Received: 16 May 2020 | Revised: 8 December 2020 | Accepted: 15 February 2021 | Published on: 9 March 2021

DOI: 10.1002/nav.21986

RESEARCH ARTICLE

WILEY

Steady-state analysis of load balancing with Coxian-2 distributed service times

Xin Liu | **Kang Gong** | **Lei Ying**

Electrical Engineering and Computer Science Department, the University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Xin Liu, 4107 Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109, USA.
Email: xinliuee@umich.edu

Funding information
Division of Computer and Network Systems, Grant/Award Numbers: 2001687, 2002608.
Division of Electrical, Communications and Cyber Systems, Grant/Award Number: 1739344.

Abstract

This paper studies load balancing for many-server (N servers) systems. Each server has a buffer of size $b-1$, and can have at most one job in service and $b-1$ jobs in the buffer. The service time of a job follows the Coxian-2 distribution. We focus on steady-state performance of load balancing policies in the heavy traffic regime such that the normalized load of system is $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$. We identify a set of policies that achieve asymptotic zero waiting. The set of policies include several classical policies such as join-the-shortest-queue (JSQ), join-the-idle-queue (JIQ), idle-one-first (I1F) and power-of-d-choices (Pod) with $d = O(N^{\alpha}\log N)$. The proof of the main result is based on Stein's method and state space collapse. A key technical contribution of this paper is the iterative state space collapse approach that leads to a simple generator approximation when applying Stein's method.

KEYWORDS

Coxian-2 service, heavy traffic regime, load balancing, state space collapse, steady-state analysis, Stein's method

1 | INTRODUCTION

The convergence of cloud computing and machine learning is transforming society in unprecedented ways, and leading to innovations in autonomous systems, healthcare, bioinformatics, social networks, online and in-store retail industry, and education. Data centers nowadays continuously process complex queries and machine learning tasks in large server farms, with tens of thousands of networked servers. Many of these queries/tasks are time sensitive such as queries for products on online retail platforms, real-time machine learning tasks such as language translation and virtual reality applications. In fact, the latency cost of a data center can be very high. In 2017, Akamai reported that 100-ms delay led to 7% drop in sales (Akamai, 2017). Therefore, it is critical for a data center to process these jobs/queries in a timely fashion, ideally without any delay. This paper focuses on the following critical question: *can we achieve almost zero-delay in large-scale data centers?* A critical step for achieving zero-delay is a good load-balancing algorithm that can balance the load across servers and assign an incoming job to an idle server immediately. Assuming exponential service times, sufficient conditions under which a load balancing algorithm achieves asymptotic zero-delay have been obtain in Xin and Lei (2020) for $0 < \alpha < 0.5$ and in Xin and Lei (2019) for $0.5 \le \alpha < 1$. The results have also been extended to parallel-jobs (Wentao & Wang, 2020), multi-server jobs (Weina et al., 2021) and jobs with data locality (Wentao et al., 2020). This paper considers jobs with Coxian-2 service times and identifies a set of load balancing algorithms that achieve zero waiting at steady-state. While Coxian-2 distribution is still a restricted service-time distribution, it has been widely used in computer systems (see, e.g., Tayfur, 1985; Miklós & Armin, 2003; Takayuki & Mor, 2003). In particular, Tayfur (1985) showed the Coxian-2 distribution can well approximate a general distribution by fitting its first three moments when the moments of the general distribution satisfies $m_3/m_1 \ge \frac{3}{2}(c+1)^2$ and $c \ge 1$, where m_1, m_3, and c are the first-order moment, third-order moment and the squared coefficient of variation, respectively (Takayuki & Mor, 2003). also showed that the Coxian-2 distribution can represent a large class of bounded Pareto distributions, which model many real-world job service times in computing and communication systems, including UNIX I/O time and the duration of HTTP and FTP transfers.

1.1 | Related work

Performance analysis of systems with distributed queues is one of the most fundamental and widely studied problems in queueing theory. Assuming exponential service time, the steady-state performance of various load balancing policies has been analyzed using the mean-field analysis (fluid-limit analysis) or diffusion-limit analysis. Among the most popular policies are: (1) join-the-shortest-queue (JSQ) (Anton, 2020; Patrick & David, 2018), which routes an incoming job to the least loaded server; (2) join-the-idle-queue (JIQ) (Alexander, 2015; Yi et al., 2011), which routes an incoming job to an idle server if possible and otherwise to a server chosen uniformly at random; (3) idle-one-first (I1F) (Varun & Neil, 2019), which routes an incoming job to an idle server if available and otherwise to a server with one job if available. If all servers have at least two jobs, the job is routed to a randomly selected server; and (4) power-of-d-choices (Pod) (Mitzenmacher, 1996; Vvedenskaya et al., 1996), which samples d servers uniformly at random and dispatches the job to the least loaded server among the d servers. With general service time distributions, performance analysis of load balancing policies with distributed queues is a much more challenging problem, and remains to be an active research area in queueing theory (Mor, 2013) (Mitzenmacher, 1996). proposed a mean-field model of the Pod policy under gamma service time distributions without proving the convergence of the stochastic system to the mean-field model (Reza et al., 2017; Thirupathaiah et al., 2019; Tim & Benny, 2018). proposed a set of PDE models to approximate load balancing polices under general service times and numerically analyzed key performance metrics (e.g., mean response time). They proved the convergence of the stochastic systems to the corresponding ODEs or PDEs at process-level (over a finite time interval instead of at steady state).

To go beyond the process-level and establish steady-state performance with general service times, a key challenge is to prove that the mean-field system (fluid-system) is stable, that is, the system converges to a unique equilibrium starting from any initial condition. Under nonexponential service time distributions, the proof of stability often relies on a so-called "monotonicity property," which requires a partial order of two mean-field systems starting from two initial conditions to be maintained over time. In particular, letting $x(t, y)$ denote the system state at time t with initial state y, given two initial conditions $y_1 \succ y_2$, where "$\succ$" is a certain partial order, "monotonicity" states that the partial order $x(t, y_1) \succ x(t, y_2)$ holds for any $t \geq 0$.

Monotonicity does hold under several load balancing policies with nonexponential service time distributions that have a decreasing hazard rate (DHR) (Alexander, 2015; Bramson et al., 2012; Foss & Stolyar, 2017). The hazard rate is defined to be $\frac{f(x)}{1-F(x)}$, where $f(x)$ is the density function of the service time and $F(x)$ is the corresponding cumulative distribution function. With DHR (Bramson et al., 2012), proved the asymptotic independence of queues in the mean-field limit under the Pod load balancing policy, and (Alexander, 2015) proved that JIQ achieves asymptotic delay optimality (Benny, 2019). proved the global stability of the mean-filed model of load balancing policies (e.g., Pod) under hyper-exponential distributions with DHR. The key step in Benny (2019) is to represent hyper-exponential distribution by a constrained Coxian distribution, where $\mu_i(1-p_i)$ is decreasing in phase i (μ_i is the service rate in phase i and p_i is the probability that a job finishing service in phase i and entering phase $i+1$). With the alternative representation, monotonicity holds in a certain partial order and the global stability is established.

When service time distributions do not satisfy DHR, only few works established the stability of mean-field systems for very limited light-traffic regimes. For example, Foss and Stolyar (2017) relaxed DHR assumption in Alexander (2015) to any general service distribution but the asymptotic optimality of JIQ only holds when the normalized load $\lambda < 0.5$. The stability of Pod with any general service time distributions with finite second moment has also been established in Bramson et al. (2012) when the load per server the normalized load $\lambda < 1/4$.

The Coxian-2 distribution considered in this paper does not necessarily satisfy DHR. Under the Coxian-2 service time distribution, each job has two phases (phase 1 and phase 2). When in service, a job finishes phase 1 with rate μ_1; and after finishing phase 1, the job leaves the system with probability $1-p$ or enters phase 2 with probability p. If the job enters phase 2, it finishes phase 2 with rate μ_2, and leaves the system. Consider a simple system with two servers. Assume the Coxian-2 service time distribution and JSQ is used for load balancing. Consider two different initial conditions for this system as shown in Figure 1, where jobs in phase 1 are in red color, jobs in phase 2 are in green color and jobs before processed by the server are in black color. The state of each server can be represented by its queue length and the expected remaining service time of the job in service. Let $Q^{(i,1)}(t)$ and $Q^{(i,2)}(t)$ denote the queue length of server i at time t with initial condition 1 and 2, respectively, and $T^{(i,1)}(t), T^{(i,2)}(t) \in \left\{\frac{1}{\mu_1}+\frac{p}{\mu_2}, \frac{1}{\mu_2}, 0\right\}$ denote the expected remaining service time of the job in service at server i with initial condition 1 and 2, respectively. At time 0, we have $Q^{(i,1)}(0) \geq Q^{(i,2)}(0)$ and $T^{(i,1)}(0) \geq T^{(i,2)}(0)$ for all $i = 1, 2$. During the time period $(0, t_1]$, two jobs arrived and were routed to servers according to JSQ, which resulted in the state shown in Figure 1. Suppose that $(1-p)\mu_1 < \mu_2$, then at time t_1, we have $T^{(2,1)}(t_1) = \frac{1}{\mu_2} < T^{(2,2)}(t_1) = \frac{1}{\mu_1}+\frac{p}{\mu_2}$, so the system does not have mononticity. Note the hazard rate of Coxian-2 distribution is $\frac{f(x)}{1-F(x)} = \frac{(1-p)\mu_1+\mu_2 e^{(1+p)\mu_1 x}}{1+e^{(1+p)\mu_1 x}}$, which is an increasing function for $(1-p)\mu_1 < \mu_2$, therefore, it does not satisfy the DHR property.

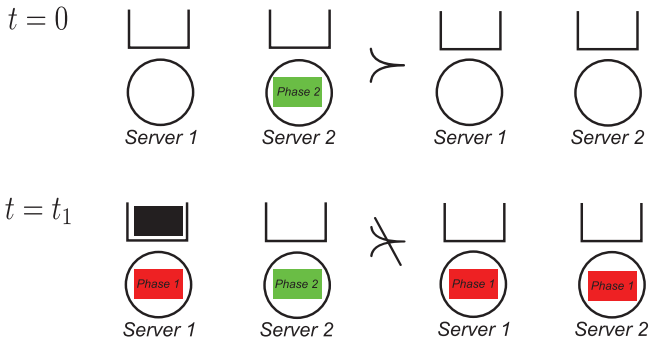

FIGURE 1 Nonmonotocity of JSQ under Coxian-2 distribution

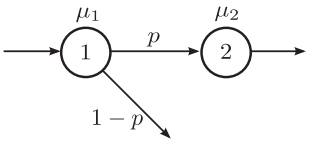

FIGURE 2 Coxian-2 distribution

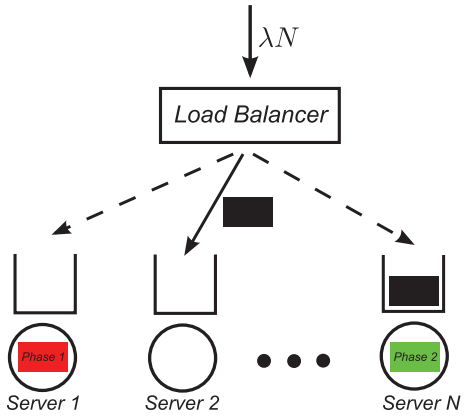

FIGURE 3 Load balancing in many-server systems

1.2 | Main contributions

In this paper, we analyze the steady-state performance of many server systems assuming Coxian service time distributions and heavy traffic regimes ($\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$). From the best of our knowledge, this is the first paper that establishes the steady-state performance of general Coxian distributions without DHR in heavy-traffic regimes. In this paper, we develop an iterative state space collapse (SSC) to show the steady-state "lives" in a restricted region (with a high probability), in which the original system is coupled with a simple system by Stein's method. With iterative SSC and Stein's method, we are able to establish several key performance metrics at steady state, including the expected queue length, the probability that a job is allocated to a busy server (waiting probability) and the waiting time. The main results include:

- For any load balancing policy in a policy set Π (the detailed definition is given in (2)), which includes join-the-shortest-queue (JSQ), join-the-idle-queue (JIQ), idle-one-first (I1F) and power-of-d-choices (Pod) with $d = O(N^\alpha \log N)$, the mean queue length is $\lambda + O\left(\frac{\log N}{\sqrt{N}}\right)$.
- For JSQ and Pod with $d = O(N^\alpha \log N)$, the waiting probability and the expected waiting time per job are both $O\left(\frac{\log N}{\sqrt{N}}\right)$.
- For JIQ and I1F, the waiting probability is $O\left(\frac{1}{N^{0.5-\alpha}\log N}\right)$.

2 | MODEL AND MAIN RESULTS

We consider a many-server system with N homogeneous servers, where job arrival follows a Poisson process with rate λN with $\lambda = 1 - N^{-\alpha}$, $0 < \alpha < 0.5$ and service times follow Coxian-2 distribution (μ_1, μ_2, p) as shown in Figure 2, where $\mu_m > 0$ is the rate a job finishes phase m when in service and $0 \le p < 1$ is the probability that a job enters phase 2 after finishing phase 1. We assume $\lambda = 1 - N^{-\alpha}$ for ease of exposition. Our results hold for $\lambda = 1 - \beta N^{-\alpha}$ with any constant $\beta > 0$, and the extension is straightforward.

Without loss of generality, we assume the mean service time to be one, that is

$$\frac{1}{\mu_1} + \frac{p}{\mu_2} = 1.$$

Under this assumption, λ is also the load of the system.

