Steady State Analysis of Load Balancing Algorithms in the Heavy Traffic Regime

by

Xin Liu

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved June 2019 by the
Graduate Supervisory Committee:

Lei Ying, Chair
Siva Theja Maguluri
Weina Wang
Junshan Zhang

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

This dissertation studies load balancing algorithms for many-server systems (with N servers) and focuses on the steady-state performance of load balancing algorithms in the heavy traffic regime such that the load of system is $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 1$. The framework of Stein's method and (iterative) state space collapse (SSC) are used to analyze three load balancing systems: 1) load balancing in the traffic regime $(0 < \alpha < 0.5)$ with exponential service time; 2) load balancing in the traffic regime $(0.5 \leq \alpha < 1)$ with exponential service time; 3) load balancing in the traffic regime $(0 < \alpha < 0.5)$ with Coxian-2 service time.

When $0 < \alpha < 0.5$, i.e. the traffic load is lighter than the Halfin-Whitt regime, the sufficient conditions are established such that any load balancing algorithm that satisfies the conditions have both asymptotic zero waiting time and zero waiting probability. Furthermore, the number of servers with more than one jobs is $o(1)$, in other words, the system collapses to a one-dimensional space. The result is proven using Stein's method and state space collapse (SSC), which are powerful mathematical tools for steady-state analysis of load balancing algorithms. The second system is in even "heavier" traffic regime $(0.5 \leq \alpha < 1)$, and an iterative refined procedure is proposed to obtain the steady-state metrics. Again, asymptotic zero delay and waiting are established for a set of load balancing algorithms. Different from the first system, the system collapses to a two-dimensional state-space instead of one-dimensional state-space. The third system is more challenging because of "non-monotonicity" with Coxian-2 service time, and an iterative state space collapse is proposed to tackle the "non-monotonicity" challenge. For these three systems, a set of load balancing algorithms is established, respectively, under which the probability that an incoming job is routed to an idle server is one asymptotically (as $N \to \infty$) at steady-state. The set of load balancing algorithms includes join-the-shortest-queue (JSQ), idle-one-first

(I1F), join-the-idle-queue (JIQ), and power-of-d-choices (Pod) with a carefully-chosen d.

To my family.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my phenomenal advisor, Professor Lei Ying for his guidance, patience, and encouragement. His enthusiasm, knowledge, and rigorous attitude toward research have deeply influenced me. He has always been a great mentor who dedicated numerous amount of time and tremendous efforts in my doctoral study. I am eternally grateful to have him as my advisor who made my Ph.D. journey rewarding.

I would like to thank Professors Siva Theja Maguluri, Weina Wang and Junshan Zhang for serving on my committee, and for their valuable insights and feedback on my thesis. Especially thanks to Weina for her inspiration on the idea of iterative state-space collapse that made Chapter 5 possible.

I have had the fortune to collaborate with Professors Jim Dai and Anton Braverman. Every meeting with them has been constructive and illuminating. I truly appreciate the opportunity to interact with and learn from them. I also thank them for their warm and considerate host during my short visit at Ithaca.

The past several years would not have been as enjoyable without the company of terrific colleagues at INLAB and friends. Thank you for all of understanding and joys brought to me in this long journey.

Finally, I would like to thank my grandparents and parents for their unconditional support and love. I own my deepest gratitude to my fiancée, Shanqi Liu for her love seven thousand miles away, scarification and waiting. Thank you for always being there for me.

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

The rapid development of cloud computing, social networking, Internet-of-things (IOT) and machine learning brings intensive volume of internet traffic to data centers. Load balancer is an indispensable component in data center to optimize resource allocation and support high quality of service (e.g. job delay). Load balancing in data centers routes incoming jobs to servers to balance the load across servers and to minimize response times to improve user experience. It has been reported in Schurman and Brutlag (2009) that an extra delay of 500 ms led to 1.2% loss of users and revenue, and low delay (i.e. short response time) is very important in modern data centers networks.

1.1 Background

Load balancing reconciles two priory objectives in many-server systems, efficiency and quality of service Leverich and Kozyrakis (2014): on one hand, data center aims to maintain high quality of service by keeping workload per server (efficiency) low; on the other hand, data center should keep high efficiency to reduce the operation cost (e.g. power and cooling down), which may sacrifice quality of service. The efficiency-quality trade-off motivates researchers to consider heavy traffic regime in large-scale server systems (N servers), where workload per server is $\lambda = 1 - N^{-\alpha}$ and α is a positive constant in $(0, 1)$. In this regime, λ (efficiency) becomes high as N increases, which even approaches to one as N becomes large. In the heavy-traffic regime, the important questions to be answered are: what load balancing algorithms should be used in this regime? how do these algorithms perform (e.g. job delay)? In this dissertation, we aim to address the two fundamental questions in the regime for $0 < \alpha < 1$. Moreover, job size (e.g. machine learning training task) in load balancing systems are highly-dynamic and the distributions of service time are general. Therefore, the conventional assumption of exponential service time needs to be relaxed. However, performance analysis of load balancing systems with non-exponential service is much more challenging Harchol-Balter (2013). This dissertation takes a first step to tackle this challenging problem by considering load balancing systems in heavy traffic regime $0 < \alpha < 0.5$ with Coxian-2 service time.

1.2 Literature Review

Steady-state analysis of many-server systems is one of the most fundamental and widely-studied problems in queueing theory. The stationary distribution of the classic $M/M/N$ system (or called Erlang-C model) is one of the earliest subjects. For systems with distributed queues where each server maintains a separate queue, it is well known that the join-the-shortest-queue (JSQ) algorithm is delay optimal Winston (1977); Weber (1978) under fairly general conditions. However, the exact stationary distribution of many-server systems under JSQ remains to be an open problem. A recent breakthrough in this area is Eschenfeldt and Gamarnik (2018), which shows that in the Halfin-Whitt regime ($\alpha = 0.5$) Halfin and Whitt (1981), the diffusion-scaled process converges to a two-dimensional diffusion limit, from which it can be shown that most servers have one job in service and $O\left(\sqrt{N}\right)$ servers have two jobs (one in service and one in buffer). This seminal work has led to several significant developments: (i) Braverman (2018) proved that the stationary distribution indeed converges to the stationary distribution of the two-dimensional diffusion limit based on Stein's method; and (ii) via stochastic coupling, Mukherjee *et al.* (2018) showed that the diffusion limit of Pod converges to that of JSQ in the Halfin-Whitt regime at the process level (over finite time) when $d = \Theta(\sqrt{N}\log N)$; and (iii) when $\alpha < 1/6$, Liu and Ying (2018) proved that the waiting probability of a job is asymptotically zero with $d = \Omega\left(\frac{\log N}{1-\lambda}\right)$ at the steady-state based on Stein's method.

Motivated by the observation in the server system (e.g. call center) that waiting time is comparable with the service time, Atar (2012) also studied a centralized server model in even heavier traffic regime with $\alpha = 1$ and proved that the total queue length with proper scaling converges to diffusion process as $N \to \infty$. He (2015) extended the results in Atar (2012) to general service time and obtained the diffusion approximation of the waiting time by joint scaling of the space and time. Braverman *et al.* (2016) provided steady-state analysis of M/M/N system in the universal regime for any $0 \leq \alpha \leq 1$ by using Stein's method. For distributed server system, Gupta and Walton (2019) considered load balancing in the regime $\alpha = 1$ as considered in Atar (2012), and it shown that the scaled total queue length weakly converges to a stochastic differential equation (diffusion process). Based on the diffusion process, various load balancing algorithms, JSQ, JIQ and I1F are compared in terms of the total queue length.

Most of previous analysis of load balancing in heavy traffic regime, e.g. Eschenfeldt and Gamarnik (2018), Mukherjee *et al.* (2018), Liu and Ying (2018) and Gupta and Walton (2019), assume exponential service time. With general service time distributions, performance analysis of load balancing algorithms with distributed queues is a much more challenging problem, and remains to be an active research area in queueing theory Harchol-Balter (2013). Mitzenmacher (1996) proposed a mean-field model of the Po*d* policy with gamma service time distributions without proving the convergence of the stochastic system to the mean-field model. Aghajani *et al.* (2017); Vasantam *et al.* (2017); Hellemans and Van Houdt (2018) proposed a set of PDE models to approximate load balancing polices with general service times and numerically analyzed key performance metrics (e.g. mean response time). They proved the convergence of the stochastic systems to the corresponding ODEs or PDEs at process-level (over a finite time interval instead of at steady state).

1.3 Summary of Contributions

In Chapter 3, we study load balancing systems (N servers) assuming exponential service time. Each server has a buffer of size $b-1$ i.e. a server can have at most one job in service and $b-1$ jobs in queue. Jobs are served in first-come-first-serve (FIFO) order. We focus on the steady-state performance of load balancing algorithms in the heavy traffic regime such that the load of system is $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$, which we call Sub-Halfin-Whitt regime ($\alpha = 0.5$ is the so-called Halfin-Whitt regime). We establish a set of load balancing algorithms under which the probability that an incoming job is routed to an idle server is one asymptotically (as $N \to \infty$) at steady-state. The set of load balancing that satisfy the condition includes JSQ, I1F, JIQ, and Pod with $d \geq N^{\alpha} \log N$. The proof of the main result is based on Stein's method and state space collapse.

In Chapter 4, we study load balancing systems (N servers) in even "heavier" traffic regime ($0.5 \leq \alpha < 1$), which we call Beyond-Halfin-Whitt regime, assuming exponential service time and finite buffer size $b-1$. By an iterative moment procedure, we obtained high-order moment bounds on a distant function of total queue length, which is used to refine the steady-state metrics (e.g. waiting time and waiting probability). Interestingly, we establish a set of "zero-delay" load balancing, which also includes JSQ, I1F, JIQ, and Pod with $d \geq N^{\alpha} \log^2 N$.

We plot our contributions in Chapter 3 and 4 in terms of the waiting jobs Ns_2 and a log-scaled version of $\log Ns_2 / \log N$ to summarize the key results in Fig. 1.1. The result in Chapter 3 shows that Ns_2 can be $O(1/N)$ for $0 < \alpha < 0.5$ at steady

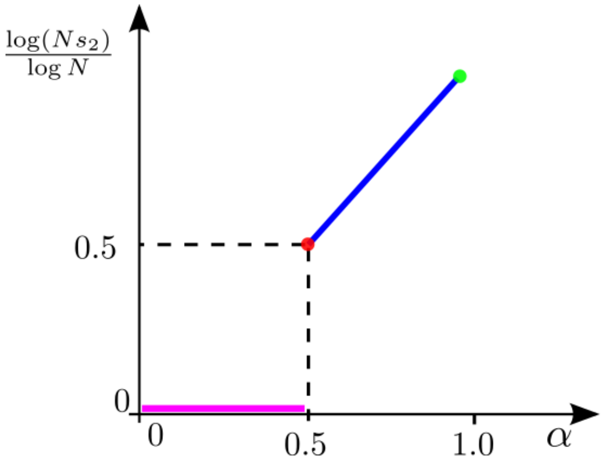

Figure 1.1: Our Contributions in Chapter 3 and 4.

state, which is purple line; Chapter 4 shows Ns_2 is $O(N^\alpha \log N)$ at steady-state for $0.5 \leq \alpha < 1$, which is blue line; Braverman (2018) shown that Ns_2 is $O(\sqrt{N})$ for Halfin-Whitt regime $\alpha = 0.5$ at steady state, which is red dot; Gupta and Walton (2019) shown that Ns_2 is $O(N)$ for $\alpha = 1$ at diffusion level, which is green dot.

In Chapter 5, we study load balancing systems (N servers) in Sub-Halfin-Whitt regime assuming Coxian-2 service time and finite buffer with size $b - 1$. We propose an iterative state space collapse to tackle “non-monotonicity” with Coxian-2 service time. We also identify a similar set of load balancing policies as in exponential service (only differs in “constant”) that achieves asymptotic zero waiting. The results suggest “insensitive” property (to service distribution) holds for load balancing system in heavy traffic regime as $N \to \infty$.

We would emphasis the analysis framework in the dissertation combines three interesting and powerful tools for steady-state analysis: Stein’s method, iterative state space collapse and iterative moment bounds. The framework helps tackle non-exponential challenges and establish high-order moments on total queue length to refine steady-state metrics (e.g. waiting time) in “heavier” traffic regime.

Chapter 2

STEIN'S METHOD AND STATE SPACE COLLAPSE

2.1 Load Balancing System

Consider a many-server system with N homogeneous servers, where job arrival follows a Poisson process with rate λN and service times are i.i.d. exponential random variables with rate one (here we take exponential service as an example to showcase our framework). We consider the traffic regime such that $\lambda = 1 - N^{-\alpha}$ for some $0 < \alpha < 1$. As shown in Figure 2.1, each server maintains a separate queue and we assume buffer size $b-1$ (i.e., each server can have one job in service and $b-1$ jobs in queue). Jobs are severed in first-in-first-come (FIFO) order.

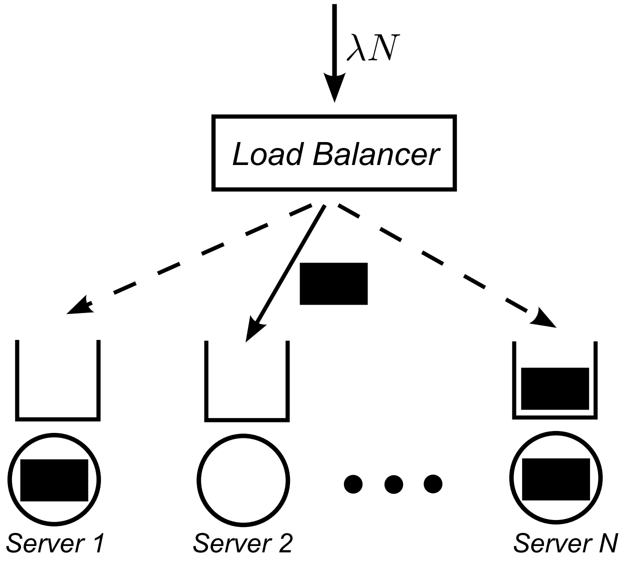

Figure 2.1: Load Balancing in Many-Server Systems.

Let $S_i(t)$ denote the fraction of servers with at least i jobs at time $t \geq 0$. Under the finite buffer assumption with buffer size $b-1$, we define $S_i(t) = 0,\ \forall i \geq b+1,\ \forall t \geq 0$

for notational convenience. Furthermore, define set $\mathbb{S} \subseteq \mathbb{R}^b$ such that

$$\mathbb{S} = \{s \in \mathbb{R}^b \mid 1 \geq s_1 \geq \cdots \geq s_b \geq 0 \text{ and } \ Ns_i \in \mathbb{N}, \ \forall i\}.$$

We then have $S(t) = [S_1(t), S_2(t), \cdots, S_b(t)]^T \in \mathbb{S}$ for any $t \geq 0$. We consider load balancing algorithms which route each incoming job to a server upon its arrival based on $S(t)$ so that $(S(t) : t \geq 0)$ is a finite-state and irreducible continuous-time Markov chain (CTMC), which implies that $(S(t) : t \geq 0)$ has a unique stationary distribution.

2.2 Generator Approximation

Define $e_i \in \mathbb{R}^b$ to be a b-dimensional vector such that the ith entry is $1/N$ and all other $b-1$ entries are zero. Furthermore, define $A_i(s)$ to be the probability that an incoming job is routed to a server with at least i jobs when the system is in state s, i.e.

$$A_i(s) = \Pr\left(\text{an incoming job is routed to a server with at least } i \text{ jobs} \middle| \, S(t) = s\right).$$

From this definition, we have $A_0(s) = 1$. Given the definition above, the CTMC transits from state s to $s + e_i$ with rate $\lambda N\left(A_{i-1}(s) - A_i(s)\right)$, which occurs when an arrival comes and is routed to an server with $i-1$ jobs; and transits from state s to $s - e_i$ with rate $N(s_i - s_{i+1})$, which occurs when a job leaves a server with i jobs.

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

$$\begin{aligned} Gf(s) = \sum_{i=1}^{b} & \lambda N(A_{i-1}(s) - A_i(s))(f(s+e_i) - f(s)) \\ & + N(s_i - s_{i+1})(f(s-e_i) - f(s)). \end{aligned} \tag{2.1}$$

For any bounded function $f : \mathbb{S} \to \mathbb{R}$,

$$E[Gf(S)] = 0, \tag{2.2}$$

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

To understand the steady-state performance of load balancing system, we will establish moment bounds on the following function:

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

where η is understood to be an "estimator" of $\sum_{i=1}^{b} S_i$. The moment bounds measure the likelihood that the total number of jobs in the system $(N\sum_{i=1}^{b} S_i)$ exceeds ηN. For example, $\eta = \lambda + \frac{k\log N}{\sqrt{N}}$ in load balancing system in the Sub-Halfin-Whitt regime in Chapter 3. The metric measures $N\sum_{i=1}^{b} S_i$ exceeds $N\lambda + k\sqrt{N}\log N$ at steady state, and can be used to bound the probability that an incoming job is routed to an idle server.

We consider a simple fluid system with arrival rate λ and departure rate $\lambda+\delta$, i.e.

$$\dot{x} = -\delta,$$

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

$$g'(x)\left(-\delta\right) = \left(\max\left\{x-\eta, 0\right\}\right)^r \ \ \forall x, \tag{2.3}$$

where $g'(x) = \frac{dg(x)}{dx}$ and r is a positive integer. The left-hand side of (2.3) is to apply the generator of the simple fluid system to function $g(x)$, i.e.

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

It is easy to verify that the solution to (2.3) is

$$g(x) = -\frac{(x-\eta)^{r+1}}{\delta(r+1)}\mathbb{I}_{x\geq\eta}, \tag{2.4}$$

and

$$g'(x) = -\frac{(x-\eta)^r}{\delta}\mathbb{I}_{x\geq\eta}. \tag{2.5}$$

Note that the simple fluid system is a one-dimensional system and the stochastic system is b-dimensional. In order to couple these two systems, we define

$$f(s) = g\left(\sum_{i=1}^{b} s_i\right), \tag{2.6}$$

and use $f(s)$ defined above in Stein's method.

Since $\sum_{i=1}^{b} s_i \leq b$ for $s \in \mathbb{S}$, $f(s)$ is a bounded for $s \in \mathbb{S}$. So

$$E[Gf(S)] = E\left[Gg\left(\sum_{i=1}^{b} S_i\right)\right] = 0. \tag{2.7}$$

Recall $\eta = \lambda + \frac{k\log N}{\sqrt{N}}$ and define

$$h_k(x) = \max\left\{x - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}.$$

Based on (2.3) and (2.7), we obtain

$$E\left[h_k^r\left(\sum_{i=1}^{b} S_i\right)\right] = E\left[g'\left(\sum_{i=1}^{b} S_i\right)(-\delta) - Gg\left(\sum_{i=1}^{b} S_i\right)\right]. \tag{2.8}$$

Note that according to the definition of $f(s)$ in (2.6) and e_j, we have

$$f(s+e_j) = g\left(\sum_{i=1}^{b} s_i + \frac{1}{N}\right)$$

and

$$f(s-e_j) = g\left(\sum_{i=1}^{b} s_i - \frac{1}{N}\right)$$

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

$$\begin{aligned}Gg\left(\sum_{i=1}^{b} s_i\right) =& N\lambda(1-A_b(s))\left(g\left(\sum_{i=1}^{b} s_i + \frac{1}{N}\right) - g\left(\sum_{i=1}^{b} s_i\right)\right)\\ &+ Ns_1\left(g\left(\sum_{i=1}^{b} s_i - \frac{1}{N}\right) - g\left(\sum_{i=1}^{b} s_i\right)\right).\end{aligned}$$

Substituting the equation above to (2.8), we have

$$\begin{aligned}
&E\left[h_k^r\left(\sum_{i=1}^{b} S_i\right)\right]\\
=&E\left[g'\left(\sum_{i=1}^{b} S_i\right)(-\delta)-N\lambda(1-A_b(S))\left(g\left(\sum_{i=1}^{b} S_i+\frac{1}{N}\right)-g\left(\sum_{i=1}^{b} S_i\right)\right)\right.\\
&\left.-NS_1\left(g\left(\sum_{i=1}^{b} S_i-\frac{1}{N}\right)-g\left(\sum_{i=1}^{b} S_i\right)\right)\right]. \qquad (2.9)
\end{aligned}$$

From the closed-forms of g and g' in (2.4) and (2.5), note that for any $x < \eta$,

$$g(x) = g'\left(x\right) = 0.$$

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

$$g'(x) = -\frac{\left(x-\eta\right)^r}{\delta}, \qquad (2.10)$$

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

$$g''(x) = -\frac{r\left(x-\eta\right)^{r-1}}{\delta}. \qquad (2.11)$$

By using mean-value theorem in the region $[\eta - \frac{1}{N}, \eta + \frac{1}{N}]$ and Taylor theorem in the region $(\eta + \frac{1}{N}, \infty)$, we have

$$\begin{aligned}
g(x+\frac{1}{N})-g\left(x\right)=&\left(g(x+\frac{1}{N})-g\left(x\right)\right)\left(1_{\eta-\frac{1}{N}\le x\le \eta+\frac{1}{N}}+1_{x>\eta+\frac{1}{N}}\right)\\
=&\frac{g'(\xi)}{N}1_{\eta-\frac{1}{N}\le x\le \eta+\frac{1}{N}}+\left(\frac{g'(x)}{N}+\frac{g''(\zeta)}{2N^2}\right)1_{x>\eta+\frac{1}{N}} \qquad (2.12)\\
g(x-\frac{1}{N})-g\left(x\right)=&\left(g(x-\frac{1}{N})-g\left(x\right)\right)\left(1_{\eta-\frac{1}{N}\le x\le \eta+\frac{1}{N}}+1_{x>\eta+\frac{1}{N}}\right)\\
=&-\frac{g'(\tilde{\xi})}{N}1_{\eta-\frac{1}{N}\le x\le \eta+\frac{1}{N}}+\left(-\frac{g'(x)}{N}+\frac{g''(\tilde{\zeta})}{2N^2}\right)1_{x>\eta+\frac{1}{N}} \qquad (2.13)
\end{aligned}$$

where $\xi, \zeta \in (x, x + \frac{1}{N})$ and $\tilde{\xi}, \tilde{\zeta} \in (x - \frac{1}{N}, x)$. Substitute (2.12) and (2.13) into the generator difference in (2.9), we summarize the generator difference the following lemma and it will be used in Chapter 3 and Chapter 4.

