# Optimal Load Balancing in Heterogeneous Server Systems

Sanidhay Bhambay and Arpan Mukhopadhyay
Department of Computer Science
University of Warwick, UK
{sanidhay.bhambay,arpan.Mukhopadhyay}@warwick.ac.uk

***Abstract*—A fundamental problem in large-scale data centers is to reduce the average response time of jobs. The Join-the-Shortest-Queue (JSQ) load balancing scheme is known to minimise the average response time of jobs for homogeneous systems consisting of identical servers. However, heterogeneous systems consisting of servers with different speeds, JSQ performs poorly. Furthermore, JSQ suffers from high communication overhead as it requires knowledge of the queue length of all servers while assigning incoming jobs to one of the destination servers. Therefore, the Join-the-Idle-Queue (JIQ) scheme was introduced which not only reduces the communication overhead but also minimises the average response time of jobs under homogeneous systems. Despite these advantages, JIQ is still known to be inefficient in minimising the average response time of jobs under heterogeneous systems. In this paper, we consider a speed-aware version of JIQ for heterogeneous systems and show that it achieves delay optimality in the fluid limit. One of the technical challenges to establishing this optimality is to show the tightness of the sequence of steady-state distributions indexed by system size. We show this tightness result by evaluating the drift of appropriate Lyapunov functions. This approach to proving tightness is different from the usual coupling approach used for homogeneous systems. Another important challenge in proving the optimality result is to establish the fluid limit which is done using the time-scale separation technique. Finally, using the monotonicity of the fluid process we have shown that the fluid limit has a unique and globally attractive fixed point.**

***Index Terms*—fluid limit, load balancing, stochastic processes, heterogeneous systems, Lypunov functions.**

## I. Introduction

A key challenge for cloud computing service providers such as Google Colab, Amazon Web Services (AWS), and Google Cloud Platform (GCP) is to satisfy user's demand for resources with minimum delay; otherwise, the user drops, leading to a loss in revenue. To overcome this challenge efficient load balancing schemes, which assign incoming user requests to servers based on their current loads, are required. The most natural way of doing so is to use the JSQ scheme in which an incoming arrival is assigned to the server having smallest number of ongoing jobs. This scheme is known to achieve the minimum average response time of jobs under a variety of settings [1], [2]. However, it is difficult to implement JSQ in large-scale data centers as they contain hundreds of thousands of servers and finding the server with minimum number of ongoing jobs incurs significant communication overhead between the job dispatcher and the servers. Therefore, many alternative load balancing schemes which require low messaging overhead have been proposed in literature. The Power-of-d (Pod) scheme [3] and the JIQ scheme [4] are some examples of schemes which require less communication between the servers and the job dispatcher. In the JIQ scheme, the dispatcher only keeps track of the idle servers in the system and once a job arrives, it is sent to one of the idle servers. If no server is idle, then the job is sent to a randomly sampled server. It has been shown that for homogeneous systems consisting of identical servers the JIQ scheme achieves the same asymptotic performance as the JSQ scheme.

The above mentioned optimality of JIQ holds only when servers are identical. However, modern data centers are equipped with different generations of CPUs, various types of acceleration devices such as GPUs, FPGAs, and ASICs, with various processing speeds [5], [6]. Therefore, modern data centers are inherently heterogeneous and for such systems, JIQ performs poorly in terms of mean response time of jobs. An efficient load balancing scheme for heterogeneous systems must use all existing servers in the system. This is because not using servers with lower processing speeds results in the wastage of expensive resources that are already installed in the system. This motivates us to consider a speed-aware load balancing schemes for heterogeneous systems. In speed-aware schemes, in addition to the current loads at the servers, the server's speed are taken into consideration while assigning incoming jobs to servers. In addition to efficiency, we aim at designing a load balancing scheme which has low communication overhead so that it can be easily implemented. To this end, we consider a speed-aware variant of the JIQ scheme, called the Speed-Aware Join-the-Idle-Queue or SA-JIQ scheme, where instead of sending arrivals to any available idle server, arrivals are sent to the fastest of the idle servers. We show that this small modification leads to significant performance enhancement for JIQ in heterogeneous systems.

### A. Main Contributions

Our main contribution in this paper is the analysis of SA-JIQ scheme in the large system limit, i.e., as the number of server $N$ in the system tends to infinity. We show that the SA-JIQ scheme achieves the minimum possible average response time of jobs in heterogeneous systems asymptotically (as $N \to \infty$). Our specific contributions are listed below:

1) Stability and Tightness: We first show that the SA-JIQ scheme is stable for all arrival rate $\lambda < 1$. We prove

ISBN 978-3-903176-49-2 © 2022 IFIP

this using Lyapunov drift method. Moreover, we show that the sequence of steady-state stationary distribution indexed by system size is tight. We take a new approach different from usual coupling approach and establish this tightness result by analysing the drift of an exponential Lyapunov function.

2) Establishing the Fluid Limit: Our next main contribution is to establish the fluid limit of the SA-JIQ scheme. The analysis of the fluid limit is more challenging due to the underlying state space and the separation of two time scale. We use martingale representation and time-scale separation approach to establish the fluid limit of SA-JIQ.
3) Characterising the Steady-State of the Fluid Limit: Finally, we have shown that the fluid limit of SA-JIQ has a unique fixed point and we prove it is globally stable using monotonicity of the fluid process . Furthermore, using the tightness result and the global stability of fixed point we have also establish the interchange of limits.

### B. Related Works

The study of efficient load balancing schemes for large scale data centers has been a hot topic of research from last 2-3 decades [7]. The JSQ scheme is the most rigorously investigated scheme in which an arrival joins the minimum queue length server. It has been shown that JSQ is optimal in stochastic order sense [1], [2], and heavy-traffic optimal in [8]. To overcome the high messaging overhead in JSQ, alternative schemes have been proposed such as Pod [3] and JIQ [4]. All these schemes are known to perform well in homogeneous systems with identical servers.

