
\int_0^\tau \Big(\sum_{j \in [M]} \mu_j \gamma_j x_{1,j}(t, \mathbf{x}(0)) - \sum_{j \in [M]} \mu_j \gamma_j x^*_{1,j} \Big) dt &= \int_0^\tau \Big(\sum_{j \in [M]} \mu_j \gamma_j x_{1,j}(t, \mathbf{x}(0)) - \lambda \Big) \\
&= - \int_0^\tau \frac{dv_1(t, \mathbf{x}(0))}{dt} dt \\
&= v_1(\mathbf{x}(0)) - v_1(\tau, \mathbf{x}(0)) \\
&\leq v_1(\mathbf{x}(0)),
\end{aligned}$$

where the first equality follows from (46), second equality follows from (80), and the last inequality follows as $v_1(t, \mathbf{x}(0))$ is uniformly bounded in t. Since the right hand side is bounded by a constant for all τ , the integral on the left hand side must converge as $\tau \to \infty$. This shows that $x_{1,j}(t, \mathbf{x}(0)) \to x^*_{1,j}$ for all $j \in [M]$ as $t \to \infty$.

Now for $i = 2$, we can write (85) as

$$\begin{aligned}\int_0^\tau \Big(\sum_{j\in[M]} \mu_j\gamma_j x_{2,j}(t, \mathbf{x}(0)) - \sum_{j\in[M]} \mu_j\gamma_j x^*_{2,j}\Big)dt &= \int_0^\tau \sum_{j\in[M]} \mu_j\gamma_j x_{2,j}(t, \mathbf{x}(0))dt \\ &= -\int_0^\tau \Big(\frac{dv_2(t, \mathbf{x}(0))}{dt} - \lambda \sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0)))\Big)dt, \\ &= \lambda \int_0^\tau \sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0)))dt + v_1(\mathbf{x}(0)) - v_1(\tau, \mathbf{x}(0)) \\ &\leq \lambda \int_0^\tau \sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0)))dt + v_1(\mathbf{x}(0)),\end{aligned} \tag{86}$$

where the first equality follows as $x^*_{2,j} = 0$ for all $j \in [M]$ and the second equality follows from (79) for $n = 2$. From the convergence $x_{1,j}(t, \mathbf{x}(0)) \to x^*_{1,j}$ for all $j \in [M]$, we know that for all $\epsilon > 0$ there exists a $t^\epsilon > 0$ such that $(x_{1,j}(t, \mathbf{x}(0)) - x^*_{1,j}) < \epsilon$ for all $t \geq t^\epsilon$ and for all $j \in [M]$. Furthermore, from (11) it follows that for $\lambda < 1$ there exists $j \in [M]$ such that $x^*_{1,j} < 1$. Hence, we can choose $\epsilon < 1 - x^*_{1,j}$ which yields

$$x_{1,j}(t, \mathbf{x}(0)) < \epsilon + x^*_{1,j} < 1, \ \forall t \geq t^\epsilon_0.$$

Hence, by properties P1, P4, and P5 of the fluid limit we have $\sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0))) = 0$ for all $t \geq t^\epsilon_0$. Therefore, for all $\tau \geq t^\epsilon$, we can write (86) as

$$\int_0^\tau \sum_{j\in[M]} \mu_j\gamma_j x_{2,j}(t, \mathbf{x}(0))dt \leq \lambda \int_0^{t^\epsilon} \sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0)))dt + v_1(\mathbf{x}(0)) \leq \lambda t^\epsilon + v_1(\mathbf{x}(0)),$$

where in the last inequality we have used the fact that $\sum_{j\in[M]}\sum_{i\geq 2} p_{i-1,j}(\mathbf{x}(t, \mathbf{x}(0))) \leq 1$ from P1. Since the right hand side of the above expression is independent of τ, the left hand side must converge as $\tau \to \infty$. This shows that $x_{2,j}(t, \mathbf{x}(0)) \to x^*_{2,j} = 0$ as $t \to \infty$ for all $j \in [M]$.

Finally, since $\mathbf{x}(t, \mathbf{x}(0)) \in S$ we have $x_{2,j}(t, \mathbf{x}(0)) \geq x_{i,j}(t, \mathbf{x}(0)) \geq 0$ for all $i \geq 3$, $j \in [M]$ and for all $t \geq 0$. Hence, $x_{i,j}(t, \mathbf{x}(0)) \to 0$ as $t \to \infty$ for all $i \geq 3$ and $j \in [M]$.

