Large-System Insensitivity of Zero-Waiting Load Balancing Algorithms

Xin Liu ShanghaiTech University liuxin7@shanghaitech.edu.cn Kang Gong University of Michigan, Ann Arbor kanggong@umich.edu

Lei Ying University of Michigan, Ann Arbor leiying@umich.edu

Abstract

This paper studies the sensitivity (or insensitivity) of a class of load balancing algorithms that achieve asymptotic zero-waiting in the sub-Halfin-Whitt regime [24], named LB-zero. Most existing results on zero-waiting load balancing algorithms assume the service time distribution is exponential. This paper establishes the *large-system insensitivity* of LB-zero for jobs whose service time follows a Coxian distribution with a finite number of phases. This result suggests that LB-zero achieves asymptotic zero-waiting for a large class of service time distributions, which is confirmed in our simulations. To prove this result, this paper develops a new technique, called "Iterative State-Space Peeling" (or ISSP for short). ISSP first identifies an iterative relation between the upper and lower bounds on the queue states and then proves that the system lives near the fixed point of the iterative bounds with a high probability. Based on ISSP, the steady-state distribution of the system is further analyzed by applying Stein's method in the neighborhood of the fixed point. ISSP, like state-space collapse in heavy-traffic analysis, is a general approach that may be used to study other complex stochastic systems.

1 Introduction

Zero-waiting load balancing refers to a load balancing algorithm under which a job is routed to an idle server to be processed immediately upon its arrival. The problem has become increasingly important as the amount of modern machine learning (ML) and artificial intelligence (AI) applications running on large-scale data centers explodes. While increasing the number of servers and the processing speed of each server is a critical step to meet the increasing demand, the design of load balancing algorithms that can efficiently utilize available resources to minimize or even eliminate the waiting time of incoming jobs is equally important, especially when a minor increase of latency (e.g. 100 milliseconds) can lead to a significant drop in a cloud-computing provider's revenue (7% drop in sales according to a recent Akamai report [2]).

Significant processes have been made over the past few years on understanding achieving asymptotic zero-waiting (as the system size approaches infinity) in a large-scale data center with distributed queues, including the classic supermarket model [14, 8, 32, 17, 3, 4, 30, 24, 25, 23, 22, 45, 9], models with data locality [40, 31] and models where each job consists of parallel tasks [39, 37, 19], etc.

However, almost all these results assume exponential service time distributions. While each of these results [14, 8, 32, 17, 29, 30, 24, 25, 23, 22, 40, 31, 39, 37, 19, 45, 9] provided important insights of achieving zero-waiting in a practical system, theoretically, it is not clear whether these principles hold for general service times. This is a very important question to answer because it is well-known that service time distributions in real-world systems are not exponential. Understanding a queueing system's performance with general service time remains one of the most important and intensively studied problems in stochastic networks. A concept that excited many theorists in the

area is "insensitivity" [12]. A queueing system is called insensitive if the steady-state distribution of queue lengths is invariant to the service time distribution. Therefore, any conclusion drawn from exponential service time distributions can be applied to general service time distributions. A result that is insensitive is robust and is expected to be widely applicable in practical systems. Unfortunately, insensitivity results are rare and often hold only under some special queueing disciplines such as processor sharing (PS) [12, 6, 21]. One of the reasons is that insensitivity, while appealing, is a very strong notion of "robustness". It requires the steady-state distribution under a general service time distribution to be *exactly the same* as that under the exponential distribution. Some recent studies started to relax it to weaker notions such as insensitivity in the heavy-traffic regime [36] or the large-system regime [7, 32], i.e. insensitivity in the limiting regimes. In the light of these recent developments, this paper addresses the following important question:

Are the zero-waiting algorithms insensitive and if so, in which notion of insensitivity?

1.1 Main Contributions

This paper provides some positive answers to the question above. First, it is well known that most of zero-waiting algorithms, such as join-the-shortest-queue (JSQ) [41] and join-the-idle-queue (JIQ) [26], are not insensitive (according to its original definition). However, we prove that in the sub-Halfin-Whitt regime, LB-zero identified in [24] in fact achieves asymptotic zero-waiting for jobs whose service time follows a Coxian distribution with a finite number of phases. This result establishes the *large-system insensitivity of LB-zero for Coxian service time distributions with a finite number of phases.* Since the Coxian family is dense in the class of positive-valued distributions, our result strongly suggests a load balancing algorithm in the LB-zero family will be able to minimize unnecessary waiting in large-scale data centers for a large class of job size or service times distributions. Our simulations further confirm it.

To prove this result, this paper develops a new technique, called "iterative state-space peeling" (or ISSP for short). ISSP first identifies an iterative relation between upper and lower bounds on the queue states. Then by iteratively "peeling off" the low-probability states, it proves that the system "lives" near the fixed point of the iterative bounds with a high probability. Based on ISSP, the steady-state distribution of the system can be further analyzed by using Stein's method in a small neighborhood of the fixed point. ISSP, like the state-space collapse in the heavy-traffic analysis, is a general technique that may be used to study other complex stochastic systems, e.g. large-system insensitivity of load balancing algorithms for other models like those studied in [29, 39, 40, 37, 38].

We remark that this paper does not establish the large-system insensitivity for an *arbitrary* service time distribution, for which we need to show that our results continue to hold for a large but finite N when the number of phases of the Coxian distribution goes to infinity. This requires an interchange limits arguments or a continuity argument and is an interesting open problem. It is also worth mentioning that a Coxian representation of a probability distribution is non-unique. The choice of the Coxian representation is out the scope of this paper.

1.2 Related Work

Steady-state analysis of distributed queueing systems has been an active research topic since the seminal work on power-of-two-choices [28, 35]. The most popular approach to study a large-scale distributed queueing system is the mean-field approach where the system is approximated using a deterministic dynamical system (a set of ordinary differential equations), called a mean-field model. In the large-system limit (as the number of servers approaches infinity), the steady-state of the stochastic system can often be shown to converge to the equilibrium point of the mean-field model using the interchange of limits (e.g. [35, 42, 44]) or Stein's method (e.g. [43, 16]).

While most studies on this topic assume exponential service time distributions for tractability, the approach has been used to study non-exponential service time distributions theoretically or numerically (see. e.g. [28, 7, 32, 1, 34, 18, 33, 21]). For example, when non-exponential service time

distributions has a decreasing hazard rate (DHR), the system often exhibits a monotonicity property such that the system starting from the empty state is dominated by the system starting from any other state. Leveraging this monotonicity, [7] proved the convergence of power-of-*d*-choices and [32] proved the convergence of JIQ to the corresponding mean-field limit, respectively. Recently, [33] studied load balancing policies under the hyper-exponential service time distribution in the light traffic regime (or in a critical traffic regime). By transforming the hyper-exponential distribution to a Coxian distribution with DHR, the monontoncity property holds in a partial order and the global stability of the mean-field model was established. [21] studied Pod with PS servers for a hyper-exponential distribution of order 2 in the light traffic regime. They also established the global stability result in a spirit similar to [33]. We note [21] considered Pod where d is a constant independent of the number of servers so the system has a non-vanishing delay in the large-system limit. For the Pod algorithm in the LB-zero family, d is a function of N and the algorithm achieves asymptotic zero waiting in the heavy-traffic regime without the DHR assumption.

Without DHR, the results are very limited. [7] proved the convergence of join the least loaded of d queues (LL(d)) for general service time distributions and that of Pod when the load of the system is small (less than 1/4). [15] proved the asymptotic optimality of JIQ under general service time distributions when the normalized load is less than 0.5. Since JIQ is an LB-zero policy, our result confirms the conjecture made in [15] that JIQ is asymptotically optimal for any load less than one, not just less than 0.5. Another significant result is [6], which identifies a set of policies that are insensitive in many-server load-balancing systems and are optimal in the class of insensitive load balancing algorithms. The asymptotic blocking probability of this class of insensitive algorithms are sensitive, so the results in [6, 20] do not apply. We also note that the waiting probability in our paper includes both blocking and being queued in the buffer, so our result implies asymptotic zero blocking of LB-zero in the sub-Halfin-Whitt regime.

[22] is the work most related to this paper, which considers the Coxian-2 distribution and shows that LB-zero achieves asymptotic zero-waiting in the sub-Halfin-Whitt regime. Inspired by [22], this paper develops the ISSP technique for general Coxian distributions with a finite number of phases and establishes its large-system insensitivity. We remark that [22] utilized a key property of Coxian-2 service time distribution that a job in the first phase (phase-1) either departs or enters the last phase (phase-2) immediately, which does not hold under a general Coxian distribution which may have many phases.

In terms of the proof, each step in ISSP utilizes the tail bound in [5] to "peel off" a low probability subspace. The tail bound is based on the Lyapunov drift analysis, and is a critical step to prove state-space collapse in the traditional heavy traffic regime with a fixed number of servers (see e.g., [13, 27, 36]). The key difference is that [13] utilizes the tail bound only once while ISSP repeatedly utilizes the tail bound guided by an iterative relation between the upper and lower bounds.

2 Model and State Representation

We consider a many-server system with N homogeneous servers, where job arrival follows a Poisson process with rate λN with $\lambda = 1 - N^{-\alpha}, 0 < \alpha < 0.5$, i.e. the system is in the sub-Halfin-Whitt regime. We assume the service times follow the Coxian distribution with M phases as shown in Figure 1, where $\mu_m > 0$ is the rate a job finishes phase m when in service and $0 \le p_i < 1, 1 \le i < M$ is the probability that a job enters phase i+1 after completing phase i and $p_M = 1$. Note we assume $\lambda = 1 - N^{-\alpha}$ for the ease of exposition and our results can be easily extended to the case that $\lambda = 1 - \beta N^{-\alpha}$ with any positive constant $\beta > 0$ independent with N. As convention, we define $\sum_{i=a}^{b} x_i = 0$ if a > b and $\prod_{i=a}^{b} x_i = 1$ if a > b for the series $\{x_i, i \ge 1\}$.



Figure 1: The Coxian-*M* Distribution: μ_m is the service rate in phase *m* and p_m is the probability entering phase m + 1 after finishing service in phase *m*.

Without loss of generality, we normalize the mean service time to be one, i.e.

$$\sum_{m=1}^{M} v_m = 1 \quad \text{with} \quad v_m = \frac{\prod_{i=1}^{m-1} p_i}{\mu_m}$$

where v_m is viewed as the average time spent in phase *m* for a job. Given the unit service rate, λ is the normalized load of the system and λv_m is the normalized load of jobs in phase *m*.

Taking Coxian-3 distribution as an example (see Figure 2), a job is colored in black if it waits in the buffer, and colored in light red, blue, and green when it is in phase 1, 2 and 3, respectively. Jobs are served with the FIFO discipline and we assume 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). The assumption of finite buffer is imposed due to a technical reason and will be explained later in the paper. Relaxing the finite-buffer assumption is not trivial technically but we conjecture our results hold without this assumption.



Figure 2: Load Balancing in Many-Server Systems with Coxian-3 : jobs colored in black, light red, blue, and green represent jobs in the buffer, in phase 1, 2, and 3, respectively.

To represent the system, define $Q_{j,m}(t)$ $(m = 1, 2, \dots, M)$ to be the fraction of servers which have j jobs at time t and the one in service is in phase m. Because an idle server does not have a phase, we define $Q_{0,1}(t)$ to be the fraction of servers that are idle at time t and $Q_{0,m}(t) = 0, \forall 2 \leq m \leq M$ for convenience. We stack $Q_{j,m}(t)$ to a matrix $Q(t) \in \mathbb{R}^{b \times M}$ such that the (j, m)th entry of the matrix is $Q_{j,m}(t)$. We further define $S_{i,m}(t) = \sum_{j \geq i} Q_{j,m}(t)$ and $S_i(t) = \sum_{m=1}^M S_{i,m}(t)$. Therefore, $S_{i,m}(t)$ 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. Stack $S_{j,m}(t)$ to be a matrix S(t) such that the (j,m)th entry of the matrix is $S_{j,m}(t)$. Since Q(t) and S(t) have a one-to-one mapping, we focus on S(t) throughout the paper. We consider load balancing policies which dispatch jobs to servers based on S(t) and under which the finite-state CTMC $\{S(t), t \geq 0\}$ is irreducible, and so it

has a unique stationary distribution. This includes well-known load balancing policies such as JSQ [41, 14, 8], JIQ [26, 32], I1F [17] and Pod [28, 35]

Let $Q_{j,m}$ be a random variable that has the distribution of $Q_{j,m}(t)$ at steady state. Correspondingly, define $S_{i,m} = \sum_{j\geq i} Q_{j,m}$ 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, both at steady state. Consider a system with 10 servers and Coxian-3 service time distribution. A realization of state representation $S_{i,m}$ is shown in Figure 3 and Table 1. Define $S \in \mathbb{R}^{b \times M}$ to be a matrix such that the (i, m)th entry is $S_{i,m}$ and $s \in \mathbb{R}^{b \times M}$ to be a realization of *S*. Define $S^{(N)}$ to be a set of *s* as follows

$$\mathcal{S}^{(N)} = \left\{ s \; \middle| \; 1 \ge s_{1,m} \ge \dots \ge s_{b,m} \ge 0, \; 1 \ge \sum_{m=1}^{M} s_{1,m}, \; N s_{i,m} \in \mathbb{N}, \; \forall i,m \right\}, \tag{1}$$

i.e., $\mathcal{S}^{(N)}$ is the set of all possible s in a system with N servers.



Figure 3: An example of the realization of $S_{i,m}$ in a system with 10 servers and Coxian-3 service time distribution

$Q_{1,1}$	$Q_{2,1}$	$Q_{3,1}$	$Q_{1,2}$	$Q_{2,2}$	$Q_{3,2}$	$Q_{1,3}$	$Q_{2,3}$	$Q_{3,3}$	$Q_{4,3}$	$Q_{5,3}$
0.2	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.1	0.1
$S_{1,1}$	$S_{2,1}$	$S_{3,1}$	$S_{1,2}$	$S_{2,2}$	$S_{3,2}$	$S_{1,3}$	$S_{2,3}$	$S_{3,3}$	$S_{4,3}$	$S_{5,3}$

Table 1: The corresponding values of $Q_{i,m}$ and $S_{i,m}$ in Figure 3: The system in Figure 3 includes four busy servers that are serving jobs in phase 1, including two servers without any waiting jobs $(Q_{1,1} = 0.2)$, one server with one waiting job $(Q_{2,1} = 0.1)$, and one server with two waiting jobs $(Q_{3,1} = 0.1)$; four busy servers that are serving jobs in phase 2, including one server without any waiting jobs $(Q_{1,2} = 0.1)$, two server with one waiting job $(Q_{2,2} = 0.2)$, and one server with two waiting jobs $(Q_{3,2} = 0.1)$; and two busy servers that are serving jobs in phase 3, including one server with three waiting jobs $(Q_{4,3} = 0.1)$ and one server with four waiting jobs $(Q_{5,3} = 0.1)$.

3 Main Results

Before introducing the main results, we first define several constants that will be used throughout the paper:

$$a_m = \frac{\mu_m}{p_1 \mu_1 + \mu_m} \qquad \qquad 2 \le m \le M$$

$$b_m = (1 - a_m) \left(1 + \sum_{r=m+1}^{M} \frac{v_r}{v_1} \right) - \frac{a_m v_m}{v_1} \qquad 2 \le m \le M$$

$$\xi = \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j$$

$$c_m = 5(1-a_m) \sum_{r=m+1}^{M} (r-1)v_r + 5a_m \sum_{r=2}^{m-1} \frac{\mu_r v_r}{\mu_m} + 5(m-2)a_m v_m + 5 - a_m \qquad 2 \le m \le M$$

$$C_M = \sum_{m=2}^{M} c_m \prod_{j=m+1}^{M} a_j.$$

These constants are positive constants and $0 < \xi < 1$. The proof can be found in Appendix A. Their values depend on the Coxian-*M* distribution but are independent of *N*.

We further define $A_1(s)$ to be the probability that an incoming job is routed to a busy server conditioned on that the system is in state $s \in \mathcal{S}^{(N)}$; i.e.