Proving the optimality of load balancing schemes is more challenging in heterogeneous systems. There are few works considering load balancing in heterogeneous systems. The Pod scheme for heterogeneous systems has been analyzed in [9] for light traffic, in [10] for heavy traffic, and in [11] for mean-field regime. Furthermore, low-complexity scheme JIQ has also been analyzed for heterogeneous systems in [12]. It has been shown that JIQ has asymptotic zero wait time as $N \to \infty$. However, this does not imply that the JIQ scheme is asymptotically delay optimal. In [13], a scheme similar to the SA-JIQ scheme has been considered for constrained heterogeneous systems with finite buffer sizes due to which tightness and stability results follow immediately . Moreover, the drift technique, applicable to finite-buffer systems, is difficult to generalise to our setting.

### C. General Notations

We use the following notations throughout the paper. We denote $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+ \bigcup \{\infty\}$. For $x, y \in \mathbb{R}$, we use $x \wedge y$, and $(x)_+$ to denote $\max(x, y)$, and $\max(x, 0)$. 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 $A$.

## II. System Model

We consider a system consisting of $N$ parallel servers, each with its own queue of infinite buffer size. The servers are assumed to be *heterogeneous* in that they can have different service rates. Specifically, we assume that there are $M$ different server types. Each type $j \in [M]$ server has service rate $\mu_j$. 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). 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. The queues are served according to the First-Come-First-Server (FCFS) scheduling discipline.

### A. Assignment Policy

Our main interest is to analyse the *Speed-Aware Join-the-Idle-Queue* or the *SA-JIQ* policy. It is defined as follows. Under the SA-JIQ policy, upon arrival of a job, it is sent to a idle server available with maximum speed. Otherwise, a server is selected from the pool $j$ with probability $p_j/N\gamma_j$, where $p_j = \mu_j\gamma_j$ is the probability of selecting $j^{\text{th}}$ pool. Ties within servers of different types are broken by choosing the server type with the maximum speed and ties between servers of the same type are broken uniformly at random.

*Remark* 1. Note that if we select a server uniformly at random (as in classical JIQ) when no idle servers are available (we call this scheme as SA-JIQ Random), then from Figure 1, it is clear that the system under SA-JIQ Random is not even stable for all $\lambda < 1$. The choice of $p_j = \gamma_j\mu_j$ is crucial in the analysis of stability and tightness. Specifically, with this choice of $p_j$, we show that the drift of appropriate Lyapunov is negative to prove stability and obtained uniform bound on stationary queue length distribution to prove tightness.

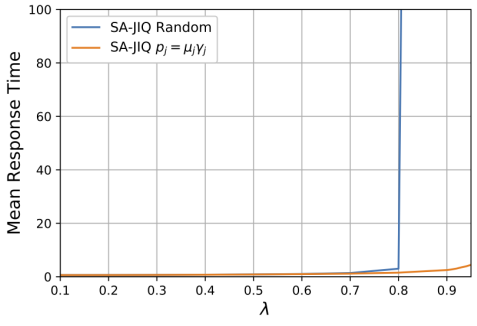


Figure 1. Both schemes are applied to a heterogeneous system consisting of two types of servers, i.e., $M = 2$. We choose $N = 10$, $\gamma_1 = 1 - \gamma_2 = 1/2$, and $\mu_1 = 4\mu_2 = 8/5$.

## III. Lower Bound on Mean Response Time of Jobs

To establish the optimality result of SA-JIQ, we need to find a lower bound on the steady state mean response time of jobs for system described in Section II for all $N$ and under any stationary job assignment policy $\Pi$. In the following proposition, we state a result from [14] which provides a lower bound on the steady-state mean response time, $\bar{T}_{N,\Pi}$, of jobs in system described in previous section under any stationary policy $\Pi$.

**Proposition 1** (Proposition 4 of [14])**.** *If $\lambda < 1$, then the steady-state mean response time, $\bar{T}_{N,\Pi}$, of jobs in the system described in Section II under any stationary policy $\Pi$ satisfies*

$$\liminf_{N\to\infty} \bar{T}_{N,\Pi} \geq \frac{z^*}{\lambda}, \tag{1}$$

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

In subsequent sections, we shall establish that the above lower bound can be achieved with equality when SA-JIQ is employed as the job assignment policy.

## IV. System State Description

The state of the system at any time $t \geq 0$ under the SA-JIQ 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 $k^{\text{th}}$ 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 for all $i \geq 1$, $j \in [M]$

$$x_{i,j}^{(N)}(t) = \frac{1}{N\gamma_j} \sum_{k\in[N\gamma_j]} \mathbb{1}\Big(Q_{k,j}^{(N)}(t) \geq i\Big). \tag{2}$$

We use both descriptors above to state and prove our results. 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 $S^{(N)} \triangleq \{\mathbf{s} = (s_{i,j}) : N\gamma_j s_{i,j} \in \mathbb{Z}_+, 1 \geq s_{i,j} \geq s_{i+1,j} \geq 0 \ \forall i \geq 1, j \in [M]\}$. 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

$$S \triangleq \{\mathbf{s} : 1 \geq s_{i,j} \geq s_{i+1,j} \geq 0, \forall i \geq 1, j \in [M], \|\mathbf{s}\|_1 < \infty\}, \tag{3}$$

where the $\ell_1$-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 $\ell_1$-norm.

## V. Main Results

In this section we state the main results for the SA-JIQ scheme. Suppose 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 the case when $\lambda < 1$.