Proof of Theorem 11.(iii): We first recall from Proposition 13 that the sequence $(\mathbf{x}^{(N)}(\infty))_N$ is tight in S under the ℓ_1-norm. Hence, by Prohorov's theorem, the sequence has convergent subsequences with limits in S. Thus, it suffices to show that all convergent subsequences has the same limit point $\mathbf{x}^*$. Let $(\mathbf{x}^{(N_k)}(\infty))_k$ be any such convergent subsequence of the sequence $(\mathbf{x}^{(N)}(\infty))_N$ with limit point $\mathbf{x}^{**} \in S$. Now, from Theorem 10, we have that the distribution of $\mathbf{x}^{**}$ must be invariant under the map $\mathbf{u} \mapsto \mathbf{x}(t, \mathbf{u})$. By the global stability result proved earlier it follows that the only measure invariant under the map $\mathbf{u} \mapsto \mathbf{x}(t, \mathbf{u})$ is δ_{x^*}. Hence, we must have $\mathbf{x}^{**} = \mathbf{x}^*$.

I Unique Characterisation of Stationary Distribution from Given State

The main idea of the proof is to show that the transition rates of the chain $(V_{l_1(\mathbf{x}),1}, V_{l_2(\mathbf{x}),2}, \dots, V_{l_M(\mathbf{x}),M})$ defined in (42) have a form similar to that of the transition rates of the Markov chain defined in (43). The proof then follows by the application of Lemma 15.

Consider the chain

$$\mathbf{C} = (V_{l_1(\mathbf{x}),1}, V_{l_2(\mathbf{x}),2}, \dots, V_{l_M(\mathbf{x}),M}).$$

We first reduce the dimension of the chain above and then rearrange the remaining components by performing the following steps sequentially:

1. **Step 1**: We first note from (42) that the stationary rate of the transition $V_{l_j(\mathbf{x}),j} \to V_{l_j(\mathbf{x}),j} - 1$ is zero (whereas the rate of the transition $V_{l_j(\mathbf{x}),j} \to V_{l_j(\mathbf{x}),j} + 1$ is strictly positive) for any component $j \in [M]$ for which either (i) $l_j(\mathbf{x}) > l_i(\mathbf{x})$ for some $i < j$ or (ii) $l_j(\mathbf{x}) - 1 > l_i(\mathbf{x})$ for some $i \neq j$. This is because for pairs $i, j \in [M]$ satisfying any of the above two conditions we have $\pi_{\mathbf{x}}(V_{l_j(\mathbf{x}),i} = \infty) = 1$ and if $V_{l_j(\mathbf{x}),i} = \infty$, then from the definition of $R_{i,j}$ we have $\mathbb{1}\Big(\mathbf{V} \in R_{l_j(\mathbf{x}),j}\Big) = 0$. Therefore, for each $j \in [M]$ which satisfies one of the above two conditions we have $\pi_{\mathbf{x}}(V_{l_j(\mathbf{x}),j} = \infty) = 1$. Hence, in the first step, we remove all these components from the chain $\mathbf{C}$.

2. **Step 2**: Next, we arrange the remaining components of the chain $\mathbf{C}$ in the increasing order of their corresponding minimum queue lengths (i.e., increasing order of $l_j(\mathbf{x})$). The components having the same minimum queue length (i.e., the same value of $l_j(\mathbf{x})$) are then arranged in the decreasing order of their service rates. It can then be easily verified from the definition of $R_{i,j}$ that the remaining components have transition rates of the form given by (43).

Let $\mathbf{Y} = (Y_{m'_1,p_1}, Y_{m'_2,p_2}, \dots, Y_{m'_H,p_H})$ denote the chain obtained after performing the steps mentioned above, where for each $i \in [H]$, $Y_{m'_i,p_i}$ is the leftover component of the chain $\mathbf{C}$ with minimum queue length m'_i in pool p_i. Note that $m'_i \leq m'_{i+1}$ and $p_i \leq p_{i+1}$ for all $i \in [H]$. Moreover, the transition rate from $Y_{m'_i,p_i} \to Y_{m'_i,p_i} + 1$ is given by $\nu_i = \gamma_{p_i}\mu_{p_i}(x_{m'_i,p_i} - x_{m'_i+1,p_i})$ and the rate of transition from $Y_{m'_i,p_i} \to Y_{m'_i,p_i} - 1$ is $\lambda\mathbb{1}\Big(0 = Y_{m'_1,p_1} = \cdots = Y_{m'_{i-1},p_{i-1}} < Y_{m'_i,p_i}\Big)$. Now it remains to verify that $\pi_{\mathbf{x}}(V_{m'_i,p_i} = \infty) = 1$ or 0 for each $i \in [H]$ using Lemma 15.