As shown in Figure 3, an arrival job is colored with black before processed by the server, and colored with red and green when it is in phase 1 and phase 2 in service, respectively. Each server has a buffer of size $b-1$, so can hold at most b jobs ($b-1$ in the buffer and one in service).

Let $Q_{j,m}(t)$ ($m = 1, 2$) denote the fraction of servers which have j jobs at time t and the one in service is in phase m. For convenience, we define $Q_{0,1}(t)$ to be the fraction of servers that are idle at time t and $Q_{0,2}(t) = 0$. Furthermore, define $Q(t)$ to be a $b \times 2$ matrix such that the (j, m)th entry of the matrix is $Q_{j,m}(t)$. Define $S_{i,m}(t) = \sum_{j \ge i} Q_{j,m}(t)$ and $S_i(t) = \sum_{m=1}^{2} S_{i,m}(t)$. In other words, $S_{i,m}(t)$ is the fraction of servers which have at least i jobs and the job in service is in phase m at time t and $S_i(t)$ is the fraction of servers with at least i jobs at time t. Furthermore define $S(t)$ to be a $b \times 2$ matrix such that the (j, m)th entry of the matrix is $S_{j,m}(t)$. Note $Q(t)$ and $S(t)$ have an one-to-one mapping. We consider load balancing policies which dispatch jobs to servers based on $Q(t)$ (or $S(t)$) and under which the finite-state CTMC $\{Q(t), t \ge 0\}$ (or $\{S(t), t \ge 0\}$) is irreducible, and so it has a unique stationary distribution. The load balancing policies include JSQ, JIQ, I1F and Pod.

Let $Q_{j,m}$ denote $Q_{j,m}(t)$ at steady state. We further define $S_{i,m} = \sum_{j \ge i} Q_{j,m}$ and $S_i = \sum_m S_{i,m}$. In other words, $S_{i,m}$ is the fraction of servers which have at least i jobs and the job in service is in phase m and S_i is the fraction of servers with at least i jobs at steady state. We illustrate the state representation $S_{i,m}$ in Figure 4 and Table D1.

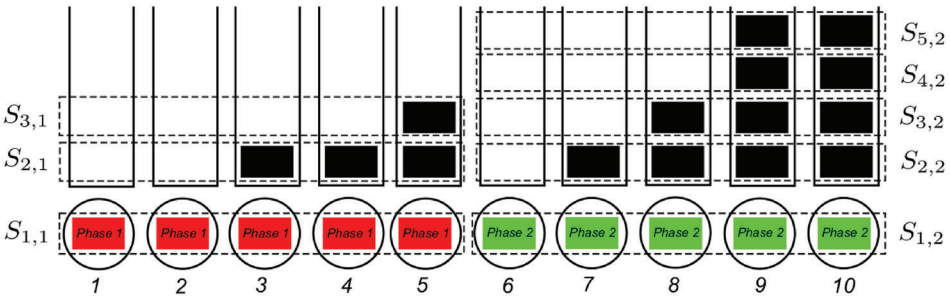

FIGURE 4 Illustrations of states $S_{i,m}$

Define S to be a $b \times 2$ random matrix such that the (i, m)th entry is $S_{i,m}$ and let $s \in \mathbb{R}^{b\times 2}$ denote a realization of S. Define $\mathcal{S}^{(N)}$ to be a set of s such that

$$\mathcal{S}^{(N)} = \left\{ s \;\middle|\; 1 \geq s_{1,m} \geq \cdots \geq s_{b,m} \geq 0, \; 1 \geq \sum_{m=1}^{2} s_{1,m}; Ns_{i,m} \in \mathbb{N}, \forall i, m \right\}. \quad (1)$$

Let $A_1(s)$ denote the probability that an incoming job is routed to a busy server conditioned on that the system is in state $s \in \mathcal{S}^{(N)}$; that is

$$A_1(s) = \mathbb{P}(\text{an incoming job is routed to a busy server} | S(t) = s).$$

Among the load balancing policies considered in this paper, define a subset

$$\Pi = \left\{ \pi \;\middle|\; \text{Under policy } \pi, A_1(s) \leq \frac{1}{\sqrt{N}} \text{ for any } s \in \mathcal{S}^{(N)} \text{ such that } s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{\min\{(1-p)\mu_1, \mu_2\}} \frac{\log N}{\sqrt{N}} \right\}. \quad (2)$$

Our main result of this paper is the following theorem.

Theorem 1 *Define* $w_u = \max\{(1-p)\mu_1, \mu_2\}$, $w_l = \min\{(1-p)\mu_1, \mu_2\}$, $\mu_{\max} = \max\{\mu_1, \mu_2\}$, *and* $k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right)$. *Under any load balancing policy in* Π, *the following bound holds*

$$\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{7\mu_{\max}}{\sqrt{N}\log N}, \quad (3)$$

when N *satisfies*

$$\frac{w_l N^{0.5-\alpha}}{1+\mu_1+\mu_2} \geq \log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)}. \quad (4)$$

Note that the condition $A_1(s) \leq \frac{1}{\sqrt{N}}$ for s such that $s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$ means that an incoming job is routed to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$ when at least $\frac{1}{N^\alpha} - \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$ fraction of servers are idle. There are several well-known policies that satisfy this condition.

- *Join-the-Shortest-Queue (JSQ)*: JSQ routes an incoming job to the least loaded server in the system. Therefore, $A_1(s) = 0$ when $s_1 < 1$.
- *Idle-One-First (I1F) (Varun & Neil, 2019)*: I1F routes an incoming job to an idle server if available; and otherwise to a server with one job if available. If all servers have at least two jobs, the job is routed to a randomly selected server. Therefore, $A_1(s) = 0$ when $s_1 < 1$.
- *Join-the-Idle-Queue (JIQ) (Yi et al., 2011)*: JIQ routes an incoming job to an idle server if possible and otherwise, routes to a server chosen uniformly at random. Therefore, $A_1(s) = 0$ when $s_1 < 1$.
- *Power-of-d-Choices (Pod) (Mitzenmacher, 1996;* Vvedenskaya et al., *1996)*: Pod samples d servers uniformly at random and dispatches the job to the least loaded server among the d servers. Ties are broken uniformly at random. When $d \geq \mu_1 N^\alpha \log N$, $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$.

A direct consequence of Theorem 1 is *asymptotic zero waiting at steady state*. Let $\mathcal{W}$ denote the event that an incoming job is routed to a busy server in a system with N servers, and $\mathbb{P}(\mathcal{W})$ denote the probability of this event at steady-state. Let $\mathcal{B}$ denote the event that an incoming job is blocked (discarded) and $\mathbb{P}(\mathcal{B})$ denote the probability of this event at steady-state. Note that the occurrence of event $\mathcal{B}$ implies the occurrence of event $\mathcal{W}$ because a job is blocked when being routed to a server with b jobs. Furthermore, let W denote the waiting time of a job (when the job is not dropped). We have the following results based on the main theorem.

Corollary 1 *The following results hold when* N *satisfies condition (4).*

- *Under JSQ and Pod with* $d \geq \mu_1 N^\alpha \log N$ *such that* $\sqrt{N} \geq \frac{8k\log N}{b-\lambda} + \frac{8bN^{0.5-\alpha}}{(b-\lambda)\mu_1}$, *we have*

$$\mathbb{E}[W] \leq \frac{2k\log N}{\sqrt{N}} + \frac{14\mu_{\max} + \frac{16\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}, \quad (5)$$

$$\mathbb{P}(\mathcal{W}) \leq \frac{1}{N} + \frac{\mu_{\max}}{\lambda}\left(\frac{k\log N}{\sqrt{N}} + \frac{7\mu_{\max} + \frac{8\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}\right). \quad (6)$$

• *Under JIQ and I1F such that* $N^{0.5-\alpha} \geq 2k \log N$,

$$\mathbb{P}(\mathcal{W}) \leq \frac{14\mu_{\max}}{N^{0.5-\alpha} \log N}. \tag{7}$$

The proof of this corollary is an application of Little's law and Markov's inequality, and can be found in Appendix D. We remark that according to (5) and (6), asymptotic zero-waiting is achieved under JSQ and Pod when $k = o\left(\frac{\sqrt{N}}{\log N}\right)$; and according to (7), asymptotic zero-waiting is achieved under JIQ and I1F when $k = O\left(\frac{N^{0.5-\alpha}}{\log N}\right)$. Since Theorem 1 assumes $k = \Theta(b)$, the buffer size has to be $O\left(\frac{N^{0.5-\alpha}}{\log N}\right)$ as well, which results in the finite-buffer assumption in this paper. This finite-buffer assumption, however, is a sufficient condition. It remains open whether such a condition is necessary.

3 | PROOF OF THEOREM 1

In this section, we present the proof of our main theorem, which is organized along the three key ingredients: generator approximation, gradient bounds, and iterative state space collapse.

3.1 | Generator approximation

Define $e_{i,m} \in \mathbb{R}^{b\times 2}$ to be a $b \times 2$-dimensional matrix such that the (i, m)th entry is $1/N$ and all other entries are zero. Furthermore, define $A_{i,m}(s)$ to be the probability that an incoming job is routed to a server with at least i jobs and the job in service in phase m, when the system is in state s, that is

$$A_{i,m}(s) = \Pr(\text{an incoming job is routed to a server with at least } i \text{ jobs and the job in service in phase } m \mid S(t) = s).$$

Given the state s of the CTMC and the corresponding q, the following events trigger a transition from state s.

• Event 1: A job arrives and is routed to a server that it has $i-1$ jobs and the job in service is in phase 1. When this occurs, $q_{i,1}$ increases by $1/N$, and $q_{i-1,1}$ decreases by $1/N$, so the CTMC has the following transition:

$$q \to q + e_{i,1} - e_{i-1,1},$$
$$s \to s + e_{i,1}.$$

This transition occurs with rate

$$\lambda N(A_{i-1,1}(s) - A_{i,1}(s)),$$

where $A_{i-1,1}(s) - A_{i,1}(s))$ is the probability that an incoming job is routing to a server with $i-1$ jobs and the job in service in phase 1. For example, under JSQ, we have $A_{i-1,1}(s) - A_{i,1}(s)) = \frac{q_{i-1,1}}{q_{i-1}} \mathbb{I}_{\{s_{i-1}=1, s_i<1\}}$, where $\frac{q_{i-1,1}}{q_{i-1}}$ is the probability that the server which receives the job is serving a job in phase 1 conditioned on the job is routed to a server with $i-1$ jobs, and $\{s_{i-1} = 1, s_i < 1\}$ implies that the shortest queue in the system has length $i-1$.

• Event 2: A job arrives and is routed to a server such that it has $i-1$ jobs and the job in service is in phase 2. When this occurs, $q_{i,2}$ increases by $1/N$, and $q_{i-1,2}$ decreases by $1/N$, so the CTMC has the following transition:

$$q \to q + e_{i,2} - e_{i-1,2},$$
$$s \to s + e_{i,2}.$$

This transition occurs with rate

$$\lambda N(A_{i-1,2}(s) - A_{i,2}(s)),$$

where $A_{i-1,2}(s) - A_{i,2}(s))$ is the probability that an incoming job is routing to a server with $i-1$ jobs and the job in service in phase 2. For example, under JSQ, we have $A_{i-1,2}(s) - A_{i,2}(s)) = \frac{q_{i-1,2}}{q_{i-1}} \mathbb{I}_{\{s_{i-1}=1, s_i<1\}}$, where $\frac{q_{i-1,2}}{q_{i-1}}$ is the probability that the server which receives the job is serving a job in phase 2 conditioned on the job is routed to a server with $i-1$ jobs, and $\{s_{i-1} = 1, s_i < 1\}$ implies that the shortest queue in the system has length $i-1$.

• Event 3: A server, which has i jobs, finishes phase 1 of the job in service. The job leaves the system without entering into phase 2. When this occurs, $q_{i,1}$ decreases by $1/N$ and $q_{i-1,1}$ increases by $1/N$, so the CTMC has the following transition:

$$q \to q - e_{i,1} + e_{i-1,1},$$
$$s \to s - e_{i,1}.$$

This transition occurs with rate

$$\mu_1 N q_{i,1}(1-p),$$

where $(1-p)$ is the probability that a job finishes phase 1 and departures without entering phase 2.

• Event 4: A server, which has i jobs, finishes phase 1 of the job in service. The job enters phase 2. When this occurs, a server in state $(i, 1)$ transits to state $(i, 2)$, so $q_{i,1}$ decreases by $1/N$ and $q_{i,2}$ increases by $1/N$. Therefore, the CTMC has the following transition:

$$q \to q - e_{i,1} + e_{i,2},$$
$$s \to s - \sum_{j=1}^{i} e_{j,1} + \sum_{j=1}^{i} e_{j,2},$$

where the transition of s can be verified based on the definition $s_{i,m} = \sum_{j\geq i} q_{j,m}$ so $s_{j,1}$ decreases by $1/N$ for any $j \leq i$ and $s_{j,2}$ increases by $1/N$ for any $j \leq i$. This event occurs with rate

$$\mu_1 N q_{i,1} p,$$

where p is the probability that a job enters phase 2 after finishing phase 1.

• Event 5: A server, which has i jobs, finishes phase 2 of the job in service. The job leaves the system. When this occurs, $q_{i,2}$ decreases by $1/N$ and $q_{i-1,1}$ increases by $1/N$ (because the server starts a new job in phase 1 and the event when $i = 1$ means the fraction of idle server increase by $1/N$), so the CTMC has the following transition:

$$q \to q - e_{i,2} + e_{i-1,1},$$
$$s \to s - \sum_{j=1}^{i} e_{j,2} + \sum_{j=1}^{i-1} e_{j,1}.$$

This transition occurs with rate

$$\mu_2 N q_{i,2}.$$

We illustrate local state transitions related to state s under JSQ in Figure 5.