**Theorem 2.** *The process $\mathbf{Q}^{(N)}$ is positive recurrent for each $\lambda < 1$ and each $N$. Furthermore, for each $j \in [M]$, each $k \in [N\gamma_j]$, and each $l \geq 1$ the following bound holds for all $\theta \in [0, -\log\lambda)$*

$$\sup_N \mathbb{P}(Q_{k,j}^{(N)}(\infty) \geq l) \leq C_j(\lambda,\theta) e^{-l\theta}, \tag{4}$$

*where $Q_{k,j}^{(N)}(\infty) = \lim_{t\to\infty} Q_{k,j}^{(N)}(t)$ and $C_j(\lambda,\theta) = (1-\lambda)/(\mu_j\gamma_j(1-\lambda e^{\theta})) > 0$.*

The theorem above implies that for $\lambda < 1$ the stationary distributions of $\mathbf{Q}^{(N)}$ and $\mathbf{x}^{(N)}$ exist and they are unique. Furthermore, the bound obtained in (4) is uniform in $N$ and essential in proving the tightness result. 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. $\mathbf{x}^{(N)}$). 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 $S$.

**Theorem 3.** *(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 $j \in [M]$*

$$x_{i,j}(t) = x_{i,j}(0) + \frac{\lambda}{\gamma_j} \int_0^t p_{i-1,j}(\mathbf{x}(u))du - \int_0^t \mu_j(x_{i,j}(u) - x_{i+1,j}(u))du, \tag{5}$$

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

$$p_{0,j}(\mathbf{s}) = \mathbb{1}(1 = s_{1,1} = \cdots = s_{1,j-1} > s_{1,j}), \tag{6}$$

$$p_{i-1,j}(\mathbf{s}) = (1 - p_0(\mathbf{s}))\mu_j\gamma_j(s_{i-1,j} - s_{i,j}), \ i \geq 2, \tag{7}$$

*where $p_0(\mathbf{s}) = \sum_{j\in[M]} p_{0,j}(\mathbf{s})$.*

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 $\mathbf{s} \in S$. Specifically, $p_{0,j}(\mathbf{s}) = 1$, when all servers in pools below $j$ are busy and there are some idle servers available in pool $j$. Furthermore, if there are no idle servers available, i.e., $p_0(\mathbf{s}) = 0$ then the pool $j$ is picked with probability $\mu_j\gamma_j$ and within pool $j$ an arrival joins the server with exactly $i-1$ jobs

with probability $(s_{i-1,j} - s_{i,j})$. Note that from (6) and (7) it can easily verified that $\sum_{j\in[M]}\sum_{i\geq 1} p_{i-1,j}(\mathbf{s}) = 1$, for any $\mathbf{s} \in S$. In our final result stated below, we show that the fixed point $\mathbf{x}^*$ is unique and globally attractive.

**Theorem 4.** *(i) If $\lambda < 1$, the fixed point $\mathbf{x}^* = (x^*_{i,j}, i \geq 1, j \in [M])$ of the fluid limit $\mathbf{x}$ described by* (5) *is unique in $S$ and is given by for all $j \in [M]$ as*

$$x^*_{1,j} = \Big(1 \wedge \frac{(\lambda - \sum_{k=1}^{j-1} \mu_k\gamma_k)_+}{\mu_j\gamma_j}\Big),\ x^*_{i,j} = 0 \quad i \geq 2. \tag{8}$$

*(ii) (Global Stability): If $\lambda < 1$, then for any $\mathbf{x}(0) \in S$ the fluid limit $\mathbf{x}$ given by* (5) *converges to $\mathbf{x}^*$ component-wise, i.e., $x_{i,j}(t, \mathbf{x}(0)) \to x^*_{i,j}$ 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 $N \to \infty$.*

It can be easily verified that the total scaled number of jobs in state $x^*$ is equal to $z^*$ as defined in (1). Thus, by Little's law, the mean response time of jobs under the SA-JIQ policy converges to $z^*/\lambda$ as $N \to \infty$, which implies the asymptotic optimality of the SA-JIQ policy.

## VI. Stability and Uniform Bounds

In this section we prove Theorem 2. To show the results of Theorem 2, we use Lypunov drift method. For any function $f : \mathbb{Z}^N_+ \to \mathbb{R}$, the drift evaluated at a state $\mathbf{Q} \in \mathbb{Z}^N_+$ is given by

$$G_{\mathbf{Q}^{(N)}} f(\mathbf{Q}) = \sum_{j\in[M]}\sum_{k\in[N\gamma_j]} [r^{+,N}_{k,j}(\mathbf{Q})(f(\mathbf{Q}+\mathbf{e}^{(N)}_{k,j}) - f(\mathbf{Q})) + r^{-,N}_{k,j}(\mathbf{Q})(f(\mathbf{Q} - \mathbf{e}^{(N)}_{k,j}) - f(\mathbf{Q}))], \tag{9}$$

where $G_{\mathbf{Q}^{(N)}}$ is the generator of $\mathbf{Q}^{(N)}$; $\mathbf{e}^{(N)}_{k,j}$ denotes the N-dimensional unit vector with one in the $(k,j)^{\text{th}}$ position; $r^{\pm,N}_{k,j}(\mathbf{Q})$ are the transition rates from the state $\mathbf{Q}$ to the states $\mathbf{Q} \pm \mathbf{e}^{(N)}_{k,j}$. Under the SA-JIQ policy, for each state $\mathbf{Q} \in \mathbb{Z}^N_+$ and each $k \in [N\gamma_j], j \in [M]$ we have

$$r^{+,N}_{k,j}(\mathbf{Q}) = \begin{cases} \frac{N\lambda}{|I_j(\mathbf{Q})|}, & \text{if } j = j^\dagger(\mathbf{Q}) \text{ and } k \in I_j(\mathbf{Q}) \\ \lambda\mu_j, & \text{if } \mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) \end{cases}, \tag{10}$$

$$r^{-,N}_{k,j}(\mathbf{Q}) = \mu_j \mathbb{1}(Q_{k,j} > 0), \tag{11}$$

where $j^\dagger(\mathbf{Q})$ is the fastest speed server pool that contains a idle server and $I_j(\mathbf{Q})$ is the set of idle servers available in pool $j$. Observe that from (10), $j^\dagger(\mathbf{Q})$ may not always exists for $\mathbf{Q} \in \mathbb{Z}^N_+$. If $j^\dagger(\mathbf{Q})$ does not exists for some $\mathbf{Q} \in \mathbb{Z}^N_+$, in this case all servers in system are busy, i.e., $\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) = 1$. Therefore, for any $\mathbf{Q} \in \mathbb{Z}^N_+$ we have $\mathbb{1}\Big(j^\dagger_E(\mathbf{Q})\Big) = 1 - \mathbb{1}(Q_{k,j} > 0 \forall k, \forall j)$, where $j^\dagger_E(\mathbf{Q})$ denotes that $j^\dagger(\mathbf{Q})$ exists.