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

We now consider the set of zero-waiting load balancing policies, named as LB-zero,

LB-zero:
$$\Pi = \left\{ \pi \mid \text{Under policy } \pi, A_1(s) \leq \frac{1}{\sqrt{N}} \text{ for any } s \in \mathcal{S}^{(N)} \text{ such that } s_1 \leq 1 - \frac{1}{N^{\alpha} \log N} \right\}$$

Note that this class of policies is similar to the one considered in [24]. Several well-known policies satisfy this condition, as summarized in Table 2.

Load Balancing Policy	Description	Condition
Join-the-Shortest-Queue	route an incoming job to the least loaded server	$A_1(s) = 0$ for $s_1 < 1$
	route an incoming job to an idle server	
Join-the-Idle-Queue	if available and otherwise, to a server	$A_1(s) = 0$ for $s_1 < 1$
	chosen uniformly at random.	
	route an incoming job to an idle server	
Idle-One-First	if available; to a server with one job if available;	$A_1(s) = 0$ for $s_1 < 1$
	and otherwise, to a randomly selected server.	
	sample d servers uniformly at random and	For sufficiently large N ,
Power-of- <i>a</i> -Choices	route the job to the least loaded	$A_1(s) \leq \frac{1}{\sqrt{N}}$
with $a \ge N \approx \log^2 N$,	server among the d servers.	for $s_1 \le 1 - \frac{\sqrt{N}}{N^{\alpha} \log N}$

Table 2: Examples of LB-Zero Policies: Join-the-Shortest-Queue, Join-the-Idle-Queue, Idle-One-First, and Power-of-*d*-Choices with a carefully chosen *d*.

To prove the large-system insensitivity of LB-zero, we first show that $S_{1,m}$ is "close" to $s_{1,m}^* = \lambda v_m$, which is the normalized load from phase-*m* of the jobs, and is also the equilibrium point of the mean-field system assuming zero-waiting (details can be found in Section 4 and 5). We call s^* the zero-waiting equilibrium. Theorem 1 shows that at the steady-state, $S_{1,m}$ concentrates around the zero-waiting equilibrium $s_{1,m}^*$ for large N. The proof of this theorem can be found in Section 6.

Theorem 1 (High Probability Bound). Define $\theta_m = \frac{6\mu_1 v_m + 5(m-1)v_m}{C}, \forall 1 \leq m \leq M$ with $C = \sqrt{\frac{2\bar{v}^2 \log(1/\xi)}{3M + (3M+4)\log(1/\xi)}}$ and $\bar{v} = \min_m v_m$. For any LB-zero policy in Π , the following bound holds

$$\mathbb{P}\left(s_{1,m}^* + \frac{\theta_m \log N}{\sqrt{N}} \le S_{1,m} \le s_{1,m}^* + \frac{1}{N^{\alpha}} - \frac{\sum_{r \ne m} \theta_r \log N}{\sqrt{N}}\right) \ge 1 - \frac{M}{N^3}$$

when N satisfies

$$\min\left\{\sum_{m=1}^{M} \theta_m, \frac{C(1-\xi)}{2\mu_1 C_M}\right\} N^{0.5-\alpha} \ge \log N \ge \max\left\{\frac{2\mu_1}{1-\xi}, \frac{C}{\mu_1}\right\}.$$
(2)

Remark 1. Theorem 1 shows that $S_{1,m}$ differs from $s_{1,m}^*$ by at most $\max\left\{\frac{\theta_m \log N}{\sqrt{N}}, \frac{1}{N^{\alpha}}\right\}$ with a probability at least $1-M/N^3$, which implies that the convergence the steady-state $S_{1,m}$, $\forall m$ to the zero-waiting equilibrium as $N \to \infty$ in probability and mean-square senses. To be best of our knowledge, this is the first result to establish such a steady-state convergence for a load balancing system under Coxian-M service time distributions in the heavy-traffic regime. Since the high probability bound holds for a large but finite N, it also provides the rate of convergence.

From Theorem 1, it is not clear whether the probability of waiting approaches zero under an LB-zero policy, which will be studied in the next theorem. Let \mathcal{W} denote the event that an incoming job is routed to a busy server in the system, and $\mathbb{P}(\mathcal{W})$ denote the probability of this event at steady-state. We have the following result on the waiting probability. The proof of Theorem 2 can be found in Section 7.

Theorem 2. Define $w_m = (1 - p_m)\mu_m$, $w_u = \max_m w_m$, $w_l = \min_m w_m$, $\mu_{\max} = \max_m \mu_m$, $\zeta = \frac{4w_u b}{w_l} [(\frac{1}{w_l} - \frac{1}{w_u}) \sum_m \theta_m w_m + \frac{1}{w_l} + 6]$, and $k = \frac{\sum_m \theta_m w_m}{w_u} + (1 + \frac{w_l}{4w_u b})\zeta - \sum_m \theta_m$. Under an LB-zero policy in II, the following result holds

$$\mathbb{P}(\mathcal{W}) \le \frac{1}{\sqrt{N}} + \frac{10\mu_{\max} + 4}{N^{0.5 - \alpha} \log N}.$$
(3)

when N satisfies

$$\min\left\{\frac{1}{2k}, \sum_{m=1}^{M} \theta_m, \frac{C(1-\xi)}{2\mu_1 C_M}\right\} N^{0.5-\alpha} \ge \log N \ge \max\left\{\log\left(\frac{1}{\xi}\right), \frac{2\mu_1}{1-\xi}, \frac{4b}{w_l\zeta}, \frac{C}{\mu_1}, 2\right\}.$$
(4)

Remark 2. Theorem 2 shows the waiting probability is $O(1/N^{0.5-\alpha})$ for a large but finite N, which implies the asymptotic zero waiting, i.e. $\mathbb{P}(W) \to 0$ as $N \to \infty$ in the sub-Halfin-Whitt regime. This asymptotic result implies LB-zero is large-system insensitive to Coxian-M distributions with a finite number of phases.

Next, we will establish these two main results. We first introduce the system dynamic of LB-zero in Section 4 and present "iterative state-space peeling" (ISSP) in Section 5, which is used to prove that the system lives near a limiting regime in Theorem 1 in Section 6 and to prove the large system insensitivity in Theorem 2 in Section 7.

4 System Dynamics

Define $e_{i,m} \in \mathbb{R}^{b \times M}$ to be a $b \times M$ -dimensional matrix with the (i, m)th entry being 1/N and all other entries being 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 is in phase m given the system state s, i.e.

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

Given state s (or the corresponding q) of the CTMC, each of the following three events triggers a state transition, which is illustrated individually in Figure 4.

• Event 1: A job arrives and is routed to a server that has i - 1 jobs and the job in service is in phase m as in the left figure in Figure 4. When this occurs, $q_{i,m}$ increases by 1/N, and $q_{i-1,m}$ decreases by 1/N (note m = 1 if i = 1 since we define the faction of idle servers to be $q_{0,1}$). So the CTMC has the following transition:

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

$$s \to s + e_{i,m},$$

where the transition of s can be verified according to the definition $s_{i,m} = \sum_{j\geq i} q_{j,m}$ so only $s_{i,m}$ increasing by 1/N. This event occurs with rate

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

where $A_{i-1,m}(s) - A_{i,m}(s)$ is the probability that an incoming job is routed to a server that has i-1 jobs and the job in service is in phase m.

• Event 2: A server with *i* jobs finishes serving a job in phase *m*. The job departs from the system without entering into the next phase as in the middle figure in Figure 4. When this event occurs, $q_{i,m}$ decreases by 1/N and $q_{i-1,1}$ increases by 1/N, so the CTMC has the following transition:

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

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

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

$$\mu_m N q_{i,m} (1 - p_m),$$

where $\mu_m Nq_{i,m}$ is the rate at which a job in phase *m* finishes the service and $(1 - p_m)$ is the probability that a job finishes phase *m* and departs from the system immediately.

• Event 3: A server with *i* jobs finishes serving a job in phase *m*, and the job enters the next phase m + 1 as shown in the right figure in Figure 4. When this event occurs, a server in state (i, m) transits to state (i, m + 1), so $q_{i,m}$ decreases by 1/N and $q_{i,m+1}$ increases by 1/N. Therefore, the CTMC has the following transition:

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

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

where the transition of s holds because $s_{i,m}$ decreases by 1/N for any $j \leq i$ and $s_{j,m+1}$ increases by 1/N for any $j \leq i$. This event occurs with rate

$\mu_m N q_{i,m} p_m,$

where $\mu_m Nq_{i,m}$ is the rate at which a job in phase *m* finishes the service and p_m is the probability that a job enters phase m + 1 after finishing phase *m*.



Figure 4: Illustrations of State Transitions on $(q_{i,m} \text{ or } s_{i,m})$ Triggered by the Three Events: 1) a job arrives to a server with i-1 jobs and the job in service in phase m; 2) a server with i jobs finishes a job in phase m, and the job departs from the system; 3) a server with i jobs finishes a job in phase m, and the job enters into the next phase.

Based on the three events above, we focus on $(S(t) : t \ge 0)$ because S(t) and Q(t) has one-to-one mapping and the dynamics of S(t) have a simpler form. Define G to be the generator of CTMC $(S(t) : t \ge 0)$. Given function $f : S^{(N)} \to \mathbb{R}$, we have

$$Gf(s) = \sum_{i=1}^{b} \sum_{m=1}^{M} \left[\lambda N(A_{i-1,m}(s) - A_{i,m}(s))(f(s+e_{i,m}) - f(s)) \right]$$
(5)

$$+(1-p_m)\mu_m Nq_{i,m} \left(f\left(s - \sum_{j=1}^i e_{j,m} + \sum_{j=1}^{i-1} e_{j,1}\right) - f(s) \right)$$
(6)

$$+p_{m}\mu_{m}Nq_{i,m}\left(f\left(s-\sum_{j=1}^{i}e_{j,m}+\sum_{j=1}^{i}e_{j,m+1}\right)-f(s)\right)\right].$$
(7)

To understand the dynamics better, we write down the mean-filed model (MFM) according to the generator:

$$\dot{s}_{i,1} = \lambda(A_{i-1,1}(s) - A_{i,1}(s)) + \sum_{m=2}^{M} (1 - p_m)\mu_m s_{i+1,m} - \mu_1 s_{i,1},$$
(8)

$$\dot{s}_{i,m} = \lambda (A_{i-1,m}(s) - A_{i,m}(s)) + p_{m-1}\mu_{m-1}s_{i,m-1} - \mu_m s_{i,m}, \ \forall m \ge 2.$$
(9)

This mean-field model is nonlinear in s because $A_{i,m}(s)$ is a nonlinear function in s, and its equilibrium point is difficult to calculate in general. However, suppose zero-waiting occurs, i.e., $s_{i,m} = 0, \forall i \geq 2$, and the faction of jobs dropped is negligible, then we can obtain the following equilibrium

$$s_{1,m}^* = \lambda \frac{\prod_{i=1}^{m-1} p_i}{\mu_m} = \lambda v_m \quad \text{and} \quad s_{i,m}^* = 0 \quad \forall i \ge 2.$$

We call s^* zero-waiting equilibrium because it is a conjectured equilibrium by assuming zerowaiting. In this following analysis, we will not solve mean-field model (8)-(9) to check whether its equilibrium is close to s^* . Instead, we will directly prove $S_{1,m}$ concentrates around $s^*_{1,m}$ and zero waiting occurs at the steady-state with a high probability.

5 Iterative State-Space Peeling (ISSP)

In this section, we illustrate the key idea of ISSP, which will be applied to prove Theorem 1. Intuitively, the original stochastic system $S_{i,m}(t)$ and the steady-state $S_{i,m}$ would be close to the MFM $s_{i,m}(t)$ and the zero-waiting equilibrium $s_{i,m}^*$, respectively. However, due to "non-monotonicity" of the system, it is extremely challenging to justify this argument. To tackle the challenge, we develop a new technique, called "Iterative State-Space Peeling" (ISSP), which first identifies an iterative relation between the upper and lower bounds on the queue states, and then proves that the system lives in a regime concentrated around the "fixed point" of the iterative bounds with a high probability.

In particular, we focus on $S_{1,m}$ the number of busy servers with the job in service in phase m because we hypothesize $S_{i,m} \to 0, \forall i \geq 2$ (in other words, the fraction of servers with any waiting jobs is negligible in a large system).

Let $L_{1,m}(n)$ denote a high probability lower bound on $S_{1,m}$ and $U_m(n)$ be a high probability upper bound on $\sum_{r=2}^{m} S_{1,r}$, established at the *n*th step of ISSP, i.e.

$$\mathbb{P}(S_{1,m} \ge L_{1,m}(n)) \text{ and } \mathbb{P}\left(\sum_{r=2}^{m} S_{1,r} \le U_m(n)\right)$$

are close to one. Our goal is to show that as n increases, $L_{1,m}(n)$ and $U_m(n)$ approach the zerowaiting equilibrium $s_{1,m}^*$. Taking Coxian-3 distribution as an example, we need to show that as nincreases $L_{1,m}(n) \to s_{1,m}^*, \forall m, U_2 \to s_{1,2}^*$ and $U_3 \to s_{1,2}^* + s_{1,3}^*$.

5.1 ISSP: Iterative lower and upper bounds

Our "iterative state-space peeling" (ISSP) is based on the following iterative relation between the upper and lower bounds on the queue states, denoted by $L_{1,m}$ and $U_m, \forall m \ge 2$:

$$L_{1,1}(n+1) \approx \min\{s_{1,1}^*, 1 - U_M(n)\}$$
(10)

$$L_{1,m}(n+1) \approx \frac{v_m}{v_{m-1}} L_{1,m-1}(n+1) \tag{11}$$

$$U_m(n+1) \approx 1 - a_m - b_m L_{1,1}(n+1) + a_m U_{m-1}(n+1)$$
(12)

where the initial condition $L_{1,m}(0) = 0, \forall m$, and $U_m(0) = 1, \forall m \geq 2$, and " \approx " is used because we ignore diminishing terms (e.g., $\frac{\log N}{\sqrt{N}}$) in the equations (10)-(12) for explaining the intuition. From(10)-(12), we can obtain an recursive equation for $L_{1,1}$:

$$L_{1,1}(n+1) \approx \min\{s_{1,1}^*, v_1 + \xi(L_{1,1}(n) - v_1)\}, \ s_{1,1}^* = \lambda v_1 \text{ and } 0 < \xi < 1.$$
(13)

Therefore, as $n \to \infty$, $L_{1,m}(n) \to s_{1,m}^*$ and $U_m(n) \to \sum_{r=2}^m s_{1,r}^*$.

To provide the intuition behind (10)-(12), we consider the mean-field model under JSQ as an example and focus on $s_{1,m}$ in (8)-(9) by ignoring $s_{i,m}, \forall i \geq 2$, that is,

$$\dot{s}_{1,1} = \lambda \mathbb{I}(s_1 < 1) - \mu_1 s_{1,1},\tag{14}$$

$$\dot{s}_{1,m} = p_{m-1}\mu_{m-1}s_{1,m-1} - \mu_m s_{1,m}, \ \forall m \ge 2,$$
(15)

where the equilibrium can be verified to be $s_{1,m}^* = \lambda v_m, \forall m$.

We next carefully analyze (14)-(15) to establish (10)-(12). To derive the lower and upper bounds on the equilibrium point of a dynamical system x(t) (or $s_{1,m}(t)$), we use the following straightforward ideas.

• If

$$\dot{x}(t) > L - x(t),\tag{16}$$

then $x(t) \ge L$ when t is sufficiently large because otherwise x(t) continues to increase.

• If

$$\dot{x}(t) < U - x(t),\tag{17}$$

then $x(t) \leq U$ when t is sufficiently large because otherwise, x(t) continues to decrease.

In the following, we will explain (10)-(12) based on the ideas above. The explanation is not a rigorous proof. The detailed proof will be presented later. We will ignore iteration index n occasionally when confusion does not arise.

The intuition to obtain (10): We start with the dynamic of $s_{1,1}$ in (14), which is

$$\dot{s}_{1,1} = \lambda \mathbb{I}(s_1 < 1) - \mu_1 s_{1,1},$$

Given $\sum_{m=2}^{M} s_{1,m} < U_M$, we have

$$\dot{s}_{1,1} = \lambda - \mu_1 s_{1,1}$$

when $s_{1,1} < 1 - U_M$, which implies $\dot{s}_{1,1} > 0$ when $s_{1,1} < \min\{1 - U_M, s_{1,1}^*\}$. Therefore, at the equilibrium point, $s_{1,1} \ge L_{1,1} \triangleq \min\{1 - U_M, s_{1,1}^*\}$ because otherwise, $s_{1,1}$ will continue to increase because $\dot{s}_{1,1} > 0$.