Let G be the generator of CTMC $(S(t) : t \geq 0)$. Given function $f : \mathcal{S}^{(N)} \to \mathbb{R}$, we have

$$Gf(s) = \sum_{i=1}^{b} [\lambda N(A_{i-1,1}(s) - A_{i,1}(s))(f(s + e_{i,1}) - f(s)) \quad (8)$$

$$+\lambda N(A_{i-1,2}(s) - A_{i,2}(s))(f(s + e_{i,2}) - f(s)) \quad (9)$$

$$+(1-p)\mu_1 N q_{i,1}(f(s - e_{i,1}) - f(s)) \quad (10)$$

$$+p\mu_1 N q_{i,1}\left(f\left(s - \sum_{j=1}^{i} e_{j,1} + \sum_{j=1}^{i} e_{j,2}\right) - f(s)\right) \quad (11)$$

$$+\mu_2 N q_{i,2}\left(f\left(s - \sum_{j=1}^{i} e_{j,2} + \sum_{j=1}^{i-1} e_{j,1}\right) - f(s)\right)\Bigg]. \quad (12)$$

For any bounded function $f : \mathcal{S}^{(N)} \to \mathbb{R}$,

$$\mathbb{E}[Gf(S)] = 0, \quad (13)$$

which can be easily verified by using the global balance equations and the fact that S represents the steady-state of the CTMC.

To understand the steady-state performance of a load balancing policy, we will establish an upper bound on the distance function in (3):

$$\max\left\{\sum_{i=1}^{b} S_i - \eta, 0\right\},$$

with

$$\eta = \lambda + \frac{k \log N}{\sqrt{N}}. \quad (14)$$

The upper bound measures the quantity that the total number of jobs in the system ($N\sum_{i=1}^{b} S_i$) exceeds $N\lambda + k\sqrt{N}\log N$ at steady state, and can be used to bound the probability that an incoming job is routed to an idle server in Corollary 1.

We consider a simple fluid system with arrival rate λ and departure rate $\lambda + \frac{\log N}{\sqrt{N}}$, that is

$$\dot{x} = -\frac{\log N}{\sqrt{N}},$$

and function $g(x)$ which is the solution of the following Stein's equation (Lei, 2016):

$$g'(x)\left(-\frac{\log N}{\sqrt{N}}\right) = \max\{x - \eta, 0\}, \forall x, \quad (15)$$

where $g'(x) = \frac{dg(x)}{dx}$. The left-hand side of (15) can be viewed as applying the generator of the simple fluid system to function $g(x)$, that is

$$\frac{dg(x)}{dt} = g'(x)\dot{x} = g'(x)\left(-\frac{\log N}{\sqrt{N}}\right).$$

It is easy to verify that the solution to (15) is

$$g(x) = -\frac{\sqrt{N}}{2\log N}(x - \eta)^2 \mathbb{I}_{x \geq \eta}, \quad (16)$$

and

$$g'(x) = -\frac{\sqrt{N}}{\log N}(x - \eta)\mathbb{I}_{x \geq \eta}. \quad (17)$$

We note that the simple fluid system is a one-dimensional system and the stochastic system is $b \times 2$-dimensional. In order to couple these two systems, we define

$$f(s) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right), \quad (18)$$

and invoke $f(s)$ in Stein's method.

Since $\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} = \sum_{i=1}^{b} s_i \leq b$ for $s \in \mathcal{S}^{(N)}$, and $f(s)$ is bounded for $s \in \mathcal{S}^{(N)}$, we have

$$\mathbb{E}[Gf(S)] = \mathbb{E}\left[Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right] = 0. \quad (19)$$

Now define

$$h(x) = \max\{x - \eta, 0\}.$$

Based on (15) and (19), we obtain

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right] = \mathbb{E}\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\left(-\frac{\log N}{\sqrt{N}}\right) - Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right]. \quad (20)$$

Note that according to the definition of $f(s)$ in (18), $e_{j,1}$ and $e_{j,2}$, we have

$$f(s + e_{j,1}) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} + \frac{1}{N}\right),$$
$$f(s + e_{j,2}) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} + \frac{1}{N}\right)$$

and

$$f(s - e_{j,1}) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} - \frac{1}{N}\right),$$
$$f(s - e_{j,2}) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} - \frac{1}{N}\right)$$

for any $1 \leq j \leq b$. Therefore,

$$Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right) = N\lambda(1 - A_b(S))$$

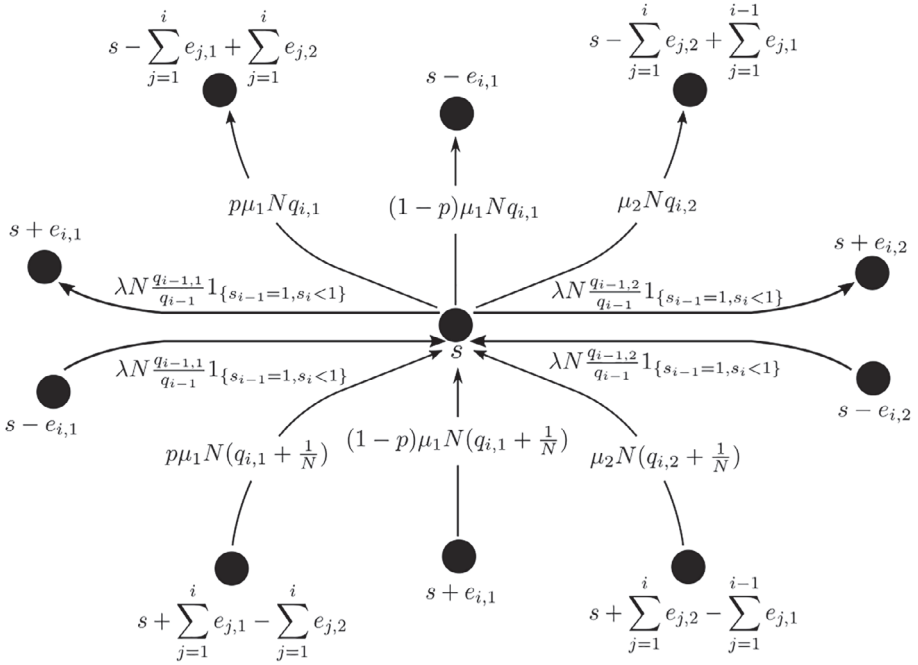

FIGURE 5 Illustrations of state transitions under JSQ for any i with $1 \le i \le b$

$$\times \left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} + \frac{1}{N} \right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} \right) \right)$$
$$+ N((1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2})$$
$$\times \left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} - \frac{1}{N} \right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} \right) \right),$$

where the first term represents the transitions when a job arrives and the second term represents the transitions when a job departures from the system. Note $(1-p)\mu_1 s_{1,1}$ and $\mu_2 s_{1,2}$ are the rates at which jobs leave the system when in phase 1 and phase 2, respectively in the state s. Therefore, $(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$ is the total departure rate. Define $d_1 = (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$ and its stochastic correspondence $D_1 = (1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}$ for simple notations.

Substituting the equation above to (20), we have

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} \right) \right]$$
$$= \mathbb{E}\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} \right) \left(-\frac{\log N}{\sqrt{N}} \right) - N\lambda(1 - A_b(S)) \right.$$
$$\times \left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} + \frac{1}{N} \right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} \right) \right)$$
$$\left. -ND_1 \left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} - \frac{1}{N} \right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m} \right) \right) \right]. \tag{21}$$

From the closed-forms of g and g' in (16) and (17), note that for any $x < \eta$,

$$g(x) = g'(x) = 0.$$

Also note that when $x > \eta + \frac{1}{N}$,

$$g'(x) = -\frac{\sqrt{N}}{\log N}(x - \eta), \tag{22}$$

so for $x > \eta + \frac{1}{N}$,

$$g''(x) = -\frac{\sqrt{N}}{\log N}. \tag{23}$$

By using mean-value theorem in the region $\mathcal{T}_1 = \left\{ x \mid \eta - \frac{1}{N} \le x \le \eta + \frac{1}{N} \right\}$ and Taylor theorem in the region $\mathcal{T}_2 = \left\{ x \mid x > \eta + \frac{1}{N} \right\}$, we have

$$g\left(x + \frac{1}{N}\right) - g(x) = \left(g\left(x + \frac{1}{N}\right) - g(x) \right) (\mathbb{I}_{x\in\mathcal{T}_1} + \mathbb{I}_{x\in\mathcal{T}_2})$$
$$= \frac{g'(\xi)}{N}\mathbb{I}_{x\in\mathcal{T}_1} + \left(\frac{g'(x)}{N} + \frac{g''(\zeta)}{2N^2} \right) \mathbb{I}_{x\in\mathcal{T}_2} \tag{24}$$

$$g\left(x - \frac{1}{N}\right) - g(x) = \left(g\left(x - \frac{1}{N}\right) - g(x) \right) (\mathbb{I}_{x\in\mathcal{T}_1} + \mathbb{I}_{x\in\mathcal{T}_2})$$
$$= -\frac{g'(\widetilde{\xi})}{N}\mathbb{I}_{x\in\mathcal{T}_1} + \left(-\frac{g'(x)}{N} + \frac{g''(\widetilde{\zeta})}{2N^2} \right) \mathbb{I}_{x\in\mathcal{T}_2} \tag{25}$$

where $\xi, \zeta \in \left(x, x + \frac{1}{N}\right)$ and $\widetilde{\xi}, \widetilde{\zeta} \in \left(x - \frac{1}{N}, x\right)$. Substitute (24) and (25) into the generator difference in (21), we have

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b} S_i \right) \right] = J_1 + J_2 + J_3, \tag{26}$$

with

$$J_1 = \mathbb{E}\left[g'\left(\sum_{i=1}^{b} S_i \right) \left(\lambda A_b(S) - \lambda - \frac{\log N}{\sqrt{N}} + D_1 \right) \mathbb{I}_{\sum_{i=1}^{b} S_i \in \mathcal{T}_2} \right], \tag{27}$$

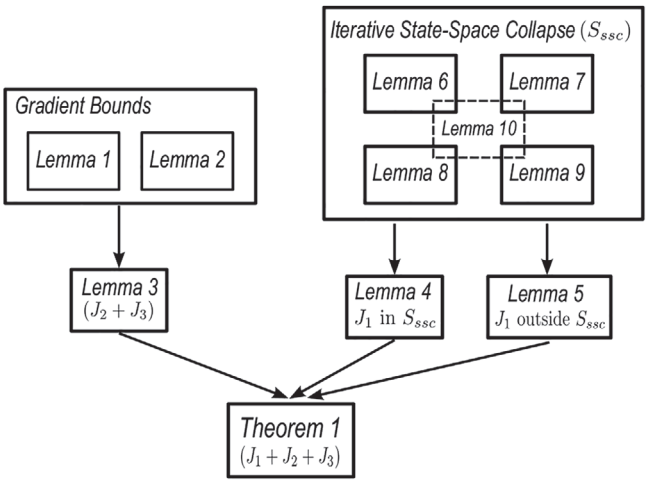

FIGURE 6 The roadmap of proving Theorem 1

$$J_2 = \mathbb{E}\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1 - A_b(S))g'(\xi) + D_1 g'(\widetilde{\xi})\right)\mathbb{I}_{\sum_{i=1}^{b} S_i \in \mathcal{T}_1}\right], \tag{28}$$

$$J_3 = -\mathbb{E}\left[\frac{1}{2N}(\lambda(1 - A_b(S))g''(\zeta) + D_1 g''(\widetilde{\zeta}))\mathbb{I}_{\sum_{i=1}^{b} S_i \in \mathcal{T}_2}\right]. \tag{29}$$

Note that in (28) and (29), we have that

$$\xi, \zeta \in \left(\sum_{i=1}^{b} S_i, \sum_{i=1}^{b} S_i + \frac{1}{N}\right) \text{ and } \widetilde{\xi}, \widetilde{\zeta} \in \left(\sum_{i=1}^{b} S_i - \frac{1}{N}, \sum_{i=1}^{b} S_i\right)$$

are random variables whose values depend on $\sum_{i=1}^{b} S_i$. We do not include $\sum_{i=1}^{b} S_i$ in the notation for simplicity.

To establish the main result in Theorem 1, we need to provide the upper bounds on (27), (28) and (29). In the following Section 3.2, we study g' and g'' to bound the terms in (28) and (29); In Section 3.3, we study SSC to bound the term in (27). We summarize the proof in a roadmap in Figure 6. Lemmas 1 and 2 establish gradient bounds, which are used to bound $J_2 + J_3$ in Lemma 3. Lemmas 6, 7, 8 and 9 are iterative SSC to show the system is in S_{ssc} with a high probability, which rely on Lemma 10 and are used to bound J_1 in Lemmas 4 and 5. We finally prove Theorem 1 by combining Lemmas 3, 4 and 5.

3.2 | Gradient bounds

To bound J_2 in (28) and J_3 in (29), we summarize bounds on g' and g'' in the following two lemmas.

Lemma 1 *Given* $x \in \left[\eta - \frac{2}{N}, \eta + \frac{2}{N}\right]$, *we have*

$$|g'(x)| \le \frac{2}{\sqrt{N}\log N}.$$

Lemma 2 *For* $x > \eta$, *we have*

$$|g''(x)| \le \frac{\sqrt{N}}{\log N}.$$

Based on the bounds on g' in Lemma 1 and g'' in Lemma 2, we provide the upper bound on $J_2 + J_3$ in the following lemma.