**Proof of Theorem 2** :

1) **Stability**: To prove stability, we define a Lyapunov function $\Phi : \mathbb{Z}^N_+ \to [0, \infty)$ as

$$\Phi(\mathbf{Q}) = \sum_{j\in[M]}\sum_{k\in[N\gamma_j]} Q^2_{k,j}. \tag{12}$$

From (9) we have $G_{\mathbf{Q}^{(N)}}\Phi(\mathbf{Q}) = \sum_{j\in[M]}\sum_{k\in[N\gamma_j]} \Big[r^{+,N}_{k,j}(\mathbf{Q})(2Q_{k,j} + 1) + r^{-,N}_{k,j}(\mathbf{Q})(-2Q_{k,j}+1)\Big]$ for any $\mathbf{Q} \in \mathbb{Z}^N_+$. Using (10) we can write

$$\sum_{j\in[M]}\sum_{k\in[N\gamma_j]} r^{+,N}_{k,j}(\mathbf{Q})(2Q_{k,j}+1) = N\lambda + (2\lambda \sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} Q_{k,j})\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j). \tag{13}$$

Similarly, using (11) we can write

$$\sum_{j\in[M]}\sum_{k\in[N\gamma_j]} r^{-,N}_{k,j}(\mathbf{Q})(-2Q_{k,j}+1) = -2\sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} Q_{k,j} + \sum_{j\in[M]}\mu_j B_j(\mathbf{Q}), \tag{14}$$

where $B_j(\mathbf{Q})$ denotes the number of busy servers in pool $j$ in state $\mathbf{Q}$. Hence, using (14) and (13) we have

$$G_{\mathbf{Q}^{(N)}}\Phi(\mathbf{Q}) \leq N\lambda + (2\lambda \sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} Q_{k,j}) \times \mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) - 2\sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} Q_{k,j} + N,$$

where the inequality follows from $B_j(\mathbf{Q}) \leq N\gamma_j$. Now suppose $\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) = 1$, then from the above expression we have

$$G_{\mathbf{Q}^{(N)}}\Phi(\mathbf{Q}) \leq N(\lambda+1) - 2(1-\lambda)\sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} Q_{k,j}.$$

Hence, it follows from the above that if $\lambda < 1$, then the drift is strictly negative whenever $\sum_j \mu_j \sum_k Q_{k,j} > \frac{N(1+\lambda)}{2(1-\lambda)}$. On the other side, if $\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) = 0$, then the drift is bounded by $N(1+\lambda)$. Thus, using the Foster-Lyapunov criterion for positive recurrence (see, e.g., Proposition D.3 of [15]) we conclude that $\mathbf{Q}^{(N)}$ is positive recurrent.

2) **Uniform Bounds**: To obtain (4), we analyse the drift of the Lyapunov function $\Psi_\theta : \mathbb{Z}^N_+ \to \mathbb{R}_+$ defined as

$$\Psi_\theta(\mathbf{Q}) \triangleq \sum_{j\in[M]}\sum_{k\in[N\gamma_j]} \exp(\theta Q_{k,j}), \tag{15}$$

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

Finally, using Chernoff bounds, we then obtain the bounds on the tail probabilities.

The steps to evaluate the drift of the above defined Lyapunov function is similar to the previous case. Due to space restriction, we write the final expression of the drift $G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q})$ as

$$\begin{aligned} &G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q}) \\ &= (e^\theta - 1)\Big[N\lambda + (\lambda \sum_{j\in[M]} \mu_j \sum_{k\in[N\gamma_j]} \exp(\theta Q_{k,j}) \\ &\qquad - N\lambda)\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) \\ &- \frac{1}{e^\theta}\sum_{j\in[M]}\mu_j \sum_{k\in[N\gamma_j]} \exp(\theta Q_{k,j}) + \frac{1}{e^\theta}\sum_j \mu_j I_j(\mathbf{Q})\Big]. \end{aligned} \tag{16}$$

Note that for all $\theta \in [0, -\log\lambda)$, we have $\lambda e^\theta - 1 < 0$. Therefore, using this fact it can be easily verified from the above expression that $\sup_{\mathbf{Q}\in\mathbb{Z}_+^N} G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q}) < \infty$. Now from the application of Proposition 1 of [16], we have

$$\mathbb{E}[G_{\mathbf{Q}^{(N)}}\Psi_\theta(\mathbf{Q}^{(N)}(\infty))] \geq 0. \tag{17}$$

Taking expectation of (16) when $\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) = 1$ with respect to the stationary distribution and applying (17) we obtain

$$\begin{aligned} &\left(1 - \lambda e^\theta\right)\mathbb{E}\left[\sum_j \mu_j \sum_k \exp(\theta Q_{k,j}^{(N)}(\infty))\right] \\ &\leq \mathbb{E}\left[\sum_j \mu_j (N\gamma_j - B_j(\mathbf{Q}^{(N)}(\infty)))\right] = N(1-\lambda), \end{aligned} \tag{18}$$

where in the equality 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 $N\lambda$. One important observation to make is that the SA-JIQ policy does not distinguish between servers of the same type. Hence, from (18) we have that for all $\theta \in [0, -\log\lambda)$