The intuition to obtain (11): Consider the dynamic of $s_{1,m}, \forall m \ge 2$ in (15):

$$\dot{s}_{1,m} = p_{m-1}\mu_{m-1}s_{1,m-1} - \mu_m s_{1,m}, \forall m \ge 2.$$

Given $s_{1,m-1} \ge L_{1,m-1}$, we have

$$\dot{s}_{1,m} \ge p_{m-1}\mu_{m-1}L_{1,m-1} - \mu_m s_{1,m}, \forall m \ge 2,$$

which implies at the equilibrium point,

$$s_{1,m} \ge L_{1,m} \triangleq \frac{p_{m-1}\mu_{m-1}}{\mu_m} L_{1,m-1} = \frac{v_m}{v_{m-1}} L_{1,m-1}$$

because otherwise, $s_{1,m}$ will continue to increase because $\dot{s}_{1,m} > 0$.

Note $\frac{v_m}{v_{m-1}} = \frac{s_{1,m}^*}{s_{1,m-1}^*}$, so we have $\frac{L_{1,m}}{L_{1,m-1}} = \frac{s_{1,m}^*}{s_{1,m-1}^*}$, which means the ratio of the lower bounds is the same as that of the corresponding equilibrium points.

The intuition to obtain (12): We focus on the dynamic of $\sum_{r=2}^{m} s_{1,r}$ that

$$\sum_{r=2}^{m} \dot{s}_{1,r} = p_1 \mu_1 s_{1,1} - \sum_{r=2}^{m-1} (1-p_r) \mu_r s_{1,r} - \mu_m s_{1,m}$$

$$\leq p_1 \mu_1 \left(1 - \sum_{r=2}^{M} s_{1,r} \right) - \sum_{r=2}^{m-1} (1-p_r) \mu_r s_{1,r} - \mu_m s_{1,m}$$

$$= p_1 \mu_1 \left(1 - \sum_{r=m+1}^{M} s_{1,r} \right) - \sum_{r=2}^{m-1} (1-p_r) \mu_r s_{1,r} - (p_1 \mu_1 + \mu_m) \sum_{r=2}^{m} s_{1,r} + \mu_m \sum_{r=2}^{m-1} s_{1,r}.$$

Given $s_{1,m} \ge L_{1,m}, \forall m \text{ and } \sum_{r=2}^{m-1} s_{1,r} \le U_{m-1}$, we have

$$\sum_{r=2}^{m} \dot{s}_{1,r} \le p_1 \mu_1 \left(1 - \sum_{r=m+1}^{M} L_{1,r} \right) - \sum_{r=2}^{m-1} (1 - p_r) \mu_r L_{1,r} - (p_1 \mu_1 + \mu_m) \sum_{r=2}^{m} s_{1,r} + \mu_m U_{m-1},$$

which implies at the equilibrium point,

$$\sum_{r=2}^{m} s_{1,r} \le U_m \triangleq \frac{p_1 \mu_1 \left(1 - \sum_{r=m+1}^{M} L_{1,r}\right) - \sum_{r=2}^{m-1} (1 - p_r) \mu_r L_{1,r} + \mu_m U_{m-1}}{p_1 \mu_1 + \mu_m}$$

because otherwise, $\sum_{r=2}^{m} s_{1,r}$ will continue to decrease because $\sum_{r=2}^{m} \dot{s}_{1,r} < 0$. By invoking $L_{1,m} = \frac{v_m}{v_{m-1}} L_{1,m-1}$, we have

$$\sum_{r=m+1}^{M} L_{1,r} = \sum_{r=m+1}^{M} \frac{v_r}{v_1} L_{1,1},$$
$$\sum_{r=2}^{m-1} (1-p_r) \mu_r L_{1,r} = p_1 \mu_1 L_{1,1} - \mu_m L_{1,m} = p_1 \mu_1 L_{1,1} - \frac{\mu_m v_m}{v_1} L_{1,1}.$$

Therefore, we have

$$U_{m} = \frac{p_{1}\mu_{1}}{p_{1}\mu_{1} + \mu_{m}} - \frac{p_{1}\mu_{1}}{p_{1}\mu_{1} + \mu_{m}} \left(1 + \sum_{r=m+1}^{M} \frac{v_{m}}{v_{1}}\right) L_{1,1} + \frac{\mu_{m}}{p_{1}\mu_{1} + \mu_{m}} \frac{v_{m}}{v_{1}} L_{1,1} + \frac{\mu_{m}}{p_{1}\mu_{1} + \mu_{m}} U_{m-1}$$
$$= \frac{p_{1}\mu_{1}}{p_{1}\mu_{1} + \mu_{m}} - \left(\frac{p_{1}\mu_{1}}{p_{1}\mu_{1} + \mu_{m}} \left(1 + \sum_{r=m+1}^{M} \frac{v_{m}}{v_{1}}\right) - \frac{\mu_{m}}{p_{1}\mu_{1} + \mu_{m}} \frac{v_{m}}{v_{1}}\right) L_{1,1} + \frac{\mu_{m}}{p_{1}\mu_{1} + \mu_{m}} U_{m-1}$$
$$= 1 - a_{m} - b_{m}L_{1,1} + a_{m}U_{m-1}.$$

where the last equality holds by the definitions of a_m and b_m .

5.2 ISSP: An illustrative example

To demonstrate ISSP, we consider JSQ with Erlang-3 distribution and no buffer, i.e., Erlang-3 with b = 1.

The mean-field model under JSQ with Erlang-3 and b = 1.

With the Erlang-3 service time distribution, we have

$$p_1 = p_2 = p_3 = 1$$
 and $\mu_1 = \mu_2 = \mu_3 = 3$.

So the MFM in this case is

$$\begin{split} \dot{s}_{1,1} &= \lambda \mathbb{I}_{\{s_1 < 1\}} - 3s_{1,1} \\ \dot{s}_{1,2} &= 3s_{1,1} - 3s_{1,2} \\ \dot{s}_{1,3} &= 3s_{1,2} - 3s_{1,3} \\ s_{i,m} &\equiv 0, \forall i \geq 2, \forall m. \end{split}$$

ISSP for JSQ with Erlang-3 and b = 1. The values of the key parameters in iterative equations in this case are $a_1 = a_2 = a_3 = \frac{1}{2}$, $b_1 = 1$, $b_2 = \frac{1}{2}$, $b_3 = 0$, and $v_1 = v_2 = v_3 = \frac{1}{3}$. The corresponding iterative equations are

$$L_{1,1}(n+1) \approx \min\left\{\frac{\lambda}{3}, 1 - U_3(n)\right\}$$
$$L_{1,2}(n+1) \approx L_{1,1}(n+1), \quad \text{and} \ L_{1,3}(n+1) \approx L_{1,2}(n+1)$$
$$U_2(n+1) \approx \frac{1}{2} - \frac{1}{2}L_{1,1}(n+1), \quad \text{and} \ U_3(n+1) \approx \frac{1}{2} + \frac{1}{2}U_2(n+1)$$

The iterative relation in terms of $L_{1,1}(n)$ is

$$L_{1,1}(n+1) = \min\left\{\frac{\lambda}{3}, \frac{1}{4} + \frac{1}{4}L_{1,1}(n)\right\}$$

which implies that $L_{1,1}(n) \to \frac{\lambda}{3}$ as $n \to \infty$, and

$$L_{1,1}(n) \rightarrow \frac{\lambda}{3}, L_{1,2}(n) \rightarrow \frac{\lambda}{3}, L_{1,3}(n) \rightarrow \frac{\lambda}{3}, U_2(n) \rightarrow \frac{\lambda}{3}, U_3(n) \rightarrow \frac{2\lambda}{3}$$

The iterative procedure has been visualized in Figure 5. In each iteration n, we first establish $L_{1,1}(n+1)$ in light red based on $U_3(n)$ from the last iteration; then obtain $L_{1,1}(n+1) \approx L_{1,2}(n+1) \approx L_{1,3}(n+1)$ in light (blue, green); and finally refine $U_2(n+1)$ and $U_3(n+1)$ in light purple, which in turn will improve $L_{1,1}$ in the next iteration.

In the following sections, we formalize the ISSP, which is then combined with Stein's method to prove Theorems 1 and 2. A roadmap can be found in Figure 6 that demonstrates the relationship of the key lemmas and theorems.



Figure 6: A Roadmap for Proving Theorems 1 and 2: Lemmas 1, 2, and 3 establish the iterative upper and lower bounds, which are used to establish Lemma 4, 5 and Theorem 1. Together with Stein's method in Lemma 7, zero-waiting is established in Lemma 6 and Theorem 2.

6 Proof of Theorem 1 based on ISSP

In this section, we present the formal statements of the iterative equations (10)-(12) as individual lemmas, which are used to prove Theorem 1.

Recall that the positive constant $C = \sqrt{\frac{2\bar{v}^2 \log(1/\xi)}{3M + (3M+4)\log(1/\xi)}}$, whose value is chosen to control the tail probability as we will see it later. Recall $a_m = \frac{\mu_m}{p_1\mu_1 + \mu_m}$, $b_m = (1-a_m)\left(1 + \sum_{r=m+1}^M \frac{v_r}{v_1}\right) - \frac{a_m v_m}{v_1}$, and $\bar{v} = \min_m v_m$. Recall $c_m = 5(1-a_m)\sum_{r=m+1}^M (r-1)v_r + 5a_m\sum_{r=2}^{m-1} \frac{\mu_r v_r}{\mu_m} + 5(m-2)a_m v_m + 5 - a_m$, $C_M = \sum_{m=2}^M c_m \prod_{j=m+1}^M a_j$, and $\Delta = \frac{\log N}{\sqrt{N}}$. We further define

$$\epsilon_1(n+1) = e^{-\frac{\log^2 N}{C^2}} + \left(\frac{C}{\Delta} + 1\right) \sigma_M(n)$$

$$\epsilon_m(n+1) = e^{-\frac{v_m^2 \log^2 N}{C^2}} + \left(\frac{C}{v_m \Delta} + 1\right) \epsilon_{m-1}(n+1)$$

$$\sigma_m(n+1) = e^{-\frac{\log^2 N}{C^2}} + \left(\frac{C}{\Delta} + 1\right) \left(\sigma_{m-1}(n+1) + \sum_{m=1}^M \epsilon_m(n+1)\right)$$

with initial values $\epsilon_m(0) = \sigma_m(0) = 0, \forall m$. These are the constants used in the tail probabilities in the following lemmas. We only state the lemmas and their proofs can be found in Appendix B.

6.1 A lower bound on $L_{1,1}(n+1)$ given $\sum_{m=2}^{M} S_{1,m} \leq U_M(n)$.

The following lemma is the rigorous statement of (10).

Lemma 1. Given

$$\mathbb{P}\left(\sum_{m=2}^{M} S_{1,m} > U_M(n)\right) \le \sigma_M(n),$$



Figure 5: Illustrations of ISSP under Erlang-3 service time distribution. The lower bounds of $s_{1,m}$, $\forall m$, keep increasing and the upper bounds of $s_{1,2}$ and $s_{1,2} + s_{1,3}$ keep decreasing until reaching the equilibriums: the initial values (at iteration 1) are $L_{1,m}(1) = 0$, $\forall m, U_2(1) = 1, U_3(1) = 1$; the lower bounds $L_{1,m}$ increase as $0 \rightarrow \frac{1}{4} \rightarrow \frac{11}{32} \rightarrow \cdots \rightarrow \frac{\lambda}{3}$; the upper bound U_2 of $s_{1,2}$ decreases as $1 \rightarrow \frac{3}{8} \rightarrow \frac{21}{64} \rightarrow \cdots \rightarrow \frac{\lambda}{3}$; the upper bound U_3 of $s_{1,2} + s_{1,3}$ decreases as $1 \rightarrow \frac{11}{16} \rightarrow \frac{85}{128} \rightarrow \cdots \rightarrow \frac{2\lambda}{3}$.

 $and \ defining$

$$L_{1,1}(n+1) = \min\left\{s_{1,1}^* - \frac{6\Delta}{C}, 1 - U_M(n) - \frac{1-\xi}{2\mu_1 N^{\alpha}} - \frac{6\Delta}{C}\right\},\$$

we have

$$\mathbb{P}(S_{1,1} < L_{1,1}(n+1)) \le \epsilon_1(n+1)$$

6.2 A lower bound on
$$S_{1,m}$$
 given $S_{1,m-1} \ge L_{1,m-1}$

The following lemma is the rigorous statement of (11).

Lemma 2. Consider $m \ge 2$. Given

$$\mathbb{P}(S_{1,m-1} < L_{1,m-1}(n+1)) \le \epsilon_{m-1}(n+1),$$

and defining

$$L_{1,m}(n+1) = \frac{v_m}{v_{m-1}} L_{1,m-1}(n+1) - \frac{5v_m}{C} \Delta,$$

we have

$$\mathbb{P}\left(S_{1,m} < L_{1,m}(n+1)\right) \le \epsilon_m(n+1).$$

6.3 An upper bound on $\sum_{r=2}^{m} S_{1,r}$ given $\sum_{r=2}^{m-1} S_{1,r} \leq U_{m-1}$ and $S_{1,m-1} \geq L_{1,m-1}$. The following lemma is the rigorous statement of (12).

Lemma 3. Consider $m \ge 2$. Given

$$\mathbb{P}\left(\sum_{r=2}^{m-1} S_{1,r} \ge U_{m-1}(n+1)\right) \le \sigma_{m-1}(n+1)$$
$$\mathbb{P}\left(S_{1,r} < L_{1,r}(n+1)\right) \le \epsilon_r(n+1) \quad 1 \le r \le M$$

and defining

$$U_m(n+1) = 1 - a_m - b_m L_{1,1}(n+1) + a_m U_{m-1}(n+1) + \frac{c_m}{C} \Delta_{n-1}(n+1) + \frac{c_m}{C}$$

we have

$$\mathbb{P}\left(\sum_{r=2}^{m} S_{1,r} \ge U_m(n+1)\right) \le \sigma_m(n+1).$$

6.4 Convergence of $L_{1,1}(n)$.

Based on Lemmas 1, 2, and 3, we will show $\{L_{1,1}(n)\}_n$ is an increasing sequence and approaches $s_{1,1}^*$.

Lemma 4. Recall that $\xi = \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j$. Given $\mathbb{P}(S_{1,1} < L_{1,1}(n)) \le \epsilon_1(n)$ and $L_{1,1}(n) \le s_{1,1}^* - \frac{6\Delta}{C}$, we have

$$\mathbb{P}(S_{1,1} < L_{1,1}(n+1)) \le \epsilon_1(n+1),$$

where

$$L_{1,1}(n+1) = \frac{1}{\mu_1} - \frac{C_M \Delta}{C(1-\xi)} - \frac{1}{2\mu_1 N^{\alpha}} + \xi \left(L_{1,1}(n) - \frac{1}{\mu_1} + \frac{C_M \Delta}{C(1-\xi)} + \frac{1}{2\mu_1 N^{\alpha}} \right)$$

and $\epsilon_1(n+1) = \epsilon_1(n)(M^2+2)(\frac{2C}{v\Delta}+1)^{2M}$. Furthermore, $L_{1,1}(n+1) > L_{1,1}(n)$ holds when $L_{1,1}(n) \le s_{1,1}^* - \frac{6\Delta}{C}$.

6.5 Proving Theorem 1

Based on the monotonicity of $L_{1,1}(n)$, we can apply Lemma 4 a sufficient number of times so that $L_{1,1}(n)$ is close to $s_{1,1}^* - \frac{6\Delta}{C}$, which is formalized in the following lemma.