Lemma 3 *For* $g(\cdot)$ *defined in (16), we have*

$$J_2 + J_3 \le \frac{6\mu_{\max}}{\sqrt{N}\log N}.$$

The proofs of the lemmas above are presented in Appendix A.

3.3 | State space collapse

In this subsection, we analyze J_1 in (27):

$$\begin{aligned}
&\mathbb{E}\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(\lambda A_b(S) - \lambda - \frac{\log N}{\sqrt{N}} + D_1\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
&= \mathbb{E}\left[\frac{\sqrt{N}}{\log N}h\left(\sum_{i=1}^{b} S_i\right)\left(-\lambda A_b(S) + \lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\right. \\
&\qquad \left.\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
&\le \mathbb{E}\left[\frac{\sqrt{N}}{\log N}h\left(\sum_{i=1}^{b} S_i\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right],
\end{aligned} \tag{30}$$

where the equality is due to Stein's Equation (15), and the inequality holds because

$$\frac{\sqrt{N}}{\log N}h\left(\sum_{i=1}^{b} S_i\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}} \ge 0.$$

We first focus on

$$\left(\lambda + \frac{\log N}{\sqrt{N}} - (1-p)\mu_1 s_{1,1} - \mu_2 s_{1,2}\right)\mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}}, \tag{31}$$

where we recall $\eta = \lambda + \frac{k\log N}{\sqrt{N}}$ and $d_1 = (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$ is the total departure rate when the system is in the state s.

We consider two cases: $s \in \mathcal{S}_{ssc}$ and $s \notin \mathcal{S}_{ssc}$, where

$$\mathcal{S}_{ssc} = \mathcal{S}_{ssc_1} \bigcup \mathcal{S}_{ssc_2},$$

and

$$\begin{aligned}
\mathcal{S}_{ssc_1} = \Bigg\{ s \;\Bigg|\; & s_1 \ge \lambda + \left(\frac{1 + \mu_1 + \mu_2}{w_l} - \mu_1\right)\frac{\log N}{\sqrt{N}} \\
& s_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}, \text{and } s_{1,2} \ge \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}} \Bigg\}, \\
\mathcal{S}_{ssc_2} = \left\{ s \;\middle|\; \sum_{i=1}^{b} s_i \le \lambda + \frac{k\log N}{\sqrt{N}} \right\}.
\end{aligned}$$

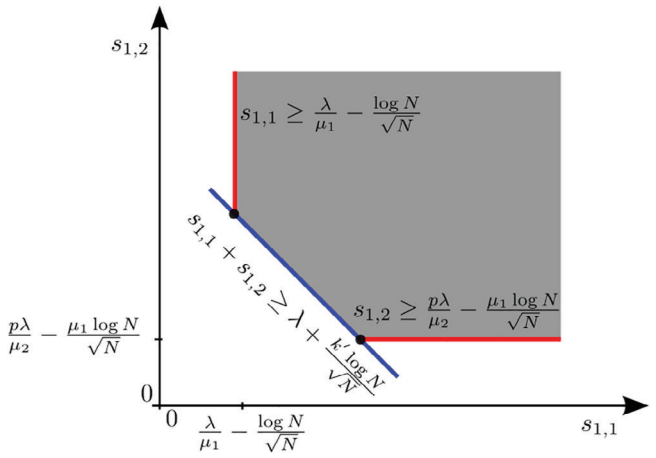

FIGURE 7 State space collapse in $\mathcal{S}_{ssc_1}$.

• **Case 1:** $\mathcal{S}_{ssc_1}$ is shown as the gray region in Figure 7. Any $s \in \mathcal{S}_{ssc_1}$ satisfies

$$(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} \geq \lambda + \frac{\log N}{\sqrt{N}},$$

so $\left(\lambda + \frac{\log N}{\sqrt{N}} - (1-p)\mu_1 s_{1,1} - \mu_2 s_{1,2}\right) \mathbb{I}_{\sum_{i=1}^b s_i > \eta + \frac{1}{N}} \leq 0$ for any $s \in \mathcal{S}_{ssc_1}$. The details are presented in Lemma 4. When $s \in \mathcal{S}_{ssc_2}$,

$$\mathbb{I}_{\sum_{i=1}^b s_i > \eta + \frac{1}{N}} = 0,$$

so $\left(\lambda + \frac{\log N}{\sqrt{N}} - (1-p)\mu_1 s_{1,1} - \mu_2 s_{1,2}\right) \mathbb{I}_{\sum_{i=1}^b s_i > \eta + \frac{1}{N}} = 0$ for any $s \in \mathcal{S}_{ssc_2}$.

• **Case 2:** We will show that

$$\mathbb{P}(S \notin \mathcal{S}_{ssc}) \leq \frac{3}{N^2}$$

in Lemma 5 using an iterative state space collapse approach.

Lemma 4 *For any* $s \in \mathcal{S}_{ssc_1}$,

$$\left(\lambda + \frac{\log N}{\sqrt{N}} - (1-p)\mu_1 s_{1,1} - \mu_2 s_{1,2}\right) \mathbb{I}_{\sum_{i=1}^b s_i > \lambda + \frac{k \log N}{\sqrt{N}} + \frac{1}{N}} \leq 0.$$

The proof of Lemma 4 can be found in Appendix B.

Lemma 5 *For a large* N *such that* $\log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)}$, *we have*

$$\mathbb{P}(S \notin \mathcal{S}_{ssc}) \leq \frac{3}{N^2}.$$

Proof The proof of Lemma 5 is based on an "iterative" procedure to establish state space collapse, which is achieved by proving a sequence of lemmas (Lemmas 6–9). The detailed proof of four lemmas can be found in Appendix C.

Define sets $\widetilde{\mathcal{S}}_1$ and $\widetilde{\mathcal{S}}_2$ such that

$$\widetilde{\mathcal{S}}_1 = \left\{ s \middle| s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}} \text{ and } s_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}} \right\} \tag{32}$$

$$\widetilde{\mathcal{S}}_2 = \left\{ s \middle| \min\left\{ \eta - s_1, \sum_{i=2}^b s_i \right\} \leq \frac{(c_1+\mu_1)\log N}{\sqrt{N}} \right\}. \tag{33}$$

According to the union bound and Lemmas 7–9, we have

$$\begin{aligned}
&\mathbb{P}(S \notin \widetilde{\mathcal{S}}_1 \cap \widetilde{\mathcal{S}}_2) \\
&\leq \frac{5}{\mu_1} \frac{\sqrt{N}}{\log N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N} \\
&\quad + \frac{16}{\mu_1\mu_2} \frac{N}{\log^2 N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}, \\
&\quad + \frac{34}{\mu_1^2\mu_2} \frac{N^{1.5}}{\log^3 N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N} \\
&\leq \frac{3}{N^2},
\end{aligned}$$

where the second inequality holds for a sufficiently large N such that

$$\log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)}.$$

We note that $\widetilde{\mathcal{S}}_1 \cap \widetilde{\mathcal{S}}_2$ is a subset of $\mathcal{S}_{ssc}$. This is because for any s which satisfies

$$\min\left\{ \eta - s_1, \sum_{i=2}^b s_i \right\} \leq \frac{(c_1+\mu_1)\log N}{\sqrt{N}},$$

we either have

$$\eta - s_1 \leq \frac{(c_1+\mu_1)\log N}{\sqrt{N}},$$

which implies

$$s_1 \geq \lambda + \left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1\right)\frac{\log N}{\sqrt{N}};$$

or

$$\sum_{i=2}^b s_i \leq \eta - s_1,$$

which implies

$$\sum_{i=1}^b s_i \leq \eta.$$

Note that

$$\widetilde{\mathcal{S}}_1 \cap \left\{ s \middle| s_1 \geq \lambda + \left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1\right)\frac{\log N}{\sqrt{N}} \right\} = \mathcal{S}_{ssc_1}$$

and

$$\widetilde{\mathcal{S}}_1 \cap \left\{ s \middle| \sum_{i=1}^b s_i \leq \eta \right\} \subseteq \mathcal{S}_{ssc_2}.$$

We, therefore, have

$$\widetilde{\mathcal{S}}_1 \cap \widetilde{\mathcal{S}}_2 \subseteq \mathcal{S}_{ssc},$$

and

$$\mathbb{P}(S \notin \mathcal{S}_{ssc}) \le \mathbb{P}(S \notin \widetilde{\mathcal{S}}_1 \cap \widetilde{\mathcal{S}}_2) \le \frac{3}{N^2},$$

so Lemma 5 holds. ∎

We present "iterative" state space collapse procedure in Lemmas 6–9.

Lemma 6 (An upper bound on $S_{1,2}$)**.**

$$\mathbb{P}\left(S_{1,2} \le \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}\right) \ge 1 - e^{-\frac{\mu_1\mu_2\log^2 N}{40\mu_{\max}}}.$$

Lemma 7 (A lower bound on $S_{1,1}$)**.**

$$\mathbb{P}\left(S_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}\right) \ge 1 - \frac{5}{\mu_1}\frac{\sqrt{N}}{\log N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}.$$

Lemma 8 (A lower bound on $S_{1,2}$)**.**

$$\mathbb{P}\left(S_{1,2} \ge \frac{p\lambda}{\mu_2} - \frac{\mu_1\log N}{\sqrt{N}}\right) \ge 1 - \frac{16}{\mu_1\mu_2}\frac{N}{\log^2 N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}.$$

Lemma 9 (A lower bound on S_1 via $\sum_{i=2}^b S_i$)**.**

$$\mathbb{P}\left(\min\left\{\lambda + \frac{k\log N}{\sqrt{N}} - S_1, \sum_{i=2}^b S_i\right\} \le \frac{(c_1+\mu_1)\log N}{\sqrt{N}}\right) \ge 1 - \frac{34}{\mu_1^2\mu_2}\frac{N^{1.5}}{\log^3 N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}$$

for $\log N \ge \frac{1}{\min\{\mu_1,\mu_2\}}$, where $k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right)$ and $c_1 = \frac{w_u b}{w_l}\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right) + 2\mu_1$.

Remark: An important contribution of this paper is the iterative state collapse method we use to prove Lemma 5. The method continues refining the state space in which the system stays with a high probability at steady-state. Figure 8 illustrates the iterative state-space collapse in Lemmas 6–8. We first show in Lemma 6 that with a high probability, $S_{1,2} \le \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}$ at steady-state. Then in the reduced state space $\left(S_{1,2} \le \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}\right)$, we further show in Lemma 7 that $S_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ with a high probability at steady state. We then further establish in Lemma 6 that $S_{1,2} \ge \frac{p\lambda}{\mu_2} - \frac{\mu_1\log N}{\sqrt{N}}$ with a high probability at steady state in the reduced state space.

3.4 | Proof of Theorem 1

Based on Lemmas 4 and 5, we can establish the following bound on (30), which is a upper bound on J_1 in (27),

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sqrt{N}}{\log N}h\left(\sum_{i=1}^b\sum_{m=1}^2 S_{i,m}\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right] \\
&= \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^b S_i - \eta\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\mathbb{I}_{S\in\mathcal{S}_{ssc}}\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right] \\
&+ \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^b S_i - \eta\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\mathbb{I}_{S\notin\mathcal{S}_{ssc}}\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right] \\
&\le \frac{3b}{N^{1.5}\log N}, \qquad (34)
\end{aligned}$$

where the last inequality holds because we have used the facts that the average total number of jobs per server is at most b and $\left(\lambda + \frac{\log N}{\sqrt{N}} - D_1\right)\mathbb{I}_{S\notin\mathcal{S}_{ssc}}\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}} < 1$.

Based on Lemma 3, we are ready to establish Theorem 1 under JSQ.

$$\begin{aligned}
&\mathbb{E}\left[\max\left\{\sum_{i=1}^b S_i - \eta, 0\right\}\right] \\
&= J_1 + J_2 + J_3 \le \frac{3b}{N^{1.5}\log N} + \frac{6\mu_{\max}}{\sqrt{N}\log N},
\end{aligned}$$

which implies

$$\mathbb{E}\left[\max\left\{\sum_{i=1}^b S_i - \eta, 0\right\}\right] \le \frac{7\mu_{\max}}{\sqrt{N}\log N}.$$

4 | CONCLUSIONS

In this paper, we considered load balancing under the Coxian-2 service time distribution in heavy traffic regimes. The Coxian-2 service time distribution does not have DHR and the system considered in this paper lacks monotonicity. We developed an iterative SSC and identified a policy set Π, in which any policy can achieve asymptotic zero delay. The set Π includes JSQ, JIQ, I1F and Pod with $d = O\left(\frac{\log N}{1-\lambda}\right)$. The proposed Stein's method with iterative SCC is a general method that can be used for steady-state analysis of other queueing systems. The key idea of this method is to use an iterative SSC to reduce the state space to a much smaller subspace, in which the system can be well approximated with a simple fluid model, and the approximation error can be quantified using Stein's method. The iterative SSC approach iteratively reduces the state space by focusing on one direction at each iteration based on the system dynamics. This provides an intuitive way to establish SSC results that may be difficult to obtain at once. For example, it remains open whether the SSC result in this paper can be proved using a single Lyapunov function. This method has already inspired and been used in recent work (Wentao et al., 2020), which developed

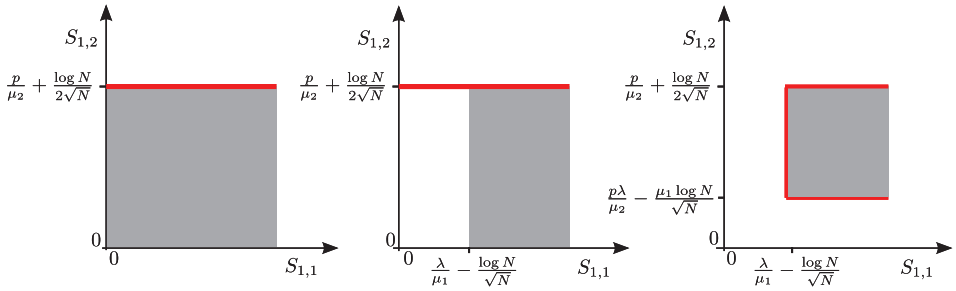

FIGURE 8 Iterative state-space collapse to show that $S_{1,1}$ and $S_{1,2}$ are in a smaller state-space (the gray region) at steady-state

zero-delay load balancing algorithms for networked servers assuming exponential service times.