Define $\rho_i = \frac{\nu_i}{\lambda}$ for each $i \in [H]$ and $\rho_{H+1} = 0$. To show that $\pi_{\mathbf{x}}(V_{m'_i,p_i} = \infty) = 1$ or 0 for each $i \in [H]$ it is suffices to show that for each $0 \leq L \leq H$ if $\sum_{i=1}^{L} \rho_i < 1$ and $\sum_{i=1}^{L+1} \rho_i \geq 1$

then we have

$$\pi_{\mathbf{x}}(V_{m'_1,p_1} = \infty) = \cdots = \pi_{\mathbf{x}}(V_{m'_L,p_L} = \infty) = 0,$$
$$\pi_{\mathbf{x}}(V_{m'_{L+1},p_{L+1}} = \infty) = \cdots = \pi_{\mathbf{x}}(V_{m'_H,p_H} = \infty) = 1.$$

Suppose $\sum_{i=1}^{L} \rho_i < 1$ and $\sum_{i=1}^{L+1} \rho_i \geq 1$. Now we use contradiction to prove that $\pi_{\mathbf{x}}(Y_{m'_i,p_i} = \infty) = 0$ for all $i \in [L]$. Assume $\pi_{\mathbf{x}}(Y_{m'_i,p_i} = \infty) = \epsilon \in (0,1]$ for $i \in [L]$. Also, assume $\bar{\pi}_{\mathbf{x}}$ to be the unique stationary distribution of $(Y_{m'_1,p_1}, \ldots, Y_{m'_i,p_i})$ given that $Y_{m'_r,p_r}$ is finite for all $r \in [i]$. Note that the stationary measure $\bar{\pi}_{\mathbf{x}}$ is same as the stationary measure π given in Lemma 15. Therefore, we can write

$$\begin{aligned}\pi_{\mathbf{x}}(R_{m'_i,p_i}) &= (1-\epsilon)\bar{\pi}_{\mathbf{x}}(0 = Y_{m'_1,p_1} = \cdots = Y_{m'_{i-1},p_{i-1}} < Y_{m'_i,p_i}) + \epsilon \\ &= (1-\epsilon)\frac{\nu_i}{\lambda} + \epsilon,\end{aligned}$$

where we use $\bar{\pi}_{\mathbf{x}}(0 = Y_{m'_1,p_1} = \cdots = Y_{m'_{i-1},p_{i-1}} < Y_{m'_i,p_i}) = \frac{\nu_i}{\lambda}$ from (44). Now substituting the above in the differential form of (10) we obtain

$$\begin{aligned}\frac{dx_{m'_i,p_i}(t)}{dt} &= \frac{\lambda}{\gamma_{p_1}}\Big((1-\epsilon)\frac{\nu_i}{\lambda} + \epsilon\Big) - \mu_{p_i}(x_{m'_1,p_1}(t) - x_{m'_i+1,p_i}(t)) \\ &= \epsilon\Big(\frac{\lambda}{\gamma_{p_i}} - \mu_{p_1}(x_{m'_i,p_i}(t) - x_{m'_i+1,p_i}(t))\Big) > 0,\end{aligned}$$

where the last step follows as $\frac{\nu_i}{\lambda} < 1$. Now, since $x_{m'_i,p_i}(t) = 1$ as m'_i is the minimum queue length in pool p_i, we must have $\frac{dx_{m'_i,p_i}(t)}{dt} < 0$ which leads to a contradiction. Therefore, we have $\pi_{\mathbf{x}}(Y_{m'_i,p_i} = \infty) = 0$ for all $i \in [L]$. Now observe that $\sum_{i=1}^{L+1} \rho_i \geq 1$ implies $\rho_{L+1} \geq 1 - \sum_{i=1}^{L} \rho_i$. Therefore, using Lemma 15, we know that the chain $(Y_{m'_1,p_1}, \ldots, Y_{m'_{L+1},p_{L+1}})$ is unstable. Moreover, again from Lemma 15, $\sum_{i=1}^{L} \rho_i < 1$ insures that the chain $(Y_{m'_1,p_1}, \ldots, Y_{m'_L,p_L})$ is stable. Hence, we have $\bar{\pi}_{\mathbf{x}}(Y_{m'_{L+1},p_{L+1}} \geq l) = 1$ for all $l \geq 0$. This shows that $\pi_{\mathbf{x}}(Y_{m'_{L+1},p_{L+1}} = \infty) = 1$. Now observe that the rate of transition from $Y_{m'_i,p_i} \to Y_{m'_i,p_i} - 1$ is $\lambda \mathbb{1}\Big(0 = Y_{m'_1,p_1} = \cdots = Y_{m'_{i-1},p_{i-1}} < Y_{m'_i,p_i}\Big) = 0$ for each $i \in \{L+2, \ldots, H\}$. Hence, with only non-zero rate of transition from $Y_{m'_i,p_i} \to Y_{m'_i,p_i} + 1$ we have $\pi_{\mathbf{x}}(Y_{m'_i,p_i} = \infty) = 1$ for each $i \in \{L+2, \ldots, H\}$.