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

Thus, for each $j \in [M]$ and $\theta \in [0, -\log\lambda)$ we have

$$\mathbb{E}\left[\exp(\theta Q_{k,j}^{(N)}(\infty))\right] \leq \frac{1}{\mu_j\gamma_j}\frac{1-\lambda}{1-\lambda e^\theta}. \tag{19}$$

Now, for each positive $\theta$ we have

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

where the first inequality follows from Markov inequality and the last inequality follows from (19). Therefore, from above set of inequality we can write

$$\mathbb{P}(Q_{k,j}^{(N)}(\infty) \geq l) = C_j(\lambda,\theta)e^{-l\theta}. \tag{20}$$

Similarly, taking expectation of (16) when $\mathbb{1}(Q_{k,j} > 0 \forall k, \forall j) = 0$ and applying (17) we obtain same bound as in (20). This completes the proof.

■

## VII. Process Convergence of SA-JIQ

One of the major step to establish the asymptotic optimality of SA-JIQ is to prove the process convergence result that is Theorem 3. In this section, we prove that the sequence of processes $(\mathbf{x}^{(N)})_{N\geq 1}$ converges weakly to the process $\mathbf{x}$. To prove this we use martingale representation approach and the time-scale separation technique from [17]. First we write the martingale representation of the evolution of $x_{i,j}^{(N)}(t)$ for all $t \geq 0$, $i \geq 1$, and $j \in [M]$ as

$$\begin{aligned} x_{i,j}^{(N)}(t) &= x_{i,j}^{(N)}(0) + \frac{\lambda}{\gamma_j}\int_0^t p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))ds \\ &\quad - \mu_j\int_0^t (x_{i,j}^{(N)}(s) - x_{i+1,j}^{(N)}(s))ds + \frac{1}{N\gamma_j}M_{i,j}^{(N)}(t), \end{aligned} \tag{21}$$

where $M_{i,j}^{(N)}(t) = (M_{i,j}^{(A,N)}(t) - M_{i,j}^{(D,N)}(t))$ with $M_{i,j}^{(A,N)}$ and $M_{i,j}^{(D,N)}$ are martingales corresponding to the arrivals and departures at the component $x_{i,j}^{(N)}$ and $p_{i-1,j}^{(N)}(\mathbf{x}^{(N)}(s))$ is the probability that an arrival joins a server in $j^{\text{th}}$ pool with exactly $i-1$ jobs for finite $N$. The detailed definitions of $M_{i,j}^{(A,N)}$ and $M_{i,j}^{(D,N)}$ are given in Appendix C of [14].

To establish the fluid limit of SA-JIQ, we define the process $\mathbf{V}^{(N)} = (\mathbf{V}^{(N)}(t) = (V_{1,j}^{(N)}(t), j \in [M]), t \geq 0)$ with $V_{1,j}^{(N)}(t) = N\gamma_j - N\gamma_j x_{1,j}^{(N)}(t)$. The component $V_{1,j}^{(N)}(t)$ counts the number of idle servers of type $j$. Note that we can write $p_{i-1,j}^{(N)}(\mathbf{x})$ for $\mathbf{x} \in S$, $j \in [M]$ in terms of the process $\mathbf{V}^{(N)}$ as

$$p_{0,j}^{(N)}(\mathbf{x}) = \mathbb{1}\Big(\mathbf{V}^{(N)} \in R_j\Big), \tag{22}$$

$$p_{i-1,j}^{(N)}(\mathbf{x}) = \mu_j\gamma_j(x_{i-1,j}^{(N)} - x_{i,j}^{(N)})\mathbb{1}\Big(\mathbf{V}^{(N)} \in H\Big)\ i \geq 2, \tag{23}$$

where

$$R_j = \left\{\mathbf{v} \in \bar{\mathbb{Z}}_+^M : 0 = v_{1,1} = \cdots = v_{1,j-1} < v_{1,j}\right\}, \tag{24}$$

$$H = \left\{\mathbf{v} \in \bar{\mathbb{Z}}_+^M : v_{1,l} = 0\ \forall l \in [M]\right\}. \tag{25}$$

The set $R_j$ represents the set of states where the pool $j$ has some idle-servers and all servers in pools below $j$ are busy. Furthermore, the set $H$ represents the set of states where all servers in each pool are busy. Moreover, from (24) and (25) we have $H = (\cup_{j\in[M]}R_j)^c$.

It can be easily verified that $\mathbf{V}^{(N)}$ is a Markov process defined on $\bar{\mathbb{Z}}_+^M$ with transition rates for $j \in [M]$ as

$$\mathbf{V}^{(N)} \to \begin{cases} \mathbf{V}^{(N)} + \mathbf{e}_j, & \text{at rate } N\gamma_j\mu_j(x_{1,j}^{(N)} - x_{2,j}^{(N)}), \\ \mathbf{V}^{(N)} - \mathbf{e}_j, & \text{at rate } N\lambda\mathbb{1}\Big(\mathbf{V}^{(N)} \in R_j\Big), \end{cases}. \tag{26}$$

Furthermore, it is important to note that there is a difference in the time-scale of the processes $\mathbf{x}^{(N)}$ and $\mathbf{V}^{(N)}$. To see this consider 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. The time-scale separation between $\mathbf{V}^{(N)}$ and $\mathbf{x}^{(N)}$ plays an important role in characterizing the limit integral involving $p^{(N)}_{i-1,j}$ term. 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 $[0,\infty) \times \bar{\mathbb{Z}}^M_+$ as

$$\beta^{(N)}(B_1 \times B_2) = \int_{A_1} \mathbb{1}\Big(\mathbf{V}^{(N)}(s) \in B_2\Big) ds.$$

for any $B_1 \in \mathscr{B}([0,\infty))$ and $B_2 \in \mathscr{B}(\bar{\mathbb{Z}}^M_+)$. Next we show 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 $[0,\infty) \times \bar{\mathbb{Z}}^M_+$ satisfying $\beta([0,t] \times \bar{\mathbb{Z}}^M_+) = t$ 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 $[0,t] \times \bar{\mathbb{Z}}^M_+$ for each $t$.