Lemma 5.

$$\mathbb{P}\left(S_{1,1} < s_{1,1}^* - \frac{6\Delta}{C}\right) \le \left(\frac{1}{\sqrt{N}}\right)^{M+6}.$$

Proof. To prove this lemma, we apply Lemma 4 n times iteratively with $n = \left\lceil \frac{\log N}{2 \log(1/\xi)} \right\rceil$ such that $\xi^n \leq \Delta$. We obtain

$$\mathbb{P}\left(S_{1,1} < \frac{1}{\mu_{1}} - \frac{C_{M}\Delta}{(1-\xi)C} - \frac{1}{2\mu_{1}N^{\alpha}} - \frac{\Delta}{\mu_{1}}\right) \leq \epsilon_{1}(0)(M^{2}+2)^{n}\left(\frac{2C}{\bar{v}\Delta} + 1\right)^{2Mn} \\ \leq \epsilon_{1}(0)(M^{2}+2)^{\frac{\log N}{\log(1/\xi)}+1}N^{\frac{2M\log N}{\log(1/\xi)}+2M} \\ \leq e^{-\frac{\bar{v}^{2}\log^{2}N}{C^{2}}}(M^{2}+2)^{\frac{\log N}{\log(1/\xi)}+1}N^{\frac{2M\log N}{\log(1/\xi)}+2M} \\ = N^{-\frac{\bar{v}^{2}\log N}{C^{2}} + \frac{2M\log N}{\log(1/\xi)}+2M}(M^{2}+2)^{\frac{\log N}{\log(1/\xi)}+1} \\ \leq N^{-\frac{\bar{v}^{2}\log N}{C^{2}} + \frac{3M\log N}{\log(1/\xi)}+2M+1},$$

where the third inequality holds because $\epsilon_1(0) \leq e^{-\frac{\tilde{v}^2 \log^2 N}{C^2}}$, and the last inequality holds because $N \geq M^2 + 2$.

Recalling $C = \sqrt{\frac{2\bar{v}^2 \log(1/\xi)}{3M + (3M+4)\log(1/\xi)}}$ and noting $N \ge 1/\xi$, we have

$$\mathbb{P}\left(S_{1,1} \leq \frac{1}{\mu_1} - \frac{C_M \Delta}{(1-\xi)C} - \frac{\Delta}{\mu_1} - \frac{1}{2\mu_1 N^{\alpha}}\right)$$
$$= \mathbb{P}\left(S_{1,1} \leq \frac{\lambda}{\mu_1} - \frac{C_M \Delta}{(1-\xi)C} - \frac{\Delta}{\mu_1} + \frac{1}{2\mu_1 N^{\alpha}}\right) \leq \left(\frac{1}{\sqrt{N}}\right)^{M+8},$$

which implies

$$\mathbb{P}\left(S_{1,1} < s_{1,m}^* - \frac{6\Delta}{C}\right) \le \left(\frac{1}{\sqrt{N}}\right)^{M+8}$$

because $N^{0.5-\alpha} \ge \frac{2\mu_1 C_M}{C(1-\xi)} \log N$.

The result above established that

$$S_{1,1} \ge s_{1,1}^* - \frac{6\Delta}{C} = \lambda v_1 - \frac{6\Delta}{C},$$

with probability $1 - \left(\frac{1}{\sqrt{N}}\right)^{M+8}$.

Combining with Lemma 2, we next prove that $S_{1,m} \ge s_{1,m}^* - \Omega(\Delta), \forall m \ge 2$ holds with a high probability. Applying Lemma 2 iteratively for $S_{1,m}, \forall m \ge 2$, we have

$$L_{1,m}(n) = \frac{v_m}{v_1} L_{1,1}(n) - \frac{5(m-1)v_m\Delta}{C}$$
$$= \frac{v_m}{v_1} \left(\lambda v_1 - \frac{6\Delta}{C}\right) - \frac{5(m-1)v_m\Delta}{C}$$
$$= \lambda v_m - \frac{6\mu_1 v_m + 5(m-1)v_m\Delta}{C},$$

	-	-	

and $S_{1,m} \geq L_{1,m}(n)$ holds with the probability $\epsilon_m(n)$. Note $\epsilon_m(n) \leq \epsilon_M(n), \forall m$ and $\epsilon_M(n)$ is bounded as follows

$$\epsilon_M(n) \le M\epsilon_1(n) \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-1} = M \left(\frac{1}{\sqrt{N}}\right)^{M+8} \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-1} \le \frac{1}{N^3}$$

Therefore, we have proved the lower bound in the theorem.

Define the event $\mathcal{K} = \{S_{1,m} \geq s_{1,m}^* + \frac{\theta_m \log N}{\sqrt{N}}, \forall m\}$. We have $\mathbb{P}(\mathcal{K}^c) \leq \frac{M}{N^3}$ according to the union bound. We now establish the upper bound in Theorem 1 as follows

$$1 = \mathbb{P}\left(S_{1,m} \le 1 - \sum_{r \ne m} S_{1,r}\right)$$
$$\leq \mathbb{P}\left(S_{1,m} \le 1 - \sum_{r \ne m} S_{1,r} \mid \mathcal{K}\right) + \mathbb{P}(\mathcal{K}^c)$$
$$\leq \mathbb{P}\left(S_{1,m} \le 1 - \sum_{r \ne m} \left(s_{1,r}^* + \frac{\theta_r \log N}{\sqrt{N}}\right)\right) + \frac{M}{N^3}$$

The proof is completed because $1 - \sum_{r \neq m}^{M} s_{1,r}^* = s_{1,m}^* + \frac{1}{N^{\alpha}}$.

7 Proof of Theorem 2

Theorem 1 shows that $S_{1,m}$ is "close" to $s_{1,m}^*$ with a high probability. However, it is not clear whether the average total queue length or the waiting probability is small under an LB-zero policy. To establish Theorem 2, we first prove an important lemma on the upper bound of the average total queue, which is used to establish the waiting probability in Theorem 2. The proof of this lemma can be found in Appendix 7.3.

Lemma 6. Define $w_m = (1 - p_m)\mu_m$, $w_u = \max_m w_m$, $w_l = \min_m w_m$, $\mu_{\max} = \max_m \mu_m$, $\zeta = \frac{4w_u b}{w_l} \left(\left(\frac{1}{w_l} - \frac{1}{w_u}\right) \sum_m \theta_m w_m + \frac{1}{w_l} + 6 \right)$ and $k = \frac{\sum_m \theta_m w_m}{w_u} + \left(1 + \frac{w_l}{4w_u b}\right)\zeta - \sum_m \theta_m$. Under a load balancing policy in LB-zero, the following bound holds

$$\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \le \frac{5\mu_{\max} + 2}{\sqrt{N}\log N}$$

when N satisfies

$$\min\left\{2k\mu_1, \sum_{m=1}^M \theta_m, \frac{C(1-\xi)}{2\mu_1 C_M}\right\} N^{0.5-\alpha} \ge \log N \ge \max\left\{\log\left(\frac{1}{\xi}\right), \frac{2\mu_1}{1-\xi}, \frac{4b}{w_l\zeta}, \frac{C}{\mu_1}\right\}.$$

Lemma 6 establishes an upper bound on the average total queue length. Recall Theorem 1 indicates the service rate under an LB-zero policy in Π is close to arrival rate λN at steady state because $\sum_{m=1}^{M} (1 - p_m) S_{1,m} \approx \sum_{m=1}^{M} (1 - p_m) s_{1,m}^* = \lambda$. Therefore, it is reasonable to couple the distributed load balancing system with a simple centralized server system with a similar arrival rate and service rate. This coupling will be done via Stein's method [11, 43, 10]. Stein's method allows us to understand the key performance metrics (e.g., average queue length) of a complicated load balancing system from the performance of a simple fluid system (to be introduced in the next section). Formally, we study the generator difference between the distributed load balancing system and a simple centralized system within a small state space identified by ISSP. This idea of coupling a simple (and almost trivial) fluid model, without ISSP, has been also used in [24, 25, 22]. We will introduce it next so the paper is self-contained.

7.1 Generator Coupling with a Single Server System

Denote $\Delta = \frac{\log N}{\sqrt{N}}$. We consider a single server queue with arrival rate λ and service rate $\lambda + \Delta$. The fluid model with respect to the queue length x is

$$\dot{x} = \frac{dx}{dt} = -\Delta. \tag{18}$$

Let function g(x) be the solution of the following Stein's equation or Poisson equation [43] for the fluid system above and distance function h(x) such that:

$$\frac{dg(x)}{dt} = g'(x)(-\Delta) = h(x), \forall x,$$
(19)

where $g'(x) = \frac{dg(x)}{dx}$. Because (19) has a very simple form, both g' and g'' can be easily solved which is different from other applications of Stein's method, e.g., [11], where establishing the gradient bounds is a key difficulty.

To analyze the total queue length at steady-state under an LB-zero policy in Π , we choose a truncated distance function:

$$h\left(\sum_{i=1}^{b} S_i\right) = \max\left\{\sum_{i=1}^{b} S_i - \eta, 0\right\}, \quad \eta = \lambda + k\Delta.$$

The distance $h\left(\sum_{i=1}^{b} S_i\right)$ can be viewed as a proxy to measure the total queue length $(N \sum_{i=1}^{b} S_i)$ at steady state.

To couple the one-dimensional fluid system in (18) with the $b \times M$ -dimensional stochastic system, we define

$$f(s) = g\left(\sum_{i=1}^{b} s_i\right) = g\left(\sum_{i=1}^{b} \sum_{m=1}^{M} s_{i,m}\right).$$
(20)

Note f(s) is bounded for $s \in \mathcal{S}^{(N)}$, we impose the generator of stochastic system G on function f and have the basic adjoint relationship for the stationary distribution S such that

$$\mathbb{E}[Gf(S)] = \mathbb{E}\left[Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right] = 0.$$
(21)

Combining (19) and (21), we connect the performance metric $h(\cdot)$ with the generator difference between the simple single-server system and G as follows

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right] = \mathbb{E}\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\left(-\Delta\right) - Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right].$$
(22)

In the following lemma, we provide an upper bound on (22) which includes two terms: the first term is from the gradient bounds and the second term is from ISSP. The proof of this lemma can be found in Appendix C.

Lemma 7. Define the regions $\mathcal{T} = \{x \mid x > \eta + \frac{1}{N}\}$ and denote the normalized service rate $D_1 = \sum_{m=1}^{M} (1-p_m) \mu_m S_{1,m}$, we have

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b} S_{i}\right)\right] \leq J_{1} + \frac{5\mu_{\max} + \lambda}{\sqrt{N}\log N},\tag{23}$$

with

$$J_1 = \mathbb{E}\left[g'\left(\sum_{i=1}^b S_i\right) \left(\lambda A_b(S) - \lambda - \Delta + D_1\right) \mathbb{I}_{\sum_{i=1}^b S_i \in \mathcal{T}}\right].$$
(24)

To establish Lemma 6 and Theorem 2 based on the lemmas above, we need to provide the upper bounds on (24), which is related to the difference between the normalized arrival rate and the normalized service rate, which will be bounded based on the ISSP result that shows the the normalized service rate at the steady-state is close to the zero-waiting equilibrium value.

7.2 State Space Peeling on $\sum_{m=1}^{M} (1-p_m) S_{1,m}$

In this subsection, we analyze J_1 in (24):

$$J_{1} = \mathbb{E}\left[\frac{1}{\Delta}h\left(\sum_{i=1}^{b}S_{i}\right)\left(-\lambda A_{b}(S)+\lambda+\Delta-D_{1}\right)\mathbb{I}_{\sum_{i=1}^{b}S_{i}>\eta+\frac{1}{N}}\right]$$
$$\leq \mathbb{E}\left[\frac{1}{\Delta}h\left(\sum_{i=1}^{b}S_{i}\right)\left(\lambda+\Delta-D_{1}\right)\mathbb{I}_{\sum_{i=1}^{b}S_{i}>\eta+\frac{1}{N}}\right],\tag{25}$$

where the first equality is due to the definition of g' in Stein's equation (19), and the inequality holds because $\frac{1}{\Delta}h\left(\sum_{i=1}^{b}S_{i}\right)\mathbb{I}_{\sum_{i=1}^{b}S_{i}>\eta+\frac{1}{N}} \geq 0$. We focus on

$$\left(\lambda + \Delta - \sum_{m=1}^{M} (1 - p_m) \mu_m s_{1,m}\right) \mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}},\tag{26}$$

where we recall $\eta = \lambda + k\Delta$ and $d_1 = \sum_{m=1}^{M} (1 - p_m)\mu_m s_{1,m}$ is the total service rate when the system is in the state s. Though we have established $S_{1,m} \approx s_{1,m}^*$ in Theorem 1, it is not sufficient for showing (25) is small enough because $\lambda + \Delta - D_1$ may be larger than Δ , which will make (25) very large. In fact, we need one more state space peeling to show (25) is non-positive.

We define two regions S_{ssp_1} and S_{ssp_2}

$$\mathcal{S}_{ssp_1} = \left\{ s \mid s_1 \ge \lambda + (k - \zeta - 6) \Delta, s_{1,m} \ge s_{1,m}^* - \theta_m \Delta \right\},$$
$$\mathcal{S}_{ssp_2} = \left\{ s \mid \sum_{i=1}^b s_i \le \lambda + k\Delta \right\},$$

where S_{ssp_1} is the region with sufficient many busy servers and S_{ssp_2} is the region with bounded total queue length. We further define a region

$$\mathcal{S}_{ssp} = \mathcal{S}_{ssp_1} \bigcup S_{ssp_2},$$

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

• Case 1: In Lemma 13 in the appendix, we show any $s \in S_{ssp_1}$ satisfies

$$\sum_{n=1}^{M} (1 - p_m) \mu_m s_{1,m} \ge \lambda + \Delta.$$

It implies $\left(\lambda + \Delta - \sum_{m=1}^{M} (1 - p_m) \mu_m s_{1,m}\right) \mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}} \leq 0$ for any $s \in \mathcal{S}_{ssp_1}$. For any $s \in \mathcal{S}_{ssp_2}$, we have $\mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}} = 0.$

It implies $\left(\lambda + \Delta - \sum_{m=1}^{M} (1 - p_m) \mu_m s_{1,m}\right) \mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}} = 0$ for any $s \in \mathcal{S}_{ssp_2}$.

• Case 2: In Lemma 14 in the appendix, we show that

$$\mathbb{P}\left(S \notin \mathcal{S}_{ssp}\right) \le \frac{2}{N^2}$$

by using an ISSP approach on S_1 and $\sum_{i=2}^{b} S_i$.