Lemma 1.

$$\begin{aligned}
&E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right]\\
=&E\left[g'\left(\sum_{i=1}^b S_i\right)\left(\lambda A_b(S)-\lambda-\delta+S_1\right)\mathbb{I}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\right] && (2.14)\\
&+E\left[\left(g'\left(\sum_{i=1}^b S_i\right)(-\delta)-\lambda(1-A_b(S))g'(\xi)+S_1g'(\tilde{\xi})\right)\mathbb{I}_{\eta-\frac{1}{N}\leq\sum_{i=1}^b S_i\leq\eta+\frac{1}{N}}\right]\\
& && (2.15)\\
&-E\left[\frac{1}{2N}\left(\lambda(1-A_b(S))g''(\zeta)+S_1g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\right]. && (2.16)
\end{aligned}$$

Note in (2.14) *and* (2.16), *we have random variables* $\xi,\zeta\in\left(\sum_{i=1}^b S_i,\sum_{i=1}^b S_i+\frac{1}{N}\right)$ *and* $\tilde{\xi},\tilde{\zeta}\in\left(\sum_{i=1}^b S_i-\frac{1}{N},\sum_{i=1}^b S_i\right)$ *whose values depend on* $\sum_{i=1}^b S_i$. □

2.3 State Space Collapse

To study the generator difference in Lemma 1, we need to understand the system behavior in the region where total queue length is larger than η, which is related to the term (2.14). In this region, we have an key observation that the system state collapses to a restricted region, called state space collapse (SSC). The SSC observation is critical to bound the term (2.14), therefore, the generator difference in (2.9). In the following, we provide an intuitive argument on SSC by using the system in Chapter 4 as an example: Load balancing in Beyond-Halfin-Whitt regime (note $\eta=1+\frac{k\log N}{N^{1-\alpha}}$ and $\delta=\frac{1}{N^\alpha}$ are chosen in Chapter 4).

For ease of exposition, we consider JSQ load balancing in a simplified system with buffer size $b=1$, where only s_1 and s_2 exists. Given the system state (s_1,s_2) and $s_1<1$, the drift of idle servers under JSQ is $1-s_1$ and that of s_2 is $-s_2$, because all arrivals are allocated to idle servers under JSQ when $s_1<1$. Specify $(s_1,s_2)=(0.5,0.5)$ and the drifts are $(0.5,-0.5)$. Therefore, s_1 increases very fast and approaches close to 1, as shown in the green region. Since most of servers are busy (s_1 is close to 1) in the green region, the total queue length per server $s_1 + s_2$ has a small (negative) drift $\lambda - s_1$, where λ is the arrival rate and s_1 is the departure rate. In other words, the process s_1 is in fast time-scale outside the green region and $s_1 + s_2$ is in slow time-scale within the green region, it implies that the system would live in the green region with a high probability.

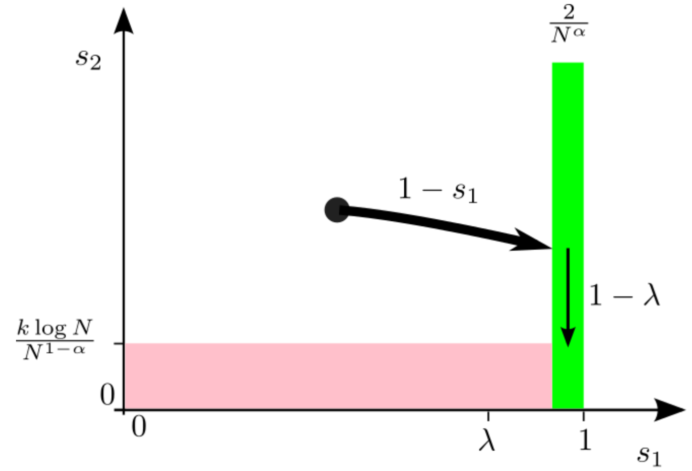

Figure 2.2: Illustration of State Space Collapse

The SSC argument above is justified in Lemma 2, where given a set Π_2 of load balancing (e.g. JSQ), either s_1 is larger than $1 - \frac{1}{2N^\alpha}$ (most of servers are busy) or $\sum_{i=1}^b s_i$ is less than $\frac{k \log N}{N^{1-\alpha}}$ (the number of waiting jobs are small). This is reasonable for JSQ-like load balancing because s_2 will not build up if idle servers exist. Lemma 2 is proved by Lyapunov drift analysis of Lyapunov function $V(s) = \min\left\{\sum_{i=2}^b s_i - \frac{k \log N}{N^{1-\alpha}}, 1 - s_1\right\}$. The details can be found in the Appendix B.3.

Lemma 2. *For any load balancing in* Π_2, *we have*

$$\Pr\left(\min\left\{\sum_{i=2}^b S_i - \frac{k \log N}{N^{1-\alpha}}, 1 - S_1\right\} \geq \frac{1}{2N^\alpha}\right) \leq e^{-\frac{(k-1)\log N}{16b}}.$$

Given SSC statement in Lemma 2, we study the system in the green region, where most of servers are busy (s_1 is close to 1), and we are able to bound the key term of (2.14) in generator difference in Lemma 1. In fact, SSC observation motivates us to approximate the original system with a simple system, where arrival rate λ and departure rate $\lambda + \delta$ with $\delta = \frac{1}{N^\alpha}$ i.e.,

$$\dot{x} = \lambda - (\lambda + \delta) = -\frac{1}{N^\alpha}.$$

We would expect that in the green region, the original system behaves "close" to the simple system, which implies the steady-state metric of these two systems is also "close" and the simple system is a good approximation.

In summary, Stein's method enables us to couple an approximated simple system with the original system associated with certain metrics. To established the metric of the original system (at steady-state), we need to study the gradient bound terms in (2.15) and (2.16) and SSC term in (2.14).

Chapter 3

STEADY-STATE ANALYSIS OF LOAD BALANCING IN SUB-HALFIN-WHITT REGIME

This chapter studies the steady-state performance of load balancing algorithms in many-server systems. We consider a system with N identical servers with buffer size $b-1$ such that $b = O\left(\sqrt{\log N}\right)$, in other words, each server can hold at most b jobs, one job in service and $b-1$ jobs in buffer. We assume jobs arrive according to a Poisson process with rate λN, where $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$ and have i.i.d. exponential service times with mean one. When a job arrives, the load balancer immediately routes the job to one of the servers. If the server's buffer is full, the job is discarded. We study a class of load balancing algorithms, which includes join-the-shortest-queue (JSQ), idle-one-first (I1F) Gupta and Walton (2019), join-the-idle-queue (JIQ) Lu *et al.* (2011); Stolyar (2015a) and power-of-d-choices (Pod) with $d \geq rN^{\alpha}\log N$ Mitzenmacher (1996); Vvedenskaya *et al.* (1996), and establish moment bounds on some function of the queue lengths. From the moment bounds, we show that under JSQ, I1F, JIQ, and Pod with $d \geq rN^{\alpha}\log N$, both the probability that a job is routed to a non-idle server and the expected waiting time per job are $O\left(\frac{b}{N^{r(0.5-\alpha)}}\right)$, where r is any positive integer such that $r \leq \frac{\log N}{72(b-1)^2}$.

Let S_i denote the fraction of servers with at least i jobs, including the one in service, at steady state. In this chapter, we prove that if a load balancing algorithm routes an incoming job to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$ when the fraction of busy servers is no more than $\eta = \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$, then the following bound holds for any positive integer $r \leq \frac{\log N}{72(b-1)^2}$,

$$E\left[\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{\bar{k}\log N}{\sqrt{N}}, 0\right\}\right)^r\right] \leq 10\left(\frac{5r(b-1)}{\sqrt{N}\log N}\right)^r, \ \bar{k} = 1 + \frac{1}{2(b-1)}. \tag{3.1}$$

This result implies that (i)

$$E\left[\sum_{i=1}^{b} S_i\right] \leq \lambda + \frac{11\lambda b + 11}{N^{r(0.5-\alpha)}}, \tag{3.2}$$

i.e, the expected queue length per server exceeds λ by at most $\frac{11\lambda b+11}{N^{r(0.5-\alpha)}}$ and (ii) under JSQ, I1F, JIQ and Pod ($d \geq rN^{\alpha}\log N$), the stationary probability that an incoming job is routed to a non-idle server is asymptotically zero (as $N \to \infty$), which will be proved in Corollary 1.

From the best of our knowledge, there are only a few papers that deal with the steady-state analysis of many-server systems with distributed queues Braverman (2018); Banerjee and Mukherjee (2019); Liu and Ying (2018). Braverman (2018); Banerjee and Mukherjee (2019) analyze the steady-state distribution of JSQ in the Halfin-Whitt regime and Liu and Ying (2018) studies the Pod with $\alpha < 1/6$. This chapter complements Braverman (2018); Banerjee and Mukherjee (2019); Liu and Ying (2018), as it applies to a class of load balancing algorithms and to any sub-Halfin-Whitt regime.

Similar to Braverman (2018); Liu and Ying (2018), the result of this chapter is proved using the mean-field approximation (fluid-limit approximation) based on Stein's method. The execution of Stein's method in this chapter, however, is quite different from Braverman (2018); Liu and Ying (2018).

In our proof, we consider a simple fluid system with arrival rate λ and departure rate $\lambda + \delta$ with $\delta = \frac{\log N}{\sqrt{N}}$ such that

$$\dot{x} = -\frac{\log N}{\sqrt{N}}. \tag{3.3}$$

x can be viewed as a fluid approximation of the normalized queue length $\sum_{i=1}^{b} S_i$ and $\dot{x}$ is the derivative of x with respect to time t. The dynamic of this fluid system (3.3) is a good approximation of the generator of the stochastic system only when the normalized service rate of the stochastic system is close to $\lambda + \frac{\log N}{\sqrt{N}}$, i.e. when $S_1 \approx \lambda + \frac{\log N}{\sqrt{N}}$. Our analysis consider three regimes of the state space:

- Regime 1: S_1 is close to $\lambda + \frac{\log N}{\sqrt{N}}$. In this regime, the simple fluid system can approximate the generator of the stochastic system. Via Stein's method, we can quantify the approximation error.

- Regime 2: $\sum_{i=2}^{b} S_i \leq \frac{c \log N}{\sqrt{N}}$ for some $c > 0$. Since $S_1 \leq 1$, in this regime, the normalized queue length is close to one.

- Regime 3: The state is not in regime 1 or regime 2. In this case, we apply the tail bound in Bertsimas *et al.* (2001) to prove that the probability it occurs is small and negligible as N increases. This is equivalent to the state-space-collapse argument, which shows that at steady-state, the system "lives" in a lower-dimensional space instead of in the full state space.

Pioneered in Stolyar (2015b) (called drift-based-fluid-limits (DFL) method) for fluid-limit analysis and in Braverman *et al.* (2016); Braverman and Dai (2017) for steady-state diffusion approximation, the power of Stein's method for steady-state approximations has been recognized in a number of recent papers Stolyar (2015b); Braverman *et al.* (2016); Ying (2016); Braverman and Dai (2017); Ying (2017); Gast (2017); Gast and Van Houdt (2018); Braverman (2018).

The surprising part of our analysis is that the simple fluid system, which only "partailly" approximates the generator of the stochastic system, is sufficient for executing Stein's method when combing with the state-space-collapse. The advantage of using such a simple fluid system is that Stein's equation can be easily solved (in explicit forms), which is often the key difficulty of applying Stein's method for complex queueing systems.

Finally, we would like to comment that all proofs in this chapter are elementary. Therefore, this chapter is another an example that demonstrates the power of Stein's method for analyzing complex queueing systems with elementary probability methods.

3.1 Main Results

Let $S_i(t)$ denote the fraction of servers with at least i jobs at time $t \geq 0$ and $S \in \mathbb{S}$ be the random variables having the stationary distribution of $(S(t) : t \geq 0)$. Let $A_1(s)$ denote the probability that an incoming job is routed to a busy server when the system is in state $s \in \mathbb{S}$; i.e.

$$A_1(s) = \Pr\left(\text{an incoming job is routed to a busy server} \middle| \, S(t) = s\right).$$

Define a set of load balancing Π_1 to be

$$\Pi_1 = \left\{ \pi \mid \text{under load balancing } \pi, A_1(s) \leq \frac{1}{\sqrt{N}} \text{ for } s \text{ such that } s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}} \right\}.$$

Our first main result of this chapter is the following theorem with respect to the first-order moment. Here the first-order moment result is for the purpose of introduction to Stein's method and state space collapse framework.

Theorem 1. *Assume $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$ and $b \leq 1 + \frac{\sqrt{\log N}}{9}$. Given a load balancing in Π_1, then for any N such that $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$, the following bound holds*

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{\bar{k}\log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{50(b-1)}{\sqrt{N}\log N},$$

where $\bar{k} = 1 + \frac{1}{2(b-1)}$. □

Note the expectation in Theorem 1 is with respect to the stationary distribution of the CTMC $(S(t) : t \geq 0)$ according to the definition of S. The condition $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$ requires the following: for any given state s in which at least $\frac{1}{N^\alpha} - \frac{\bar{k}\log N}{\sqrt{N}}$ fraction of servers are idle, an incoming job should be routed to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$. Note $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{1}{N^\alpha} \geq \frac{4\bar{k}\log N}{\sqrt{N}}$, which guarantee that $\lambda + \frac{\bar{k}\log N}{\sqrt{N}} < 1$ and $\frac{1}{N^\alpha} > \frac{\bar{k}\log N}{\sqrt{N}}$. There are several well-known policies that satisfy this condition.

- **J**oin-the-Shortest-Queue (JSQ): JSQ routes an incoming job to the least loaded server in the system, so $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$.

- **I**dle-One-First (I1F): I1F routes an incoming job to an idle server if available and else to a server with one job if available. Otherwise, the job is routed to a randomly selected server. Therefore, $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$.

- **J**oin-the-Idle-Queue (JIQ): JIQ routes an incoming job to an idle server if possible and otherwise, routes the job to a server chosen uniformly at random. Therefore, $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$.

- **P**ower-of-d-Choices (Pod): Pod samples d servers uniformly at random and dispatches the job to the least loaded server among the d servers. Ties are broken uniformly at random. Given $d \geq N^\alpha \log N$, $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$.

3.2 Proof of Theorem 1

In this section, we present the proof of our main theorem. As modularized in Chapter 2, we study gradient bounds and state space collapse (SSC).

Let $\eta = \lambda + \frac{k \log N}{\sqrt{N}}$ and $\delta = \frac{\log N}{\sqrt{N}}$ in Lemma 1, we have

$$
\begin{aligned}
& E\left[h_k\left(\sum_{i=1}^b S_i\right)\right] \\
& = E\left[g'\left(\sum_{i=1}^b S_i\right)\left(\lambda A_b(S) - \lambda - \frac{\log N}{\sqrt{N}} + S_1\right)\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right] \qquad (3.4) \\
& + E\left[\left(g'\left(\sum_{i=1}^b S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1 - A_b(S))g'(\xi) + S_1 g'(\tilde{\xi})\right)\mathbb{I}_{\eta - \frac{1}{N} \le \sum_{i=1}^b S_i \le \eta + \frac{1}{N}}\right] \\
& \qquad (3.5) \\
& - E\left[\frac{1}{2N}\left(\lambda(1 - A_b(S))g''(\zeta) + S_1 g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right]. \qquad (3.6)
\end{aligned}
$$

Note in (3.4) and (3.6), we have random variables $\xi, \zeta \in \left(\sum_{i=1}^b S_i, \sum_{i=1}^b S_i + \frac{1}{N}\right)$ and $\tilde{\xi}, \tilde{\zeta} \in \left(\sum_{i=1}^b S_i - \frac{1}{N}, \sum_{i=1}^b S_i\right)$ whose values depend on $\sum_{i=1}^b S_i$.

Next, we study g' and g'' to bound the terms (3.5) and (3.6), and SSC to bound the term (3.4).

3.2.1 Gradient Bounds

Let $\eta = \lambda + \frac{k \log N}{\sqrt{N}}$ and $\delta = \frac{\log N}{\sqrt{N}}$ in Lemma 21 and Lemma 22. We have the following two lemmas.

Lemma 3. *For any* $x \in \left[\lambda + \frac{k \log N}{\sqrt{N}} - \frac{2}{N}, \lambda + \frac{k \log N}{\sqrt{N}} + \frac{2}{N}\right]$, *we have*

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

Lemma 4. *For* $x > \lambda + \frac{k \log N}{\sqrt{N}}$, *we have*

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

Based on Lemma 3 and Lemma 4, we have the following lemma.

Lemma 5.

$$(3.5) + (3.6) \le \frac{5}{\sqrt{N} \log N}$$

Proof. Based on Lemma 3, we bound the term (3.5)

$$E\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1 - A_b(S))g'(\xi) + S_1 g'(\tilde{\xi})\right)\mathbb{I}_{\eta-\frac{1}{N}\leq\sum_{i=1}^{b} S_i\leq\eta+\frac{1}{N}}\right]$$
$$\leq\left(\lambda + \frac{\log N}{\sqrt{N}} + 1\right)\frac{2}{\sqrt{N}\log N},$$

where $\lambda + \frac{\log N}{\sqrt{N}} \leq 1$ in the first inequality according to the assumption that $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ in Theorem 1.

Based on Lemma 4, we bound the term (3.6)

$$\begin{aligned}&-E\left[\frac{1}{2N}\left(\lambda(1 - A_b(S))g''(\zeta) + S_1 g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^{b} S_i>\eta+\frac{1}{N}}\right]\\ \leq &E\left[\frac{1}{2N}\left(\lambda|g''(\zeta)| + S_1|g''(\tilde{\zeta})|\right)\mathbb{I}_{\sum_{i=1}^{b} S_i>\eta+\frac{1}{N}}\right] \leq \frac{1+\frac{1}{N}}{\sqrt{N}\log N}.\end{aligned}$$

These two terms collectively prove Lemma 5. □

3.2.2 State Space Collapse (SSC)

In this section, we study the SSC term in (3.4). As mentioned in Chapter 2, we proved SSC by Lyapunov drift analysis. Define

$$V(s) = \min\left\{\sum_{i=2}^{b} s_i, \lambda + \frac{k\log N}{\sqrt{N}} - s_1\right\}, \tag{3.7}$$

where $\bar{k} - \frac{r}{\sqrt{N}\log N} \leq k \leq \bar{k}$. We have the following lemma on state space collapse.

Lemma 6. *Given the Lyapunov function defined in* (3.7) *and denote* $\tilde{k} = 1 + \frac{1}{4(b-1)}$, *we have*

$$\Pr\left(V(S) \geq \frac{\tilde{k}\log N}{\sqrt{N}}\right) \leq e^{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}}.$$

Based on Lemma 6, we establish an upper bound on (3.4) by splitting two regions: Ω and its complementary $\bar{\Omega}$, where

$$\Omega = \left\{s \mid V(s) \leq \frac{\tilde{k}\log N}{\sqrt{N}}\right\},$$

and have the following lemma.