**Lemma 5.** *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 for $t \geq 0$ and $j \in [M]$ satisfies*

$$x_{1,j}(t) = x_{1,j}(0) + \frac{\lambda}{\gamma_j}\beta([0,t] \times R_j) - \mu_j \int_0^t (x_{1,j}(s) - x_{2,j}(s))ds, \quad (27)$$

$$x_{i,j}(t) = x_{i,j}(0) + \lambda \int_{[0,t]\times H} \mu_j(x_{i-1,j}(s) - x_{i,j}(s))d\beta - \mu_j \int_0^t (x_{i,j}(s) - x_{i+1,j}(s))ds, \ i \geq 2. \quad (28)$$

Due to lack of space the detailed proof of above lemma follows from Lemma 14 of [14]. The final step in proving Theorem 3 is the characterisation the limit $\beta([0,t], R_j)$ appearing in (27). To do so, we define for any $\mathbf{x} \in S$ a Markov process $\mathbf{V}_\mathbf{x}$ on $\bar{\mathbb{Z}}^M_+$ with transition rates for $j \in [M]$ as

$$\mathbf{V}_\mathbf{x} \to \begin{cases} \mathbf{V}_\mathbf{x} + \mathbf{e}_j, & \text{at rate } \gamma_j\mu_j(x_{1,j} - x_{2,j}) \\ \mathbf{V}_\mathbf{x} - \mathbf{e}_j, & \text{at rate } \lambda \mathbb{1}(\mathbf{V}_\mathbf{x} \in R_j) \end{cases}. \quad (29)$$

From Lemma 2 and Theorem 3 of [17], it follows that the limit $\beta([0,t] \times R_j)$ satisfies

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

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

$$\pi_\mathbf{x}(\{\mathbf{V} \in \bar{\mathbb{Z}}^M_+ : V_{1,j} = \infty\}) = 1, \text{ if } x_{1,j} < 1. \quad (30)$$

Furthermore, we can write $\pi_\mathbf{x}(H) = 1 - \sum_{j\in[M]} \pi_\mathbf{x}(R_j)$, as sets $R_j$'s are mutually exclusive. We set $p_{0,j}(\mathbf{x}) = \pi_\mathbf{x}(R_j)$. To prove Theorem 3, it is remains to show that $\mathbf{x}$ uniquely determines the stationary measure $\pi_\mathbf{x}$ and has form of (6)-(7). First, note that the Markov chain given by (29) is defined on $\bar{\mathbb{Z}}^M_+$ and has $2^M$ communicating classes. Among these communicating classes, there is only one restricted strictly to $\mathbb{Z}^M_+$; all other $2^M - 1$ classes have at least one infinite component. To show the uniqueness $\pi_\mathbf{x}$ we need to show that it is concentrated only on a single communicating class among these $2^M$ classes. For this it is sufficient to show $\pi_\mathbf{x}(V_{1,j} = \infty) = 0$ or 1 for all $j \in [M]$. To show this, we use the result of the next lemma which characterises the stationary distribution of the Markov chain $\mathbf{V}_\mathbf{x}$.

**Lemma 6.** *Define $\rho_j = \frac{\gamma_j\mu_j(x_{1,j} - x_{2,j})}{\lambda}$. If $\sum_{j\in[M]} \rho_j < 1$, then the Markov chain $\mathbf{V}_\mathbf{x}$ is positive recurrent. Furthermore, if $\pi$ denotes the stationary distribution of the chain, then we have $\pi\{0 = V_{1,1} = \cdots = V_{1,j-1} < V_{1,j}\} = \rho_j, \ \forall j \in [K]$.*

The proof of above lemma follows from Lemma 15 of [14]. Now suppose $\rho_1 < 1$, then the component $V_{1,1}$ is stable. We show that $\pi_\mathbf{x}(V_{1,1} = \infty) = 0$ with contradiction. Assume $\pi_\mathbf{x}(V_{1,1} = \infty) = \epsilon \in (0,1]$. Also, assume $\bar{\pi}_\mathbf{x}$ to be a stationary distribution of the Markov chain $\mathbf{V}_\mathbf{x}$. Therefore, we can write

$$\pi_\mathbf{x}(R_1) = (1-\epsilon)\bar{\pi}_\mathbf{x}(V_{1,1} > 0) + \epsilon = (1-\epsilon)\rho_1 + \epsilon,$$

where we get $\bar{\pi}_\mathbf{x}(V_{1,1} > 0) = \rho_1$ from Lemma 6. Now substitute this in differential form of (5) for $i = 1$ at time $t$, we get

$$\frac{dx_{1,1}(t)}{dt} = \epsilon\Big(\frac{\lambda}{\gamma_1} - \mu_1(x_{1,1}(t) - x_{2,1}(t))\Big) > 0,$$

where the last step follows as $\rho_1 < 1$. But if $x_{1,1}(t) = 1$, we must have $\frac{dx_{1,1}(t)}{dt} < 0$ which leads to a contradiction. Therefore, we have $\pi_\mathbf{x}(V_{1,1} = \infty) = 0$. Suppose $\rho_1 \geq 1$, then the component $V_{1,1}$ is unstable and we have $\bar{\pi}_\mathbf{x}(V_{1,1} \geq l) = 1$ for all $l \geq 0$. This shows that $\pi_\mathbf{x}(V_{1,1} = \infty) = 1$.