7.3 Proving Lemma 6

Based on the two cases above, we split (25) into two regions $s \in S_{ssp}$ and $s \notin S_{ssp}$ and obtain

$$(25) = \mathbb{E}\left[\frac{1}{\Delta}\left(\sum_{i=1}^{b} S_{i} - \eta\right)\left(\lambda + \Delta - D_{1}\right)\mathbb{I}_{S\in\mathcal{S}_{ssp}}\mathbb{I}_{\sum_{i=1}^{b} S_{i} > \eta + \frac{1}{N}}\right] \\ + \mathbb{E}\left[\frac{1}{\Delta}\left(\sum_{i=1}^{b} S_{i} - \eta\right)\left(\lambda + \Delta - D_{1}\right)\mathbb{I}_{S\notin\mathcal{S}_{ssp}}\mathbb{I}_{\sum_{i=1}^{b} S_{i} > \eta + \frac{1}{N}}\right] \\ \leq \mathbb{E}\left[\frac{1}{\Delta}\left(\sum_{i=1}^{b} S_{i} - \eta\right)\left(\lambda + \Delta - D_{1}\right)\mathbb{I}_{S\notin\mathcal{S}_{ssp}}\mathbb{I}_{\sum_{i=1}^{b} S_{i} > \eta + \frac{1}{N}}\right] \\ \leq \frac{2b}{N^{1.5}\log N}, \tag{27}$$

where the first inequality holds because of Lemma 13 and the second inequality holds because the average total number of jobs per server is at most b and $(\lambda + \Delta - D_1) \mathbb{I}_{S \notin S_{ssp}} \mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}} < 1.$

By combining (23) and (27), we can now establish Lemma 6 in the following

$$\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \eta, 0\right\}\right] \le \frac{2b}{N^{1.5}\log N} + \frac{5\mu_{\max} + \lambda}{\sqrt{N}\log N} \le \frac{5\mu_{\max} + 2}{\sqrt{N}\log N}.$$

7.4 Proving Theorem 2

Once we have Lemma 6, we can prove Theorem 2 with the property of LB-zero and the Markov inequality. For an LB-zero policy in Π , the waiting probability satisfies

$$\mathbb{P}(\mathcal{W}) = \mathbb{P}\left(\mathcal{W}|\sum_{i=1}^{b} S_i \le 1 - \frac{1}{N^{\alpha} \log N}\right) \mathbb{P}\left(\sum_{i=1}^{b} S_i \le 1 - \frac{1}{N^{\alpha} \log N}\right) \\ + \mathbb{P}\left(\mathcal{W}|\sum_{i=1}^{b} S_i > 1 - \frac{1}{N^{\alpha} \log N}\right) \mathbb{P}\left(\sum_{i=1}^{b} S_i > 1 - \frac{1}{N^{\alpha} \log N}\right) \\ \le \mathbb{P}\left(\mathcal{W}|\sum_{i=1}^{b} S_i < 1 - \frac{1}{N^{\alpha} \log N}\right) + \mathbb{P}\left(\sum_{i=1}^{b} S_i > 1 - \frac{1}{N^{\alpha} \log N}\right)$$

where the first term is bounded by $\frac{1}{\sqrt{N}}$ because of the definition of LB-zero. For the second term, we have

$$\begin{split} \mathbb{P}\left(\sum_{i=1}^{b} S_i > 1 - \frac{1}{N^{\alpha} \log N}\right) &\leq \mathbb{P}\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k \log N}{\sqrt{N}}, 0\right\} > \frac{1}{N^{\alpha}} \left(1 - \frac{1}{\log N}\right) - \frac{k \log N}{\sqrt{N}}\right) \\ &\leq \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k \log N}{\sqrt{N}}, 0\right\}\right]}{\frac{1}{N^{\alpha}} \left(1 - \frac{1}{\log N}\right) - \frac{k \log N}{\sqrt{N}}} \\ &\leq \frac{\mathbb{E}\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k \log N}{\sqrt{N}}, 0\right\}\right]}{\frac{1}{2N^{\alpha}}} \\ &\leq \frac{10\mu_{\max} + 4}{N^{0.5 - \alpha} \log N} \end{split}$$

where the second inequality holds because of the Markov inequality; the third inequality holds because $\log N \ge 2$ and $\frac{N^{0.5-\alpha}}{\log N} \ge 2k$; and the last inequality holds because of Lemma 6.

Finally, we remark that we choose $k = \Omega(b)$ to prove Lemma 15, which is the technical reason we assumed b is finite. We, however, believe our results hold even for $b = \infty$.

8 Simulations

In this section, we confirm our theoretical results of ISSP in Theorem 1 and the large system insensitivity in Theorem 2 with simulations. We considered two policies, JSQ and JIQ, and $\lambda = 1 - N^{\alpha}$ We conjecture that our results hold even for $\alpha = 0.5$ because it holds for any $\alpha < 0.5$. To confirm this, we used $\alpha = 0.5$ in our simulations.

8.1 Large System Insensitivity under JSQ and JIQ

We first studied the average total queue length (per server) $\mathbb{E}[\sum_{i=1}^{b} S_i]$ and the waiting probability $\mathbb{P}(\mathcal{W})$ under a Coxian-4 service time distribution with the parameters p = [0.5, 0.5, 0.5, 1.0] and $\mu = [1.875, 1.875, 1.875, 1.875]$. We plotted $\mathbb{E}[\sum_{i=1}^{b} S_i]$ and $\mathbb{P}(\mathcal{W})$ versus the number of servers N. The results are obtained with 10 trials and each trial has 10^7 steps. From Figure 7, the waiting probability tends to zero when N increases as we expected and JIQ almost has the identical performance as JSQ.



(a) Average total queue length under Coxian 4

(b) Waiting probability under Coxian 4

Figure 7: Asymptotic zero waiting under JSQ and JIQ



(a) Average total queue length under Coxian M

(b) Waiting probability under Coxian M

Figure 8: Large System Insensitivity under JSQ and JIQ

We then investigated $\mathbb{E}[\sum_{i=1}^{b} S_i]$ and $\mathbb{P}(\mathcal{W})$ versus Coxian-*M* with various number of phases *M* and fixed $N = 10^4$. In particular, we consider Coxian-*M* with $p = [p_1, p_2, \cdots, 1]$ and $\mu =$

 $[\mu_1, \mu_2, \cdots, \mu_M]$, where $p_m = 0.5, \forall 1 \leq m \leq M - 1$ and $\bar{\mu} = \mu_m, \forall m$ (identical service times). We plotted $\mathbb{E}[\sum_{i=1}^b S_i]$ and $\mathbb{P}(\mathcal{W})$ versus M. From Figure 8, we can observe that the average queue length and the waiting probability remain roughly the same under different Ms, which confirms the insensitivity.

8.2 ISSP under JSQ and JIQ

In this section, we investigated ISSP by studying the trajectory of $S_{1,m}(t)$, $\forall m$ under JSQ and JIQ. We considered N = 10,000 and a Coxian-4 service time distribution with p = [0.5, 0.5, 0.5, 1.0] and $\mu = [1.875, 1.875, 1.875, 1.875]$. We plotted $S_{1,m}(t)$ of JSQ and JIQ in Figure 9. The results are obtained with 10 trials and each trial has 10^6 steps. We observed that $S_{1,m(t)}$ under both policies concentrates around dash lines $s_{1,m}^*, \forall m$, which confirms the high probability bounds in 1. Note that the system was initialized with the zero-waiting equilibrium, instead of the empty state, in our simulations.



Figure 9: The evolution of the system states under JSQ and JIQ

9 Conclusions and Discussions

In this paper, we studied a distributed queueing system under Coxian service time distributions in the sub-Halfin-Whitt regime. We established that a set of load balancing policies, named LB-zero, achieves asymptotic zero-waiting, i.e. insensitive in the large-system regime. To tackle the nonmonotonicity under general service time distributions, we developed a technique, called iterative state-space peeling (ISSP), which iteratively removes the low-probability states, and results in a small state-space that can be analyzed using a simple mean-field model. This ISSP approach may be used for other problems as well. One possible application is to study load-balancing in many server systems with heterogeneous servers or jobs belonging to multiple priority classes. For heterogeneous servers, we can use ISSP to identify the "typical" load of each type of the servers; and for jobs with different priorities, we can use ISSP to identify the "typical" distribution of job types in the system. In the reduced state space based on the typical load of the typical distribution at the steady-state, the steady-state performance of the many server system may become tractable like in this paper.

References

 R. Aghajani, X. Li, and K. Ramanan. The PDE method for the analysis of randomized load balancing networks. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 1(2):38:1–38:28, 2017.

- [2] Akamai. The state of online retail performance report, 2017.
- [3] S. Banerjee and D. Mukherjee. Join-the-Shortest Queue diffusion limit in Halfin–Whitt regime: Tail asymptotics and scaling of extrema. Ann. Appl. Probab., 29(2):1262 – 1309, 2019.
- [4] S. Banerjee and D. Mukherjee. Join-the-Shortest Queue diffusion limit in Halfin–Whitt regime: Sensitivity on the heavy-traffic parameter. Ann. Appl. Probab., 30(1):80 – 144, 2020.
- [5] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. Ann. Appl. Probab., 11(4):1384 1428, 2001.
- [6] T. Bonald, M. Jonckheere, and A. Proutiere. Insensitive load balancing. ACM SIGMETRICS Performance Evaluation Review, page 367–377, June 2004.
- [7] M. Bramson, Y. Lu, and B. Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Syst.*, 71(3):247–292, 2012.
- [8] A. Braverman. Steady-state analysis of the Join-the-Shortest-Queue model in the Halfin-Whitt regime. *Mathematics of Operations Research*, 45(3):1069–1103, 2020.
- [9] A. Braverman. The Join-the-Shortest-Queue system in the Halfin-Whitt regime: rates of convergence to the diffusion limit. Arxiv preprint arXiv:2202.02889, 2022.
- [10] A. Braverman. The prelimit generator comparison approach of Stein's method. *Stoch. Syst.*, 2022.
- [11] A. Braverman, J. G. Dai, and J. Feng. Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.*, 6:301–366, 2016.
- [12] D. Y. Burman. Insensitivity in queueing systems. Adv. in Appl. Probab., 13(4):846–859, 1981.
- [13] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst.*, 72(3-4):311–359, Dec. 2012.
- [14] P. Eschenfeldt and D. Gamarnik. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Mathematics of Operations Research*, 43(3):867–886, 2018.
- [15] S. Foss and A. L. Stolyar. Large-scale join-idle-queue system with general service times. J. Appl. Probab., 54(4):995–1007, 2017.
- [16] N. Gast. Expected values estimated via mean-field approximation are 1/n-accurate. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 1(1):17:1–17:26, 2017.
- [17] V. Gupta and N. Walton. Load balancing in the nondegenerate slowdown regime. Operations Research, 67(1):281–294, 2019.
- [18] T. Hellemans and B. Van Houdt. On the power-of-d-choices with least loaded server selection. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 2(2):27:1–27:22, 2018.
- [19] Y. Hong and W. Wang. Sharp waiting-time bounds for multiserver jobs. arXiv preprint arXiv:2109.05343, 2021.
- [20] M. Jonckheere and B. Prabhu. Asymptotics of insensitive load balancing and blocking phases. Queueing Syst., 2018.
- [21] G. Kielanski and B. Van Houdt. On the asymptotic insensitivity of the supermarket model in processor sharing systems. Proc. Ann. ACM SIGMETRICS Conf., 2021.

- [22] X. Liu, K. Gong, and L. Ying. Steady-state analysis of load balancing with Coxian-2 distributed service times. Naval Research Logistics (NRL), 2021.
- [23] X. Liu and L. Ying. On achieving zero delay with power-of-d-choices load balancing. In Proc. IEEE Int. Conf. Computer Communications (INFOCOM), Honolulu, Hawaii, 2018.
- [24] X. Liu and L. Ying. Steady-state analysis of load balancing algorithms in the sub-Halfin-Whitt regime. J. Appl. Probab., 57(2):578 – 596, June 2020.
- [25] X. Liu and L. Ying. Universal scaling of distributed queues under load balancing in the super-Halfin-Whitt regime. *IEEE/ACM Trans. Netw.*, 2021.
- [26] 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.*, 68(11):1056–1071, 2011.
- [27] S. T. Maguluri and R. Srikant. Heavy traffic queue length behavior in a switch under the maxweight algorithm. Stoch. Syst., 6(1):211–250, 2016.
- [28] M. Mitzenmacher. The Power of Two Choices in Randomized Load Balancing. PhD thesis, University of California at Berkeley, 1996.
- [29] D. Mukherjee, S. C. Borst, and J. S. van Leeuwaarden. Asymptotically optimal load balancing topologies. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 2(1):14:1–14:29, Apr. 2018.
- [30] D. Mukherjee, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting. Universality of power-of-d load balancing in many-server systems. *Stoch. Syst.*, 8(4):265–292, 2018.
- [31] D. Rutten and D. Mukherjee. Load balancing under strict compatibility constraints. *Proc.* Ann. ACM SIGMETRICS Conf., 2021.
- [32] A. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. Queueing Syst., 80(4):341–361, 2015.
- [33] B. Van Houdt. Global attraction of ODE-based mean field models with hyperexponential job sizes. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 3(2), 2019.
- [34] T. Vasantam, A. Mukhopadhyay, and R. R. Mazumdar. Insensitivity of the mean field limit of loss systems under SQ(d) routeing. Adv. in Appl. Probab., 51(4):1027–1066, 2019.
- [35] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20– 34, 1996.
- [36] W. Wang, S. T. Maguluri, R. Srikant, and L. Ying. Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies. *Mathematics of Operations Research*, 2021.
- [37] W. Wang, Q. Xie, and M. Harchol-Balter. Zero queueing for multi-server jobs. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 2021.
- [38] W. Weng and R. Srikant. An algorithm for improved delay-scaling in input-queued switches. *Queueing Syst.*, 2022.
- [39] W. Weng and W. Wang. Achieving zero asymptotic queueing delay for parallel jobs. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 2020.
- [40] W. Weng, X. Zhou, and R. Srikant. Optimal load balancing with locality constraints. Proc. ACM Meas. Anal. Comput. Syst. (POMACS), 2020.

- [41] W. Winston. Optimality of the shortest line discipline. J. Appl. Probab., 14(1):181–189, 1977.
- [42] Q. Xie, X. Dong, Y. Lu, and R. Srikant. Power of d choices for large-scale bin packing: A loss model. In Proc. Ann. ACM SIGMETRICS Conf., June 2015.
- [43] L. Ying. On the approximation error of mean-field models. In Proc. Ann. ACM SIGMETRICS Conf., Antibes Juan-les-Pins, France, June 2016.
- [44] L. Ying, R. Srikant, and X. Kang. The power of slightly more than one sample in randomized load balancing. In Proc. IEEE Int. Conf. Computer Communications (INFOCOM), Hong Kong, 2015.
- [45] Z. Zhao, S. Banerjee, and D. Mukherjee. Many-server asymptotics for Join-the-Shortest Queue in the super-Halfin-Whitt scaling window. *Arxiv preprint arXiv:2106.00121*, 2021.

A The Properties of the Constants

The following two lemmas show that the constants a_m , b_m , c_m , $(\forall m \ge 2)$, and C_M are positive, and $0 < \xi < 1$.

Lemma 8. The constants $a_m, b_m, c_m, \forall m \ge 2$, and C_M are positive.

Proof. $1 < a_m < 1, \forall 2 \le m \le M$ holds by the definition. Therefore, it is easy to verify $c_m, \forall m \ge 2$, and C_M are positive.

Next, we prove $b_m, \forall m \ge 2$, is positive as follows:

$$b_m = (1 - a_m) \left(1 + \sum_{r=m+1}^M \frac{v_r}{v_1} \right) - \frac{a_m v_m}{v_1}$$

= $1 - a_m - \frac{a_m v_m}{v_1} + (1 - a_m) \sum_{r=m+1}^M \frac{v_r}{v_1}$
= $\frac{p_1 \mu_1}{p_1 \mu_1 + \mu_m} \left(1 - \prod_{i=2}^{m-1} p_i \right) + (1 - a_m) \sum_{r=m+1}^M \frac{v_r}{v_1}$
> 0

where the first equality holds by the definition of b_m ; the third equality by substituting the definition of a_m and v_m ; the last inequality holds because $\prod_{i=2}^{m-1} p_i \leq 1, \forall m \geq 2$ and $0 < a_m < 1, \forall m \geq 2$. \Box

Lemma 9. $1 - \xi = \mu_1 \prod_{m=2}^{M} a_m$.

Proof. Recall the definition of $\xi = \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j$. We have

$$\begin{split} 1 - \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j &= 1 - \sum_{m=2}^{M} \left((1 - a_m) \left(1 + \sum_{r=m+1}^{M} \frac{v_r}{v_1} \right) - a_m \frac{v_m}{v_1} \right) \prod_{j=m+1}^{M} a_j \\ &= 1 - \sum_{m=2}^{M} (1 - a_m) \prod_{j=m+1}^{M} a_j - \sum_{m=2}^{M} \left((1 - a_m) \sum_{r=m+1}^{M} \frac{v_r}{v_1} - a_m \frac{v_m}{v_1} \right) \prod_{j=m+1}^{M} a_j \\ &= \prod_{j=2}^{M} a_j - \sum_{m=2}^{M} \left(\sum_{r=m+1}^{M} \frac{v_r}{v_1} - a_m \sum_{r=m}^{M} \frac{v_r}{v_1} \right) \prod_{j=m+1}^{M} a_j \\ &= \prod_{j=2}^{M} a_j - \sum_{m=2}^{M} \sum_{r=m+1}^{M} \frac{v_r}{v_1} \prod_{j=m+1}^{M} a_j + \sum_{m=2}^{M} \sum_{r=m}^{M} \frac{v_r}{v_1} \prod_{j=m}^{M} a_j \\ &= \prod_{j=2}^{M} a_j + \sum_{m=2}^{M} \frac{v_m}{v_1} \prod_{j=2}^{M} a_j = \prod_{j=2}^{M} a_j + \frac{1 - v_1}{v_1} \prod_{j=2}^{M} a_j = \mu_1 \prod_{m=2}^{M} a_m. \end{split}$$

B Proof of the Lemmas for Theorem 1

We first prove Lemmas 1, 2, 3, and 4 used in ISSP.