Lemma 7.

$$(3.4) \leq \left(1 - \frac{1}{5(b-1)}\right) E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]$$
$$+ \frac{b\sqrt{N}}{\log N} e^{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}}.$$

The proofs of two lemmas are in Appendix B.2

3.2.3 Proving Theorem 1

Based on Lemma 5 and Lemma 7, we have

$$\begin{aligned}
&E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \\
\leq &\left(1 - \frac{1}{5(b-1)}\right) E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \\
&+ \frac{b\sqrt{N}}{\log N} e^{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}} + \frac{5}{\sqrt{N}\log N}, \\
\leq &\left(1 - \frac{1}{5(b-1)}\right) E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] + \frac{10}{\sqrt{N}\log N},
\end{aligned}$$

where the last inequality holds because $b \leq 1 + \frac{\sqrt{\log N}}{9}$ implies

$$\frac{b\sqrt{N}}{\log N} e^{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}} \leq \frac{5}{\sqrt{N}\log N}.$$

Therefore, we have

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{50(b-1)}{\sqrt{N}\log N},$$

which proves Theorem 1.

So far, we already see the application of Stein's method and SSC in establishing Theorem 1. In fact, by using an iterative refined procedure in Chapter 4, we are able to obtain a high-order moment bound in the following theorem (here we only state the theorem and the proof is clear after introducing the iterative refined procedure in Chapter 4), which helps establish "zero-delay" results in Corollary 1.

Theorem 2. *Assume* $\lambda = 1 - N^{-\alpha}$ *for* $0 < \alpha < 0.5$ *and* $b \leq 1 + \frac{\sqrt{\log N}}{9}$. *Given a load balancing in* Π_1, *then for any integers* N *and* r *such that* $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ *and* $1 \leq r \leq \frac{\log N}{72(b-1)^2}$, *the following bound holds*

$$E\left[\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{\bar{k}\log N}{\sqrt{N}}, 0\right\}\right)^r\right] \leq 10\left(\frac{5r(b-1)}{\sqrt{N}\log N}\right)^r,$$

where $\bar{k} = 1 + \frac{1}{2(b-1)}$. □

Note JSQ, JIQ, I1F and Pod with $d \geq rN^{\alpha}\log N$ are all in Π_1. A direct consequence of Theorem 2 is asymptotic zero waiting ($N \to \infty$) at steady-state. Let $\mathcal{W}$ denote the event that an incoming job is routed to a busy server, and $p_{\mathcal{W}}$ denote the probability of this event at steady-state. Let $\mathcal{B}$ denote the event that an incoming job is blocked (discarded) and $p_{\mathcal{B}}$ denote the probability of this event at steady-state. Note the $\mathcal{B} \subseteq \mathcal{W}$ because an incoming job is blocked when being routed to a busy server with b jobs. Furthermore, let W denote the waiting time of jobs, which are not blocked, at steady-state. We have the following results based on the main theorem.

Corollary 1. *Assume* $\lambda = 1 - N^{-\alpha}$ *for* $0 < \alpha < 0.5$ *and* $b \leq 1 + \frac{\sqrt{\log N}}{9}$. *Given a load balancing in* Π_1, *then the following results hold for any integers* N *and* r *such that* $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ *and* $1 \leq r \leq \frac{\log N}{72(b-1)^2}$:

$$\textit{Waiting Time per Job:} \quad E[W] \leq \frac{11b}{N^{r(0.5-\alpha)}} \tag{3.8}$$

$$\textit{Waiting Probability:} \quad p_{\mathcal{W}} \leq \frac{11}{N^{r(0.5-\alpha)}} \tag{3.9}$$

$$\textit{Fraction of Busy Servers:} \quad \lambda - \frac{11}{N^{r(0.5-\alpha)}} \leq E[S_1] \leq \lambda \tag{3.10}$$

$$\textit{Number of Buffered Jobs per Server:} \quad E\left[\sum_{i=2}^{b} S_i\right] \leq \frac{11\lambda b + 11}{N^{r(0.5-\alpha)}}. \tag{3.11}$$

The proof of this lemma is an application of the Markov inequality and Little's Law, which can be found in Section 3.3. We remark that the corollary above requires

$A_1(s) \leq \frac{1}{N^{0.5r}}$ for any $s \in \mathbb{S}$ such that $s_1 \leq \lambda + \frac{k \log N}{\sqrt{N}}$, which is more restrictive than the assumption in the theorem which only requires $A_1(s) \leq 1/\sqrt{N}$ for the same s. However, it is easy to verify that JSQ, I1F, JIQ and Pod with $d \geq rN^{\alpha} \log N$ also satisfy this condition. We further remark that the probability of waiting and the expected waiting time are both $O\left(\frac{b}{N^{r(0.5-\alpha)}}\right)$. Under the assumption that $b = O\left(\sqrt{\log N}\right)$, for any positive integer r, we can find a sufficiently large N such that r satisfies the condition in the corollary. The significance of this is that it implies that the waiting probability and the mean waiting time decay faster than any polynomial function of $1/N$ in the sub-Halfin-Whitt regime. Furthermore, from (3.11), we have

$$E\left[\sum_{i=2}^{b} NS_i\right] \leq \frac{11\lambda b + 11}{N^{r(0.5-\alpha)-1}}.$$

Note that $\sum_{i=2}^{b} NS_i$ is the total number of jobs in the buffers at steady state, so our result shows that for sufficiently large N, not only the expected number of buffered jobs per server is almost zero, but also the total number of buffered jobs in all N servers is almost zero.

3.3 Proof of Corollary 1

Based on the moment bound in Theorem 2, we study waiting probability $p_{\mathcal{W}}$, waiting time $E[W]$, $E[S_1]$ and $E[\sum_{i=2}^{b} S_i]$ for JSQ, I1F, and JIQ. The analysis for

Pod is similar and will be provided later. We begin with the waiting probability $p_{\mathcal{W}}$

$$\begin{aligned}
p_{\mathcal{W}} =& \Pr\left(S_1 = 1\right) \leq \Pr\left(\sum_{i=1}^{b} S_i \geq 1\right) \\
\leq& \Pr\left(h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right) \geq \left(\frac{1}{N^{\alpha}} - \frac{\bar{k}\log N}{\sqrt{N}}\right)_{\bar{k}}^{r}\right) \\
\leq& \frac{E\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{1}{N^{\alpha}} - \frac{\bar{k}\log N}{\sqrt{N}}\right)^{r}} && (3.12) \\
\leq& \frac{E\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{1}{2N^{\alpha}}\right)^{r}} && (3.13) \\
\leq& 10\left(\frac{10r(b-1)}{N^{0.5-\alpha}\log N}\right)^{r} && (3.14) \\
\leq& \frac{10}{N^{r(0.5-\alpha)}} && (3.15)
\end{aligned}$$

where (3.12) is from Markov's inequality, (3.13) holds because $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{\bar{k}\log N}{\sqrt{N}} \leq \frac{1}{2N^{\alpha}}$, (3.14) holds by substituting Theorem 2, and (3.15) holds because $r \leq \frac{\log N}{72(b-1)^2}$ implies $\log N \geq 10r(b-1)$.

From $p_{\mathcal{W}}$, we can obtain an upper bound of $E[W]$:

$$E[W] = E[W|\text{ a job routed to busy servers}] \times p_{\mathcal{W}} \leq bp_{\mathcal{W}}$$

where the last inequality holds because the expected waiting time for a job routed to a busy server is at most $b-1$.

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

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

Therefore, we have

$$\begin{aligned}
E\left[\sum_{i=1}^{b} S_i\right] =& \lambda\left(1-p_{\mathcal{B}}\right)\left(E[W]+1\right) \leq \lambda E[W] + \lambda \\
\leq& \lambda b \cdot p_{\mathcal{W}} + \lambda \leq \frac{10\lambda b}{N^{r(0.5-\alpha)}} + \lambda.
\end{aligned}$$

Further, according to the work conservation law, we have the following lower bound on $E[S_1]$

$$E[S_1] = \lambda(1-p_{\mathcal{B}}) \geq \lambda(1-p_{\mathcal{W}}) \geq \lambda - \frac{10}{N^{r(0.5-\alpha)}}$$

which yields an upper bound on $E\left[\sum_{i=2}^{b} S_i\right]$:

$$E\left[\sum_{i=2}^{b} S_i\right] \leq \frac{10\lambda b + 10}{N^{r(0.5-\alpha)}}.$$

The analysis for Pod$d$ with $d \geq rN^{\alpha}\log N$ is similar, except the waiting probability $p_{\mathcal{W}}$ in the first step becomes

$$\begin{aligned}
p_{\mathcal{W}} =& \Pr\left(\mathcal{W}\middle| S_1 \leq 1-\frac{1}{2N^{\alpha}}\right)\Pr\left(S_1 \leq 1-\frac{1}{2N^{\alpha}}\right) \\
&+ \Pr\left(\mathcal{W}\middle| S_1 > 1-\frac{1}{2N^{\alpha}}\right)\Pr\left(S_1 > 1-\frac{1}{2N^{\alpha}}\right) \\
\leq& \Pr\left(\mathcal{W}\middle| S_1 \leq 1-\frac{1}{2N^{\alpha}}\right) + \Pr\left(S_1 > 1-\frac{1}{2N^{\alpha}}\right) \\
\leq& \left(1-\frac{1}{2N^{\alpha}}\right)^{rN^{\alpha}\log N} + \Pr\left(\sum_{i=1}^{b} S_i > 1-\frac{1}{2N^{\alpha}}\right) \\
\leq& N^{-\frac{r}{2}} + \Pr\left(h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right) \geq \left(\frac{1}{2N^{\alpha}} - \frac{\bar{k}\log N}{\sqrt{N}}\right)^{r}\right) && (3.16) \\
\leq& \frac{1}{N^{0.5r}} + \frac{E\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{1}{2N^{\alpha}} - \frac{\bar{k}\log N}{\sqrt{N}}\right)^{r}} && (3.17) \\
\leq& \frac{1}{N^{0.5r}} + \frac{E\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{1}{4N^{\alpha}}\right)^{r}} && (3.18) \\
\leq& \frac{1}{N^{0.5r}} + 10\left(\frac{20r(b-1)}{N^{0.5-\alpha}\log N}\right)^{r} && (3.19) \\
\leq& \frac{1}{N^{0.5r}} + \frac{10}{N^{r(0.5-\alpha)}} && (3.20) \\
\leq& \frac{11}{N^{r(0.5-\alpha)}} && (3.21)
\end{aligned}$$

where (3.16) holds because $(1-\frac{1}{x})^x \leq \frac{1}{e}$ for $x \geq 1$, (3.17) is a result of the Markov inequality; (3.18) holds because $N \geq \left(4\bar{k}\log N\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{1}{4N^{\alpha}} \geq \frac{\bar{k}\log N}{\sqrt{N}}$; (3.19) holds by substituting Theorem 2; (3.21) holds because $r \leq \frac{\log N}{72(b-1)^2}$ implies $\log N \geq 20r(b-1)$. The remaining analysis to obtain $E[W]$, $E[S_1]$ and $E[\sum_{i=1}^{b} S_i]$ are the same as analysis in JSQ.

3.4 Summary

In this chapter, we studied the steady-state performance of load balancing systems in the Sub-Halfin-Whitt regime ($\alpha < 0.5$). We showcase Stein's method and SSC are powerful tools to obtain the upper bound on a distance function of total queue length and we established a set of load balancing algorithms, where waiting probability and waiting time are asymptotic zero.

This chapter studied the Sub-Halfin-Whitt regime ($\alpha < 0.5$), one interesting extension is to consider a "heavier" traffic regimes where $0.5 \leq \alpha < 1$. In such a regime, the state space collapse result in this chapter does not hold. It would require a different fluid model and a different state-space collapse analysis and we study this regime in Chapter 4.

Chapter 4

STEADY-STATE ANALYSIS OF LOAD BALANCING IN BEYOND-HALFIN-WHITT REGIME

In Chapter 3, we already have "zero-delay" load balancing in the Sub-Halfin-Whitt regime ($0 < \alpha < 0.5$). This chapter studies the steady-state performance of load balancing in the Beyond-Halfin-Whitt regime for $0.5 \leq \alpha < 1$, and we have several main results:

- We first obtained a high-order moment bound at steady-state on a distant function of total queue length (per server) under a set Π_2 of load balancing algorithms (including JSQ, JIQ, I1F and Pod).

- We then established under any load balancing in Π_2, the waiting probability and the expected waiting time is $O\left(\frac{\log N}{N^{1-\alpha}}\right)$, which approaches to zero asymptotically as N increases.

- We also proved under any load balancing in $\widetilde{\Pi}_2$, only busy servers and servers with only exactly two jobs exist in the system. Interestingly and surprisingly, the result coincides with Eschenfeldt and Gamarnik (2018) in Halfin-Whitt regime ($\alpha = 0.5$), which suggests, the proper scaled version of idle servers and servers with exactly two jobs are very likely to converge into a two-dimensional stochastic process as in Eschenfeldt and Gamarnik (2018).

We use Fig. 1.1 as in Fig. 4.1 to explain our main contribution. Our results show Ns_2 is $O(N^\alpha \log N)$ at steady-state for $0.5 \leq \alpha < 1$, which is blue line; Braverman (2018) shown that Ns_2 is $O(\sqrt{N})$ for Halfin-Whitt regime $\alpha = 0.5$ at steady state,

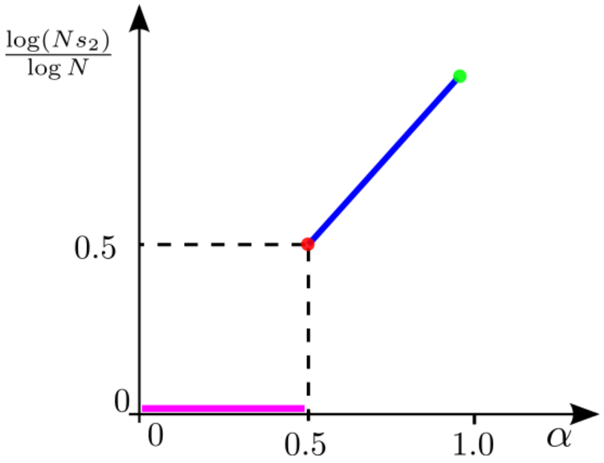

Figure 4.1: Illustration of Our Contribution and Related Work.

which is red dot; Gupta and Walton (2019) shown that Ns_2 is $O(N)$ for $\alpha = 1$ at diffusion level, which is green dot; Chapter 3 shown that Ns_2 can be $O(1/N)$ for $0 < \alpha < 0.5$ at steady state, which is purple line; In Fig. 4.1, we observed an interesting phase transition phenomenon at $\alpha = 0.5$, where Ns_2 vanishes for $\alpha < 0.5$ and scaled with N^α for $0.5 \leq \alpha \leq 1$. The intuitive explain is as follows: we approximate load balancing system to be M/M/1 system with arrival rate λN and service rate N, where the number of "waiting" jobs in M/M/1 is $\lambda/(1-\lambda) = O(N^\alpha)$. In load balancing systems, we compare the number of waiting jobs $O(N^\alpha)$ with the number of "idle" servers $N(1-\lambda) = O(N^{1-\alpha})$ when $0 < \alpha < 1$:

- For $\alpha < 0.5$, we have $O(N^\alpha) \ll O(N^{1-\alpha})$, and "waiting" jobs are close to zero;
- For $\alpha = 0.5$, we have $O(N^\alpha) = O(N^{1-\alpha})$, and "waiting" jobs are $O(\sqrt{N})$;
- For $0.5 < \alpha < 1$, we have $O(N^\alpha) \gg O(N^{1-\alpha})$, and "waiting" jobs are $O(N^\alpha)$.

4.1 Main Results

Let $S_i(t)$ denote the fraction of servers with at least i jobs at time $t \geq 0$ and $S \in \mathbb{S}$ be the random variables having the stationary distribution of $(S(t) : t \geq 0)$.

Let $A_1(s)$ denote the probability that an incoming job is routed to a busy server when the system is in state $s \in \mathbb{S}$; i.e.

$$A_1(s) = \Pr\left(\text{an incoming job is routed to a busy server} \middle| S(t) = s\right).$$

Define a set of load balancing Π_2 to be

$$\Pi_2 = \left\{ \pi \mid \text{under load balancing } \pi, A_1(s) \leq \frac{1}{\sqrt{N}} \text{ for } s \text{ such that } s_1 \leq 1 - \frac{1}{4N^\alpha} \right\}.$$

A load balancing algorithm in Π implies that for any given state s in which at least $\frac{1}{4N^\alpha}$ fraction of servers are idle, an incoming job should be routed to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$. There are several well-known algorithms that satisfy this condition.

- **J**oin-the-Shortest-Queue (JSQ): JSQ routes an incoming job to the least loaded server in the system, so $A_1(s) = 0$ when $s_1 \leq 1 - \frac{1}{4N^\alpha}$.

- **I**dle-One-First (I1F): I1F routes an incoming job to an idle server if available and else to a server with one job if available. Otherwise, the job is routed to a randomly selected server. Therefore, $A_1(s) = 0$ when $s_1 \leq 1 - \frac{1}{4N^\alpha}$.

- **P**ower-of-d-Choices (Pod): Pod samples d servers uniformly at random and dispatches the job to the least loaded server among the d servers. Ties are broken uniformly at random. Given $d \geq N^\alpha \log^2 N$, $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq 1 - \frac{1}{4N^\alpha}$.

- **J**oin-the-Idle-Queue (JIQ): JIQ routes an incoming job to an idle server if possible and otherwise, routes the job to a server chosen uniformly at random. Therefore, $A_1(s) = 0$ when $s_1 \leq 1 - \frac{1}{4N^\alpha}$.

We first have the following moment bounds which are instrumental for establishing the main results of this paper.

Theorem 3. *Assume $\lambda = 1 - N^{-\alpha}$ for $0.5 \leq \alpha < 1$ and buffer size b. For any load balancing algorithms in Π_2, the following bound holds at steady-state*

$$E\left[\left(\max\left\{\sum_{i=1}^{b} S_i - 1 - \frac{\bar{k}\log N}{N^{1-\alpha}}, 0\right\}\right)^r\right] \leq 10\left(\frac{2r}{N^{1-\alpha}}\right)^r,$$

where r is a positive integer and $\bar{k} = 32rb + 1$.

Note the expectation in Theorem 3 is with respect to the stationary distribution of the CTMC $(S(t) : t \geq 0)$ according to the definition of S. Based on Theorem 3, we have the *universal* scaling results and asymptotic zero waiting results in Corollary 1. These results hold for load balancing algorithms that satisfy additional conditions, defined below. Define a set of load balancing $\widetilde{\Pi}_2$ to be

$$\widetilde{\Pi}_2 = \left\{\pi \in \Pi_2, \middle| A_2(s) \leq 10\left(\frac{2r}{N^{1-\alpha}}\right)^r \forall s \in \mathbb{S} \text{ such that } s_2 \leq 0.95, \right.$$
$$\left. \text{and } A_b(s) \leq s_b, \forall s \in \mathbb{S}\right\}.$$

The additional conditions require: (i) when at least 5% servers have one job or less, the probability an incoming job is routed to a server with at least two jobs should be no more than $10\left(\frac{2r}{N^{1-\alpha}}\right)^r$, and (ii) given state s, the probability that a job is dropped because of being routed to a server with full buffer is is upper bounded by that under a random routing algorithm, which is s_b. It is easy to see that JSQ, I1F and Pod with $d \geq N^\alpha \log^2 N$ are in $\widetilde{\Pi}_2$, but JIQ is not.

To establish the universal scaling results, in Corollary 2, we first show that almost no server has more than two jobs under a load balancing algorithm in $\widetilde{\Pi}_2$. Though JIQ is not in $\widetilde{\Pi}_2$, a weaker result is presented.

Corollary 2. *Assume $\lambda = 1 - N^{-\alpha}$ for $0.5 \leq \alpha < 1$. The following results hold for any N such that $\frac{N^{1-\alpha}}{k\log N} \geq 5$,*

- *Under any load balancing algorithm in $\widetilde{\Pi}_2$,*

$$E[S_3] \leq 20\left(\frac{3r}{N^{1-\alpha}}\right)^r.$$

- *Under JIQ,*

$$E[S_3] \leq \frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{16r}{N^{\frac{r(1-\alpha)}{r+1}}}.$$

□

.

Next, we analyze the waiting time, waiting probability for algorithms in $\widetilde{\Pi}_2$, and the steady-state queues. Let $\mathcal{W}$ denote the event that an incoming job is routed to a busy server, and $p_{\mathcal{W}}$ denote the probability of this event at steady-state. Let $\mathcal{B}$ denote the event that an incoming job is blocked (discarded) and $p_{\mathcal{B}}$ denote the probability of this event at steady-state. Note the $\mathcal{B} \subseteq \mathcal{W}$ because an incoming job is blocked when being routed to a busy server with b jobs. Furthermore, let W denote the steady-state waiting time of those jobs that are not blocked.