We also would like to remark it is nontrivial to extend the results in this paper beyond Coxian-2. The analysis in this paper utilized some simple yet critical properties of the Coxian-2 distribution: a job in phase-1 either departs or enters phase-2 immediately, and a job always starts its service from phase-1. In a Coxian-M distribution or a general phase-type distribution, the dependence between jobs in different phases becomes more involved. In particular, it becomes more challenging to establish a result similar to Lemma 6. Recall that for a Coxian-2 distribution, $s_{1,2}$ decreases when its value is large because a large $s_{1,2}$ implies $s_{1,1}$ is small (because $s_{1,1} + s_{1,2} \leq 1$) so the rate at which jobs move from phase-1 to phase-2 is small. However, for a Coxian-M distribution, a large $s_{1,M}$ is not sufficient to guarantee that $s_{1,M-1}$ is small enough so that $s_{1,M}$ will decrease. For a general phase-type distribution, jobs in the queues can be in any phase, not necessarily in phase-1, which makes it difficult to show that $S_{1,M}$ will be close to its "equilibrium". However, we believe if a proper Lyapunov function could be found to establish a "good" upper bound on $S_{1,M}$, then we may apply the iterative approach in this paper to establish SSC and to extend the results in this paper to more general service distributions.

ACKNOWLEDGMENTS

The authors are very grateful to Prof. Jim Dai for his insightful comments. The discussions with Jim had continuously stimulated the authors during the writing of this paper. This work was supported in part by NSF ECCS 1739344, CNS 2002608 and CNS 2001687.

DATA AVAILABILITY STATEMENT

one Data sharing not applicable – no new data generated

ORCID

Xin Liu https://orcid.org/0000-0001-5869-3186

REFERENCES

Akamai. (2017). *The state of online retail performance report*.

Alexander, S. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing System*, *80*(4), 341–361.

Anton, B. (2020). Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Mathematics of Operations Research*, *45*(3), 1069–1103.

Benny, V. H. (2019). Global attraction of ODE-based mean field models with hyperexponential job sizes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, *3*(2), 1–23.

Bertsimas, D., Gamarnik, D., & Tsitsiklis, J. N. (2001). Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Advances in Applied Probability*, *11*, 1384–1482.

Bramson, M., Lu, Y., & Prabhakar, B. (2012). Asymptotic independence of queues under randomized load balancing. *Queueing Systems*, *71*(3), 247–292.

Eryilmaz Atilla, S. R. (2012). Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing System*, *72*(3–4), 311–359.

Foss, S., & Stolyar, A. L. (2017). Large-scale join-idle-queue system with general service times. *Journal of Applied Probability*, *54*(4), 995–1007.

Lei, Y. (2016). *On the approximation error of mean-field models*. In Proceedings of the 2016 ACM SIGMETRICS International Conference. France: Antibes Juan-les-Pins.

Miklós, T., & Armin, H. (2003). Matching moments for acyclic discrete and continuous phase-type distributions of second order. *International Journal of Simulation Systems, Science & Technology*, *3*, 47–57.

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

Mor, H.-B. (2013). Performance modeling and design of computer systems: Queueing theory in action. Cambridge University Press.

Patrick, E., & David, G. (2018). Join the shortest queue with many servers. The heavy-traffic asymptotics. *Mathematics of Operations Research*, *43*, 867–886.

Reza, A., Xingjie, L., & Kavita, R. (2017). The PDE method for the analysis of randomized load balancing networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, *1*(2), 38:1–38:28.

Takayuki, O., & Mor, H.-B. (2003). *Necessary and sufficient conditions for representing general distributions by Coxians*. In International Conference on Modelling Techniques and Tools for Computer Performance Evaluation (pp. 182–199). Berlin, Heidelberg: Springer.

Tayfur, A. (1985). On the phase-type approximations of general distributions. *IIE Transactions*, *17*(2), 110–116.

Thirupathaiah, V., Arpan, M., & Mazumdar Ravi, R. (2019). Insensitivity of the mean field limit of loss systems under SQ(d) routeing. *Advances in Applied Probability*, *51*(4), 1027–1066.

Tim, H., & Benny, V. H. (2018). On the power-of-*d*-choices with least loaded server selection. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, *2*(2), 27:1–27:22.

Varun, G., & Neil, W. (2019). Load balancing in the nondegenerate slowdown regime. *Operations Research*, *67*(1), 281–294.

Vvedenskaya, N. D., Dobrushin, R. L., & Karpelevich, F. I. (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, *32*(1), 20–34.

Wang, W., Theja, M. S., Srikant, R., & Lei, Y. (2018). Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *ACM SIGMETRICS Performance Evaluation Review*, *45*(3), 232–245.

Weina, W., Qiaomin, X., & Mor, H.-B. (2021). *Zero queueing for multi-server jobs*. Arxiv preprint arXiv:2011.10521.

Wentao, W., & Wang, W. (2020). Achieving zero asymptotic queueing delay for parallel jobs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, *42*, 1–36.

Wentao, W., Xingyu, Z., & Srikant, R. (2020). Optimal load balancing with locality constraints. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, *45*, 1–37.

Xin, L., & Lei, Y. (2019). *On universal scaling of distributed queues under load balancing*. arXiv preprint arXiv:1912.11904.

Xin, L., & Lei, Y. (2020). Steady-state analysis of load balancing algorithms in the sub-Halfin–Whitt regime. *Journal of Applied Probability*, *57*(2), 578–596.

Yi, L., Qiaomin, X., Gabriel, K., Alan, G., Larus James, R., & Albert, G. (2011). Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, *68*(11), 1056–1071.

How to cite this article: Liu X, Gong K, Ying L. Steady-state analysis of load balancing with Coxian-2 distributed service times. *Naval Research Logistics* 2022;69:57–75. https://doi.org/10.1002/nav.21986

APPENDIX A: GRADIENT BOUNDS

A.1 | Proof of Lemma 1

Proof From the definition of g function in (15), we have

$$g'(x) = \frac{\max\{x-\eta, 0\}}{-\frac{\log N}{\sqrt{N}}}.$$

Hence, for any $x \in \left[\eta - \frac{2}{N}, \eta + \frac{2}{N}\right]$, we have

$$|g'(x)| \le \frac{|x-\eta|}{\frac{\log N}{\sqrt{N}}} \le \frac{\frac{2}{N}}{\frac{\log N}{\sqrt{N}}} = \frac{2}{\sqrt{N}\log N}.$$

■

A.2 | Proof of Lemma 2

Proof From the definition of g function in (15), we have

$$g'(x) = \frac{\max\{x-\eta, 0\}}{-\frac{\log N}{\sqrt{N}}}.$$

For $x > \eta$, we have

$$g'(x) = \frac{x-\eta}{-\frac{\log N}{\sqrt{N}}},$$

which implies

$$|g''(x)| = \left|\frac{1}{-\frac{\log N}{\sqrt{N}}}\right| = \frac{\sqrt{N}}{\log N}.$$

■

A.3 | Proof of Lemma 3

Proof Note $(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} \le \mu_{\max} s_1 \le \mu_{\max}$, then we have

$$J_2 + J_3 \le \mathbb{E}\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) + \lambda|g'(\xi)| + \mu_{\max}|g'(\widetilde{\xi})|\right)\mathbb{I}_{\sum_{i=1}^{b} S_i \in \mathcal{T}_1}\right] \tag{A1}$$

$$+\mathbb{E}\left[\frac{1}{N}(\lambda|g''(\eta)| + \mu_{\max}|g''(\widetilde{\eta})|)\mathbb{I}_{\sum_{i=1}^{b} S_i \in \mathcal{T}_2}\right] \tag{A2}$$

$$\le \frac{4\mu_{\max}}{\sqrt{N}\log N} + \frac{\lambda + \mu_{\max}}{N}\frac{\sqrt{N}}{\log N} \tag{A3}$$

$$\le \frac{6\mu_{\max}}{\sqrt{N}\log N} \tag{A4}$$

■

APPENDIX B: PROOF OF LEMMA 4

We consider the following problem

$$\min_{(s_{1,1}, s_{1,2}) \in \mathcal{S}_{ssc_1}} (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2},$$

which is a linear programming in terms of variables $s_{1,1}$ and $s_{1,2}$. Therefore, we only need to consider the extreme points of set $\mathcal{S}_{ssc_1}$. In fact, from Figure 7, it is clear that we only need to consider the following two extreme points.

- Case 1: $s_{1,1} = \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ and $s_{1,2} = \lambda + \left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1\right)\frac{\log N}{\sqrt{N}} - s_{1,1} = \frac{p\lambda}{\mu_2} + \left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1 + 1\right)\frac{\log N}{\sqrt{N}}$, where we use the fact $\frac{1}{\mu_1} + \frac{p}{\mu_2} = 1$. In this case,

$$(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$$

$$= \lambda + \left(-(1-p)\mu_1 + \mu_2\left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1 + 1\right)\right)\frac{\log N}{\sqrt{N}} \tag{B5}$$

$$\geq \lambda + (-(1-p)\mu_1 + (1+\mu_1-\mu_1\mu_2+2\mu_2))\frac{\log N}{\sqrt{N}} \tag{B6}$$

$$= \lambda + (1+\mu_2)\frac{\log N}{\sqrt{N}} \tag{B7}$$

$$\geq \lambda + \frac{\log N}{\sqrt{N}}, \tag{B8}$$

where (B6) holds because $w_l = \min\{(1-p)\mu_1, \mu_2\}$ and (B7) holds because $\frac{1}{\mu_1} + \frac{p}{\mu_2} = 1$.

• Case 2: $s_{1,1} = \lambda + \left(\frac{1+\mu_1+\mu_2}{w_l} - \mu_1\right)\frac{\log N}{\sqrt{N}} - s_{1,2} = \frac{\lambda}{\mu_1} + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$ and $s_{1,2} = \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$. At this extreme point, we have

$$(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$$
$$= \lambda + \left((1-p)\mu_1\left(\frac{1+\mu_1+\mu_2}{w_l}\right) - \mu_1\mu_2\right)\frac{\log N}{\sqrt{N}} \tag{B9}$$

$$\geq \lambda + (1+\mu_1+\mu_2-\mu_1\mu_2)\frac{\log N}{\sqrt{N}} \tag{B10}$$

$$\geq \lambda + \frac{\log N}{\sqrt{N}}, \tag{B11}$$

where (B10) holds because $w_l = \min\{(1-p)\mu_1, \mu_2\}$ and (B11) holds because $\mu_1 + \mu_2 \geq p\mu_1 + \mu_2 = \mu_1\mu_2$.

APPENDIX C: PROOF OF ITERATIVE STATE SPACE COLLAPSE

We present the iterative SSC approach for proving Lemmas 6–9. The first three lemmas are on the upper and lower bounds on $S_{1,1}$ and $S_{1,2}$, illustrated in Figure C1, which shows that both $S_{1,1}$ and $S_{1,2}$ are close to its equilibrium values, in particular, with a high probability, $S_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ and $S_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$. However, these two low bounds do not guarantee the total departure rate, which is $(1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}$, is larger than the arrival rate λ. Therefore, we need Lemma 9 to guarantee sufficient fraction of busy servers S_1 such that the total departure rate is "larger than" the arrival rate λ. We therefore need Lemma 9 to further establish a lower bound on S_1 unless the total normalized queue length $\sum_{i=1}^{b} S_i$ is small.

C.1 | A tail bound from Wang et al. (2018)

To prove the space collapse results, we first introduce Lemma 10, which will be repeatedly used to obtain probability tail bounds. Lemma 10 allows us to apply Lyapunov-drift-based heavy traffic analysis (Eryilmaz Atilla, 2012) to reduced state spaces instead of to the entire state space. The lemma is an extension of the tail bound in Bertsimas et al. (2001). This Lyapunov drift analysis on reduced state space enables us to iteratively refine the state space at steady state. The lemma was proven in Wang et al. (2018). We include the proof to make the paper self-contained.

Lemma 10 *Let $(S(t): t \geq 0)$ be a continuous-time Markov chain over a finite state space $\mathcal{S}$ and is irreducible, so it has a unique stationary distribution π. Consider a Lyapunov function $V: \mathcal{S} \to R^+$ and define the drift of V at a state $s \in \mathcal{S}$ as*

$$\nabla V(s) = \sum_{s' \in \mathcal{S}: s' \neq s} q_{s,s'}(V(s') - V(s)),$$

where $q_{s,s'}$ is the transition rate from s to s'. Assume

$$\nu_{\max} := \max_{s,s' \in \mathcal{S}: q_{s,s'} > 0} |V(s') - V(s)| < \infty \text{ and}$$

$$\overline{q} := \max_{s \in \mathcal{S}}(-q_{s,s}) < \infty$$

and define

$q_{\max} := \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}: V(s) < V(s')} q_{s,s'}$.