B.1 A tail bound from [36]

We introduce Lemma 10 from [36], which is an extension of the tail bound in [5] and is the key to establish ISSP. Lemma 10 allows us to apply the Lyapunov drift analysis to iteratively reduce the state space.

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

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

and define

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

Assume there exists a set \mathcal{E} with B > 0, $\gamma > 0$, $\delta \ge 0$ such that the following conditions hold

- (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

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

with

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

According to Lemma 10, the critical step in establishing the tail bound is to construct proper Lyapunov functions. In the following sections, we construct a sequence of Lyapunov functions and apply Lemma 10 to prove Lemmas 1, 2, 3, and 4. In the following proofs, we ignore the iteration number n for a clean notation.

B.2 Proof of Lemma 1: A lower bound on $L_{1,1}(n+1)$ given $\sum_{m=2}^{M} S_{1,m} \leq U_M(n)$ Lemma 1. *Given*

$$\mathbb{P}\left(\sum_{m=2}^{M} S_{1,m} > U_M(n)\right) \le \sigma_M(n)$$

and defining

$$L_{1,1}(n+1) = \min\left\{s_{1,1}^* - \frac{6\Delta}{C}, 1 - U_M(n) - \frac{1-\xi}{2\mu_1 N^{\alpha}} - \frac{6\Delta}{C}\right\},\$$

we have

$$\mathbb{P}(S_{1,1} < L_{1,1}(n+1)) \le \epsilon_1(n+1).$$

To prove Lemma 1 using Lemma 10, we consider the following Lyapunov function

$$V(s) = \tilde{L}_{1,1} - s_{1,1},\tag{28}$$

where $\tilde{L}_{1,1} = \min\left\{1 - \frac{1-\xi}{2\mu_1 N^{\alpha}} - U_M, s_{1,1}^*\right\}$ and define

$$\mathcal{E} = \left\{ s \mid \sum_{m=2}^{M} s_{1,m} \le U_M \right\} \text{ and } B = \frac{2\Delta}{C}$$

When $V(s) = \tilde{L}_{1,1} - s_{1,1} \ge B$ and $s \in \mathcal{E}$, we have

$$s_1 = \sum_{m=1}^M s_{1,m} \le U_M + \tilde{L}_{1,1} - \frac{2\Delta}{C} = 1 - \frac{1-\xi}{2\mu_1 N^{\alpha}} - \frac{2\Delta}{C} \le 1 - \frac{1}{N^{\alpha} \log N},$$

where the last inequality holds due to $\log N \geq \frac{2\mu_1}{1-\xi}$. Therefore, the drift of V(s) satisfies

$$\nabla V(s) = -\lambda(1 - A_1(s)) + \mu_1 s_{1,1} - \sum_m (1 - p_m) \mu_m s_{2,m}$$

$$\stackrel{(a)}{\leq} \frac{1}{\sqrt{N}} - \lambda + \mu_1 s_{1,1}$$

$$\stackrel{(b)}{\leq} \frac{1}{\sqrt{N}} - \lambda + \mu_1 \left(\tilde{L}_{1,1} - \frac{2\Delta}{C}\right)$$

$$= \frac{1}{\sqrt{N}} - (\lambda - \mu_1 \tilde{L}_{1,1}) - \frac{2\mu_1 \Delta}{C}$$

$$\stackrel{(c)}{\leq} \frac{1}{\sqrt{N}} - \frac{2\mu_1 \Delta}{C}$$

$$\stackrel{(d)}{\leq} - \frac{\mu_1 \Delta}{C}$$

where

• (a) holds because $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq 1 - \frac{1}{N^{\alpha} \log N}$ for a LB-zero policy in Π and $s_{2,m} \geq 0$;

- (b) holds because $V(s) = \tilde{L}_{1,1} s_{1,1} \ge \frac{2\Delta}{C}$;
- (c) holds because $\tilde{L}_{1,1} \leq s_{1,1}^* = \frac{\lambda}{\mu_1};$
- and (d) holds because $\log N \ge \frac{C}{\mu_1}$.

Moreover, we have

$$\nabla V(s) = -\lambda(1 - A_1(s)) + \mu_1 s_{1,1} - \sum_m (1 - p_m) \mu_m s_{2,m} \le \mu_1 s_{1,1} - \sum_m (1 - p_m) \mu_m s_{2,m} = \mu_1 s_{1,1} - \sum_m (1 - p_m) \mu_1 s_{1,1} - \sum_$$

Define $\gamma = \frac{\mu_1 \Delta}{C}$ and $\delta = \mu_1$. We now apply Lemma 10 with $j = \frac{2\sqrt{N}\log N}{C}$. Since $q_{\max} = \mu_1 N$ and $\nu_{\max} = \frac{1}{N}$, we have

$$\alpha = \frac{1}{1 + \frac{\Delta}{C}}$$
 and $\beta = \frac{C}{\Delta} + 1$,

and

$$\mathbb{P}\left(S_{1,1} < L_{1,1}(n+1)\right) \leq \mathbb{P}\left(S_{1,1} \leq L_{1,1}(n+1)\right)$$

$$\stackrel{(a)}{=} \mathbb{P}\left(V(S) \geq B + 2\nu_{\max}j\right)$$

$$\stackrel{(b)}{\leq} \left(\frac{1}{1+\frac{\Delta}{C}}\right)^{\frac{2\sqrt{N}\log N}{C}} + \beta \mathbb{P}\left(S \notin \mathcal{E}\right)$$

$$\stackrel{(c)}{\leq} \left(1 - \frac{\Delta}{2C}\right)^{\frac{2\sqrt{N}\log N}{C}} + \beta\sigma_M$$

$$\leq e^{-\frac{\log^2 N}{C^2}} + \beta\sigma_M$$

where

- (a) holds by substituting $B = \frac{2\Delta}{C}$, $\nu_{\text{max}} = \frac{1}{N}$ and $j = \frac{2\sqrt{N}\log N}{C}$;
- (b) holds based on Lemma 10;
- and (c) holds because $\frac{1}{C} \leq \frac{1}{\Delta}$ and the assumption of the lemma on $\mathbb{P}(S \notin \mathcal{E})$.

B.3 Proof of Lemma 2: A lower bound on $S_{1,m}$ given $S_{1,m-1} \ge L_{1,m-1}$ for $m \ge 2$ Lemma 2. Consider $m \ge 2$. Given

$$\mathbb{P}\left(S_{1,m-1} < L_{1,m-1}(n+1)\right) \le \epsilon_{m-1}(n+1),$$

and defining

$$L_{1,m}(n+1) = \frac{v_m}{v_{m-1}} L_{1,m-1}(n+1) - \frac{5v_m}{C} \Delta,$$

we have

$$\mathbb{P}\left(S_{1,m} < L_{1,m}(n+1)\right) \le \epsilon_m(n+1)$$

To prove Lemma 2, consider Lyapunov function

$$V(s) = \frac{v_m}{v_{m-1}} L_{1,m-1} - s_{1,m}.$$
(29)

and define

$$\mathcal{E} = \{ s \mid s_{1,m-1} \ge L_{1,m-1} \}.$$

Given $V(s) \geq \frac{v_m}{C} \Delta$, we have

$$s_{1,m} \le \frac{v_m}{v_{m-1}} L_{1,m-1} - \frac{v_m}{C} \Delta.$$

Therefore, the drift of V(s) when $V(s) \geq \frac{v_m}{C} \Delta$ and $s \in \mathcal{E}$ is

$$\nabla V(s) = \mu_m s_{1,m} - p_{m-1}\mu_{m-1}s_{1,m-1}$$

$$\stackrel{(a)}{=} \mu_m \left(s_{1,m} - \frac{v_m}{v_{m-1}}s_{1,m-1}\right)$$

$$\stackrel{(b)}{\leq} \mu_m \left(s_{1,m} - \frac{v_m}{v_{m-1}}L_{m-1}\right)$$

$$\stackrel{(c)}{\leq} - \frac{\mu_m v_m}{C}\Delta$$

where

- (a) holds according to the definition of $v_m = \frac{\prod_{i=1}^{m-1} p_i}{\mu_m}$;
- (b) holds because $s_{1,m-1} \ge L_{1,m-1}$;
- and (c) holds because $s_{1,m} \leq \frac{v_m}{v_{m-1}}L_{1,m-1} \frac{v_m}{C}\Delta$.

Moreover, we have

$$\nabla V(s) = \mu_m s_{1,m} - p_{m-1} \mu_{m-1} s_{1,m-1} \le \mu_m.$$

Define $B = \frac{v_m}{C}\Delta$, $\gamma = \frac{\mu_m v_m}{C}\Delta$, and $\delta = \mu_m$. Combining $q_{\text{max}} = \mu_m N$ and $\nu_{\text{max}} = \frac{1}{N}$, we have

$$\alpha = \frac{1}{1 + \frac{v_m}{C}\Delta}$$
 and $\beta = \frac{C}{v_m\Delta} + 1.$

Applying Lemma 10 with $j = \frac{2v_m \sqrt{N} \log N}{C}$, we have

$$\begin{split} \mathbb{P}\left(S_{1,m} < L_{1,m}(n+1)\right) &\stackrel{(a)}{\leq} \mathbb{P}\left(V(S) \ge B + 2\nu_{\max}j\right) \\ &\stackrel{(b)}{\leq} \left(\frac{\mu_m}{\mu_m + \frac{\mu_m v_m}{C}\Delta}\right)^{\frac{2v_m \sqrt{N}\log N}{C}} + \beta \mathbb{P}\left(S \notin \mathcal{E}\right) \\ &\stackrel{(c)}{\leq} \left(1 - \frac{v_m}{2C}\Delta\right)^{\frac{2v_m \sqrt{N}\log N}{C}} + \beta \epsilon_{m-1} \\ &\stackrel{\leq}{\leq} e^{-\frac{v_m^2 \log^2 N}{C^2}} + \beta \epsilon_{m-1}, \end{split}$$

where

- (a) holds by substituting $B = \frac{v_m}{C} \Delta$, $\nu_{\max} = \frac{1}{N}$ and $j = \frac{2v_m \sqrt{N} \log N}{C}$;
- (b) holds based on Lemma 10;
- and (c) holds because $\frac{v_m}{C} \leq \frac{1}{\Delta}$.

B.4 Proof of Lemma 3: An upper bound on $\sum_{r=2}^{m} S_{1,r}$ given $\sum_{r=2}^{m-1} S_{1,r} \leq U_{m-1}, \forall m \geq 2$ and $S_{1,m} \geq L_m, \forall m \geq 1$

Lemma 3. Consider $m \ge 2$. Given

$$\mathbb{P}\left(\sum_{r=2}^{m-1} S_{1,r} \ge U_{m-1}(n+1)\right) \le \sigma_{m-1}(n+1)$$
$$\mathbb{P}\left(S_{1,r} < L_{1,r}(n+1)\right) \le \epsilon_r(n+1) \quad 1 \le r \le M$$

and defining

$$U_m(n+1) = 1 - a_m - b_m L_{1,1}(n+1) + a_m U_{m-1}(n+1) + \frac{c_m}{C}\Delta,$$

 $we\ have$

$$\mathbb{P}\left(\sum_{r=2}^{m} S_{1,r} \ge U_m(n+1)\right) \le \sigma_m(n+1).$$

Consider Lyapunov function

$$V(s) = \sum_{r=2}^{m} s_{1,r} - B_m,$$
(30)

where

$$B_m = \frac{p_1 \mu_1 (1 - \sum_{r=m+1}^M L_{1,r}) - \sum_{r=2}^{m-1} (1 - p_r) \mu_r L_{1,r} + \mu_m U_{m-1}}{p_1 \mu_1 + \mu_m},$$

and define

$$\mathcal{E} = \left\{ s \; \left| \; \sum_{r=2}^{m-1} s_{1,r} \le U_{m-1}, \; \text{and} \; \; s_{1,r} \ge L_{1,r}, \forall r \ge 1 \right\}. \right.$$

Given $V(s) \geq \frac{p_1\mu_1}{p_1\mu_1 + \mu_m} \frac{\Delta}{C}$ and $s \in \mathcal{E}$, we have

$$\begin{aligned} \nabla V(s) &= \sum_{r=2}^{m} \left(p_{r-1} \mu_{r-1} s_{1,r-1} - \mu_{r} s_{1,r} \right) \\ &= p_{1} \mu_{1} s_{1,1} - \mu_{m} s_{1,m} - \sum_{r=2}^{m-1} (1 - p_{r}) \mu_{r} s_{1,r} \\ &\stackrel{(a)}{\leq} p_{1} \mu_{1} - p_{1} \mu_{1} \sum_{r=2}^{M} s_{1,r} - \mu_{m} s_{1,m} - \sum_{r=2}^{m-1} (1 - p_{r}) \mu_{r} s_{1,r} \\ &= p_{1} \mu_{1} - p_{1} \mu_{1} \sum_{r=2}^{M} s_{1,r} - \mu_{m} \sum_{r=2}^{m} s_{1,r} + \mu_{m} \sum_{r=2}^{m-1} s_{1,r} - \sum_{r=2}^{m-1} (1 - p_{r}) \mu_{r} s_{1,r} \\ &= p_{1} \mu_{1} - (p_{1} \mu_{1} + \mu_{m}) \sum_{r=2}^{m} s_{1,r} - p_{1} \mu_{1} \sum_{r=m+1}^{M} s_{1,r} - \sum_{r=2}^{m-1} (1 - p_{r}) \mu_{r} s_{1,r} + \mu_{m} \sum_{r=2}^{m-1} s_{1,r} \\ &\stackrel{(b)}{\leq} p_{1} \mu_{1} \left(1 - \sum_{r=m+1}^{M} L_{1,r} \right) - (p_{1} \mu_{1} + \mu_{m}) B_{m} - \sum_{r=2}^{m-1} (1 - p_{r}) \mu_{r} L_{1,r} + \mu_{m} U_{m-1} - \frac{p_{1} \mu_{1} \Delta}{C} \\ &\stackrel{(c)}{=} - \frac{p_{1} \mu_{1} \Delta}{C}, \end{aligned}$$

where

- (a) holds because $s_{1,1} = s_1 \sum_{r=2}^{M} s_{1,r}$ and $s_1 \le 1$;
- (b) holds because $s_{1,r} \ge L_{1,r}$ for any $1 \le r \le M$, $\sum_{r=2}^{m-1} s_{1,r} \le U_{m-1}$ and $\sum_{r=2}^{m} s_{1,r} \ge B_m + \frac{p_1\mu_1}{p_1\mu_1 + \mu_m} \frac{\Delta}{C}$ implied by $V(s) \ge \frac{p_1\mu_1}{p_1\mu_1 + \mu_m} \frac{\Delta}{C}$;

• and (c) holds by the definition of $B_m = \frac{p_1 \mu_1 (1 - \sum_{r=m+1}^M L_{1,r}) - \sum_{r=2}^{m-1} (1-p_r) \mu_r L_{1,r} + \mu_m U_{m-1}}{p_1 \mu_1 + \mu_m}$. Moreover, we have

$$\nabla V(s) = p_1 \mu_1 s_{1,1} - \mu_m s_{1,m} - \sum_{r=2}^{m-1} (1 - p_r) \mu_r s_{1,r} \le p_1 \mu_1 s_{1,1} \le p_1 \mu_1$$

We now apply Lemma 10 with $j = \frac{2\sqrt{N}\log N}{C}$. Define $B = \frac{p_1\mu_1}{p_1\mu_1 + \mu_r}\frac{\Delta}{C}$, $\gamma = \frac{p_1\mu_1\Delta}{C}$, and $\delta = p_1\mu_1$. Since $q_{\max} = p_1\mu_1N$ and $\nu_{\max} = \frac{1}{N}$, we have

$$\alpha = \frac{1}{1 + \frac{\Delta}{C}}$$
 and $\beta = \frac{C}{\Delta} + 1$,

and

$$\mathbb{P}\left(V(S) \ge B + 2\nu_{\max}j\right) \stackrel{(a)}{=} \mathbb{P}\left(\sum_{m=2}^{r} S_{1,m} - B_r \ge \frac{p_1\mu_1}{p_1\mu_1 + \mu_m} \frac{\Delta}{C} + \frac{4\Delta}{C}\right)$$

$$\stackrel{(b)}{\le} \left(\frac{1}{1 + \frac{\Delta}{C}}\right)^{\frac{2\sqrt{N}\log N}{C}} + \beta \mathbb{P}\left(S \notin \mathcal{E}\right)$$

$$\stackrel{(c)}{\le} \left(1 - \frac{\Delta}{2C}\right)^{\frac{2\sqrt{N}\log N}{C}} + \beta \left(\sigma_{m-1} + \sum_{m=1}^{M} \epsilon_m\right)$$

$$\le e^{-\frac{\log^2 N}{C^2}} + \beta \left(\sigma_{m-1} + \sum_{m=1}^{M} \epsilon_m\right)$$

where

- (a) holds by substituting $B = \frac{p_1 \mu_1}{p_1 \mu_1 + \mu_m} \frac{\Delta}{C}$, $\nu_{\max} = \frac{1}{N}$ and $j = \frac{2\sqrt{N} \log N}{C}$;
- (b) holds based on Lemma 10;
- and (c) holds because $\frac{1}{C} \leq \frac{1}{\Delta}$ and union bounds on $\mathbb{P}(S \notin \mathcal{E})$.