Corollary 3. *Assume $\lambda = 1 - N^{-\alpha}$ for $0.5 \leq \alpha < 1$. Given any positive constant r, the following results hold for a sufficiently large N*

- *Under load balancing algorithm in $\widetilde{\Pi}_2$ and assume $N^{1-\alpha} \geq 3(40)^{\frac{r}{2}}r$, we have*

$$E\left[W\right] \leq \frac{4\bar{k}\log N}{N^{1-\alpha}},$$

and

$$p_{\mathcal{W}} \leq 20\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} + \frac{2\bar{k}\log N}{N^{1-\alpha}}.$$

We furthermore have

$$\lambda N - 10N\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} \leq E\left[NS_1\right] \leq \lambda N,$$

and

$$E\left[NS_2\right] \leq 10N\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} + 2\bar{k}N^{\alpha}\log N = O(N^{\alpha}\log N).$$

- *Consider JIQ and assume* $\log N \geq \frac{5b(2r)^r}{\bar{k}}$*. We have*

$$E\left[W\right] \leq \frac{7\bar{k}\log N}{N^{\frac{r(1-\alpha)}{r+1}}},$$

and

$$p_{\mathcal{W}} \leq \frac{12\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} + \frac{2\bar{k}\log N}{N^{1-\alpha}}.$$

We furthermore have

$$\lambda - \frac{6\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} \leq E\left[S_1\right] \leq \lambda,$$

and

$$E\left[\sum_{i=2}^{b} S_i\right] \leq \frac{6\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} + \frac{2\bar{k}\log N}{N^{1-\alpha}}.$$

□

In the M/M/N system where a centralized queue is maintained for complete resource pooling, the average waiting time per job is $O\left(\frac{1}{N^{1-\alpha}}\right)$. In load balancing systems, Corollary 1 suggests the waiting time to be $O\left(\frac{\bar{k}\log N}{N^{1-\alpha}}\right)$. Therefore, the expected waiting of a load balancing algorithms in $\widetilde{\Pi}_2$ is close to that in the M/M/N system when N is large. Therefore, load balancing algorithms in $\widetilde{\Pi}_2$ have near optimal delay performance since the mean waiting time of the M/M/N system is a lower bound on that of any many-server systems with distributed queues. We conjecture that the average waiting time of load balancing algorithms in $\widetilde{\Pi}_2$ is $\Theta\left(\frac{1}{N^{1-\alpha}}\right)$ as in the M/M/N system. The additional term $\bar{k}\log N$, howerver, is needed in establishing a state-space-collapse result due to technical reasons.

4.2 Proof of Theorem 3

In this section, we present the proof of Theorem 3. As modularized in Chapter 2, we study gradient bounds and state space collapse (SSC).

Let $\eta = 1 + \frac{k \log N}{N^{1-\alpha}}$ and $\delta = \frac{1}{N^\alpha}$ in Lemma 1, where $\bar{k} - \frac{r}{N^\alpha \log N} \leq k \leq \bar{k}$. The generator difference is summarized in the following lemma.

Lemma 8.

$$
\begin{aligned}
&E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right]\\
=&E\left[g'\left(\sum_{i=1}^b S_i\right)\left(\lambda A_b(S) - \lambda - \frac{1}{N^\alpha} + S_1\right)\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right] \qquad (4.1)\\
&+ E\left[\left(g'\left(\sum_{i=1}^b S_i\right)\left(-\frac{1}{N^\alpha}\right) - \lambda(1 - A_b(S))g'(\xi) + S_1 g'(\tilde{\xi})\right)\mathbb{I}_{\eta - \frac{1}{N} \leq \sum_{i=1}^b S_i \leq \eta + \frac{1}{N}}\right]\\
&\qquad (4.2)\\
&- E\left[\frac{1}{2N}\left(\lambda(1 - A_b(S))g''(\zeta) + S_1 g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^b S_i > \eta + \frac{1}{N}}\right]. \qquad (4.3)
\end{aligned}
$$

Here $\xi, \zeta \in \left(\sum_{i=1}^b S_i, \sum_{i=1}^b S_i + \frac{1}{N}\right)$ *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* $\sum_{i=1}^b S_i$.

The following sections provide upper bounds on (4.1), (4.2) and (4.3).

4.2.1 Gradient Bounds

Let $\eta = 1 + \frac{k \log N}{N^{1-\alpha}}$ and $\delta = \frac{1}{N^\alpha}$ in Lemma 21 and Lemma 22. We have the following two lemmas.

Lemma 9. *For any* $x \in \left[1 + \frac{k \log N}{N^{1-\alpha}} - \frac{2}{N}, 1 + \frac{k \log N}{N^{1-\alpha}} + \frac{2}{N}\right]$, *we have*

$$|g'(x)| \leq \frac{2^r}{N^{r-0.5} \log N}.$$

Lemma 10. *For* $x > 1 + \frac{k \log N}{N^{1-\alpha}}$, *we have*

$$|g''(x)| \leq \frac{r\sqrt{N}}{\log N} h^{r-1}(x).$$

Based on Lemma 9 and Lemma 10, we bound the term (4.2) and (4.3), respectively, and have the following lemma.

Lemma 11.

$$(4.2) + (4.3) \leq \frac{2^{r+1}}{N^{r-\alpha}} + \frac{rE\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right]}{N^{1-\alpha}}.$$

Proof. Based on Lemma 9, the term (4.2) is bounded by

$$\begin{aligned}&E\left[\left(g'\left(\sum_{i=1}^b S_i\right)\left(-\frac{1}{N^\alpha}\right) - \lambda(1-A_b(S))g'(\xi) + S_1 g'(\tilde{\xi})\right)\mathbb{I}_{\eta-\frac{1}{N}\leq\sum_{i=1}^b S_i\leq\eta+\frac{1}{N}}\right]\\ \leq&\left(\lambda + \frac{1}{N^\alpha} + 1\right)\frac{2^r}{N^{r-\alpha}} \leq \frac{2^{r+1}}{N^{r-\alpha}};\end{aligned}$$

Based on Lemma 10, the term (4.3) is bounded by

$$\begin{aligned}&-E\left[\frac{1}{2N}\left(\lambda(1-A_b(S))g''(\zeta) + S_1 g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\right]\\ \leq&E\left[\frac{1}{2N}\left(\lambda|g''(\zeta)| + S_1|g''(\tilde{\zeta})|\right)\mathbb{I}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\right] \leq \frac{rE\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right]}{N^{1-\alpha}}.\end{aligned}$$

The two terms collectively prove Lemma 11. □

4.2.2 State Space Collapse

In this section, we study SSC term in (4.1). As discussed in Chapter 2, we proved state space collapse by Lyapunov drift analysis. Define

$$V(s) = \min\left\{\sum_{i=2}^b s_i - \frac{k\log N}{N^{1-\alpha}}, 1-s_1\right\}, \tag{4.4}$$

we have the following lemma on state space collapse.

Lemma 12. *For any load balancing in* Π_2*, we have*

$$\Pr\left(\min\left\{\sum_{i=2}^b S_i - \frac{k\log N}{N^{1-\alpha}}, 1-S_1\right\} \geq \frac{1}{2N^\alpha}\right) \leq e^{-\frac{(k-1)\log N}{16b}}.$$

According to Lemma 12, we split SSC term (4.1) into two regions, Ω and its complementary $\bar{\Omega}$, where

$$\Omega = \left\{s \;\middle|\; \min\left\{\sum_{i=2}^b s_i - \frac{k\log N}{N^{1-\alpha}}, 1-S_1\right\} \leq \frac{1}{2N^\alpha}\right\}.$$

For (4.1) in the region Ω, we have the key observation $1 - S_1 \leq \frac{1}{2N^\alpha}$; For (4.1) in the region $\bar{\Omega}$ (outside Ω), we apply the probability tail in Lemma 12; the upper bounds in the two regions collectively provide the following lemma.

Lemma 13.

$$(4.1) \leq \frac{1}{2}E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] + N^\alpha b^r e^{-\frac{(\bar{k}-1)\log N}{16b}}.$$

The proofs of Lemma 12 and Lemma 13 are in Appendix B.3.

4.2.3 Iterative Moment Bounds

From Lemma 11 and Lemma 13, we have the upper bound on $E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right]$ that

$$E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] \leq (4.2) + (4.3) + (4.1),$$

where $E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i\right)\right]$ in Lemma 13 gives an iterative relation. We specify the iterative relation between $E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right]$ and $E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i\right)\right]$ in the following lemma, which is used to prove Theorem 3.

Lemma 14. *Assume $\lambda = 1 - N^{-\alpha}$, $0.5 \leq \alpha < 1$. The following bound holds at steady-state for any positive integer r such that:*

$$E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] \leq \frac{2^{r+2}}{N^{r-\alpha}} + \frac{2r}{N^{1-\alpha}}E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right].$$

Proof. Given Lemma 11 and Lemma 13, we have

$$\begin{aligned} E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] \leq & \frac{1}{2}E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] + N^\alpha b^r e^{-\frac{(\bar{k}-1)\log N}{16b}} \\ & + \frac{2^{r+1}}{N^{r-\alpha}} + \frac{r}{N^{1-\alpha}}E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right] \\ \leq & \frac{1}{2}E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] + \frac{2^{r+1}+1}{N^{r-\alpha}} + \frac{r}{N^{1-\alpha}}E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right] \end{aligned} \tag{4.5}$$

where the second inequality holds because that

$$N^{\alpha} b^{r} e^{-\frac{(\bar{k}-1)\log N}{16b}} \leq \frac{1}{N^{r-\alpha}},$$

under the assumption of $\bar{k} = 1 + 32rb$. Finally, by moving $\frac{1}{2}E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right]$ in (4.5) to the left-hand side, we have

$$E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] \leq \frac{2^{r+2}+2}{N^{r-\alpha}} + \frac{2r}{N^{1-\alpha}} E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right].$$

□

4.2.4 Proving Theorem 3

Base on Lemma 14, we carefully expand the iteration and establish Theorem 3. Denote $w_r = \frac{2r}{N^{1-\alpha}}$ and $z_r = \frac{2^{r+2}+2}{N^{r-\alpha}}$ in Lemma 14, and we have

$$E\left[h_k^r\left(\sum_{i=1}^b S_i\right)\right] \leq w_r \cdot E\left[h_k^{r-1}\left(\sum_{i=1}^b S_i + \frac{1}{N}\right)\right] + z_r.$$

Expand $E\left[h^r\left(\sum_{i=1}^b S_i\right)\right]$ iteratively until $r = 1$ that

$$\begin{aligned} E\left[h^r\left(\sum_{i=1}^b S_i\right)\right] &\leq \prod_{j=1}^r w_j + \sum_{i=1}^{r-1} z_i \prod_{j=i+1}^r w_j + z_r \\ &\leq \prod_{j=1}^r w_j + r z_1 \prod_{j=2}^r w_j \\ &\leq (r+1) z_1 \prod_{j=2}^r w_j \\ &\leq (r+1) z_1 (w_r)^{r-1} \\ &\leq 10\left(\frac{2r}{N^{1-\alpha}}\right)^r \end{aligned}$$

where the second inequality holds because $z_i \prod_{j=i+1}^r w_j$ is decreasing for $2 \leq i \leq r$ that

$$\frac{z_i \prod_{j=i+1}^r w_j}{z_{i-1} \prod_{j=i}^r w_j} = \frac{z_i}{z_{i-1} w_i} = \frac{\frac{2^{i+2}+2}{N^{i-\alpha}}}{\frac{2^{i+1}+2}{N^{i-1-\alpha}} \frac{2i}{N^{1-\alpha}}} \leq \frac{1}{2iN^{\alpha}} \leq 1;$$

the third inequality holds because $w_1 = \frac{2}{N^{1-\alpha}} \leq z_1 = \frac{10}{N^{1-\alpha}}$ implies

$$\prod_{j=1}^{r} w_j \leq z_1 \prod_{j=2}^{r} w_j;$$

the forth inequality holds because w_r is increasing in r.

4.3 Proof of Corollary 3

We prove Corollary 1 by following the main steps: i) bound the blocking probability $p_{\mathcal{B}}$; ii) study the expected waiting time $E[W]$ based on $p_{\mathcal{B}}$; iii) study the waiting probability $p_{\mathcal{W}}$ based on $p_{\mathcal{B}}$ and $E[W]$.

4.3.1 Load Balancing Algorithms in $\widetilde{\Pi}_2$

Let $\delta_b = \sqrt{10}\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}}$, we study $p_{\mathcal{B}}$ by splitting into two regions:

$$\begin{aligned} p_{\mathcal{B}} =& \Pr\left(\mathcal{B} \left| S_b \leq \delta_b\right.\right) \Pr\left(S_b \leq \delta_b\right) \\ &+ \Pr\left(\mathcal{B} \left| S_b > \delta_b\right.\right) \Pr\left(S_b > \delta_b\right) \\ \leq& \Pr\left(\mathcal{B} \left| S_b \leq \delta_b\right.\right) + \Pr\left(S_b > \delta_b\right). \end{aligned}$$

For load balancing in $\widetilde{\Pi}_2$, we have

$$\begin{aligned} p_{\mathcal{B}} \leq& \delta_b + \Pr\left(S_b > \delta_b\right) \\ \leq& \delta_b + \Pr\left(S_3 > \delta_b\right) \\ \leq& \delta_b + \frac{E[S_3]}{\delta_b} \\ \leq& 10\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} \end{aligned}$$

where the first inequality holds because $A_b(s) \leq s_b$ for load balancing in $\widetilde{\Pi}_2$; the third inequality holds by Markov inequality; the last inequality holds because of the upper bound on $E[S_3]$ in $\widetilde{\Pi}_2$ in Corollary 2.

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

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

Therefore, the average waiting time is

$$\begin{aligned}
E[W] =& \frac{E\left[\sum_{i=1}^{b} S_i\right]}{\lambda(1-p_{\mathcal{B}})} - 1 \\
\leq& \frac{1 + \frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{20}{N^{1-\alpha}}}{\lambda(1-p_{\mathcal{B}})} - 1 \\
=& \frac{\frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{20}{N^{1-\alpha}} + \frac{1}{N^{\alpha}} + \lambda p_{\mathcal{B}}}{\lambda(1-p_{\mathcal{B}})} \\
\leq& \frac{3\bar{k}\log N}{\lambda(1-p_{\mathcal{B}})} \leq \frac{4\bar{k}\log N}{N^{1-\alpha}}
\end{aligned}$$

where the first inequality holds by letting $r = 1$ in Theorem 3 ; the last inequality holds because the upper bound of $p_{\mathcal{B}}$ for a large N such that $N^{1-\alpha} \geq 3(40)^{\frac{2}{r}} r$.

From the work conservation law, we have

$$E[S_1] = \lambda(1-p_{\mathcal{B}}),$$

which implies

$$\lambda - 10\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} \leq E[S_1] \leq \lambda.$$

The bound on $E[S_2]$ is established

$$E[S_2] \leq E\left[\sum_{i=2}^{b} S_i\right] \leq 10\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} + \frac{(\bar{k}+20)\log N}{N^{1-\alpha}},$$

by the fact

$$E\left[\sum_{i=1}^{b} S_i\right] \leq 1 + \frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{20}{N^{1-\alpha}}, \quad \bar{k} = 32rb.$$

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

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

where T_Q is the waiting time for the jobs waiting in the buffer. Since $E[T_Q]$ is lower bounded by one, we have

$$p_{\overline{\mathcal{W}}} \leq \frac{E\left[\sum_{i=2}^{b} S_i\right]}{\lambda}.$$

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

$$\begin{aligned} p_{\mathcal{W}} =& p_{\mathcal{B}} + p_{\overline{\mathcal{W}}} \\ \leq& p_{\mathcal{B}} + \frac{E\left[\sum_{i=2}^{b} S_i\right]}{\lambda} \\ \leq& 20\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{2}} + \frac{2\bar{k}\log N}{N^{1-\alpha}}. \end{aligned}$$

4.3.2 JIQ

For JIQ, it requires more steps to establish the upper bound on $p_{\mathcal{B}}$. Let $\epsilon = N^{-\frac{r(1-\alpha)}{r+1}}$ and $\delta_b = \frac{3\bar{k}\epsilon\log N}{b-1}$ (the reason to define the two quantities will become clear as going forward). We study $p_{\mathcal{B}}$ by splitting into two terms as above:

$$p_{\mathcal{B}} \leq \delta_b + \Pr\left(S_b > \delta_b\right).$$

Consider $\Pr\left(S_b > \delta_b\right)$ by splitting state space S according to state space collapse in Lemma 12 and the high-moment bound in Theorem 3 as follows

$$\begin{aligned}
\Pr\left(S_b > \delta_b\right) =& \Pr\left(S_b > \delta_b, V(S) \geq \frac{1}{2N^\alpha}\right) \\
&+ \Pr\left(S_b > \delta_b, h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) \geq \epsilon, V(S) < \frac{1}{2N^\alpha}\right) \\
&+ \Pr\left(S_b > \delta_b, h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) < \epsilon, V(S) < \frac{1}{2N^\alpha}\right) \\
\leq& \Pr\left(V(S) \geq \frac{1}{2N^\alpha}\right) + \Pr\left(h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) \geq \epsilon\right) \\
&+ \Pr\left(S_b > \delta_b, h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) < \epsilon, V(S) < \frac{1}{2N^\alpha}\right) \\
\leq& \frac{1}{N^{2r}} + \frac{10(2r)^r}{N^{\frac{r(1-\alpha)}{r+1}}}
\end{aligned}$$

where the last inequality holds because

- the first term yields from

$$\Pr\left(V(S) \geq \frac{1}{2N^\alpha}\right) \leq \frac{1}{N^{2r}},$$

according to Lemma 12.

- the second term yields from

$$\begin{aligned}
\Pr\left(h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) \geq \epsilon\right) =& \Pr\left(h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right) \geq \epsilon^r\right) \\
\leq& \frac{E\left[h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right)\right]}{\epsilon^r} \\
\leq& 10\left(\frac{2r}{\epsilon N^{1-\alpha}}\right)^r = 10\left(\frac{2r}{N^{\frac{1-\alpha}{r+1}}}\right)^r,
\end{aligned}$$

according to Theorem 3.

- The third term is 0 because

  $$\mathcal{Z} = \left\{ s \mid s_b > \delta_b, h_{\bar{k}}\left(\sum_{i=1}^{b} s_i\right) < \epsilon, V(s) < \frac{1}{2N^\alpha} \right\}$$

  is an empty set as we will show in the following.

  Since $V(S) < \frac{1}{2N^\alpha}$ implies that

  $$S_1 > 1 - \frac{1}{2N^\alpha} \quad \text{or} \quad \sum_{i=2}^{b} S_i < \frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{1}{2N^\alpha},$$

  we have $\mathcal{Z} \subseteq \mathcal{Z}_1$ with

  $$\mathcal{Z}_1 = \left\{ s \mid S_b > \delta_b, \sum_{i=1}^{b} S_i \le \frac{\bar{k}\log N}{N^{1-\alpha}} + \epsilon + \frac{1}{2N^\alpha} \right\},$$

  because

  $$h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) < \epsilon \quad \text{and} \quad S_1 > 1 - \frac{1}{2N^\alpha}$$

  implies

  $$\sum_{i=1}^{b} S_i \le \frac{\bar{k}\log N}{N^{1-\alpha}} + \epsilon + \frac{1}{2N^\alpha}.$$

  However, we also have

  $$\begin{aligned}\sum_{i=2}^{b} S_i >&(b-1)S_b \ge (b-1)\delta_b \ge 3\bar{k}\epsilon\log N \\ \ge&\frac{\bar{k}\log N}{N^{1-\alpha}} + \epsilon + \frac{1}{2N^\alpha}.\end{aligned}$$

  which implies $\mathcal{Z}_1 = \emptyset$.

Finally, we have upper bound on $p_{\mathcal{B}}$ that

$$\begin{aligned}p_{\mathcal{B}} \le&\frac{3\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} + \frac{1}{N^{2r}} + \frac{10(2r)^r}{N^{\frac{r(1-\alpha)}{r+1}}} \\ \le&\frac{3\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} + \frac{1}{N^{2r}} + \frac{2\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}} \\ \le&\frac{6\bar{k}}{b}\frac{\log N}{N^{\frac{r(1-\alpha)}{r+1}}},\end{aligned}$$

where the second inequality holds because $\log N \ge \frac{5b(2r)^r}{\bar{k}}$.

4.4 Proof of Corollary 2

Let the test function $f(s)=\sum_{i=3}^{b} s_i$ in

$$E[Gf(S)]=0,$$

and we have $E[S_3]$ under JSQ, I1F, and Pod, respectively.