If there exists a set $\mathcal{E}$ with $B > 0$, $\gamma > 0$, $\delta \geq 0$ such that the following conditions satisfy:

(i) $\nabla V(s) \leq -\gamma$ when $V(s) \geq B$ and $s \in \mathcal{E}$.

(ii) $\nabla V(s) \leq \delta$ when $V(s) \geq B$ and $s \notin \mathcal{E}$.

Then

$$\mathbb{P}(V(S) \geq B + 2\nu_{\max} j) \leq \alpha^j + \beta \mathbb{P}(S \notin \mathcal{E}), \forall j \in \mathbb{N},$$

with

$$\alpha = \frac{q_{\max}\nu_{\max}}{q_{\max}\nu_{\max} + \gamma} \text{ and } \beta = \frac{\delta}{\gamma} + 1.$$

Proof Let $C \geq B - \nu_{\max}$ and consider Lyapunov function.

$$\widehat{V}(s) = \max\{C, V(s)\}.$$

At steady state, we have

$$0 = \sum_{V(s) \leq C - \nu_{\max}} \pi(s) \sum_{s' \neq s} q_{s,s'}(\widehat{V}(s') - \widehat{V}(s))$$
$$+ \sum_{C - \nu_{\max} < V(s) \leq C + \nu_{\max}} \pi(s) \sum_{s' \neq s} q_{s,s'}(\widehat{V}(s') - \widehat{V}(s))$$
$$+ \sum_{V(s) > C + \nu_{\max}} \pi(s) \sum_{s' \neq s} q_{s,s'}(\widehat{V}(s') - \widehat{V}(s)). \tag{C12}$$

Note $\nabla \widehat{V}(s) = \sum_{s' \neq s} q_{s,s'}(\widehat{V}(s') - \widehat{V}(s))$. We consider three terms in (C12) as follows:

• The first term is 0 because $V(s) \leq C - \nu_{\max}$ and $V(s') \leq C$ imply $\widehat{V}(s) = \widehat{V}(s') = C$.

• The second term is bounded

$$\sum_{C - \nu_{\max} < V(s) \leq C + \nu_{\max}} \pi(s) \sum_{s' \neq s} q_{s,s'}(\widehat{V}(s') - \widehat{V}(s))$$

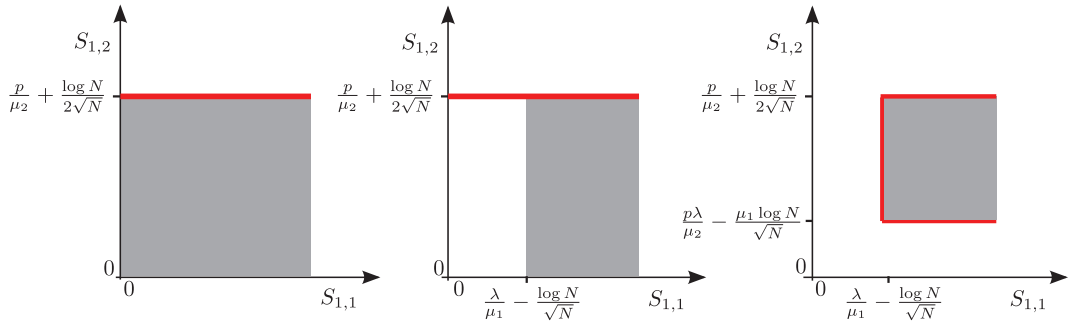

FIGURE C1 Bounds (red lines) on $S_{1,1}$ and $S_{1,2}$

$$\leq \sum_{C-\nu_{\max}<V(s)\leq C+\nu_{\max}} \pi(s) q_{\max}\nu_{\max}$$
$$\leq q_{\max}\nu_{\max}(\mathbb{P}(V(S)>C-\nu_{\max})-\mathbb{P}(V(S)>C+\nu_{\max}))$$

- The third term is divided into two regions $s\in\mathcal{E}$ and $s\notin\mathcal{E}$

$$\begin{aligned}
&\sum_{V(s)>C+\nu_{\max}} \pi(s)\sum_{s'\neq s} q_{s,s'}(\widehat{V}(s')-\widehat{V}(s)) \\
&= \sum_{\substack{V(s)>C+\nu_{\max}\\ s\in\mathcal{E}}} \pi(s)\sum_{s'\neq s} q_{s,s'}(\widehat{V}(s')-\widehat{V}(s)) \\
&\quad + \sum_{\substack{V(s)>C+\nu_{\max}\\ s\notin\mathcal{E}}} \pi(s)\sum_{s'\neq s} q_{s,s'}(\widehat{V}(s')-\widehat{V}(s)) \\
&\leq -\gamma\mathbb{P}(V(S)>C+\nu_{\max}, s\in\mathcal{E}) \\
&\quad + \delta\mathbb{P}(V(S)>C+\nu_{\max}, s\notin\mathcal{E}) \\
&= -\gamma\mathbb{P}(V(S)>C+\nu_{\max}) \\
&\quad + (\delta+\gamma)\mathbb{P}(V(S)>C+\nu_{\max}, s\notin\mathcal{E})
\end{aligned}$$

where the inequality holds because of two conditions (i) and (ii).

Combining three terms above, we have

$$\begin{aligned}
&(q_{\max}\nu_{\max}+\gamma)\mathbb{P}(V(S)>C+\nu_{\max}) \\
&\leq q_{\max}\nu_{\max}\mathbb{P}(V(S)>C-\nu_{\max}) \\
&\quad + (\delta+\gamma)\mathbb{P}(V(S)>C+\nu_{\max}, S\notin\mathcal{E})
\end{aligned}$$

which implies.

$$\begin{aligned}
&\mathbb{P}(V(S)>C+\nu_{\max}) \\
&\leq \frac{q_{\max}\nu_{\max}}{q_{\max}\nu_{\max}+\gamma}\mathbb{P}(V(S)>C-\nu_{\max}) \\
&\quad + \frac{\delta+\gamma}{q_{\max}\nu_{\max}+\gamma}\mathbb{P}(V(S)>C+\nu_{\max}, S\notin\mathcal{E}) \\
&\leq \frac{q_{\max}\nu_{\max}}{q_{\max}\nu_{\max}+\gamma}\mathbb{P}(V(S)>C-\nu_{\max}) \\
&\quad + \frac{\delta+\gamma}{q_{\max}\nu_{\max}+\gamma}\mathbb{P}(S\notin\mathcal{E}) \\
&= \alpha\mathbb{P}(V(S)>C-\nu_{\max})+\kappa\mathbb{P}(S\notin\mathcal{E}),
\end{aligned}$$

where

$$\alpha=\frac{q_{\max}\nu_{\max}}{q_{\max}\nu_{\max}+\gamma} \text{ and } \kappa=\frac{\delta+\gamma}{q_{\max}\nu_{\max}+\gamma}.$$

Let $C=B+(2j-1)\nu_{\max}, \forall j\in\mathbb{N}$ and we have

$$\begin{aligned}
&\mathbb{P}(V(S)>B+2\nu_{\max}j) \\
&\leq \alpha\mathbb{P}(V(S)>B+2(j-1)\nu_{\max})+\kappa\mathbb{P}(S\notin\mathcal{E}) \qquad \text{(C13)}
\end{aligned}$$

By recursively using the inequality (C13), we have

$$\begin{aligned}
\mathbb{P}(V(S)>B+2\nu_{\max}j) &\leq \alpha^j+\kappa\mathbb{P}(S\notin\mathcal{E})\sum_{i=0}^{j}\alpha^i \\
&\leq \alpha^j+\frac{\kappa}{1-\alpha}\mathbb{P}(S\notin\mathcal{E}) \\
&= \alpha^j+\beta\mathbb{P}(S\notin\mathcal{E})
\end{aligned}$$

■

As mentioned above, Lemma 10 is an extension of Theorem 1 in Bertsimas et al. (2001), where $\mathcal{E}=S^{(N)}$ is the entire state space and $\mathbb{P}(S\notin\mathcal{E})=0$. As suggested in Lemma 10, constructing proper Lyapunov functions are critical to establish the tail bounds. In the following lemmas, we construct a sequence of Lyapunov functions and apply Lemma 10 to establish SSC results.

C.2 | Proof of Lemma 6: An upper bound on $S_{1,2}$

To prove Lemma 6, we first establish a Lyaponuv drift analysis for $\mathcal{E}=S^{(N)}$ (the entire state space) in Lemma 11.

Lemma 11 *Consider Lyapunov function* $V(s)=s_{1,2}-\frac{p}{\mu_2}$. *When* $V(s)\geq\frac{\log N}{4\sqrt{N}}$, *we have*

$$\nabla V(s)\leq -\frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}}.$$

Proof When $V(s)=s_{1,2}-\frac{p}{\mu_2}\geq\frac{\log N}{4\sqrt{N}}$, we have

$$\nabla V(s)=p\mu_1 s_{1,1}-\mu_2 s_{1,2} \qquad \text{(C14)}$$
$$\leq p\mu_1-(p\mu_1+\mu_2)s_{1,2} \qquad \text{(C15)}$$
$$=\mu_1(p-\mu_2 s_{1,2})\leq -\frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}} \qquad \text{(C16)}$$

(C14) and (C15) holds because $s_{1,1}=s_1-s_{1,2}\leq 1-s_{1,2}$ (note this structure is simple yet

critical in proving Lemma 11 and driving iterative SSC in Figure 6); (C15) and (C16) holds because $\frac{1}{\mu_1}+\frac{p}{\mu_2}=1$ implies $p\mu_1+\mu_2=\mu_1\mu_2$. ∎

From Lemma 11, we know $B=\frac{\log N}{4\sqrt{N}}$ and $\gamma=\frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}}$. According to the definition of $q_{\max}$ and $\nu_{\max}$, we have $q_{\max}=\mu_{\max}N$ and $\nu_{\max}=\frac{1}{N}$. Since $\mathcal{E}=\mathcal{S}^{(N)}$ is the entire space, then $\mathbb{P}(S\notin\mathcal{E})=0$, we use Lemma 10 (or Theorem 1 in Bertsimas et al. (2001)) to obtain the following tail bound with $j=\frac{\sqrt{N}\log N}{8}$,.

$$\mathbb{P}(V(S)\geq B+2\nu_{\max}j)=\mathbb{P}\left(S_{1,2}-\frac{p}{\mu_2}\geq\frac{\log N}{2\sqrt{N}}\right) \tag{C17}$$

$$\leq\left(\frac{1}{1+\frac{\mu_1\mu_2}{4\mu_{\max}}\frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{8}} \tag{C18}$$

$$\leq\left(1-\frac{\mu_1\mu_2}{5\mu_{\max}}\frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{8}} \tag{C19}$$

$$\leq e^{-\frac{\mu_1\mu_2\log^2 N}{40\mu_{\max}}} \tag{C20}$$

• (C17) holds by substituting $B=\frac{\log N}{4\sqrt{N}}$, $\nu_{\max}=\frac{1}{N}$ and $j=\frac{\sqrt{N}\log N}{8}$;

• (C17) and (C18) holds based on Lemma 12;

• (C18) and (C19) holds because $\frac{\mu_1\mu_2}{\mu_{\max}}\leq\frac{\sqrt{N}}{\log N}$ for a large N satisfying (4).

C.3 | Proof of Lemma 7: A lower bound on $S_{1,1}$

To prove Lemma 7, we first establish a Lyaponuv drift analysis in Lemma 12.

Lemma 12 *Consider Lyapunov function*

$$V(s)=\frac{\lambda}{\mu_1}-s_{1,1}. \tag{C21}$$

We have

• $\nabla V(s)\leq-\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}$, *when*

$$V(s)\geq\frac{\log N}{2\sqrt{N}}\ \text{ and }\ s_{1,2}\leq\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}};$$

• $\nabla V(s)\leq 1$, *when*

$$V(s)\geq\frac{\log N}{2\sqrt{N}}\ \text{ and }\ s_{1,2}\geq\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}.$$

Proof Assuming $s_{1,2}\leq\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}$ and $\frac{\lambda}{\mu_1}-s_{1,1}\geq\frac{\log N}{2\sqrt{N}}$, we have

$$s_1=s_{11}+s_{12}\leq\frac{p}{\mu_2}+\frac{\lambda}{\mu_1}$$
$$=1-\frac{1}{\mu_1 N^{\alpha}}\leq\lambda+\frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}<1.$$

Therefore, the drift of $V(s)$ is.

$$\nabla V(s)=-\lambda(1-A_1(s))+\mu_1 s_{1,1}-(1-p)\mu_1 s_{2,1}-\mu_2 s_{2,2} \tag{C22}$$

$$\leq\frac{1}{\sqrt{N}}-\lambda+\mu_1 s_{1,1}-(1-p)\mu_1 s_{2,1}-\mu_2 s_{2,2} \tag{C23}$$

$$\leq\frac{1}{\sqrt{N}}-\lambda+\mu_1 s_{1,1} \tag{C24}$$

$$\leq\frac{1}{\sqrt{N}}-\frac{\mu_1}{2}\frac{\log N}{\sqrt{N}} \tag{C25}$$

$$\leq-\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}, \tag{C26}$$

where

• (C22) and (C23) holds because $A_1(s)\leq\frac{1}{\sqrt{N}}$ under any policy in Π;

• (C24) and (C25) holds because $s_{1,1}\leq\frac{\lambda}{\mu_1}-\frac{\log N}{2\sqrt{N}}$.