Similarly, for general $j \in [M]$, if $\rho_j < 1 - \sum_{k=1}^{j-1} \rho_k$ then the component $Y_{1,j}$ is stable and we get $\pi_\mathbf{x}(V_{1,j} = \infty) = 0$ using the same contradiction as above, else $\pi_\mathbf{x}(V_{1,j} = \infty) = 1$.

Next we show that $\pi_\mathbf{x}$ has form of (6)-(7). From (24), we can write $\pi_\mathbf{x}(R_1) = \pi_\mathbf{x}(V_{1,1} > 0) = \mathbb{1}(x_{1,1} < 1)$, where the last equality follows from (30). Proceeding iteratively, we can easily verify that $\pi_\mathbf{x}(R_j) = \pi_\mathbf{x}(0 = V_{1,1} = \cdots = V_{1,j-1} < V_{1,j}) = \mathbb{1}(1 = x_{1,1} = x_{1,2} = \cdots = x_{1,j-1} > x_{1,j})$. Moreover, using (28) we can write $p_{i-1,j}(\mathbf{x}) = \pi_\mathbf{x}(H)\mu_j\gamma_j(x_{i-1,j} - x_{i,j}), = (1 - \sum_{j\in[M]} p_{0,j}(\mathbf{x}))\mu_j\gamma_j(x_{i-1,j} - x_{i,j})$ for $i \geq 2$ and $j \in [M]$.

## VIII. Properties of Fluid Limit

In this section, we prove Theorem 4. We first prove that the fluid limit $\mathbf{x}$ described by (5) has a unique fixed point.

### A. Fixed Point

From (5) it follows that for $\mathbf{x}^* \in S$ to be a fixed point of the fluid limit $\mathbf{x}$, we must have

$$\frac{\lambda}{\gamma_j} p_{i-1,j}(\mathbf{x}^*) = \mu_j(x^*_{i,j} - x^*_{i+1,j}), \ i \geqslant 1, \ j \in [M]. \quad (31)$$

Summing (31) over all $i \geq 1$ and for all $j \in [M]$, we get $\lambda \sum_{i\geq 1} \sum_{j\in[M]} p_{i-1,j}(\mathbf{x}^*) = \sum_{j\in[M]} \mu_j\gamma_j x^*_{1,j}$. This implies that

$$\lambda = \sum_{j\in[M]} \mu_j\gamma_j x^*_{1,j}. \tag{32}$$

To prove that fixed point has form (8), we consider different cases based on the interval in which $\lambda$ belongs

1) *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 $x^*_{1,1} < 1$, from (6) this means that $p_{0,1}(\mathbf{x}^*) = 1$. Hence, summing (31) over all $i \geq 1$ and for $j = 1$, we get $x^*_{1,1} = \lambda/\mu_1\gamma_1$. Similarly, summing (31) 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 $j \in \{2, \ldots, M\}$. Now, suppose $x^*_{1,1} = 1$. Then from (32), 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 $\mathbf{x}^* \in S$.
2) *If* $\sum_{i=1}^{j-1} \mu_i\gamma_i \leq \lambda < \sum_{i=1}^{j} \mu_i\gamma_i$, *for* $j \in \{2, \ldots, M\}$: 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 $l \geq 2$. It can be easily verified using induction that $x^*_{1,k} = 1$ for all $k \in [j-1]$. For this an argument similar to the previous case has to be used iteratively. Next, we prove that $x^*_{1,j} = (\lambda - \sum_{i=1}^{j-1} \mu_i\gamma_i)/\mu_j\gamma_j$. Suppose $x^*_{1,j} = 1$. Therefore, using (32), 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 $x^*_{1,j} < 1$. So far we have proved that $x^*_{1,k} = 1$ for all $k \in [j-1]$ and $x^*_{1,j} < 1$. Therefore, using (6) and equation (31), we can easily get $x^*_{1,k} = 0$ for all $k \geq j+1$. Now using (32), we obtain $x^*_{1,j} = (\lambda - \sum_{i=1}^{j-1} \mu_i\gamma_i)\mu_j\gamma_j$. Similarly, using (6) and (31), we can easily verify that $x^*_{l,k} = 0$ for all $k \in [M]$ and for all $l \geq 2$.

### B. Global Stability

Next we prove that the fixed point $\mathbf{x}^*$ is globally stable. We start the proof by proving that the fluid limit process $\mathbf{x}$ is monotone.

**Lemma 7.** *Let* $\mathbf{x}(\cdot, \mathbf{u}) = (\mathbf{x}(t, \mathbf{u}), t \geq 0)$ *denote a solution to* (5) *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* $t \geq 0$, *where the inequality is defined component-wise.*

*Proof.* We first write the differential form of (5) for $i \geq 1$, $j \in [M]$ as

$$\frac{dx_{i,j}(t)}{dt} = \frac{\lambda}{\gamma_j} p_{i-1,j}(\mathbf{x}(t)) - \mu_j(x_{i,j}(t) - x_{i+1,j}(t)). \tag{33}$$

Next, we show that $\frac{dx_{i,j}(t)}{dt} = f_{i,j}(\mathbf{x})$ is non-decreasing in $x_{k,l}$ for $k \neq i$ and for $l \neq j$. Consider the above equation for $i = 1$, i.e., $f_{1,j}(\mathbf{x})$. Suppose for some $\mathbf{x} \in S$, we have $p_{0,j}(\mathbf{x}) = \mathbb{1}(1 = x_{1,1} = x_{1,2} = \cdots = x_{1,j-1} > x_{1,j}) = 1$ for $j \in [M]$ and we know that $x_{1,j} \in [0,1]$. Therefore, $x_{1,k}$ for $k \in [j-1]$ still be 1 after increment. This implies that $p_{0,j}$ remains 1 after increasing $x_{1,k}$ for $k \in [j-1]$. Now, suppose for some $\mathbf{x} \in S$, we have $p_{0,j}(\mathbf{x}) = \mathbb{1}(1 = x_{1,1} = x_{1,2} = \cdots = x_{1,j-1} > x_{1,j}) = 0$ for $j \in [M]$. Increasing $x_{1,k}$ for $k \in [j-1]$ can change $p_{0,j}$ from 0 to 1. Hence, $f_{1,j}(\mathbf{x})$ is non-decreasing in all components except $x_{1,j}$. Using the similar argument as above it can be easily verify that $f_{i,j}(\mathbf{x})$ for $i \geq 2$ and $j \in [M]$ is non-decreasing in all components except $x_{i,j}$. ■