- For JSQ,

$$E[S_3]=\lambda E\left[1(S_2=1)-1(S_b=1)\right].$$

- For I1F,

$$E[S_3]=\lambda E\left[1(S_2=1)(S_2-S_3)\right].$$

- For Pod,

$$E[S_3]=\lambda E\left[S_2^d-S_b^d\right].$$

We then provide the upper bound of $E[S_3]$ under load balancing algorithms in $\widetilde{\Pi}_2$.

$$\begin{aligned}E[S_3]\leq& E\left[A_2(S)\right]\\ =&E\left[A_2(S)|S_2\geq 0.95\right]\Pr(S_2\geq 0.95)\\ &+E\left[A_2(S)|S_2< 0.95\right]\Pr(S_2< 0.95)\\ \leq&\Pr(S_2\geq 0.95)+E\left[A_2(S)|S_2<0.95\right].\end{aligned} \tag{4.6}$$

The probability in (4.6) is bounded

$$\begin{aligned}\Pr(S_2 \geq 0.95) \leq & \Pr(S_1 + S_2 \geq 1.9) \\ \leq & \Pr\left(h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) \geq 0.9 - \frac{\bar{k}\log N}{N^{1-\alpha}}\right) \\ = & \Pr\left(h_k^r\left(\sum_{i=1}^{b} S_i\right) \geq \left(0.9 - \frac{\bar{k}\log N}{N^{1-\alpha}}\right)^r\right) \\ \leq & \frac{E\left[h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(0.9 - \frac{\bar{k}\log N}{N^{1-\alpha}}\right)^r} \\ \leq & 10\left(\frac{3r}{N^{1-\alpha}}\right)^r\end{aligned}$$

where the last inequality holds because $\frac{N^{1-\alpha}}{\bar{k}\log N} \geq 5$.

The conditional expectation in (4.6) is bounded

$$E\left[A_2(S)|S_2 < 0.95\right] \leq 10\left(\frac{2r}{N^{1-\alpha}}\right)^r$$

for any load balancing algorithms in $\widetilde{\Pi}_2$.

Next, we show JSQ, I1F, and Pod are in $\widetilde{\Pi}_2$, respectively.

- For JSQ,

  $$A_2(s) = 1_{\{s_2=1, s_3<1\}} = 0,$$

  for $s_2 < 0.95$.

- For I1F,

  $$A_2(s) = \mathbb{I}_{\{s_2=2\}} s_2 = 0,$$

  for $s_2 < 0.95$.

- For Pod with $d \geq N^\alpha \log^2 N$,

  $$\begin{aligned}A_2(s) = & s_2^d \leq (0.95)^d \\ \leq & 10\left(\frac{2r}{N^{1-\alpha}}\right)^r.\end{aligned}$$

For JIQ,

$$A_2(s) = \mathbb{I}_{\{s_1=1\}} s_2,$$

which might not be in $\widetilde{\Pi}_2$. However, we can still study $E[S_3]$ by using Theorem 3. Let $\epsilon = 3\left(\frac{3r}{N^{1-\alpha}}\right)^{\frac{r}{r+1}}$ and we have

$$\begin{aligned}
E[S_3] \leq & E\left[\mathbb{I}_{\{S_1=1\}} S_2\right] \\
\leq & E\left[\mathbb{I}_{\{S_1=1\}} S_2 \mid h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) \leq \epsilon\right] + \Pr\left(h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) > \epsilon\right) \\
\leq & \frac{\bar{k}\log N}{N^{1-\alpha}} + \epsilon + \Pr\left(h_{\bar{k}}\left(\sum_{i=1}^{b} S_i\right) > \epsilon\right) \\
\leq & \frac{\bar{k}\log N}{N^{1-\alpha}} + \epsilon + \frac{E\left[h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right)\right]}{\epsilon^r} \\
\leq & \frac{\bar{k}\log N}{N^{1-\alpha}} + \frac{16r}{N^{\frac{r(1-\alpha)}{r+1}}}
\end{aligned}$$

where the first inequality holds by substituting $A_2(s) = \mathbb{I}_{\{s_1=1\}} s_2$ in JIQ; the third inequality holds because of the definition of $h(\cdot)$ and $\mathbb{I}_{\{s_1=1\}} = 1$; the fourth inequality holds by Markov inequality.

4.5 Summary

In this chapter, we studied the steady-state performance of load balancing balancing systems in the Beyond-Halfin-Whitt regime ($0.5 \leq \alpha < 1$). We established high-order moments on a distance function of total queue length for a set of load balancing algorithm Π_2. Based on the high-order moments, the waiting probability and waiting time of incoming job under JSQ, JIQ, I1F and Pod are proved to be asymptotic zero. Further, under JSQ, I1F and Pod, only servers with one or two jobs exist asymptotically.

Chapter 5

STEADY-STATE ANALYSIS OF LOAD BALANCING WITH COXIAN-2 SERVICE

The exponential service has a nice “monotonicity property”, which states a partial order of two mean-field systems starting from two initial conditions to be maintained over time. In particular, letting $x(t, y)$ denote the system state at time t with initial state y, given two initial conditions $y_1 \succ y_2$, where ”$\succ$” is a certain partial order, “monotonicity” states that the partial order $x(t, y_1) \succ x(t, y_2)$ holds for any $t \geq 0$.

Monotonicity does hold under several load balancing algorithms with some non-exponential service time distributions. Typically, it holds when the service time distribution has a decreasing hazard rate (DHR) Bramson *et al.* (2012); Stolyar (2015a); Foss and Stolyar (2017), where the hazard rate is defined to be $\frac{f(x)}{1-F(x)}$ and $f(x)$ is the density function of the service time and $F(x)$ is the corresponding cumulative distribution function.

With monotonicity, Bramson *et al.* (2012) studied Pod load balancing assuming the arrival load per server $\lambda < 1/4$ under any general service time (the second order moment exists) and shown the servers are asymptotic independent as the number of servers $N \to \infty$ to obtain the steady-state performance, where the proof of asymptotic independence relies heavily on the monotonicity. Stolyar (2015a) shown JIQ achieved asymptotic optimality under the service time with decreasing hazard rate (DHR) for any $\lambda < 1$, where monoticity holds for JIQ under DHR assumption. Foss and Stolyar (2017) relaxed DHR assumption in Stolyar (2015a) to any general service distribution and proved the asymptotic optimality of JIQ when the average load per server $\lambda < 0.5$. Houdt (2018) proved the global stability of the mean-filed model of load balancing policies (e.g. Pod) under hyper-exponential distribution. The key step in Houdt (2018) is to represent hyper-exponential distribution by a constrained Coxian distribution, where $\mu_i(1-p_i)$ is decreasing in phase i (μ_i is the service rate in phase i and p_i is the probability that a job finishing service in phase i and entering phase $i+1$). With the alternative representation, monotonicity holds in a certain partial order and the global stability is established.

The Coxian-2 distribution considered in this chapter does not necessarily satisfy DHR. Each job has two phases (phase 1 and phase 2) under the Coxian-2 service time distribution. When in service, a job finishes phase 1 with rate μ_1; and after finishing phase 1, the job leaves the system with probability $1-p$ or enters phase 2 with probability p. If the job enters phase 2, it finishes phase 2 with rate μ_2, and leaves the system.

Now consider a simple system with two servers. Assume the Coxian-2 service time distribution and JSQ is used for load balancing. Consider the system states as shown in Figure 5.1, where jobs in phase 1 are in red color and jobs in phase 2 are in green color. The state of each server can be represented by its queue length and the expected remaining service time of the job in service. Let $Q^{(i,j)}(t)$ denote the queue length of server i at time t in system j, and $T^{(i,j)}(t) \in \left\{\frac{1}{\mu_1}+\frac{p}{\mu_2}, \frac{1}{\mu_2}, 0\right\}$ denotes the expected remaining service time of the job in service at server i in system j. At time 0, we have $Q^{(i,1)}(0) \geq Q^{(i,2)}(0)$ and $T^{(i,1)}(0) \geq T^{(i,2)}(0)$ for all i. During the time period $(t_0, t_1]$, two jobs arrive and were routed to servers according to JSQ, which resulted in the state shown in Figure 5.1. Suppose that $(1-p)\mu_1 < \mu_2$, then at time t_1, we have $T^{(2,1)}(t_1) = \frac{1}{\mu_2} < T^{(2,2)}(t_1) = \frac{1}{\mu_1} + \frac{p}{\mu_2}$, so the system does not have mononticity. This is because Coxian-2 distribution does not satisfy the DHR property when $(1-p)\mu_1 < \mu_2$.

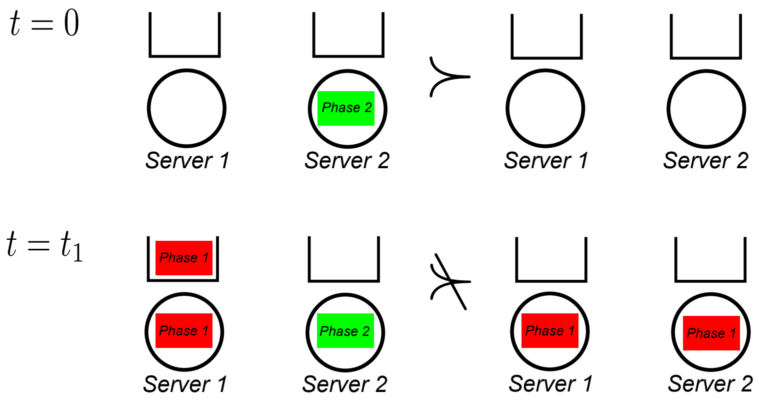

Figure 5.1: Non-Monotocity of JSQ under Coxian-2 Distribution.

Due to the non-monotonicity challenge, there are only a few papers that deal with the steady-state analysis of load balancing systems under non-exponential service time distribution. In this chapter, we analyzed the steady-state performance of load balancing algorithms in the heavy traffic regime, where $\lambda = 1 - N^{-\alpha}$ for $0 < \alpha < 0.5$, under Coxian-2 service time. To overcome the non-monotonicity challenge, we develop an iterative state space collapse (SSC) to show the steady-state "lives" in a restricted region (with a high probability), in which the original system is coupled with a simple system by Stein's method. With iterative SSC and Stein's method, we are able to establish several key performance metrics at steady state, the expected queue length, the probability that a job is allocated to a busy server (waiting probability) and the waiting time. We summarize our results as follows:

- For any load balancing policy in Π_3 (please refer to (5.2) for the formal definition), including JSQ, JIQ, I1F and Pod with $d = O(N^{\alpha} \log N)$, the mean queue length is $\lambda + O\left(\frac{\log N}{\sqrt{N}}\right)$.
- For JSQ and Pod with $d = O(N^{\alpha} \log N)$, the waiting probability and the expected waiting time per job are both $O\left(\frac{\log N}{\sqrt{N}}\right)$.
- For JIQ and I1F, the waiting probability is $O\left(\frac{1}{N^{0.5-\alpha} \log N}\right)$.

.

5.1 Model and Main Results

We consider load balancing system with N homogeneous servers, where job arrival follows a Poisson process with rate λN with $\lambda = 1 - N^{-\alpha}, 0 < \alpha < 0.5$ and service times follow Coxian-2 distribution (μ_1, μ_2, p) as shown in Figure 5.2, where $\mu_m > 0$ is the rate a job finishes phase m when in service and $0 \leq p < 1$ is the probability that a job enters phase 2 after finishing phase 1. Without loss of generality, we assume the mean service time to be one, i.e.

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

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). Jobs are served in FIFO order.

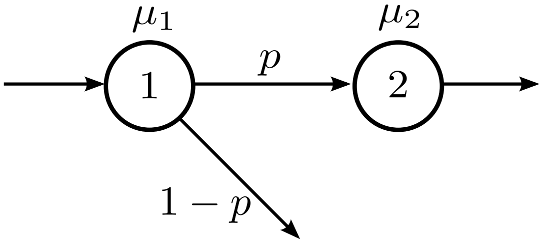

Figure 5.2: Coxian-2 Distribution.

Let $Q_{j,m}(t)$ ($m = 1, 2$) denote the fraction of servers which have j jobs at time t and the one in service is in phase m. For convenience, we define $Q_{0,1}(t)$ to be the fraction of servers that are idle at time t and $Q_{0,2}(t) = 0$. Furthermore define $Q(t)$ to be a $b \times 2$ matrix such that the (j, m)th entry of the matrix is $Q_{j,m}(t)$. Define $S_{i,m}(t) = \sum_{j \geq i} Q_{j,m}(t)$ and $S_i(t) = \sum_{m=1}^{2} S_{i,m}(t)$. In other words, $S_{i,m}(t)$ is the fraction of servers which have at least i jobs and the job in service is in phase m at time t and $S_i(t)$ is the fraction of servers with at least i jobs at time t. Furthermore define

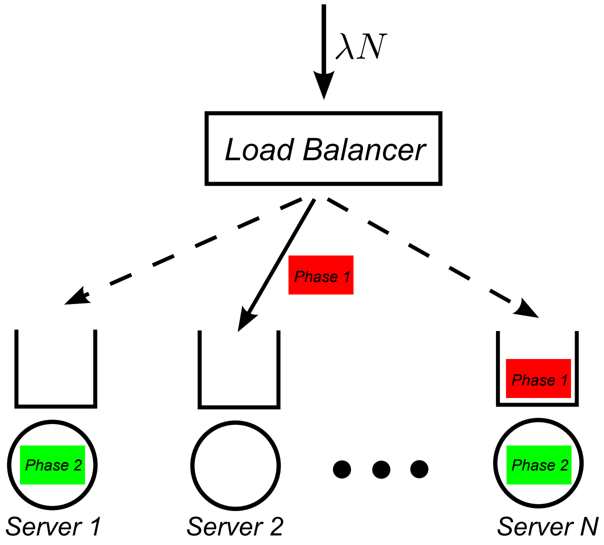

Figure 5.3: Load Balancing in Many-Server Systems under Coxian-2.

$S(t)$ to be a $b\times 2$ matrix such that the (j,m)th entry of the matrix is $S_{j,m}(t)$. Note $Q(t)$ and $S(t)$ have one-to-one mapping. We consider a load balancing algorithm which dispatches jobs to servers based on $Q(t)$ (or $S(t)$) and under which, $\{Q(t), t \geq 0\}$ (or $\{S(t), t \geq 0\}$), which is a finite-state CTMC, is irreducible so has a unique stationary distribution. This class of algorithms include JSQ, JIQ, I1F and Pod.

Let $Q_{j,m}$ denote $Q_{j,m}(t)$ in steady state. We further 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 at steady state. We illustrate the state representation $S_{i,m}$ in Figure 5.4.

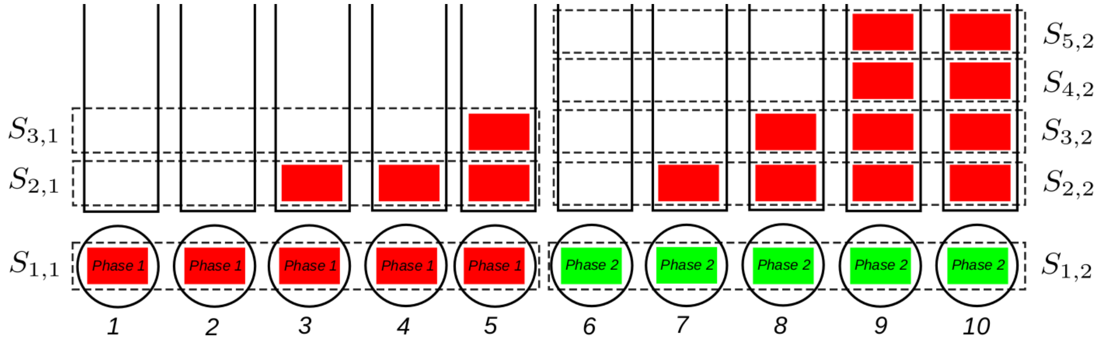

Figure 5.4: Illustrations of States $S_{i,m}$.

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

$$\mathbb{S} = \left\{ s \;\middle|\; 1 \geq s_{1,m} \geq \cdots \geq s_{b,m} \geq 0,\; 1 \geq \sum_{m=1}^{2} s_{1,m};\; Ns_{i,m} \in \mathbb{N},\; \forall i, m \right\}. \tag{5.1}$$

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

$$A_1(s) = \Pr\left(\text{an incoming job is routed to a busy server} \middle| S(t) = s\right).$$

Among load balancing policy (or algorithm) considered in this chapter, define a subset

$$\Pi_3 = \left\{ \pi \;\middle|\; \text{under } \pi, A_1(s) \leq \frac{1}{\sqrt{N}} \;\forall s \in \mathbb{S}, s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{\min\{(1-p)\mu_1, \mu_2\}} \frac{\log N}{\sqrt{N}} \right\}. \tag{5.2}$$

Our main result of this chapter is the following theorem.

Theorem 4. *Define* $w_u = \max\{(1-p)\mu_1, \mu_2\}$, $w_l = \min\{(1-p)\mu_1, \mu_2\}$, $\mu_{\max} = \max\{\mu_1, \mu_2\}$, *and* $k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right)$. *Under any load balancing policy in* Π_3, *the following bound holds when a large* N *satisfying* $\frac{w_l N^{0.5-\alpha}}{1+\mu_1+\mu_2} \geq \log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1\mu_2}{40}\right)}$

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k \log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{7\mu_{\max}}{\sqrt{N}\log N}. \tag{5.3}$$

□

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

- **J**oin-the-Shortest-Queue (JSQ): JSQ routes an incoming job to the least loaded server in the system. Therefore, $A_1(s) = 0$ when $s_1 < 1$.

- **I**dle-One-First (I1F) Gupta and Walton (2019): I1F routes an incoming job to an idle server if available; and otherwise to a server with one job if available. If all servers have at least two jobs, the job is routed to a randomly selected server. Therefore, $A_1(s) = 0$ when $s_1 < 1$.

- **J**oin-the-Idle-Queue (JIQ) Lu *et al.* (2011): JIQ routes an incoming job to an idle server if possible and otherwise, routes a server chosen uniformly at random. Therefore, $A_1(s) = 0$ when $s_1 < 1$.

- **P**ower-of-d-Choices (Pod) Mitzenmacher (1996); Vvedenskaya *et al.* (1996): Pod samples d servers uniformly at random and dispatches the job to the least loaded server among the d servers. Ties are broken uniformly at random. When $d \geq \mu_1 N^\alpha \log N$, $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$.

A direct consequence of Theorem 4 is *asymptotic zero waiting at steady state.* Let $\mathcal{W}$ denote the event that an incoming job is routed to a busy server in a system with N servers, and $p_{\mathcal{W}}$ denote the probability of this event at steady-state. Let $\mathcal{B}$ denote the event that an incoming job is blocked (discarded) and $p_{\mathcal{B}}$ denote the probability of this event at steady-state. Note event $\mathcal{W}$ occurs implies $\mathcal{B}$ occurs because a job is blocked when being routed to a server with b jobs. Furthermore, let W denote the waiting time of a job (when the job is not dropped). We have the following results based on the main theorem.

Corollary 4. *The following results hold when a large N satisfying $\frac{w_l N^{0.5-\alpha}}{1+\mu_1+\mu_2} \geq \log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1\mu_2}{40}\right)}$,*

- *Under JSQ and Pod with $d = \mu_1 N^\alpha \log N$, we have*

$$
\begin{aligned}
E\left[W\right] \leq & \frac{2k\log N}{\sqrt{N}} + \frac{14\mu_{\max} + \frac{16\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}, \\
p_{\mathcal{W}} \leq & \frac{\mu_{\max}}{\lambda}\left(\frac{k\log N}{\sqrt{N}} + \frac{7\mu_{\max} + \frac{8\mu_{\max}}{b-\lambda}}{\sqrt{N}\log N}\right).
\end{aligned}
$$

- *Under JIQ and I1F,*

$$
p_{\mathcal{W}} \leq \frac{14\mu_{\max}}{N^{0.5-\alpha}\log N}.
$$

□

The proof of this corollary is an application of Little's law and Markov's inequality, and can be found in Section 5.4.

5.2 Proof of Theorem 4 under JSQ

In this section, we present the proof of our main theorem for JSQ, which is organized along the three key ingredients: 1) generator approximation; 2) gradient bounds; 3) state space collapse. The proof for other load balancing algorithms is similar and will be discussed in Section 5.3. Since load balancing under Coxian-2 service is more complex than load balancing under exponential service, we derive the generator approximation from the beginning.

5.2.1 Generator Approximation

Define $e_{i,m} \in \mathbb{R}^{b\times 2}$ to be a $b \times 2$-dimensional matrix such that the (i,m)th entry is $1/N$ and all other entries are zero.

Given the state of the CTMC s and q, there are possible events under JSQ as listed below.

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

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

  This transition occurs with rate

  $$\lambda N \frac{q_{i-1,1}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}},$$

  where $\frac{q_{i-1,1}}{q_{i-1}}$ is the probability that the server which receives the job is serving a job in phase 1 conditioned on the job is routed to a server with $i-1$ jobs, and $\{s_{i-1}=1, s_i<1\}$ implies that the shortest queue in the system has length $i-1$.

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

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

This transition occurs with rate

$$\lambda N \frac{q_{i-1,2}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}},$$

where $\frac{q_{i-1,2}}{q_{i-1}}$ is the probability that the server which receives the job is serving a job in phase 2 conditioned on the job is routed to a server with $i-1$ jobs, and $\{s_{i-1}=1, s_i<1\}$ implies that the shortest queue in the system has length $i-1$.