Assuming $s_{12}>\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}$ and $s_{1,1}\leq\frac{\lambda}{\mu_1}-\frac{\log N}{2\sqrt{N}}$, we have

$$\nabla V(s)=-\lambda(1-A_1(s))+\mu_1 s_{1,1}-(1-p)\mu_1 s_{2,1}-\mu_2 s_{2,2}$$
$$\leq\mu_1 s_{1,1}<1.$$

∎

Let $\mathcal{E}=\left\{s\mid s\leq\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}\right\}$. we have $V(s)=\frac{\lambda}{\mu_1}-s_{1,1}$ satisfying two conditions:

• $\nabla V(s)\leq-\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}$ when $V(s)\geq\frac{\log N}{2\sqrt{N}}$ and $s_{1,2}\in\mathcal{E}$.

• $\nabla V(s)\leq 1$ when $V(s)\geq\frac{\log N}{2\sqrt{N}}$ and $s_{1,2}\notin\mathcal{E}$.

Define $B=\frac{\log N}{2\sqrt{N}}$, $\gamma=\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}$, and $\delta=1$. Combining $q_{\max}\leq\mu_{\max}N$ and $\nu_{\max}\leq\frac{1}{N}$, we have

$$\alpha\leq\frac{1}{1+\frac{\mu_1}{3\mu_{\max}}\frac{\log N}{\sqrt{N}}}\ \text{ and }\ \beta=\frac{1}{\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}}+1.$$

Based on Lemma 10 with $j=\frac{\sqrt{N}\log N}{4}$, we have

$$\mathbb{P}(V(S)\geq B+2\nu_{\max}j)=\mathbb{P}\left(\frac{\lambda}{\mu_1}-S_{1,1}\geq\frac{\log N}{\sqrt{N}}\right) \tag{C27}$$

$$\leq\left(\frac{1}{1+\frac{\mu_1}{3\mu_{\max}}\frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{4}}+\beta\mathbb{P}(S_{1,2}\notin\mathcal{E}) \tag{C28}$$

$$\leq\left(1-\frac{\mu_1}{4\mu_{\max}}\frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{4}}+\frac{4}{\mu_1}\frac{\sqrt{N}}{\log N}e^{-\frac{\mu_1\mu_2\log^2 N}{40\mu_{\max}}} \tag{C29}$$

$$\leq e^{-\frac{\mu_1\log^2 N}{16\mu_{\max}}}+\frac{4}{\mu_1}\frac{\sqrt{N}}{\log N}e^{-\frac{\mu_1\mu_2\log^2 N}{40\mu_{\max}}} \tag{C30}$$

$$\leq\frac{5}{\mu_1}\frac{\sqrt{N}}{\log N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}},\frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}, \tag{C31}$$

where

• (C27) holds by substituting $B = \frac{\log N}{2\sqrt{N}}$, $\nu_{\max} = \frac{1}{N}$ and $j = \frac{\sqrt{N}\log N}{4}$;

• (C27) and (C28) holds based on Lemma 7.3;

• (C28) and (C29) holds because (i) in the first term in (C28), $\frac{\mu_1}{\mu_{\max}} \le \frac{\sqrt{N}}{\log N}$ for a large N satisfying (4), and (ii) the second term in (C28) can be bounded by applying Lemma 6.

C.4 | Proof of Lemma 8: A lower bound on $S_{1,2}$

Lemma 13 Consider Lyapunov function $V(s) = \frac{p\lambda}{\mu_2} - s_{1,2}$.

We have

• $\nabla V(s) \le -\frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}$, when

$$V(s) \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}} \text{ and } s_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}};$$

• $\nabla V(s) \le 1$, when

$$V(s) \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}} \text{ and } s_{1,1} \le \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}.$$

Proof Assuming $V(s) = \frac{p\lambda}{\mu_2} - s_{1,2} \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$ and $s_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$, we have

$$\nabla V(s) = -(p\mu_1 s_{1,1} - \mu_2 s_{1,2}) \tag{C32}$$

$$\le -\left(p\lambda - \frac{p\mu_1 \log N}{\sqrt{N}} - \mu_2 s_{1,2}\right) \tag{C33}$$

$$\le -\frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}, \tag{C34}$$

where

• (C32) and (C33) holds because $s_{1,1} \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$;

• (C33) and (C34) holds because $s_{1,2} \le \frac{p\lambda}{\mu_2} - \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$.

Next, assuming $\frac{p\lambda}{\mu_2} - s_{1,2} \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$ and $s_{1,1} < \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$, we have

$$\nabla V(s) = -(p\mu_1 s_{1,1} - \mu_2 s_{1,2}) \le \mu_2 s_{1,2} \le p\lambda \le 1. \tag{C35}$$ ∎

Defining $\mathcal{E} = \left\{s \mid s \ge \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}\right\}$, we have $V(s) = \frac{p\lambda}{\mu_2} - s_{1,2}$ satisfying two conditions:

• $\nabla V(s) \le -\frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}$ when $V(s) \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$ and $s_{1,1} \in \mathcal{E}$.

• $\nabla V(s) \le 1$ when $V(s) \ge \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$ and $s_{1,1} \notin \mathcal{E}$.

Define $B = \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$, $\gamma = \frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}$ and $\delta = 1$. Combining $q_{\max} = \mu_{\max} N$ and $\nu_{\max} = \frac{1}{N}$, we have

$$\alpha \le \frac{1}{1 + \frac{\mu_2}{2\mu_{\max}}\frac{\log N}{\sqrt{N}}} \text{ and } \beta = \frac{2}{\mu_2}\frac{\sqrt{N}}{\log N} + 1.$$

Based on Lemma 10 with $j = \frac{\sqrt{N}\log N}{4}$, we have

$$\mathbb{P}(V(S) \ge B + 2\nu_{\max} j) = \mathbb{P}\left(\frac{p\lambda}{\mu_2} - S_{1,2} \ge \left(\frac{p\mu_1}{\mu_2} + 1\right)\frac{\log N}{\sqrt{N}}\right) \tag{C36}$$

$$\le \left(\frac{1}{1 + \frac{\mu_2}{2\mu_{\max}}\frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{4}} + \frac{2}{\mu_2}\frac{\sqrt{N}}{\log N}\mathbb{P}(S_{1,1} \notin \mathcal{E}) \tag{C37}$$

$$\le \left(1 - \frac{\mu_2}{3\mu_{\max}}\frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{4}} + \frac{3}{\mu_2}\frac{\sqrt{N}}{\log N}\mathbb{P}(S_{1,1} \notin \mathcal{E}) \tag{C38}$$

$$\le e^{-\frac{\mu_2 \log^2 N}{12\mu_{\max}}} + \frac{15}{\mu_1\mu_2}\frac{N}{\log^2 N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N} \tag{C39}$$

$$\le \frac{16}{\mu_1\mu_2}\frac{N}{\log^2 N}e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}, \tag{C40}$$

where

• (C36) holds by substituting B, $\nu_{\max}$ and j;

• (C36) and (C37) holds due to Lemma 13;

• (C37) and (C38) holds because $\frac{\mu_2}{\mu_{\max}} \le \frac{\sqrt{N}}{\log N}$ for N satisfying (4) in the first term of (C38);

• (C38) and (C39) holds by Lemma 7 to obtain the tail bound in the second term of (C39).

Recall $\frac{p\mu_1}{\mu_2} + 1 = \mu_1$ and the proof is completed.

C.5 | Proof of Lemma 9: SSC on S_1 and $\sum_{i=2}^{b} S_i$

Define $L_{1,1} = \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ and $L_{1,2} = \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$. Recall $w_u = \max((1-p)\mu_1, \mu_2)$, $w_l = \min((1-p)\mu_1, \mu_2)$, $k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right)$ and $c_1 = \frac{w_u b}{w_l}\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right) + 2\mu_1$.

Lemma 14 Consider Lyapunov function

$$V(s) = \min\left\{\lambda + \frac{k\log N}{\sqrt{N}} - s_1, \sum_{i=2}^{b} s_i\right\}. \tag{C41}$$

We have

• $\nabla V(s) \le -\frac{w_u \mu_1 \log N}{\sqrt{N}}$, when

$$V(s) \ge \frac{c_1 \log N}{\sqrt{N}} \text{ with } s_{1,1} \ge L_{1,1} \text{ and } s_{1,2} \ge L_{1,2};$$

• $\nabla V(s) \le w_u$, when

$$V(s) \ge \frac{c_1 \log N}{\sqrt{N}} \text{ with } s_{1,1} \le L_{1,1} \text{ or } s_{1,2} \le L_{1,2}.$$

Proof When $V(s) \ge \frac{c_1 \log N}{\sqrt{N}}$, the following two inequalities hold.

$$s_1 \le \lambda + \frac{(k-c_1)\log N}{\sqrt{N}} = \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}, \quad \text{(C42)}$$

$$\sum_{i=2}^{b} s_i \ge \frac{c_1 \log N}{\sqrt{N}}. \quad \text{(C43)}$$

We have two observations based on (C42) and (C43):

• (C42) implies $A_1(s) \le \frac{1}{\sqrt{N}}$ under any policy in Π;

• (C43) implies $s_2 \ge \frac{c_1}{b}\frac{\log N}{\sqrt{N}}$ because $s_2 \ge s_3 \ge \cdots \ge s_b$, and we have

$$(1-p)\mu_1 s_{2,1} + \mu_2 s_{2,2} \ge w_l s_2 \ge \frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}}, \quad \text{(C44)}$$

where a finite buffer size is required such that the lower bound $w_l s_2 \ge \frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}}$ is meaningful.

We study the Lyapunov drift and consider two cases:

• Suppose $\lambda + \frac{k\log N}{\sqrt{N}} - s_1 \ge \sum_{i=2}^{b} s_i \ge \frac{c_1 \log N}{\sqrt{N}}$.

In this case, $V(s) = \sum_{i=2}^{b} s_i$, and

$$\nabla V(s) \le \lambda(A_1(s) - A_b(s)) - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \quad \text{(C45)}$$

$$\le \frac{1}{\sqrt{N}} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \quad \text{(C46)}$$

$$\le \frac{1}{\sqrt{N}} - \frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}} \quad \text{(C47)}$$

$$\le \frac{1}{\sqrt{N}} - \frac{2 w_u \mu_1 \log N}{\sqrt{N}} \quad \text{(C48)}$$

$$\le -\frac{w_u \mu_1 \log N}{\sqrt{N}}, \quad \text{(C49)}$$

where

– (C45) and (C46) holds because $A_1(s) \le \frac{1}{\sqrt{N}}$ under any policy in Π;.

– (C46) and (C47) holds because (C44);

– (C47) and (C48) holds because $c_1 \ge \frac{w_u b}{w_l} 2\mu_1$.

• Suppose $\sum_{i=2}^{b} s_i > \lambda + \frac{k\log N}{\sqrt{N}} - s_1 \ge \frac{c_1 \log N}{\sqrt{N}}$.

In this case, $V(s) = \lambda + \frac{k\log N}{\sqrt{N}} - s_1$, and

$$\nabla V(s) \quad \text{(C50)}$$

$$\le -\lambda(1 - A_1(s)) + (1-p)\mu_1 s_{1,1} \quad \text{(C51)}$$

$$+ \mu_2 s_{1,2} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \quad \text{(C51)}$$

$$\le \frac{1}{\sqrt{N}} - \lambda + w_u s_1 - (w_u - (1-p)\mu_1)s_{1,1}$$
$$-(w_u - \mu_2)s_{1,2} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}$$

$$\le \frac{1}{\sqrt{N}} - \lambda + w_u(s_1 - L_{1,1} - L_{1,2})$$
$$+ ((1-p)\mu_1 L_{1,1} + \mu_2 L_{1,2})$$
$$- (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \quad \text{(C53)}$$

$$= \frac{1}{\sqrt{N}} + (w_u(k - c_1 + 1 + \mu_1) - (1-p)\mu_1 - \mu_1\mu_2)\frac{\log N}{\sqrt{N}}$$
$$- (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \quad \text{(C54)}$$

$$\le \frac{1}{\sqrt{N}} + (w_u(k - c_1 + 1 + \mu_1)$$
$$- (1-p)\mu_1 - \mu_1\mu_2)\frac{\log N}{\sqrt{N}} - \frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}} \quad \text{(C55)}$$

$$= w_u\left(k - \left(1 + \frac{w_l}{w_u b}\right)c_1 + \mu_1\right)\frac{\log N}{\sqrt{N}}$$
$$+ \frac{1}{\sqrt{N}} - ((1-p)\mu_1 + \mu_1\mu_2 - w_u)\frac{\log N}{\sqrt{N}} \quad \text{(C56)}$$

$$\le w_u\left(k - \left(1 + \frac{w_l}{w_u b}\right)c_1 + \mu_1\right)\frac{\log N}{\sqrt{N}} \quad \text{(C57)}$$

$$\le -\frac{w_u \mu_1 \log N}{\sqrt{N}}, \quad \text{(C58)}$$

where

– (C51) and (C52) holds by adding and substructing $w_u s_1 = w_u(s_{1,1} + s_{1,2})$;

– (C52) and (C53) holds because $s_{1,1}$ and $s_{1,2}$ taking the lower bounds at $L_{1,1}$ and $L_{1,2}$ gives an upper bound;

– (C53) and (C54) holds by substituting $L_{1,1} = \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$, $L_{1,2} = \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$ and $s_1 \le \lambda + \frac{(k-c_1)\log N}{\sqrt{N}}$. We have $s_1 - L_{1,1} - L_{1,2} = (k - c_1 + 1 + \mu_1)\frac{\log N}{\sqrt{N}}$ and $(1-p)\mu_1 L_{1,1} + \mu_2 L_{1,2} = \lambda - ((1-p)\mu_1 + \mu_1\mu_2)\frac{\log N}{\sqrt{N}}$.