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 8.** *For* $\mathbf{u} \in S$ *let* $\mathbf{x}(\mathbf{u}, \cdot)$ *denote a solution of* (5) *in* $S$. *Then for all* $t \geq 0$ *and for all* $n \geq 1$ *we have*

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

*Proof.* Multiplying (33) by $\gamma_j$ and summing first over $i \geq n$ and then over $j \in [M]$ we obtain the above expression. ■

From the monotonicity of the fluid process $\mathbf{x}$ it follows that for any $\mathbf{x}(0) \in S$ and any $t \geq 0$ we have $\mathbf{x}(t, \min(\mathbf{x}(0), \mathbf{x}^*)) \leq \mathbf{x}(t, \mathbf{x}(0)) \leq \mathbf{x}(t, \max(\mathbf{x}(0), \mathbf{x}^*))$. 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 need to 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 $n \geq 1$. Then 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 by showing

$$\int_0^\infty |x_{i,j}(t, \mathbf{x}(0)) - x^*_{i,j}| dt < \infty. \tag{34}$$

The proofs of uniform boundedness of $v_n(t, \mathbf{x}(0))$ and verifying (34) are similarly followed from Appendix H of [14].

### C. Interchange of Limits

To show interchange of limits, we first prove that the sequence of stationary measures $(\mathbf{x}^{(N)}(\infty))_{N\geq 1}$ is tight in $S$. The necessary and sufficient criterion for tightness of the sequence $(\mathbf{x}^{(N)}(\infty))_N$ in $S$ under the $\ell_1$-norm is proved in Lemma 22 of [14] and is given by

$$\lim_{l\to\infty} \limsup_{N\to\infty} \mathbb{P}\Big( \max_{j\in[M]} \sum_{i\geq l} x^{(N)}_{i,j}(\infty) > \epsilon \Big) = 0, \ \forall \epsilon > 0. \tag{35}$$

Using (4), we show that the sequence $(\mathbf{x}^{(N)}(\infty))_N$ satisfies (35). Fix any $\epsilon > 0$ and $l \geq 1$. Using Markov inequality, we obtain

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

Since $\left(x_{i,j}^{(N)}(\infty)\right)_i$ is a sequence ofnon-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

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

where the last equality follows from (2). Now, from (4) we know that for any $\theta \in [0, -\log\lambda)$ we have

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

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 $N$, therefore the condition of tightness (35) is verified by fixing some $\theta \in (0, -\log\lambda)$ and letting $l \to \infty$. Now, the sequence $(\mathbf{x}^{(N)}(\infty))_N$ is tight in $S$ under the $\ell_1$-norm. Hence, the interchange of limits follows immediately using Prohorov's theorem and using the global stability of fixed point $\mathbf{x}^*$. For more details see Appendix H of [14].

## IX. 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 $3 \times 10^7$. In Figure 2, we have plotted the mean response time of jobs for different schemes as a function of the normalised arrival rate $\lambda$. We see that with SA-JIQ we obtain up to $60\%$ reduction in average response time of jobs compared to classical JIQ. As expected, the performance of SQ$(2,2)$ lies in between classical JIQ and SA-JIQ. Detailed definition of SQ$(2,2)$ is given in [18].

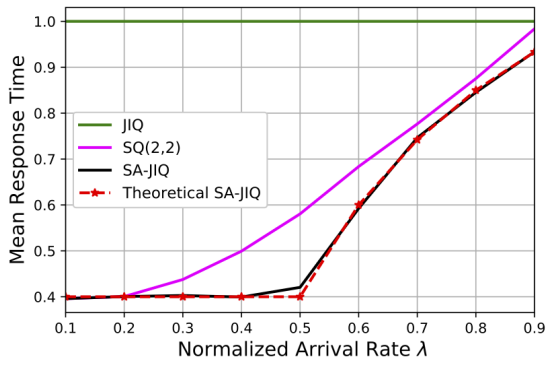


Figure 2. Mean response time as a function of normalized arrival rate $\lambda$ with $N = 1000$ servers. We set $\mu_1 = 4\mu_2 = 20/8$, $\gamma_1 = 1 - \gamma_2 = 1/5$, and $d_1 = d_2 = 2$.

## X. Conclusion

In this work, we have analysed the SA-JIQ scheme for the heterogeneous systems. We have shown the delay optimality of SA-JIQ under the fluid limit while maintaining low communication overhead between the dispatchers and servers. To obtain this optimality result we proved that the sequence of steady-state distributions is tight using drift methods. Furthermore, the fluid limit of the SA-JIQ scheme has been established using the time-scale separation technique. In last, we have show that the fluid limit has a unique and globally stable fixed point.

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

[4] 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.

[5] 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*, 2016, pp. 58–72.

[6] 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.

[7] 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.

[8] G. Foschini and J. Salz, "A basic dynamic routing problem and diffusion," *IEEE Transactions on Communications*, vol. 26, no. 3, pp. 320–327, 1978.

[9] 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.

[10] 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.

[11] 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.

[12] A. L. Stolyar, "Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers," *Queueing Systems*, vol. 85, no. 1-2, pp. 31–65, 2017.

[13] 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.

[14] S. Bhambay and A. Mukhopadhyay, "Asymptotic optimality of speed-aware jsq for heterogeneous systems," *arXiv preprint arXiv:2203.01721*, 2022.

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

[16] 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.

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

[18] 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.