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

  $$\begin{aligned} q &\to q - e_{i,1} + e_{i-1,1}, \\ s &\to s - e_{i,1}. \end{aligned}$$

  This transition occurs with rate

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

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

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

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

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

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

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

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

$$\begin{aligned} q &\rightarrow q - e_{i,2} + e_{i-1,1}, \\ s &\rightarrow s - \sum_{j=1}^{i} e_{j,2} + \sum_{j=1}^{i-1} e_{j,1}. \end{aligned}$$

This transition occurs with rate

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

We illustrate local state transitions related to state s in Fig. 5.5.

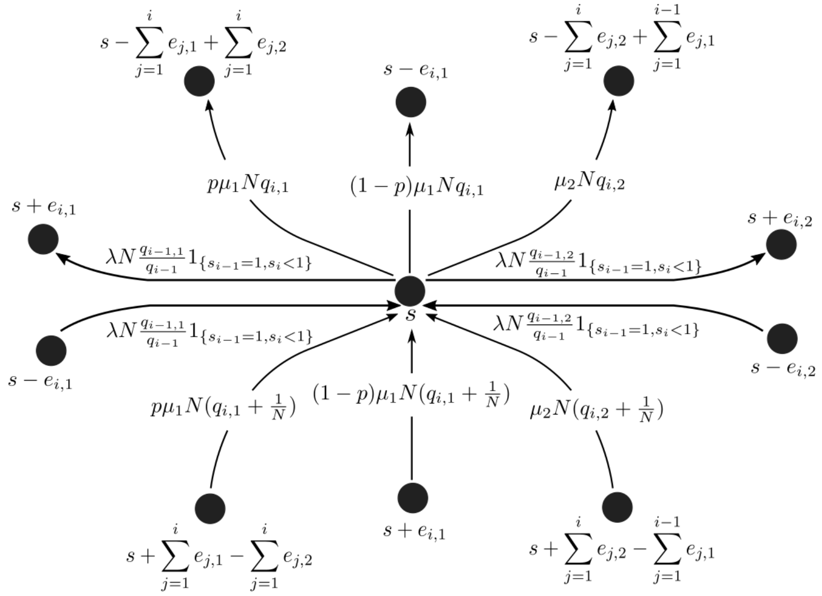
$s - \sum_{j=1}^{i} e_{j,1} + \sum_{j=1}^{i} e_{j,2}$
$s - e_{i,1}$
$s - \sum_{j=1}^{i} e_{j,2} + \sum_{j=1}^{i-1} e_{j,1}$
$p\mu_1 N q_{i,1}$
$(1-p)\mu_1 N q_{i,1}$
$\mu_2 N q_{i,2}$
$s + e_{i,1}$
$s + e_{i,2}$
$\lambda N \frac{q_{i-1,1}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}}$
$\lambda N \frac{q_{i-1,2}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}}$
s
$\lambda N \frac{q_{i-1,1}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}}$
$\lambda N \frac{q_{i-1,2}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}}$
$s - e_{i,1}$
$s - e_{i,2}$
$p\mu_1 N(q_{i,1} + \frac{1}{N})$
$(1-p)\mu_1 N(q_{i,1} + \frac{1}{N})$
$\mu_2 N(q_{i,2} + \frac{1}{N})$
$s + \sum_{j=1}^{i} e_{j,1} - \sum_{j=1}^{i} e_{j,2}$
$s + e_{i,1}$
$s + \sum_{j=1}^{i} e_{j,2} - \sum_{j=1}^{i-1} e_{j,1}$

Figure 5.5: Illustrations of State Transitions for any i with $1 \leq i \leq b$.

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

$$Gf(s) = \sum_{i=1}^{b} \Bigg[\lambda N \frac{q_{i-1,1}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}} (f(s + e_{i,1}) - f(s)) \tag{5.4}$$

$$+ \lambda N \frac{q_{i-1,2}}{q_{i-1}} 1_{\{s_{i-1}=1, s_i<1\}} (f(s + e_{i,2}) - f(s)) \tag{5.5}$$

$$+ (1-p)\mu_1 N q_{i,1} (f(s - e_{i,1}) - f(s)) \tag{5.6}$$

$$+ p\mu_1 N q_{i,1} \left(f\left(s - \sum_{j=1}^{i} e_{j,1} + \sum_{j=1}^{i} e_{j,2} \right) - f(s) \right) \tag{5.7}$$

$$+ \mu_2 N q_{i,2} \left(f\left(s - \sum_{j=1}^{i} e_{j,2} + \sum_{j=1}^{i-1} e_{j,1} \right) - f(s) \right) \Bigg] \tag{5.8}$$

For any bounded function $f : \mathbb{S} \to \mathbb{R}$,

$$E[Gf(S)] = 0, \tag{5.9}$$

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

To understand the steady-state performance of a load balancing algorithm, we will establish an upper bound on the following function:

$$\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}.$$

The upper bounds measure the quantity that the total number of jobs in the system ($N\sum_{i=1}^{b} S_i$) exceeds $N\lambda + k\sqrt{N}\log N$ at steady state, and can be used to bound the probability that an incoming job is routed to an idle server in Corollary 4.

We consider a simple fluid system with arrival rate λ and departure rate $\lambda+\delta$ with $\delta = \frac{\log N}{\sqrt{N}}$, i.e.

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

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

$$g'(x)\left(-\frac{\log N}{\sqrt{N}}\right) = \max\left\{x - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}, \forall x, \tag{5.10}$$

where $g'(x) = \frac{dg(x)}{dx}$. The left-hand side of (5.10) can be viewed as applying the generator of the simple fluid system to function $g(x)$, i.e.

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

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

$$f(s) = g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right), \tag{5.11}$$

and use $f(s)$ defined above in Stein's method.

Since $\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} = \sum_{i=1}^{b} s_i \le b$ for $s\in\mathbb{S}$, and $f(s)$ is a bounded for $s\in\mathbb{S}$. So

$$E[Gf(S)] = E\left[Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right] = 0. \tag{5.12}$$

Now define

$$h(x) = \max\left\{x - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}.$$

Based on (5.10) and (5.12), we obtain

$$E\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right] = E\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\left(-\frac{\log N}{\sqrt{N}}\right) - Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} S_{i,m}\right)\right]. \tag{5.13}$$

Note that according to the definition of $f(s)$ in (5.11), $e_{j,1}$ and $e_{j,2}$, we have

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

and

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

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

$$\begin{aligned}
&Gg\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right)\\
=&N\lambda\left(1 - 1_{\{s_b=1\}}\right)\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} + \frac{1}{N}\right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right)\right)\\
&+ N\left((1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}\right)\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m} - \frac{1}{N}\right) - g\left(\sum_{i=1}^{b}\sum_{m=1}^{2} s_{i,m}\right)\right),
\end{aligned}$$

where the first term represents the transitions when a job arrives and the second term represents the transitions when a job leaves the system.

Substituting the equation above to (5.13), we have

$$\begin{aligned}
&E\left[h\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}\right)\right]\\
=&E\left[g'\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}\right)\left(-\frac{\log N}{\sqrt{N}}\right)\right.\\
&-N\lambda(1-1_{\{S_b=1\}})\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}+\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}\right)\right)\\
&\left.-N\left((1-p)\mu_1 S_{1,1}+\mu_2 S_{1,2}\right)\left(g\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}-\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}\sum_{m=1}^{2}S_{i,m}\right)\right)\right].
\end{aligned} \tag{5.14}$$

Define $\eta = \lambda + \frac{k\log N}{\sqrt{N}}$ to simplify notation. From the definition of g and g', for any $x < \eta$,

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

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

$$g'(x) = -\frac{\sqrt{N}}{\log N}\left(x - \lambda - \frac{k\log N}{\sqrt{N}}\right), \tag{5.15}$$

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

$$g''(x) = -\frac{\sqrt{N}}{\log N}. \tag{5.16}$$

By using mean-value theorem in the region $[\eta - \frac{1}{N}, \eta + \frac{1}{N}]$ and Taylor theorem in the region $(\eta + \frac{1}{N}, \infty)$, we have

$$\begin{aligned}
g(x+\frac{1}{N}) - g(x) =& \left(g(x+\frac{1}{N}) - g(x)\right)\left(1_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}} + 1_{x>\eta+\frac{1}{N}}\right)\\
=&\frac{g'(\xi)}{N}1_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}} + \left(\frac{g'(x)}{N} + \frac{g''(\zeta)}{2N^2}\right)1_{x>\eta+\frac{1}{N}}
\end{aligned} \tag{5.17}$$

$$\begin{aligned}
g(x-\frac{1}{N}) - g(x) =& \left(g(x-\frac{1}{N}) - g(x)\right)\left(1_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}} + 1_{x>\eta+\frac{1}{N}}\right)\\
=&-\frac{g'(\tilde{\xi})}{N}1_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}} + \left(-\frac{g'(x)}{N} + \frac{g''(\tilde{\zeta})}{2N^2}\right)1_{x>\eta+\frac{1}{N}}
\end{aligned} \tag{5.18}$$

where $\xi, \zeta \in (x, x+\frac{1}{N})$ and $\tilde{\xi}, \tilde{\zeta} \in (x-\frac{1}{N}, x)$. Substitute (5.17) and (5.18) into the generator difference in (5.14), we have

$$
\begin{aligned}
& E\left[h\left(\sum_{i=1}^{b} S_i\right)\right] \\
=& E\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(\lambda 1_{\{S_b=1\}} - \lambda - \frac{\log N}{\sqrt{N}} + (1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
& \qquad (5.19) \\
& + E\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1-1_{\{S_b=1\}})g'(\xi)\right.\right. \\
& \qquad \left.\left. + ((1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2})g'(\tilde{\xi})\right)\mathbb{I}_{\eta - \frac{1}{N} \le \sum_{i=1}^{b} S_i \le \eta + \frac{1}{N}}\right] \qquad (5.20) \\
& - E\left[\frac{1}{2N}\left(\lambda(1-1_{\{S_b=1\}})g''(\zeta) + ((1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2})g''(\tilde{\zeta})\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]. \\
& \qquad (5.21)
\end{aligned}
$$

Note in (5.20) and (5.21), we have random variables $\xi, \zeta \in \left(\sum_{i=1}^{b} S_i, \sum_{i=1}^{b} S_i + \frac{1}{N}\right)$ and $\tilde{\xi}, \tilde{\zeta} \in \left(\sum_{i=1}^{b} S_i - \frac{1}{N}, \sum_{i=1}^{b} S_i\right)$ whose values depend on $\sum_{i=1}^{b} S_i$.

To establish the main result in Theorem 4, we need to provide the upper bounds on (5.19), (5.20) and (5.21). In the following subsection 5.2.2, we study g' and g'' to bound the terms in (5.20) and (5.21); In the subsection 5.2.3, we study SSC to bound the term in (5.19).

5.2.2 Gradient Bounds

Let $\eta = \lambda + \frac{k \log N}{\sqrt{N}}$ and $\delta = \frac{\log N}{\sqrt{N}}$ in Lemma 21 and Lemma 22. We have the following two lemmas.

Lemma 15. *Given* $x \in \left[\lambda + \frac{k \log N}{\sqrt{N}} - \frac{2}{N}, \lambda + \frac{k \log N}{\sqrt{N}} + \frac{2}{N}\right]$, *we have*

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

□

Lemma 16. *For $x > \lambda + \frac{k \log N}{\sqrt{N}}$, we have*

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

□

Based on the bounds on g' in Lemma 21 and g'' in Lemma 22, we provide the upper bound on (5.20) + (5.21) in the following lemma.

Lemma 17. *For $g(\cdot)$ defined in* (5.10)*, we have*

$$(5.20) + (5.21) \leq \frac{6\mu_{\max}}{\sqrt{N}\log N}.$$

□

Proof. Note $((1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}) \leq \mu_{\max} S_1 \leq \mu_{\max}$, then we have

$$\begin{aligned}
(5.20) + (5.21) \leq & E\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) + \lambda|g'(\xi)| + \mu_{\max}|g'(\tilde{\xi})|\right)\mathbb{I}_{\sum_{i=1}^{b} S_i \in \Omega_R}\right] \\
& + E\left[\frac{1}{N}\left(\lambda|g''(\eta)| + \mu_{\max}|g''(\tilde{\eta})|\right)\mathbb{I}_{\sum_{i=1}^{b} S_i \in \Omega_L}\right] \\
\leq & \frac{4\mu_{max}}{\sqrt{N}\log N} + \frac{\lambda + \mu_{\max}}{N}\frac{\sqrt{N}}{\log N} \\
\leq & \frac{6\mu_{max}}{\sqrt{N}\log N}
\end{aligned}$$

□

5.2.3 State Space Collapse (SSC)

In this subsection, we analyze (5.19):

$$\begin{aligned}
& E\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(\lambda 1_{\{S_b=1\}} - \eta + (1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
= & E\left[\frac{\sqrt{N}}{\log N} h\left(\sum_{i=1}^{b} S_i\right)\left(-\lambda 1_{\{S_b=1\}} + \eta - (1-p)\mu_1 S_{1,1} - \mu_2 S_{1,2}\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
\leq & E\left[\frac{\sqrt{N}}{\log N} h\left(\sum_{i=1}^{b} S_i\right)\left(\eta - (1-p)\mu_1 S_{1,1} - \mu_2 S_{1,2}\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right], \qquad (5.22)
\end{aligned}$$

where the equality is due to Stein's equation (5.10), and the inequality holds because

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

We first focus on

$$(\eta - (1-p)\mu_1 s_{1,1} - \mu_2 s_{1,2})\, \mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}}, \tag{5.23}$$

where $(1-p)\mu_1 s_{1,1}$ and $\mu_2 s_{1,2}$ are the rates at which jobs leave the system when in phase 1 and phase 2, respectively. Therefore, $(1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2}$ is the total departure rate when the system in state $S = s$.

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

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

and

$$\mathcal{S}_{ssc_1} = \left\{ s \,\middle|\, s_1 \geq \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}, s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}, s_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}} \right\},$$

$$\mathcal{S}_{ssc_2} = \left\{ s \,\middle|\, \sum_{i=1}^{b} s_i \leq \lambda + \frac{k \log N}{\sqrt{N}} \right\}.$$

- **C**ase 1: $\mathcal{S}_{ssc_1}$ is shown as the gray region in Fig. 5.6. Any $s \in \mathcal{S}_{ssc_1}$ satisfies

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

  so (5.23) ≤ 0 for any $s \in \mathcal{S}_{ssc_1}$. The details are presented in Lemma 18. When $s \in \mathcal{S}_{ssc_2}$,

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

  so (5.23) $= 0$ for any $s \in \mathcal{S}_{ssc_2}$.

- **C**ase 2: We will show that

  $$\Pr\left(S \notin \mathcal{S}_{ssc}\right) \leq \frac{3}{N^2}$$

  in Lemma 19 using an iterative state space collapse approach.

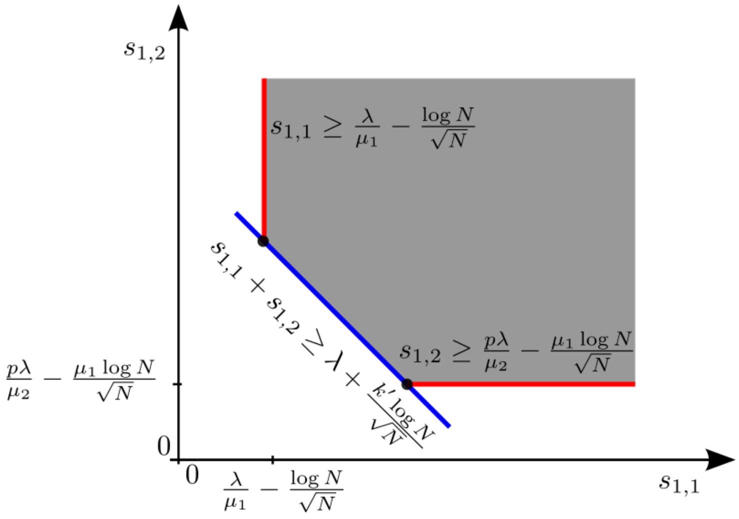
$s_{1,2}$
$s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$
$s_{1,1} + s_{1,2} \geq \lambda + \frac{k' \log N}{\sqrt{N}}$
$s_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$
$\frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$
0
0
$\frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$
$s_{1,1}$

Figure 5.6: State Space Collapse in $\mathcal{S}_{ssc_1}$.

Lemma 18. *For any $s \in \mathcal{S}_{ssc_1}$,*

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

□

Proof. We consider the following problem

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

which is a linear programming in terms of variables $s_{1,1}$ and $s_{1,2}$. Therefore, we only need to consider the extreme points of set $\mathcal{S}_{ssc_1}$. In fact, from Figure 5.6, it is clear that we only need to consider the following two extreme points.

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

$$\begin{aligned}(1-p)\mu_1 s_{1,1}+\mu_2 s_{1,2} =&\lambda+\left(-(1-p)\mu_1+\mu_2\left(\frac{1+\mu_1+\mu_2}{w_l}+1\right)\right)\frac{\log N}{\sqrt{N}}\\ \geq&\lambda+\left(-(1-p)\mu_1+(1+\mu_1+2\mu_2)\right)\frac{\log N}{\sqrt{N}} \qquad (5.24)\\ \geq&\lambda+(1+p\mu_1+2\mu_2)\frac{\log N}{\sqrt{N}}\\ \geq&\lambda+\frac{\log N}{\sqrt{N}},\end{aligned}$$

where (5.24) holds because $w_l=\min\{(1-p)\mu_1,\mu_2\}$.

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

  At this extreme point, we have

$$\begin{aligned}(1-p)\mu_1 s_{1,1}+\mu_2 s_{1,2} =&\lambda+\left((1-p)\mu_1\left(\frac{1+\mu_1+\mu_2}{w_l}+\mu_1\right)-\mu_2\mu_1\right)\frac{\log N}{\sqrt{N}}\\ \geq&\lambda+\left(1+\mu_1+\mu_2+(1-p)\mu_1^2-\mu_2\mu_1\right)\frac{\log N}{\sqrt{N}} \qquad (5.25)\\ \geq&\lambda+\frac{\log N}{\sqrt{N}}, \qquad (5.26)\end{aligned}$$

where (5.25) holds because $w_l=\min\{(1-p)\mu_1,\mu_2\}$ and (5.26) holds because $\mu_1+\mu_2\geq p\mu_1+\mu_2=\mu_1\mu_2$.

□

Lemma 19. *For a large N such that $\log N\geq\frac{3.5}{\min\left(\frac{\mu_1}{16},\frac{\mu_2}{12},\frac{\mu_1\mu_2}{40}\right)}$, we have*

$$\Pr\left(S\notin\mathcal{S}_{ssc}\right)\leq\frac{3}{N^2}.$$

□

Based on Lemma 18 and Lemma 19, we can establish the following bound on (5.19), in the following lemma.

Lemma 20. *Under JSQ, we have*

$$(5.19) \leq \frac{3b}{N^{1.5}\log N}$$

for a sufficiently large N such that $\log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1\mu_2}{40}\right)}$. □

Proof.

$$\begin{aligned}
&(5.19) \leq (5.22)\\
=&E\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \eta\right)(\lambda + \delta - (1-p)\mu_1 S_{1,1} - \mu_2 S_{1,2})\, \mathbb{I}_{S\in\mathcal{S}_{ssc}} \mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]\\
&+ E\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \eta\right)(\lambda + \delta - (1-p)\mu_1 S_{1,1} - \mu_2 S_{1,2})\, \mathbb{I}_{S\notin\mathcal{S}_{ssc}} \mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]\\
\leq&\frac{3b}{N^{1.5}\log N} \qquad (5.27)
\end{aligned}$$

where (5.27) holds because of Lemma 18 on $S \in \mathcal{S}_{ssc}$ and Lemma 19 on $S \notin \mathcal{S}_{ssc}$. □

5.2.4 Proving Theorem 4 under JSQ

Based on Lemma 17 and Lemma 20, we are ready to establish Theorem 4 under JSQ.

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \eta, 0\right\}\right] = (5.19) + (5.20) + (5.21) \leq \frac{3b}{N^{1.5}\log N} + \frac{6\mu_{\max}}{\sqrt{N}\log N},$$

which implies

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

Remark: An important contribution of this chapter is the iterative state collapse method we use to prove Lemma 19. The method continues refining the state space in which the system stays at steady-state with a high probability. Figure 5.7 illustrates the first few steps of the iterative state-space collapse proof. We first show that with a high probability, $S_{1,2} \leq \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}$ at steady-state. Then in the reduced state space

$\left(S_{1,2} \leq \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}\right)$, we further show $S_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ with a high probability at steady state. We then further establish $S_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$ with a high probability at steady state in the reduced state space. Similar steps are taken to finally prove that $S \in \mathcal{S}_{ssc}$ with a high probability at steady state.