– (C54) and (C55) holds by substituting the lower bound of $(1-p)\mu_1 s_{2,1} + \mu_2 s_{2,2}$ in (C44);

– (C55) and (C56) holds by combining the terms with c_1;.

– (C56) and (C57) holds because $((1-p)\mu_1 + \mu_1\mu_2 - w_u)\log N = (\mu_1 + \mu_2 - w_u)\log N \ge 1$ when N satisfies (4);

– (C57) and (C58) holds because $k - \left(1 + \frac{w_l}{w_u b}\right)c_1 \le -2\mu_1$.

Next, we show $\nabla V(s) \le w_u$ based on the upper bounds (C45) and (C51).

- Consider the upper bound in (C45). We have

$$\nabla V(s) \le \lambda(A_1(s) - A_b(s)) - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \le 1 \le w_u,$$

where $1 \le w_u$ holds because $\frac{1}{\mu_1} + \frac{p}{\mu_2} = 1$.

- Consider the upper bound in (C51). We have

$$\begin{aligned}\nabla V(s) &\le -\lambda(1 - A_1(s)) + (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} \\ &\quad - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \\ &\le (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} \le w_u,\end{aligned}$$

where the last inequality holds because $s_{1,1} + s_{1,2} = s_1 \le 1$. ∎

Let $\mathcal{E} = \{s \mid s_{1,1} \ge L_{1,1}, s_{1,2} \ge L_{1,2}\}$. We have $V(s) = \min\left\{\lambda + \frac{k\log N}{\sqrt{N}} - s_1, \sum_{i=2}^{b} s_i\right\}$ satisfying the following two conditions based on Lemma 14:

- $\nabla V(s) \le -\frac{w_u\mu_1 \log N}{\sqrt{N}}$ when $V(s) \ge \frac{c_1 \log N}{\sqrt{N}}$ and $s \in \mathcal{E}$.
- $\nabla V(s) \le w_u$ when $V(s) \ge \frac{c_1 \log N}{\sqrt{N}}$ and $s \notin \mathcal{E}$.

Define $B = \frac{c_1 \log N}{\sqrt{N}}$, $\gamma = \frac{w_u\mu_1 \log N}{\sqrt{N}}$ and $\delta = w_u$. Combining $q_{\max} = \mu_{\max} N$ and $\nu_{\max} = \frac{1}{N}$, we have

$$\alpha \le \frac{1}{1 + \frac{w_u\mu_1 \log N}{\mu_{\max}\sqrt{N}}} \quad \text{and} \quad \beta = \frac{\sqrt{N}}{\mu_1 \log N} + 1.$$

Based on Lemma 10 with $j = \frac{\mu_1\sqrt{N}\log N}{2}$, we have

$$\mathbb{P}(V(S) \ge B + 2\nu_{\max} j) \tag{C59}$$

$$= \mathbb{P}\left(V(S) \ge \frac{c_1 \log N}{\sqrt{N}} + \frac{\mu_1 \log N}{\sqrt{N}}\right) \tag{C60}$$

$$\le \left(\frac{1}{1 + \frac{w_u\mu_1 \log N}{\mu_{\max}\sqrt{N}}}\right)^{\frac{\mu_1\sqrt{N}\log N}{2}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right)\mathbb{P}(s \notin \mathcal{E}) \tag{C61}$$

$$\le \left(1 - \frac{w_u\mu_1}{2\mu_{\max}}\frac{\log N}{\sqrt{N}}\right)^{\frac{\mu_1\sqrt{N}\log N}{2}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right)\mathbb{P}(s \notin \mathcal{E}) \tag{C62}$$

$$\le e^{-\frac{w_u\mu_1^2 \log^2 N}{4\mu_{\max}}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right)\frac{32}{\mu_1\mu_2}\frac{N}{\log^2 N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N} \tag{C63}$$

$$\le \frac{34}{\mu_1^2\mu_2}\frac{N^{1.5}}{\log^3 N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}, \tag{C64}$$

where

- (C60) holds by substituting B, $\nu_{\max}$ and j;
- (C60) and (C61) holds based on Lemma 14;
- (C61) and (C62) holds $\frac{w_u\mu_1}{\mu_{\max}} \le \frac{\sqrt{N}}{\log N}$ for a large N for the first term in (C62);
- (C62) and (C63) holds by applying the union bound on $\mathbb{P}(S \notin \mathcal{E})$ such that

$$\begin{aligned}\mathbb{P}(s \notin \mathcal{E}) &\le \mathbb{P}(s_{1,1} < L_{1,1}) + \mathbb{P}(s_{1,2} < L_{1,2}) \\ &\le \frac{32}{\mu_1\mu_2}\frac{N}{\log^2 N} e^{-\min\left(\frac{\mu_1}{16\mu_{\max}}, \frac{\mu_2}{12\mu_{\max}}, \frac{\mu_1\mu_2}{40\mu_{\max}}\right)\log^2 N}.\end{aligned}$$

APPENDIX D: PROOF OF THE COROLLARY

Under JSQ, a job is discarded or blocked only if all buffers are full, that is, when $N\sum_{i=1}^{b} S_i = Nb$. From Theorem 1, we have

$$\mathbb{P}(\mathcal{B}) = \mathbb{P}\left(N\sum_{i=1}^{b} S_i = Nb\right) = \mathbb{P}\left(\sum_{i=1}^{b} S_i \ge b\right) \tag{D65}$$

$$\le \mathbb{P}\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\} \ge b - \lambda - \frac{k\log N}{\sqrt{N}}\right) \tag{D66}$$

$$\le \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]}{b - \lambda - \frac{k\log N}{\sqrt{N}}} \tag{D67}$$

$$\le \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N} \tag{D68}$$

where (D66) to (D67) holds due to the Markov inequality; and (D67) to (D68) holds because of Theorem 1 and $b - \lambda \ge \frac{8k\log N}{\sqrt{N}}$.

For jobs that are not discarded, the average queueing delay according to Little's law is.

$$\frac{\mathbb{E}\left[\sum_{i=1}^{b} S_i\right]}{\lambda(1 - \mathbb{P}(\mathcal{B}))}.$$

Therefore, the average waiting time is.

$$\begin{aligned}\mathbb{E}[W] &= \frac{\mathbb{E}\left[\sum_{i=1}^{b} S_i\right]}{\lambda(1 - \mathbb{P}(\mathcal{B}))} - 1 \\ &\le \frac{\frac{k\log N}{\sqrt{N}} + \frac{7\mu_{\max}}{\sqrt{N}\log N} + \lambda\mathbb{P}(\mathcal{B})}{\lambda(1 - \mathbb{P}(\mathcal{B}))} \\ &\le \frac{2k\log N}{\sqrt{N}} + \frac{14\mu_{\max} + \frac{16\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N},\end{aligned}$$

where the last inequality holds because $\lambda(1 - \mathbb{P}(\mathcal{B})) \ge 0.5$ under $b - \lambda \ge \frac{8k\log N}{\sqrt{N}}$.

Next, we study the waiting probability $\mathbb{P}(\mathcal{W})$. Define $\overline{\mathcal{W}}$ to be the event that a job entered into the system (not blocked) and waited in the buffer and $\mathbb{P}(\overline{\mathcal{W}})$ is the steady-state probability of $\overline{\mathcal{W}}$. Applying Little's law to the jobs waiting in the buffer,

$$\lambda\mathbb{P}(\overline{\mathcal{W}})\mathbb{E}[T_Q] = \mathbb{E}\left[\sum_{i=2}^{b} S_i\right],$$

where T_Q is the waiting time for the jobs waiting in the buffer. Since $\mathbb{E}[T_Q]$ is lower bounded by $\overline{T}_Q = \min\left\{\frac{1}{\mu_1}, \frac{1}{\mu_2}\right\}$, we have

$$\mathbb{P}(\overline{\mathcal{W}}) \le \frac{\mathbb{E}\left[\sum_{i=2}^{b} S_i\right]}{\lambda \overline{T}_Q}.$$

We now provide a bound on $\mathbb{E}\left[\sum_{i=2}^{b} S_i\right]$. From the work-conserving law, we have

$$\mathbb{E}[S_1] = \lambda(1 - \mathbb{P}(\mathcal{B})) \ge \lambda\left(1 - \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N}\right).$$

Therefore, we have

$$\mathbb{E}[S_1] \ge \lambda - \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N}.$$

From Theorem 1, one has.

$$\mathbb{E}\left[\sum_{i=1}^{b} S_i\right] \le \lambda + \frac{k\log N}{\sqrt{N}} + \frac{7\mu_{\max}}{\sqrt{N}\log N}.$$

The above two inequalities give the following bound on $\mathbb{E}\left[\sum_{i=2}^{b} S_i\right]$:

$$\mathbb{E}\left[\sum_{i=2}^{b} S_i\right] \le \frac{k\log N}{\sqrt{N}} + \frac{7\mu_{\max} + \frac{8\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}.$$

Finally, a job not routed to an idle server is either blocked or waited in the buffer.

$$\mathbb{P}(\mathcal{W}) = \mathbb{P}(\mathcal{B}) + \mathbb{P}(\overline{\mathcal{W}}) \le \mathbb{P}(\mathcal{B}) + \frac{\mathbb{E}\left[\sum_{i=2}^{b} S_i\right]}{\lambda \overline{T}_Q}$$
$$\le \frac{1}{\lambda \overline{T}_Q}\frac{k\log N}{\sqrt{N}} + \frac{1}{\lambda \overline{T}_Q}\frac{7\mu_{\max} + \frac{8\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}.$$

The analysis for Pod is similar, except that.

$$\mathbb{P}(\mathcal{B}) = \mathbb{P}\left(\mathcal{B}\middle| S_b \le 1 - \frac{1}{\mu_1 N^\alpha}\right)\mathbb{P}\left(S_b \le 1 - \frac{1}{\mu_1 N^\alpha}\right) \tag{D69}$$

$$+\mathbb{P}\left(\mathcal{B}\middle| S_b > 1 - \frac{1}{\mu_1 N^\alpha}\right)\mathbb{P}\left(S_b > 1 - \frac{1}{\mu_1 N^\alpha}\right) \tag{D70}$$

$$\le \mathbb{P}\left(\mathcal{B}\middle| S_b \le 1 - \frac{1}{\mu_1 N^\alpha}\right) + \mathbb{P}\left(S_b > 1 - \frac{1}{\mu_1 N^\alpha}\right) \tag{D71}$$

$$\le \left(1 - \frac{1}{\mu_1 N^\alpha}\right)^{\mu_1 N^\alpha \log N} + \mathbb{P}\left(\sum_{i=1}^{b} S_i > b - \frac{b}{\mu_1 N^\alpha}\right) \tag{D72}$$

$$\le \frac{1}{N} + \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]}{b - \lambda - \frac{k\log N}{\sqrt{N}} - \frac{b}{\mu_1 N^\alpha}}. \tag{D73}$$

$$\le \frac{1}{N} + \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N}. \tag{D74}$$

Table D1 Values of $Q_{i,m}$ and $S_{i,m}$ in Figure 4

$Q_{1,1}$	$Q_{2,1}$	$Q_{3,1}$	$Q_{1,2}$	$Q_{2,2}$	$Q_{3,2}$	$Q_{4,2}$	$Q_{5,2}$
0.2	0.2	0.1	0.1	0.1	0.1	0	0.2
$S_{1,1}$	$S_{2,1}$	$S_{3,1}$	$S_{1,2}$	$S_{2,2}$	$S_{3,2}$	$S_{4,2}$	$S_{5,2}$
0.5	0.3	0.1	0.5	0.4	0.3	0.2	0.2

(D71) and (D72) holds because it denotes the probability of the event all sampled d servers have b jobs; (D72) to (D73) holds because $\left(1 - \frac{1}{x}\right)^x \le \frac{1}{e}$ for $x \ge 1$ and the Markov inequality; (D73) to (D74) holds because of Theorem 1 and $b - \lambda \ge \frac{8k\log N}{\sqrt{N}} + \frac{8b}{\mu_1 N^\alpha}$. The remaining analysis is the same.

Finally, for JIQ and I1F, we have not been able to bound $\mathbb{P}(\mathcal{B})$. However,

$$\mathbb{P}(\mathcal{W}) = \mathbb{P}(S_1 = 1) \le \mathbb{P}\left(\sum_{i=1}^{b} S_i \ge 1\right) \tag{D75}$$

$$\le \mathbb{P}\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\} \ge \frac{1}{N^\alpha} - \frac{k\log N}{\sqrt{N}}\right) \tag{D76}$$

$$\le \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]}{\frac{1}{N^\alpha} - \frac{k\log N}{\sqrt{N}}} \tag{D77}$$

$$\le \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]}{\frac{1}{2N^\alpha}} \tag{D78}$$

$$\le \frac{14\mu_{\max}}{N^{0.5-\alpha}\log N}. \tag{D79}$$

(D76) and (D77) holds because of the Markov inequality; (D77) and (D78) holds because $2k \le \frac{N^{0.5-\alpha}}{\log N}$; (D78)–(D79) holds because of Theorem 1. Given the choice of $k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right)$ in Theorem 1, we need the buffer size b to be at the same order, which leads to the finite-buffer assumption.