Now we prove $U_m = B_m + \left(\frac{p_1\mu_1}{p_1\mu_1 + \mu_m} + 4\right) \frac{\Delta}{C}$, which serves the upper bound on $\sum_{r=2}^m S_{1,r}$ and we represent U_m with $L_{1,1}$. Recall the definition of $L_{1,m}$ from the previous subsection that

$$L_{1,m} = \frac{v_m}{v_{m-1}} L_{1,m-1} - \frac{5v_m}{C} \Delta, \forall m \ge 2,$$
(31)

which implies that

$$L_{1,m} = \frac{v_m}{v_1} L_{1,1} - \frac{5(m-1)v_m}{C} \Delta, \forall m \ge 2.$$
(32)

Therefore, we have

$$\sum_{r=m+1}^{M} L_{1,r} = \sum_{r=m+1}^{M} \frac{v_r}{v_1} L_{1,1} - \sum_{r=m+1}^{M} \frac{5(r-1)v_r}{C} \Delta,$$
(33)

and

$$\sum_{r=2}^{m-1} (1-p_r)\mu_r L_{1,r} = \sum_{r=2}^{m-1} p_{r-1}\mu_{r-1}L_{1,r-1} - p_r\mu_r L_{1,r} - \frac{5\mu_r v_r}{C}\Delta$$
$$= p_1\mu_1 - \mu_m L_{1,m} + \frac{5\mu_m v_m}{C}\Delta - \sum_{r=2}^{m-1} \frac{5\mu_r v_r}{C}\Delta$$
$$= \left(p_1\mu_1 - \frac{\mu_m v_m}{v_1}\right)L_{1,1} - \sum_{r=2}^{m-1} \frac{5\mu_r v_r}{C}\Delta + \frac{5(m-2)\mu_m v_m}{C}\Delta$$
(34)

where the first and second equalities hold by substituting (31) and the last equality holds by substituting (32). Finally we have

$$B_m + \left(\frac{p_1\mu_1}{p_1\mu_1 + \mu_m} + 4\right)\frac{\Delta}{C} \tag{35}$$

$$=\frac{p_{1}\mu_{1}(1-\sum_{r=m+1}^{M}L_{r})-\sum_{r=2}^{m-1}(1-p_{r})\mu_{r}L_{1,r}+\mu_{m}U_{m-1}}{p_{1}\mu_{1}+\mu_{m}}+\left(\frac{p_{1}\mu_{1}}{p_{1}\mu_{1}+\mu_{m}}+4\right)\frac{\Delta}{C}$$
(36)

$$=\frac{p_{1}\mu_{1} - \left(p_{1}\mu_{1} - \frac{\mu_{m}v_{m}}{v_{1}} + p_{1}\mu_{1}\sum_{r=m+1}^{M}\frac{v_{r}}{v_{1}}\right)L_{1,1} + \mu_{m}U_{m-1}}{p_{1}\mu_{1} + \mu_{m}} + \frac{c_{m}}{C}\Delta$$
(37)

$$=\frac{p_{1}\mu_{1}}{p_{1}\mu_{1}+\mu_{m}}-\left(\frac{p_{1}\mu_{1}}{p_{1}\mu_{1}+\mu_{m}}\left(1+\sum_{r=m+1}^{M}\frac{v_{r}}{v_{1}}\right)-\frac{\mu_{m}}{p_{1}\mu_{1}+\mu_{m}}\frac{v_{m}}{v_{1}}\right)L_{1,1}+\frac{\mu_{m}}{p_{1}\mu_{1}+\mu_{m}}U_{m-1}+\frac{c_{m}}{C}\Delta$$
(38)

$$=1 - a_m - b_m L_{1,1} + a_m U_{m-1} + \frac{c_m}{C} \Delta$$
(39)
= U_m (40)

where

$$c_m = \frac{5p_1\mu_1\sum_{r=m+1}^{M}(r-1)v_r}{p_1\mu_1 + \mu_m} + \frac{5\sum_{r=2}^{m-1}\mu_r v_r - 5(m-2)\mu_m v_m}{p_1\mu_1 + \mu_m} + \frac{p_1\mu_1}{p_1\mu_1 + \mu_m} + 4$$

B.5 Proof of Lemma 4: Convergence of $L_{1,1}(n)$

Lemma 4. Recall that $\xi = \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j$. Given $\mathbb{P}(S_{1,1} < L_{1,1}(n)) \le \epsilon_1(n)$ and $L_{1,1}(n) \le s_{1,1}^* - \frac{6\Delta}{C}$, we have

$$\mathbb{P}(S_{1,1} < L_{1,1}(n+1)) \le \epsilon_1(n+1),$$

where

$$L_{1,1}(n+1) = \frac{1}{\mu_1} - \frac{C_M \Delta}{C(1-\xi)} - \frac{1}{2\mu_1 N^{\alpha}} + \xi \left(L_{1,1}(n) - \frac{1}{\mu_1} + \frac{C_M \Delta}{C(1-\xi)} + \frac{1}{2\mu_1 N^{\alpha}} \right)$$

and $\epsilon_1(n+1) = \epsilon_1(n)(M^2+2)(\frac{2C}{\bar{v}\Delta}+1)^{2M}$. Furthermore, $L_{1,1}(n+1) > L_{1,1}(n)$ holds when $L_{1,1}(n) \leq s_{1,1}^* - \frac{6\Delta}{C}$.

Starting from $L_{1,1}(n) \leq s_{1,1}^* - \frac{6\Delta}{C}$, we can apply Lemma 11 to obtain lower bounds $L_{1,m}(n)$ for $m \geq 2$ and then apply Lemma 12 to obtain upper bounds $U_m(n)$ for all $m \geq 2$, including $U_M(n)$. Then from $U_M(n)$, we obtain new lower bound $L_{1,1}(n+1)$. This iterative process implies that $U_M(n)$ and $L_{1,1}(n+1)$ are both a function of $L_{1,1}(n)$, as shown below. Recall that in Lemma 3, we obtained

$$U_m = 1 - a_m - b_m L_{1,1} + a_m U_{m-1} + \frac{c_m \Delta}{C}.$$

By recursively substituting U_m , we can write U_M as a function of $L_{1,1}$ as follows:

$$U_M = \sum_{m=2}^M (1 - a_m) \prod_{j=m+1}^M a_j - L_{1,1} \sum_{m=2}^M b_m \prod_{j=m+1}^M a_j + \frac{\Delta}{C} \sum_{m=2}^M c_m \prod_{j=m+1}^M a_j$$

Let us consider

$$L_{1,1}(n+1) = 1 - \frac{1-\xi}{2\mu_1 N^{\alpha}} - U_M(n) - \frac{6\Delta}{C}$$

= $\prod_{m=2}^M a_m + L_{1,1} \sum_{m=2}^M b_m \prod_{j=m+1}^M a_j - \frac{\Delta}{C} \left(\sum_{m=2}^M c_m \prod_{j=m+1}^M a_j + 6 \right) - \frac{1-\xi}{2\mu_1 N^{\alpha}}$

where we use $\sum_{m=2}^{M} (1-a_m) \prod_{j=m+1}^{M} a_j = \sum_{m=2}^{M} (\prod_{j=m+1}^{M} a_j - \prod_{j=m}^{M} a_j) = 1 - \prod_{m=2}^{M} a_m$. In Lemma 9, we will show $1 - \xi = \mu_1 \prod_{m=2}^{M} a_m$. We now center $L_{1,1}$ around $\frac{1}{\mu_1} - \frac{C_M \Delta}{C(1-\xi)} - \frac{1}{2\mu_1 N^{\alpha}}$ and have

$$L_{1,1}(n+1) - \frac{1}{\mu_1} + \frac{C_M \Delta}{C(1-\xi)} + \frac{1}{2\mu_1 N^{\alpha}} = \xi \left(L_{1,1}(n) - \frac{1}{\mu_1} + \frac{C_M \Delta}{C(1-\xi)} + \frac{1}{2\mu_1 N^{\alpha}} \right),$$

where $\xi = \sum_{m=2}^{M} b_m \prod_{j=m+1}^{M} a_j$ and $C_M = \sum_{m=2}^{M} c_m \prod_{j=m+1}^{M} a_j + 6$. Next we study the probability of $\epsilon_1(n+1)$ given $\epsilon_1(n)$. From Lemma 2, we have

$$\epsilon_m = e^{-\frac{v_m^2 \log^2 N}{C^2}} + \left(\frac{C}{v_m \Delta} + 1\right) \epsilon_{m-1}$$
$$\leq e^{-\frac{\bar{v}^2 \log^2 N}{C^2}} + \left(\frac{C}{\bar{v}\Delta} + 1\right) \epsilon_{m-1}$$

By expanding the above inequality from ϵ_M until ϵ_1 , it implies that

$$\epsilon_M \leq \sum_{m=2}^{M} e^{-\frac{\bar{v}^2 \log^2 N}{C^2}} \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-m} + \epsilon_1 \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-1}$$
$$\leq \left(\sum_{m=2}^{M} e^{-\frac{\bar{v}^2 \log^2 N}{C^2}} + \epsilon_1\right) \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-1}$$
$$\leq M \epsilon_1 \left(\frac{C}{\bar{v}\Delta} + 1\right)^{M-1}$$

where the inequality holds because $\epsilon_1 \ge e^{-\frac{\bar{v}^2 \log^2 N}{C^2}}$. From Lemma 3, we have

$$\sigma_m = e^{-\frac{\log^2 N}{C^2}} + \left(\frac{C}{\Delta} + 1\right) \left(\sigma_{m-1} + \sum_{m=1}^M \epsilon_m\right)$$
$$\leq e^{-\frac{\log^2 N}{C^2}} + M\epsilon_M + \left(\frac{C}{\Delta} + 1\right)\sigma_{m-1}$$

which implies that

$$\sigma_{M} \leq \left(e^{-\frac{\log^{2} N}{C^{2}}} + M\epsilon_{M}\right) \left(\frac{C}{\Delta} + 1\right)^{M}$$
$$\leq \left(e^{-\frac{\log^{2} N}{C^{2}}} + M^{2}\epsilon_{1}\left(\frac{C}{\bar{v}\Delta} + 1\right)^{M}\right) \left(\frac{C}{\Delta} + 1\right)^{M-1}$$
$$\leq \epsilon_{1}(M^{2} + 1)\left(\frac{C}{\bar{v}\Delta} + 1\right)^{2M-1}$$

Therefore, we have

$$e^{-\frac{\log^2 N}{C^2}} + \left(\frac{C}{\Delta} + 1\right)\sigma_M \leq e^{-\frac{\log^2 N}{C^2}} + \epsilon_1(M^2 + 1)\left(\frac{C}{\bar{v}\Delta} + 1\right)^{2M}$$
$$\leq \epsilon_1(M^2 + 2)\left(\frac{C}{\bar{v}\Delta} + 1\right)^{2M} = \epsilon_1(n+1)$$

Lastly, we prove the "monontocity" improvement of $\{L_{1,1}(n)\}_n$ by studying

$$L_{1,1}(n+1) - L_{1,1}(n) = (1-\xi) \left(\frac{1}{\mu_1} - \frac{C_M \Delta}{C(1-\xi)} - \frac{1}{2\mu_1 N^{\alpha}} - L_{1,1}(n) \right),$$

which is positive for $L_{1,1}(n) < \frac{\lambda}{\mu_1} - \frac{6\Delta}{C} < \frac{1}{\mu_1} - \frac{C_M \Delta}{C(1-\xi)} - \frac{1}{2\mu_1 N^{\alpha}}$.

C Proof of Lemma 7

According to the definition of $e_{j,m}$ and f(s) in (20), we have

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

and

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

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

$$Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{M}s_{i,m}\right)$$

= $N\lambda\left(1-A_{b}(S)\right)\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}s_{i,m}+\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}s_{i,m}\right)\right)$
+ $N\left(\sum_{m=1}^{M}(1-p_{m})\mu_{m}s_{1,m}\right)\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}s_{i,m}-\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}s_{i,m}\right)\right),$

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 $(1 - p_m)\mu_m s_{1,m}$ is the rates at which jobs leave the system when in phase m in the state s. Therefore, $\sum_{m=1}^{M} (1 - p_m)\mu_m s_{1,m}$ is the total departure rate. Define $d_1 = \sum_{m=1}^{M} (1 - p_m)\mu_m s_{1,m}$ and its stochastic correspondence $D_1 = \sum_{m=1}^{M} (1 - p_m)\mu_m S_{1,m}$ for simple notations.

Substituting the generator equation above to (22), we have

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right] \\
=\mathbb{E}\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\left(-\Delta\right) \\
-N\lambda(1-A_{b}(S))\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}+\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right) \\
-ND_{1}\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}-\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{M}S_{i,m}\right)\right)\right].$$
(41)

According to (19), it is easy to verify

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

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

$$g'(x) = -\frac{x-\eta}{\Delta},\tag{42}$$

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

$$g''(x) = -\frac{1}{\Delta}.$$
(43)

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

$$g(x + \frac{1}{N}) - g(x) = \left(g\left(x + \frac{1}{N}\right) - g(x)\right) \left(\mathbb{I}_{x \in \mathcal{T}_1} + \mathbb{I}_{x \in \mathcal{T}_2}\right)$$
$$= \frac{g'(\xi)}{N} \mathbb{I}_{x \in \mathcal{T}_1} + \left(\frac{g'(x)}{N} + \frac{g''(\zeta)}{2N^2}\right) \mathbb{I}_{x \in \mathcal{T}_2}$$
(44)

$$g(x - \frac{1}{N}) - g(x) = \left(g\left(x - \frac{1}{N}\right) - g(x)\right) \left(\mathbb{I}_{x \in \mathcal{T}_1} + \mathbb{I}_{x \in \mathcal{T}_2}\right)$$
$$= -\frac{g'(\tilde{\xi})}{N} \mathbb{I}_{x \in \mathcal{T}_1} + \left(-\frac{g'(x)}{N} + \frac{g''(\tilde{\zeta})}{2N^2}\right) \mathbb{I}_{x \in \mathcal{T}_2}$$
(45)

where $\xi, \zeta \in (x, x + \frac{1}{N})$ and $\tilde{\xi}, \tilde{\zeta} \in (x - \frac{1}{N}, x)$. Substitute (44) and (45) into the generator difference in (41), we have

$$\mathbb{E}\left[h\left(\sum_{i=1}^{b} S_i\right)\right] = J_1 + J_2 + J_3,\tag{46}$$

with

$$J_1 = \mathbb{E}\left[g'\left(\sum_{i=1}^b S_i\right)\left(\lambda A_b(S) - \lambda - \Delta + D_1\right)\mathbb{I}_{\sum_{i=1}^b S_i \in \mathcal{T}_2}\right],\tag{47}$$

$$J_2 = \mathbb{E}\left[\left(g'\left(\sum_{i=1}^b S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1 - A_b(S))g'(\xi) + D_1g'(\tilde{\xi})\right)\mathbb{I}_{\sum_{i=1}^b S_i \in \mathcal{T}_1}\right],\tag{48}$$

$$J_3 = -\mathbb{E}\left[\frac{1}{2N}\left(\lambda(1-A_b(S))g''(\zeta) + D_1g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^b S_i \in \mathcal{T}_2}\right].$$
(49)

Note that in (48) and (49), we have that

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

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. The proof of Lemma 7 is completed by upper bounding J_2 and J_3 , for which, we establish gradient bounds on g' and g'' in Lemma 11 and Lemma 12.