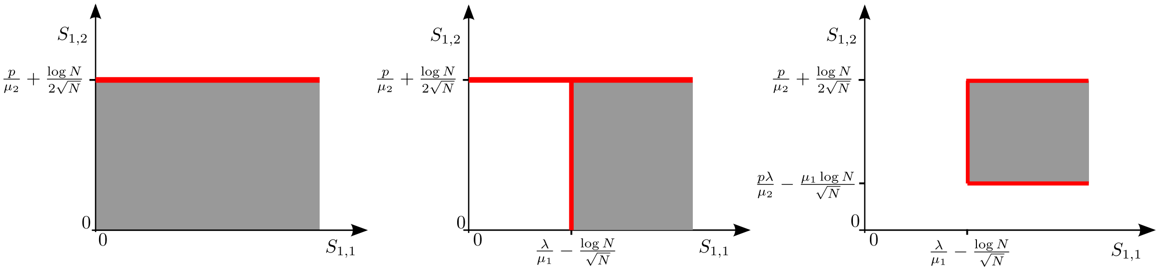
$S_{1,2}$
$\frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}$
0
0
$S_{1,1}$
$S_{1,2}$
$\frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}$
0
0
$\frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$
$S_{1,1}$
$S_{1,2}$
$\frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}$
$\frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$
0
0
$\frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$
$S_{1,1}$

Figure 5.7: Iterative State-Space Collapse to Show that $S_{1,1}$ and $S_{1,2}$ are in a Smaller State-Space (Gray Region) at Steady-State

5.3 Extension to Policy Set Π_3

In this section, we extend the analysis of JSQ to any policy in Π_3. Most steps are the same for a policy in Π_3 as for JSQ, except minor differences in proving Lemma 28 and Lemma 30. We next list the places where minor changes are needed.

In Lemma 28, under the condition $s_1 \leq \lambda + \frac{1+\mu_1+\mu_2}{w_l} \frac{\log N}{\sqrt{N}}$,

- For JSQ,

$$\begin{aligned}\nabla V(s) =& -\lambda 1_{\{S_1<1\}} + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \\ =& -\lambda + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\end{aligned}$$

- For a policy in Π_3,

$$\begin{aligned}\nabla V(s) =& -\lambda\left(1 - A_1(s)\right) + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \\ \leq& \frac{1}{\sqrt{N}} - \lambda + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\end{aligned}$$

Therefore, Lemma 28 still holds for policies in Π_3.

In Lemma 30, under the condition $s_1 \le \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}$,

- For JSQ,
  - If $\lambda + \frac{k\log N}{\sqrt{N}} - s_1 \ge \sum_{i=2}^{b} s_i$,

$$\begin{aligned}\nabla V(s) \le & -\lambda 1_{\{S_1=1\}} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\\ = & -(1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\end{aligned}$$

  - If $\sum_{i=2}^{b} s_i > \lambda + \frac{k\log N}{\sqrt{N}} - s_1$,

$$\begin{aligned}\nabla V(s) \le & -\lambda 1_{\{S_1<1\}} + (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\\ = & -\lambda + (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\end{aligned}$$

- For a policy in Π_3,
  - If $\lambda + \frac{k\log N}{\sqrt{N}} - s_1 \ge \sum_{i=2}^{b} s_i$,

$$\begin{aligned}\nabla V(s) \le & -\lambda(A_1(s) - 1_{\{s_b=1\}}) - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\\ \le & \frac{1}{\sqrt{N}} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2}\end{aligned}$$

  - If $\sum_{i=2}^{b} s_i > \lambda + \frac{k\log N}{\sqrt{N}} - s_1$,

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

Therefore, Lemma 30 still holds for policies in Π_3.

5.4 Proof of Corollary 4

Under JSQ, a job is discarded or blocked only if all buffers are full, i.e. when $N\sum_{i=1}^b S_i = Nb$. From Theorem 3, we have

$$
\begin{aligned}
p_{\mathcal{B}} =& \Pr\left(N\sum_{i=1}^b S_i = Nb\right) = \Pr\left(\sum_{i=1}^b S_i \geq b\right) && (5.28)\\
\leq& \Pr\left(\max\left\{\sum_{i=1}^b S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\} \geq b - \lambda - \frac{k\log N}{\sqrt{N}}\right) && (5.29)\\
\leq& \frac{E\left[\max\left\{\sum_{i=1}^b S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]}{b-\lambda-\frac{k\log N}{\sqrt{N}}} && (5.30)\\
\leq& \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N} && (5.31)
\end{aligned}
$$

where (5.29) to (5.30) holds due to the Markov inequality; and (5.30) to (5.31) holds because of Thereom 3 and $b - \lambda \geq \frac{k\log N}{\sqrt{N}}$;

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

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

Therefore, the average waiting time is

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

where the last inequality holds because $\lambda(1-p_{\mathcal{B}_N}) \geq 0.5$.

From the work-conserving law, we have

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

Therefore, we have

$$\lambda - \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N} \leq E[S_1] \leq \lambda,$$

which implies

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

due to the fact

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

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

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

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

$$p_{\overline{\mathcal{W}}} \leq \frac{E\left[\sum_{i=2}^{b} S_i\right]}{\lambda \overline{T}_Q}.$$

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

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

The analysis for Pod is similar, except that

$$\begin{aligned}
p_{\mathcal{B}} =& \Pr\left(\mathcal{B}\middle| S_b \leq 1-\frac{1}{\mu_1 N^{\alpha}}\right)\Pr\left(S_b \leq 1-\frac{1}{\mu_1 N^{\alpha}}\right)\\
&+\Pr\left(\mathcal{B}\middle| S_b > 1-\frac{1}{\mu_1 N^{\alpha}}\right)\Pr\left(S_b > 1-\frac{1}{\mu_1 N^{\alpha}}\right)\\
\leq& \Pr\left(\mathcal{B}\middle| S_b \leq 1-\frac{1}{\mu_1 N^{\alpha}}\right)+\Pr\left(S_b > 1-\frac{1}{\mu_1 N^{\alpha}}\right)\\
\leq& \left(1-\frac{1}{\mu_1 N^{\alpha}}\right)^{\mu_1 N^{\alpha}\log N}+\Pr\left(\sum_{i=1}^{b} S_i > b-\frac{b}{\mu_1 N^{\alpha}}\right)\\
\leq& \frac{8\mu_{\max}}{b-\lambda}\frac{1}{\sqrt{N}\log N}.
\end{aligned}$$

The remaining analysis is the same.

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

$$\begin{aligned}
p_{\mathcal{W}} =& \Pr\left(S_1 = 1\right) \leq \Pr\left(\sum_{i=1}^{b} S_i \geq 1\right)\\
\leq& \Pr\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right\} \geq \frac{1}{N^{\alpha}}-\frac{k\log N}{\sqrt{N}}\right).
\end{aligned}$$

The result follows from the Markov inequality.

5.5 Summary

In this chapter, we considered load balancing under Coxian-2 service time in the Sub-Halfin-Whitt regime. We developed an iterative SSC to overcome the non-monotonicity challenge and establish a policy set Π_3, in which any policy can achieve zero delay asymptotically. The set Π_3 includes JSQ, JIQ, I1F and Pod with $d \geq \mu_1 N^{\alpha}\log N$.

Chapter 6

CONCLUSION

In this dissertation, we studied steady-state performance of load balancing algorithms for many-server systems (N servers) in heavy traffic regime. We developed Stein's method and (iterative) state space collapse (SSC) framework to analyze load balancing systems in various traffic regime (Sub-Halfin-Whitt and Beyound-Halfin-Whitt regime) and service assumption (exponential service and Coxian-2 service).

Chapter 3 studied load balancing in the Sub-Halfin-Whitt regime under exponential service. Stein's method and state space collapse (SSC) are introduced in this chapter and demonstrated to be a potential framework in steady-state analysis of load balancing. With the framework, a set of "zero-delay" load balancing are established, under which the waiting time and waiting probability achieve zero asymptotically (as $N \to \infty$). JSQ, JIQ, I1F, and Pod belong to "zero-delay" load balancing.

Chapter 4 studied load balancing in the Beyond-Halfin-Whitt regime under exponential service. Though in "heavier" traffic regime, state space collapse is still proved by Lyapunov drift analysis with a carefully-designed Lyapunov function. Combine with an iterative refined procedure, high-order bound on total queue length are obtained, and a set of "zero-delay" load balancing is also established.

Chapter 5 studied load balancing in the Sub-Halfin-Whitt regime under Coxian-2 service. Load balancing under Coxian-2 service is challenging because of "non-monotonicity". To tackle the "non-monotonicity" challenge, an interesting iterative state space collapse is proposed to reduce state space step by step, and it helps establish a similar set of "zero-delay" load balancing as in exponential service.

REFERENCES

Aghajani, R., X. Li and K. Ramanan, “The pde method for the analysis of randomized load balancing networks”, Proc. ACM Meas. Anal. Comput. Syst. **1**, 2, 38:1–38:28, URL `http://doi.acm.org/10.1145/3154497` (2017).

Atar, R., “A diffusion regime with nondegenerate slowdown”, Oper. Res. **60**, 2, 490–500, URL `https://doi.org/10.1287/opre.1110.1030` (2012).

Banerjee, S. and D. Mukherjee, “Join-the-shortest queue diffusion limit in halfinwhitt regime: Tail asymptotics and scaling of extrema”, Ann. Appl. Probab. **29**, 2, 1262–1309, URL `https://doi.org/10.1214/18-AAP1436` (2019).

Bertsimas, D., D. Gamarnik and J. N. Tsitsiklis, “Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions”, Adv. in Appl. Probab. (2001).

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

Braverman, A., “Steady-state analysis of the join the shortest queue model in the halfin-whitt regime”, arXiv:1801.05121 (2018).

Braverman, A. and J. G. Dai, “Steins method for steady-state diffusion approximations of $m/Ph/n + m$ systems”, Ann. Appl. Probab. **27**, 1, 550–581, URL `https://doi.org/10.1214/16-AAP1211` (2017).

Braverman, A., J. G. Dai and J. Feng, “Stein’s method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models”, Stochastic Systems **6**, 301–366 (2016).

Eschenfeldt, P. and D. Gamarnik, “Join the shortest queue with many servers. the heavy-traffic asymptotics”, Mathematics of Operations Research (2018).

Foss, S. and A. L. Stolyar, “Large-scale join-idle-queue system with general service times”, Journal of Applied Probability **54**, 4, 9951007 (2017).

Gast, N., “Expected values estimated via mean-field approximation are 1/n-accurate”, Proc. ACM Meas. Anal. Comput. Syst. **1**, 1, 17:1–17:26 (2017).

Gast, N. and B. Van Houdt, “A refined mean field approximation”, in “Proc. Ann. ACM SIGMETRICS Conf.”, (Irvien, CA, 2018).

Gupta, V. and N. Walton, “Load balancing in the nondegenerate slowdown regime”, Operations Research **67**, 1, 281–294, URL `https://doi.org/10.1287/opre.2018.1768` (2019).

Halfin, S. and W. Whitt, “Heavy-traffic limits for queues with many exponential servers”, Operations Research **29**, 3, 567–588 (1981).

Harchol-Balter, M., *Performance Modeling and Design of Computer Systems: Queueing Theory in Action* (Cambridge University Press, 2013).

He, S., "Diffusion approximation for efficiency-driven queues: A space-time scaling approach", arXiv:1506.06309 (2015).

Hellemans, T. and B. Van Houdt, "On the power-of-d-choices with least loaded server selection", Proc. ACM Meas. Anal. Comput. Syst. **2**, 2, 27:1–27:22, URL `http://doi.acm.org/10.1145/3224422` (2018).

Houdt, B. V., "Global attraction of ode-based mean field models with hyperexponential job sizes", CoRR **abs/1811.05239**, URL `http://arxiv.org/abs/1811.05239` (2018).

Leverich, J. and C. Kozyrakis, "Reconciling high server utilization and submillisecond quality-of-service", in "Proceedings of the Ninth European Conference on Computer Systems", EuroSys '14, pp. 4:1–4:14 (ACM, New York, NY, USA, 2014), URL `http://doi.acm.org/10.1145/2592798.2592821`.

Liu, X. 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).

Lu, Y., Q. Xie, G. Kliot, A. Geller, J. R. Larus and A. Greenberg, "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services", Performance Evaluation **68**, 11, 1056–1071 (2011).

Mitzenmacher, M., *The Power of Two Choices in Randomized Load Balancing*, Ph.D. thesis, University of California at Berkeley (1996).

Mukherjee, D., S. C. Borst, J. S. H. van Leeuwaarden and P. A. Whiting, "Universality of power-of-d load balancing in many-server systems", Stochastic Systems **8**, 4, 265–292, URL `https://doi.org/10.1287/stsy.2018.0016` (2018).

Schurman, E. and J. Brutlag, "The user and business impact of server delays, additional bytes, and http chunking in web search", in "O'Reilly Velocity Web Performance and Operations Conf.", (2009).

Stolyar, A., "Pull-based load distribution in large-scale heterogeneous service systems", Queueing Syst. **80**, 4, 341–361 (2015a).

Stolyar, A., "Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime", Stoch. Syst. **5**, 2, 239–267 (2015b).

Vasantam, T., A. Mukhopadhyay and R. R. Mazumdar, "Insensitivity of the mean-field limit of loss systems under power-of-d routing", CoRR **abs/1708.09328**, URL `http://arxiv.org/abs/1708.09328` (2017).

Vvedenskaya, N. D., 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).

Wang, W., S. T. Maguluri, R. Srikant and L. Ying, “Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing”, in “IFIP Performance”, (New York City, 2017).

Weber, R. R., “On the optimal assignment of customers to parallel servers”, J. Appl. Probab. **15**, 2, 406–413 (1978).

Winston, W., “Optimality of the shortest line discipline”, J. Appl. Probab. **14**, 1, 181–189 (1977).

Ying, L., “On the approximation error of mean-field models”, in “Proc. Ann. ACM SIGMETRICS Conf.”, (Antibes Juan-les-Pins, France, 2016).

Ying, L., “Stein’s method for mean field approximations in light and heavy traffic regimes”, Proc. ACM Meas. Anal. Comput. Syst. **1**, 1, 12:1–12:27 (2017).

APPENDIX A

GRADIENT BOUNDS

A.1 The First-Order Gradient g'

Lemma 21. *For any* $x \in \left[\eta - \frac{2}{N}, \eta + \frac{2}{N}\right]$, *we have*

$$|g'(x)| \leq \frac{2^r}{\delta N^r}.$$

Proof. For any $x \in \left[\eta - \frac{2}{N}, \eta + \frac{2}{N}\right]$, we have from the closed-form expression of g',

$$\begin{aligned} |g'(x)| \leq & \frac{|x-\eta|^r}{\delta} \\ \leq & \frac{\left(\frac{2}{N}\right)^r}{\delta} = \frac{2^r}{\delta N^r} \end{aligned}$$

□

A.2 The Second-Order Gradient g''

Lemma 22. *For* $x > \eta$, *we have*

$$|g''(x)| \leq \frac{r}{\delta} \left(\max\left\{x-\eta, 0\right\}\right)^{r-1}.$$

Proof. For $x > \eta$, we have

$$g'(x) = \frac{\left(x-\eta\right)^r}{-\delta},$$

which implies

$$g''(x) = \frac{r\left(x-\eta\right)^{r-1}}{-\delta}.$$

and

$$|g''(x)| = \left|\frac{r\left(x-\eta\right)^{r-1}}{-\delta}\right| \leq \frac{r}{\delta}\left(\max\left\{x-\eta, 0\right\}\right)^{r-1}.$$

□

APPENDIX B

STATE SPACE COLLAPSE

B.1 A Tail Bound from Bertsimas *et al.* (2001)

First, we present the following result from Bertsimas *et al.* (2001). The following version of the lemma is from Wang *et al.* (2017), but the result was proven in Bertsimas *et al.* (2001).

Lemma 23. *Let $(X(t) : t \geq 0)$ be a continuous-time Markov chain over a countable state space $\mathbb{X}$. Suppose that it is irreducible, nonexplosive and positive-recurrent, and X denotes the steady state of $(X(t) : t \geq 0)$. Consider a Lyapunov function $V : \mathbb{X} \to \mathbb{R}^+$ and define the drift of V at a state $i \in \mathbb{X}$ as*

$$\Delta V(i) = \sum_{i' \in \mathcal{X}: i' \neq i} q_{ii'}(V(i') - V(i)),$$

where $q_{ii'}$ is the transition rate from i to i'. Suppose that the drift satisfies the following conditions:

(i) There exists constants $\gamma > 0$ and $B > 0$ such that $\Delta V(i) \leq -\gamma$ for any $i \in \mathbb{X}$ with $V(i) > B$.

(ii) $\nu_{\max} := \sup\limits_{i,i' \in \mathbb{X}: q_{ii'} > 0} |V(i') - V(i)| < \infty$.

(iii) $\bar{q} := \sup\limits_{i \in \mathbb{X}} (-q_{ii}) < \infty$.

Then for any non-negative integer j, we have

$$\Pr\left(V(X) > B + 2\nu_{\max} j\right) \leq \left(\frac{q_{\max}\nu_{\max}}{q_{\max}\nu_{\max} + \gamma}\right)^{j+1},$$

where

$$q_{\max} = \sup_{i \in \mathbb{X}} \sum_{i' \in \mathbb{X}: V(i) < V(i')} q_{ii'}.$$

B.2 SSC in Sub-Haffin-Whitt Regime

B.2.1 Proof of Lemma 6

Consider a Lyapunov function in (3.7)

$$V(s) = \min\left\{\sum_{i=2}^{b} s_i, \lambda + \frac{k \log N}{\sqrt{N}} - s_1\right\}. \tag{B.1}$$

Lemma 24. *Under any load balancing algorithm such that $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$, we have for $N \geq \left(\frac{4\bar{k}\log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ that*

$$\triangledown V(s) \leq -\frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} + \frac{1}{\sqrt{N}},$$

for any state $s \in \mathbb{S}$ such that $V(s) \geq \frac{\log N}{\sqrt{N}}$.

Proof. For the Lyapunov function defined in (B.1), the Lyapunov drift is

$$\begin{aligned}&\nabla V(s) = E\left[GV(S)|S=s\right]\\ =&\sum_{i=1}^{b} \lambda N(A_{i-1}(s)-A_i(s))(V(s+e_i)-V(s)) + N(s_i - s_{i+1})(V(s-e_i)-V(s)).\end{aligned}$$

Given $V(s) \geq \frac{\log N}{\sqrt{N}}$, we consider the following two cases.

- Case 1: Assume $\sum_{i=2}^{b} s_i \leq \lambda + \frac{k \log N}{\sqrt{N}} - s_1$. Note that

$$\begin{aligned}&V(s+e_1) \leq \sum_{i=2}^{b} s_i,\ V(s-e_1) = \sum_{i=2}^{b} s_i,\\ &V(s+e_j) \leq \sum_{i=2}^{b} s_i + \frac{1}{N},\ V(s-e_j) = \sum_{i=2}^{b} s_i - \frac{1}{N},\ \forall\ 2 \leq j \leq b.\end{aligned}$$

Furthermore, $V(s) = \sum_{i=2}^{b} s_i \geq \frac{\log N}{\sqrt{N}}$, which implies $s_2 \geq \frac{1}{b-1}\frac{\log N}{\sqrt{N}}$ because $s_2 \geq s_3 \geq \cdots \geq s_b$. Therefore, we have

$$\nabla V(s) \leq \lambda(A_1(s) - A_b(s)) - s_2 \leq -\frac{1}{b-1}\frac{\log N}{\sqrt{N}} + \frac{1}{\sqrt{N}},$$

where the last inequality holds because $\sum_{i=1}^{b} s_i \leq \lambda + \frac{k \log N}{\sqrt{N}}$ implies that $s_1 \leq \lambda + \frac{k \log N}{\sqrt{N}}$ which further implies that $A_1(s) \leq \frac{1}{\sqrt{N}}$.

- Case 2: Assume $\sum_{i=2}^{b} s_i > \lambda + \frac{k \log N}{\sqrt{N}} - s_1$. Note that

$$\begin{aligned}&V(s+e_1) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1 - \frac{1}{N},\ V(s-e_1) \leq \lambda + \frac{k \log N}{\sqrt{N}} - s_1 + \frac{1}{N},\\ &V(s+e_j) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1,\ V(s-e_j) \leq \lambda + \frac{k \log N}{\sqrt{N}} - s_1,\ \forall\ 2 \leq j \leq b.\end{aligned}$$

In this case $\sum_{i=2}^{b} s_i \geq V(s) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1 \geq \frac{\log N}{\sqrt{N}}$, which also implies $s_2 \geq \frac{1}{b-1}\frac{\log N}{\sqrt{N}}$. Therefore, we have

$$\begin{aligned}\nabla V(s) \leq& -\lambda(1 - A_1(s)) + (s_1 - s_2)\\ =& s_1 - s_2 - \lambda + \lambda A_1(s)\\ \leq& (k-1)\frac{\log N}{\sqrt{N}} - s_2 + \lambda A_1(s)\\ \leq& \left(k - 1 - \frac{1}{b-1}\right)\frac{\log N}{\sqrt{N}} + \frac{1}{\sqrt{N}}\\ \leq& -\frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} + \frac{1}{\sqrt{N}}\end{aligned}$$

where the second inequality holds because $s_1 \leq \lambda + (k-1)\frac{\log N}{\sqrt{N}}$ and it implies $A_1(s) \leq \frac{1}{\sqrt{N}}$; the last inequality holds because $s_2 \geq \frac{1}{b-1}\frac{\log N}{\sqrt{N}}$; the last equality holds because $\bar{k} - \frac{r}{\sqrt{N}\log N} \leq k \leq \bar{k}$.