Lemma 11. Given $x \in \left[\eta - \frac{2}{N}, \eta + \frac{2}{N}\right]$, we have

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

Proof. From the definition of g function in (19), 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

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

Lemma 12. For $x > \eta$, we have

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

Proof. From the definition of g function in (19), 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

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

Based on gradient bounds in Lemma 11 and 12 and note $\sum_{m} (1 - p_m) \mu_m s_{1,m} \leq \mu_{\max} s_1 \leq \mu_{\max}$, then we have

$$J_{2} + J_{3} \leq \mathbb{E} \left[\left(g'\left(\sum_{i=1}^{b} S_{i}\right) \left(-\frac{\log N}{\sqrt{N}}\right) + \lambda |g'(\xi)| + \mu_{\max}|g'(\tilde{\xi})| \right) \mathbb{I}_{\sum_{i=1}^{b} S_{i} \in \mathcal{T}_{1}} \right] \\ + \mathbb{E} \left[\frac{1}{N} \left(\lambda |g''(\eta)| + \mu_{\max}|g''(\tilde{\eta})| \right) \mathbb{I}_{\sum_{i=1}^{b} S_{i} \in \mathcal{T}_{2}} \right] \\ \leq \frac{4\mu_{\max}}{\sqrt{N} \log N} + \frac{\lambda + \mu_{\max}}{N} \frac{\sqrt{N}}{\log N} \\ = \frac{5\mu_{\max} + \lambda}{\sqrt{N} \log N}$$

D Lemma 13 and the Proof

Lemma 13. For any $s \in S_{ssp_1}$,

$$\left(\lambda + \Delta - \sum_{m=1}^{M} (1 - p_m) \mu_m s_{1,m}\right) \mathbb{I}_{\sum_{i=1}^{b} s_i > \lambda + k\Delta + \frac{1}{N}} \le 0.$$

Г		
L		
L		

Proof. We consider the following linear programming problem

$$\min_{s_{1,m} \in \mathcal{S}_{ssp_1}} \sum_{m=1}^{M} (1 - p_m) \mu_m s_{1,m},$$

with \mathcal{S}_{ssp_1} defined by

$$S_{ssp_1} = \{ s \mid s_1 \ge \lambda + (k - \zeta - 6) \Delta, \ s_{1,m} \ge s_{1,m}^* - \theta_m \Delta \}.$$

Recall $w_m = (1 - p_m)\mu_m$. The minimum value is achieved when the maximum mass is allocated to m^* such that $w_{m^*} = w_l = \min_m w_m$. Therefore, we have

$$\sum_{m=1}^{M} w_m s_{1,m} \stackrel{(a)}{\geq} \sum_{m \neq m^*}^{M} w_m (s_{1,m}^* - \theta_m \Delta) + w_{m^*} \left(s_{1,m^*}^* + \left(k - \zeta - 6 + \sum_{m \neq m^*}^{M} \theta_m \right) \Delta \right)$$
$$\stackrel{(b)}{=} \lambda + w_l \left(k - \zeta - 6 + \sum_m \theta_m \right) \Delta - \sum_m w_m \theta_m \Delta$$
$$\stackrel{(c)}{=} \lambda + \Delta$$

where

- (a) holds because $s_1 \ge \lambda + (k \zeta 6)\Delta$ and $s_{1,m}, \forall m \ne m^*$ takes $L_{1,m} = s_{1,m}^* \theta_m \Delta;$
- (b) holds because $\sum_{m} w_m s_{1,m}^* = \lambda;$
- and (c) holds because $w_l(k \zeta 6 + \sum_m \theta_m) \sum_m w_m \theta_m = 1$ given carefully chosen $\zeta = \frac{4w_u b}{w_l} [(\frac{1}{w_l} \frac{1}{w_u}) \sum_m \theta_m w_m + \frac{1}{w_l} + 6]$ and $k = \frac{\sum_m \theta_m w_m}{w_u} + (1 + \frac{w_l}{4w_u b})\zeta \sum_m \theta_m$.

E Lemma 14 and the Proof

Lemma 14. For a large N such that , we have

$$\mathbb{P}\left(S \notin \mathcal{S}_{ssp}\right) \leq \frac{2}{N^2}.$$

Proof. The proof of Lemma 14 again relies on iterative state space peeling, which is based on Theorem 1 and Lemma 15 below.

Lemma 15 (A Lower Bound on S_1 via $\sum_{i=2}^{b} S_i$).

$$\mathbb{P}\left(\min\left\{\lambda+k\Delta-S_1,\sum_{i=2}^b S_i\right\} \le (\zeta+6)\Delta\right) \ge 1-\frac{1}{N^2},$$

where $\zeta = \frac{4w_u b}{w_l} [(\frac{1}{w_l} - \frac{1}{w_u}) \sum_m w_m \theta_m + \frac{1}{w_l} + 6]$ and $k = \frac{\sum_m w_m \theta_m}{w_u} + (1 + \frac{w_l}{4bw_u})\zeta - \sum_m \theta_m$.

Based on Theorem 1 and Lemma 15, we define sets $\tilde{\mathcal{S}}_1$ and $\tilde{\mathcal{S}}_2$ such that

$$\tilde{\mathcal{S}}_1 = \left\{ s \mid s_{1,m} \ge s_{1,m}^* - \theta_m \Delta \right\}$$
(50)

$$\tilde{\mathcal{S}}_2 = \left\{ s \mid \min\left\{\eta - s_1, \sum_{i=2}^b s_i\right\} \le (\zeta + 6)\Delta \right\}.$$
(51)

According to the union bound and Theorem 1 and Lemma 15, we have

$$\mathbb{P}\left(S \notin \tilde{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_2\right) \le \frac{M}{N^3} + \frac{1}{N^2} \le \frac{2}{N^2}.$$

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

$$\min\left\{\lambda + k\Delta - s_1, \sum_{i=2}^b s_i\right\} \le (\zeta + 6)\Delta,$$

we either have

$$\lambda + k\Delta - s_1 \le (\zeta + 6)\Delta,$$

which implies

$$s_1 \ge \lambda + (k - \zeta - 6)\,\Delta_2$$

or

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

which implies

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

Note that

$$\mathcal{S}_1 \cap \{s \mid s_1 \ge \lambda + (k - \zeta - 6) \Delta\} = \mathcal{S}_{ssp_1}$$

and

$$\tilde{\mathcal{S}}_1 \cap \left\{ s \mid \sum_{i=1}^b s_i \leq \eta \right\} \subseteq \mathcal{S}_{ssp_2}.$$

 $\tilde{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_2 \subseteq \mathcal{S}_{ssp},$

We, therefore, have

and

$$\mathbb{P}\left(S \notin \mathcal{S}_{ssp}\right) \leq \mathbb{P}\left(S \notin \tilde{\mathcal{S}}_{1} \cap \tilde{\mathcal{S}}_{2}\right) \leq \frac{2}{N^{2}},$$

so Lemma 14 holds.

Next, we prove Lemma 15.

E.1 Proof of Lemma 15

Recall $w_u = \max_{1 \le m \le M} (1 - p_m) \mu_m$, $w_l = \min_{1 \le m \le M} (1 - p_m) \mu_m$, and $L_m = s_{1,m}^* - \theta_m \Delta$. Let $\zeta = \frac{4w_u b}{w_l} \left(\left(\frac{1}{w_l} - \frac{1}{w_u} \right) \sum_m w_m \theta_m + \frac{1}{w_l} + 6 \right)$ and $k = \frac{\sum_m w_m \theta_m}{w_u} + \left(1 + \frac{w_l}{4bw_u} \right) \zeta - \sum_m \theta_m$. Consider Lyapunov function

$$V(s) = \min\left\{\lambda + k\Delta - s_1, \sum_{i=2}^{b} s_i\right\}$$
(52)

and define

$$\mathcal{E} = \left\{ s \mid s_{1,m} \ge L_m, \forall 1 \le m \le M \right\}.$$

When $V(s) \ge \zeta \Delta$, the following two inequalities hold

$$s_1 \le \lambda + (k - \zeta)\Delta \le 1 - \frac{1 - \xi}{2\mu_1 N^{\alpha}},\tag{53}$$

$$\sum_{i=2}^{b} s_i \ge \zeta \Delta. \tag{54}$$

We have two observations based on (53) and (54):

- (53) implies that $A_1(s) \leq \frac{1}{\sqrt{N}}$ under any policy in Π ;
- (54) implies that $s_2 \ge \frac{\zeta \Delta}{b}$ because $s_2 \ge s_3 \ge \cdots \ge s_b$, and we have

$$\sum_{m=1}^{M} (1 - p_m) \mu_m s_{2,m} \ge w_l s_2 \ge \frac{w_l c_1 \Delta}{b},$$
(55)

where a finite buffer size is required such that the lower bound $w_l s_2 \geq \frac{w_l c_1 \Delta}{b}$ is meaningful. We next study the Lyapunov drift when $V(s) \geq \zeta \Delta$ and $s \in \mathcal{E}$ by considering two cases:

• Suppose $\lambda + k\Delta - s_1 \ge \sum_{i=2}^b s_i \ge \zeta \Delta$. In this case, $V(s) = \sum_{i=2}^b s_i$ and

$$\nabla V(s) \leq \lambda (A_1(s) - A_b(s)) - \sum_{m=1}^M (1 - p_m) \mu_m s_{2,m}$$

$$\stackrel{(a)}{\leq} \frac{1}{\sqrt{N}} - \sum_{m=1}^M (1 - p_m) \mu_m s_{2,m}$$

$$\stackrel{(b)}{\leq} \frac{1}{\sqrt{N}} - \frac{w_l c_1 \Delta}{b}$$

$$\stackrel{(c)}{\leq} - \frac{w_l c_1 \Delta}{2b}$$
(56)

where

- (a) holds because $A_1(s) \leq \frac{1}{\sqrt{N}}$ under any policy in Π ;
- (b) holds because of (55);
- and (c) holds because $\log N \ge \frac{4b}{w_l c_1}$.

• Suppose $\sum_{i=2}^{b} s_i > \lambda + k\Delta - s_1 \ge \zeta \Delta$. In this case, $V(s) = \lambda + k\Delta - s_1$ and

$$\nabla V(s) \leq -\lambda(1-A_{1}(s)) + \sum_{m=1}^{M} (1-p_{m})\mu_{m}s_{1,m} - \sum_{m=1}^{M} (1-p_{m})\mu_{m}s_{2,m}$$

$$\leq \frac{1}{\sqrt{N}} - \lambda + w_{u}s_{1} - \sum_{m=1}^{M} (w_{u} - w_{m})s_{1,m} - \sum_{m=1}^{M} w_{m}s_{2,m}$$

$$\stackrel{(a)}{\leq} \frac{1}{\sqrt{N}} - \lambda + w_{u} \left(s_{1} - \sum_{m=1}^{M} L_{m}\right) + \sum_{m=1}^{M} w_{m}L_{m} - \sum_{m=1}^{M} w_{m}s_{2,m}$$

$$\stackrel{(b)}{=} \frac{1}{\sqrt{N}} + \left(w_{u} \left(k - \zeta + \sum_{m=1}^{M} \theta_{m}\right) - \sum_{m=1}^{M} w_{m}\theta_{m}\right)\Delta - \sum_{m=1}^{M} w_{m}\mu_{m}s_{2,m}$$

$$\stackrel{(c)}{\leq} \frac{1}{\sqrt{N}} + \left(w_{u} \left(k - \zeta + \sum_{m=1}^{M} \theta_{m}\right) - \sum_{m=1}^{M} w_{m}\theta_{m}\right)\Delta - \frac{w_{l}c_{1}\Delta}{b}$$

$$= \frac{1}{\sqrt{N}} - \frac{3w_{l}c_{1}\Delta}{4b}$$

$$\stackrel{(c)}{\leq} - \frac{w_{l}c_{1}\Delta}{2b}$$

$$(58)$$

where

- (a) holds because $s_{1,m} \ge L_m, \forall m \ge 1;$
- (b) holds because $s_1 \leq \lambda + (k \zeta)\Delta$ and $L_m = s_{1,m}^* \theta_m \Delta, \forall m \geq 1;$
- (c) holds because $w_u(k-\zeta+\sum_{m=1}^M\theta_m)-\sum_{m=1}^Mw_m\theta_m=-\frac{w_lc_1}{4b}$ given k and ζ ;
- and (d) holds because $\log N \ge \frac{4b}{w_l c_1}$.

Next, we further show $\nabla V(s) \leq w_u$ based on the upper bounds (56) and (57).

• Consider the upper bound in (56). We have

$$\nabla V(s) \le \lambda (A_1(s) - A_b(s)) - \sum_{m=1}^M (1 - p_m) \mu_m s_{2,m} \le 1 \le w_u,$$

where $1 \le w_u$ holds because $\sum_{m=1}^{M} v_m = 1$.

• Consider the upper bound in (57). We have

$$\nabla V(s) \le -\lambda(1 - A_1(s)) + \sum_{m=1}^M (1 - p_m)\mu_m s_{1,m} - \sum_{m=1}^M (1 - p_m)\mu_m s_{2,m}$$
$$\le \sum_{m=1}^M (1 - p_m)\mu_m s_{1,m} \le w_u,$$

where the last inequality holds because $\sum_{m} s_{1,m} = s_1 \leq 1$.

We now apply Lemma 10. Define $B = \zeta \Delta$, $\gamma = \frac{w_l \zeta \Delta}{2b}$ and $\delta = w_u$. Combining $q_{\max} = w_u N$ and $\nu_{\max} = \frac{1}{N}$, we have

$$\alpha = \frac{w_u}{w_u + \frac{w_l \zeta \Delta}{2b}}$$
 and $\beta = \frac{2w_u b}{w_l \zeta \Delta} + 1.$

Choosing $j = 3\sqrt{N} \log N$, we have

$$\mathbb{P}\left(V(S) \ge B + 2\nu_{\max}j\right) \stackrel{(a)}{=} \mathbb{P}\left(V(S) \ge (\zeta + 6)\Delta\right)$$

$$\stackrel{(b)}{\le} \left(\frac{1}{1 + \frac{w_l\zeta\Delta}{2w_ub}}\right)^{3\sqrt{N}\log N} + \beta \mathbb{P}\left(S \notin \mathcal{E}\right)$$

$$\le \left(1 - \frac{w_lc_1\Delta}{3w_ub}\right)^{3\sqrt{N}\log N} + \beta \mathbb{P}\left(S \notin \mathcal{E}\right)$$

$$\stackrel{(d)}{\le} e^{-\frac{w_lc_1}{w_ub}\log^2 N} + \left(\frac{2w_ub}{w_l\zeta\Delta} + 1\right) \mathbb{P}\left(S \notin \mathcal{E}\right)$$

$$\stackrel{(d)}{\le} e^{-\frac{w_lc_1}{w_ub}\log^2 N} + \left(\frac{2w_ub}{w_l\zeta\Delta} + 1\right) \frac{M}{N^3}$$

$$\stackrel{(e)}{\le} \frac{1}{N^2}$$

where

- (a) holds by substituting $B = \zeta \Delta$, $\nu_{\max} = \frac{1}{N}$ and $j = 3\sqrt{N} \log N$;
- (b) holds based on Lemma 10;
- (c) holds because $\frac{w_l c_1}{w_u b} \leq \frac{1}{\Delta}$;
- (d) holds by union bounds on $\mathbb{P}\left(S \notin \mathcal{E}\right)$;
- and (e) holds because $\frac{w_l c_1}{w_u b} \ge 24$ and $2\left(\frac{2Mbw_u}{w_l \zeta \Delta} + M\right) \le N$.