□

From Lemma 24, we have

$$B = \frac{\log N}{\sqrt{N}} \text{ and } \gamma = \frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} - \frac{1}{\sqrt{N}},$$

and it is easy to verify

$$q_{\max} \leq N \text{ and } v_{\max} \leq \frac{1}{N}.$$

Based on Lemma 23 with $j = \frac{\sqrt{N}\log N}{8(b-1)}$, we have

$$\begin{aligned}\Pr\left(V(S) \geq \frac{\tilde{k}\log N}{\sqrt{N}}\right) \leq & \left(\frac{1}{1+\frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} - \frac{1}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{8(b-1)}+1} \\ \leq & \left(1 - \frac{1}{4(b-1)}\frac{\log N}{\sqrt{N}} + \frac{1}{2\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{8(b-1)}} \\ \leq & e^{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}}.\end{aligned}$$

where the second inequality holds because $\frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} \leq 1 + \frac{1}{\sqrt{N}}$ for a large N.

B.2.2 Proof of Lemma 7

Given the SSC result in Lemma 6, we now bound (3.4) by considering two regimes, $V(s) \leq \frac{\tilde{k}\log N}{\sqrt{N}}$ and $V(s) > \frac{\tilde{k}\log N}{\sqrt{N}}$, as follows

$$\begin{aligned}& E\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\ =& E\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbb{I}_{V(S)\leq\frac{\tilde{k}\log N}{\sqrt{N}}}\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \qquad \text{(B.2)} \\ +& E\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)\left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbb{I}_{V(S)>\frac{\tilde{k}\log N}{\sqrt{N}}}\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]. \qquad \text{(B.3)}\end{aligned}$$

To bound (B.2), we consider state s such that $\mathbb{I}_{V(s)\le \frac{\tilde{k}\log N}{\sqrt{N}}} = 1$ and $\mathbb{I}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}} = 1$ because otherwise (B.2) $= 0$. For any state s such that $\sum_{i=1}^{b} s_i > \eta + \frac{1}{N} = \lambda + \frac{k\log N}{\sqrt{N}} + \frac{1}{N}$, we have

$$V(s) = \lambda + \frac{k\log N}{\sqrt{N}} - s_1. \tag{B.4}$$

Given (B.4), $V(s) \le \frac{\tilde{k}\log N}{\sqrt{N}}$ means

$$\begin{aligned}\lambda + \frac{\log N}{\sqrt{N}} - s_1 \le& \left(\tilde{k} - k + 1\right)\frac{\log N}{\sqrt{N}}\\ \le& \left(1 - \frac{1}{4(b-1)}\right)\frac{\log N}{\sqrt{N}}.\end{aligned}$$

Therefore, we have

$$\text{(B.2)} \le \left(1 - \frac{1}{4(b-1)}\right) E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]. \tag{B.5}$$

To bound (B.3), we have

$$\begin{aligned}\text{(B.3)} \le& \frac{b\sqrt{N}}{\log N} E\left[\mathbb{I}_{V(S) > \frac{\tilde{k}\log N}{\sqrt{N}}}\right]\\ \le& \frac{b\sqrt{N}}{\log N} e^{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}}\end{aligned} \tag{B.6}$$

where the first inequality holds because $\sum_{i=1}^{b} s_i - \lambda - \frac{k\log N}{\sqrt{N}} \le \sum_{i=1}^{b} s_i \le b$ and $\lambda + \frac{\log N}{\sqrt{N}} - s_1 \le 1$ for a large N; and the second inequality holds due to Lemma 6.

Based on (B.5) and (B.6), we obtain the following upper bound on (3.4):

$$\begin{aligned}&E\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(\lambda B(S) - \lambda - \frac{\log N}{\sqrt{N}} + S_1\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta}\right]\\ \le& \left(1 - \frac{1}{4(b-1)}\right) E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right]\\ &+ \frac{b\sqrt{N}}{\log N} e^{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}},\end{aligned} \tag{B.7}$$

B.3 SSC in Beyond-Haffin-Whitt Regime

B.3.1 Proof of Lemma 12

Consider Lyapunov function in (4.4)

$$V(s) = \min\left\{\sum_{i=2}^{b} s_i - \frac{k\log N}{N^{1-\alpha}}, 1 - s_1\right\},$$

to study its drift in the following lemma.

Lemma 25. *Given any load balancing in* Π_2, *we have*

$$\nabla V(s) \leq \frac{2}{\sqrt{N}} - \frac{\bar{k}}{b}\frac{\log N}{N^{1-\alpha}},$$

for any state $s \in \mathbb{S}$ *such that*

$$V(s) \geq \frac{1}{4N^{\alpha}}.$$

Proof. Given $V(s) \geq \frac{1}{4N^{\alpha}}$, we have two cases.

- Case 1: $1 - s_1 \geq V(s) = \sum_{i=2}^{b} s_i - \frac{k \log N}{N^{1-\alpha}} \geq \frac{1}{4N^{\alpha}}$, we have

$$\begin{aligned}\nabla V(s) \leq & \lambda(A_1(s) - A_b(s)) - s_2 \\ \leq & \frac{1}{\sqrt{N}} - s_2 \\ \leq & \frac{1}{\sqrt{N}} - \frac{1}{4bN^{\alpha}} - \frac{k}{b}\frac{\log N}{N^{1-\alpha}} \\ \leq & \frac{1}{\sqrt{N}} - \frac{\bar{k}}{b}\frac{\log N}{N^{1-\alpha}}\end{aligned}$$

  where the second inequality holds because we consider load balancing in Π; the third inequality holds because $s_2 \geq \frac{\sum_{i=2}^{b} s_i}{b} \geq \frac{1}{4bN^{\alpha}} + \frac{k}{b}\frac{\log N}{N^{1-\alpha}}$.

- Case 2: $\sum_{i=2}^{b} s_i - \frac{k \log N}{N^{1-\alpha}} \geq V(s) = 1 - s_1 \geq \frac{1}{4N^{\alpha}}$, we have

$$\begin{aligned}\nabla V(s) \leq & -\lambda(1 - A_1(s)) + (s_1 - s_2) \\ = & s_1 - s_2 - \lambda + \lambda A_1(s) \\ \leq & \frac{3}{4N^{\alpha}} - s_2 + \lambda A_1(s) \\ \leq & \frac{1}{\sqrt{N}} + \frac{3}{4N^{\alpha}} - \frac{1}{4bN^{\alpha}} - \frac{k}{b}\frac{\log N}{N^{1-\alpha}} \\ \leq & \frac{2}{\sqrt{N}} - \frac{\bar{k}}{b}\frac{\log N}{N^{1-\alpha}}\end{aligned}$$

where the second inequality holds because $s_1 \leq 1 - \frac{1}{4N^{\alpha}}$; the third inequality holds because we consider load balancing in Π and $s_2 \geq \frac{\sum_{i=2}^{b} s_i}{b} \geq \frac{1}{4bN^{\alpha}} + \frac{k}{b}\frac{\log N}{N^{1-\alpha}}$. □

From Lemma 25, we have

$$B = \frac{1}{4N^{\alpha}},\ \gamma = \frac{\bar{k} - 1}{b}\frac{\log N}{N^{1-\alpha}},\ q_{\max} \leq N \text{ and } v_{\max} \leq \frac{1}{N}.$$

Based on Lemma 23 with $j = \frac{N^{1-\alpha}}{8}$, we have

$$\begin{aligned}\Pr\left(V(S) \geq \frac{1}{2N^\alpha}\right) \leq & \left(\frac{1}{1 - \frac{\bar{k}-1}{b}\frac{\log N}{N^{1-\alpha}}}\right)^{\frac{N^{1-\alpha}}{8}} \\ \leq & \left(1 - \frac{\bar{k}-1}{2b}\frac{\log N}{N^{1-\alpha}}\right)^{\frac{N^{1-\alpha}}{8}} \\ \leq & e^{-\frac{(\bar{k}-1)\log N}{16b}}.\end{aligned}$$

B.3.2 Proof of Lemma 13

According to Lemma 12, we split SSC term (4.1) into two regions, Ω and its complementary $\bar{\Omega}$ as follows

$$\begin{aligned} & E\left[N^\alpha\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{N^{1-\alpha}}\right)^r \left(\lambda + \frac{1}{N^\alpha} - S_1\right)\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\ =& E\left[N^\alpha\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{N^{1-\alpha}}\right)^r \left(\lambda + \frac{1}{N^\alpha} - S_1\right)\mathbb{I}_{V(S)\leq\frac{1}{2N^\alpha}}\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \qquad \text{(B.8)} \\ +& E\left[N^\alpha\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{N^{1-\alpha}}\right)^r \left(\lambda + \frac{1}{N^\alpha} - S_1\right)\mathbb{I}_{V(S)>\frac{1}{2N^\alpha}}\mathbb{I}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]. \qquad \text{(B.9)}\end{aligned}$$

The term (B.8) is related to the region in Ω, where $V(s) \leq \frac{1}{2N^\alpha}$. Given $\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}$, then $V(s) = 1 - s_1$ and $V(s) \leq \frac{1}{2N^\alpha}$ implies $s_1 \geq 1 - \frac{1}{2N^\alpha}$. Therefore, we have

$$\text{(B.8)} \leq \frac{1}{2}E\left[\left(\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right)^r\right].$$

The term (B.9) is related to the region in $\bar{\Omega}$, where we use Lemma 12 and have

$$\text{(B.9)} \leq N^\alpha b^r e^{-\frac{(k-1)\log N}{16b}}.$$

These two terms give Lemma 13.

APPENDIX C

ITERATIVE STATE SPACE COLLAPSE

C.1 A Conditional Tail Bound

To prove the space space collapse results, we first introduce Lemma 26, which will be repeatedly used to obtain probability tail bounds. Lemma 26 allows us to apply Lyapunov drift analysis to in reduced state spaces instead of the complete state space. The lemma is an extension of the tail bound in Bertsimas *et al.* (2001). This Lyapunov drift analysis on reduced state space enables us to iteratively refine the state space in which the system stays at steady state. The lemma was proven in Wang *et al.* (2017). We include the proof so the dissertation is self-contained.

Lemma 26. *Let $(S(t) : t \geq 0)$ be a continuous-time Markov chain over a finite state space $\mathcal{S}$ and is irreducible, so it has a unique stationary distribution π. Consider a Lyapunov function $V : \mathcal{S} \to R^+$ and define the drift of V at a state $s \in \mathcal{S}$ as*

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

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

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

and define

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

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

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

Then

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

with

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

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

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

At steady state, we have

$$\begin{aligned} 0 = & \sum_{V(s) \leq C - \nu_{max}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s) \right) \\ & + \sum_{C - \nu_{max} < V(s) \leq C + \nu_{max}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s) \right) \\ & + \sum_{V(s) > C + \nu_{max}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s) \right). \end{aligned} \tag{C.1}$$

Note $\nabla \hat{V}(s) = \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s)\right)$. We consider three terms in (C.1) as follows:

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

$$\begin{aligned}
&\sum_{C-\nu_{max}<V(s)\leq C+\nu_{max}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s)\right) \\
\leq &\sum_{C-\nu_{max}<V(s)\leq C+\nu_{max}} \pi(s) q_{max} \nu_{max} \\
\leq &q_{max}\nu_{max} \left(\Pr(V(s) > C - \nu_{max}) - \Pr(V(s) > C + \nu_{max})\right)
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{V(s)>C+\nu_{max}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s)\right) \\
= &\sum_{\substack{V(s)>C+\nu_{max} \\ s \in \mathcal{E}}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s)\right) \\
&+ \sum_{\substack{V(s)>C+\nu_{max} \\ s \notin \mathcal{E}}} \pi(s) \sum_{s' \neq s} q_{s,s'} \left(\hat{V}(s') - \hat{V}(s)\right) \\
\leq &-\gamma \Pr\left(V(s) > C + \nu_{max}, s \in \mathcal{E}\right) + \delta \Pr\left(V(s) > C + \nu_{max}, s \notin \mathcal{E}\right) \\
= &-\gamma \Pr\left(V(s) > C + \nu_{max}\right) + (\delta + \gamma) \Pr\left(V(s) > C + \nu_{max}, s \notin \mathcal{E}\right)
\end{aligned}$$

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

Combine three terms above, we have

$$\begin{aligned}
&(q_{max}\nu_{max} + \gamma) \Pr(V(s) > C + \nu_{max}) \\
\leq &q_{max}\nu_{max} \Pr(V(s) > C - \nu_{max}) + (\delta + \gamma) \Pr\left(V(s) > C + \nu_{max}, s \notin \mathcal{E}\right)
\end{aligned}$$

which implies

$$\begin{aligned}
&\Pr(V(s) > C + \nu_{max}) \\
\leq &\frac{q_{max}\nu_{max}}{q_{max}\nu_{max} + \gamma} \Pr(V(s) > C - \nu_{max}) + \frac{\delta + \gamma}{q_{max}\nu_{max} + \gamma} \Pr\left(V(s) > C + \nu_{max}, s \notin \mathcal{E}\right) \\
\leq &\frac{q_{max}\nu_{max}}{q_{max}\nu_{max} + \gamma} \Pr(V(s) > C - \nu_{max}) + \frac{\delta + \gamma}{q_{max}\nu_{max} + \gamma} \Pr\left(s \notin \mathcal{E}\right) \\
= &\alpha \Pr(V(s) > C - \nu_{max}) + \kappa \Pr\left(s \notin \mathcal{E}\right)
\end{aligned}$$

where

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

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

$$\begin{aligned} &\Pr\left(V(s) > B + 2\nu_{max}j\right) \\ \leq &\alpha \Pr\left(V(s) > B + 2(j-1)\nu_{max}\right) + \kappa \Pr\left(s \notin \mathcal{E}\right) \end{aligned} \tag{C.2}$$

By recursively using the inequality (C.2), we have

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

□

As mentioned above, Lemma 26 is an extension of Theorem 1 in Bertsimas *et al.* (2001), where $\mathcal{E} = \mathcal{S}$ is the entire state space and $\Pr\left(s \notin \mathcal{E}\right) = 0$. As suggested in Lemma 26, constructing proper Lyapunov functions are critical to establish the tail bounds. In the following lemmas, we construct a sequence of Lyapunov functions and apply Lemma 26 to establish SSC results.

C.2 SSC under Coxian-2

The proof of this lemma is based on an "iterative" procedure to establish SSC, which is achieved by proving a sequence of four lemmas.

Lemma 27 (An Upper Bound on $S_{1,2}$)**.**

$$\Pr\left(S_{1,2} \leq \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}\right) \geq 1 - e^{-\frac{\mu_1\mu_2 \log^2 N}{40}}.$$

□

Lemma 28 (A Lower Bound on $S_{1,1}$)**.**

$$\Pr\left(S_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}\right) \geq 1 - \frac{5}{\mu_1}\frac{\sqrt{N}}{\log N} e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_1\mu_2}{40}\right)\log^2 N}.$$

□

Lemma 29 (A Lower Bound on $S_{1,2}$)**.**

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

□

Lemma 30 (A Lower Bound on S_1 via $\sum_{i=2}^{b} S_i$)**.**

$$\begin{aligned}&\Pr\left(\min\left\{\lambda + \frac{k\log N}{\sqrt{N}} - S_1, \sum_{i=2}^{b} S_i\right\} \leq \frac{(c_1+\mu_1)\log N}{\sqrt{N}}\right)\\ &\geq 1 - \frac{34}{\mu_1^2\mu_2}\frac{N^{1.5}}{\log^3 N}e^{-\min\left(\frac{\mu_1}{16},\frac{\mu_2}{12},\frac{\mu_1\mu_2}{40}\right)\log^2 N}\end{aligned}$$

for $\min\{\mu_1,\mu_2\} \geq \frac{1}{\log N}$*, where*

$$k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right) \text{ and } c_1 = \frac{w_u b}{w_l}\left(\frac{1+\mu_1+\mu_2}{w_l} + 2\mu_1\right) + 2\mu_1.$$

□

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

$$\begin{aligned}\tilde{\mathcal{S}}_1 &= \left\{s \middle| s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}} \text{ and } s_{1,2} \geq \frac{p\lambda}{\mu_2} - \frac{\mu_1\log N}{\sqrt{N}}\right\}\\ \tilde{\mathcal{S}}_2 &= \left\{s \middle| \min\left\{\lambda + \frac{k\log N}{\sqrt{N}} - s_1, \sum_{i=2}^{b} s_i\right\} \leq \frac{(c_1+\mu_1)\log N}{\sqrt{N}}\right\}.\end{aligned}$$

According to the union bound and Lemmas 28-30, we have

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

where the second inequality holds for a large N such that $\log N \geq \frac{3.5}{\min\left(\frac{\mu_1}{16},\frac{\mu_2}{12},\frac{\mu_1\mu_2}{40}\right)}$.

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

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

we either have

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

which implies

$$s_1 \geq \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}} \text{ or } \sum_{i=2}^{b} s_i \leq \lambda + \frac{k\log N}{\sqrt{N}} - s_1,$$

which implies

$$\sum_{i=1}^{b} s_i \leq \lambda + \frac{k\log N}{\sqrt{N}}.$$

Note that

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

and

$$\tilde{\mathcal{S}}_1 \cap \left\{ s \,\middle|\, \sum_{i=1}^{b} s_i \leq \lambda + \frac{k\log N}{\sqrt{N}} \right\} \subseteq \mathcal{S}_{ssc_1}.$$

We, therefore, have

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

and

$$\Pr\left(S \notin \mathcal{S}_{ssc}\right) \leq \Pr\left(S \notin \tilde{\mathcal{S}}_1 \cap \mathcal{S}_2\right) \leq \frac{3}{N^2},$$

so Lemma 19 holds.

We next present the iterative SSC approach for proving Lemma 27-Lemma 30. The first three lemmas are on the upper and lower bounds on $S_{1,1}$ and $S_{1,2}$, illustrated in Fig. C.1, which shows that both $S_{1,1}$ and $S_{1,2}$ are close to its equilibrium values, in particular, with a high probability, $S_{1,1} \geq \lambda/\mu_1 - \frac{\log N}{\sqrt{N}}$ and $S_{1,2} \geq p\lambda/\mu_2 - \frac{\mu_1 \log N}{\sqrt{N}}$. However, these two low bounds do not guarantee the total departure rate, which is $(1-p)\mu_1 S_{1,1} + \mu_2 S_{1,2}$, is larger than the arrival rate λ. Therefore, we need Lemma 30 to guarantee sufficient fraction of busy servers S_1 such that the total departure rate is "larger than" the arrival rate λ. We therefore need Lemma 30 to further establish a lower bound on S_1 unless the total normalized queue length $\sum_{i=1}^{b} S_i$ is small.

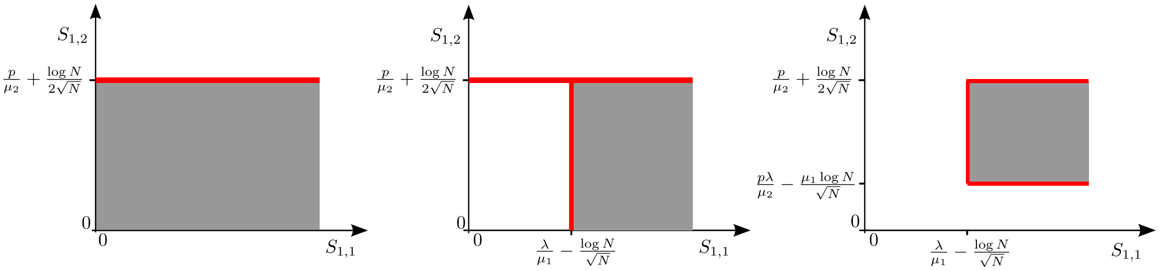
$S_{1,2}$
$\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}$
0
0
$S_{1,1}$
$S_{1,2}$
$\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}$
0
0
$\frac{\lambda}{\mu_1}-\frac{\log N}{\sqrt{N}}$
$S_{1,1}$
$S_{1,2}$
$\frac{p}{\mu_2}+\frac{\log N}{2\sqrt{N}}$
$\frac{p\lambda}{\mu_2}-\frac{\mu_1\log N}{\sqrt{N}}$
0
0
$\frac{\lambda}{\mu_1}-\frac{\log N}{\sqrt{N}}$
$S_{1,1}$

Figure C.1: Bounds (Red Lines) on $S_{1,1}$ and $S_{1,2}$.

C.2.1 Proof of Lemma 27: An Upper Bound on $S_{1,2}$.

To prove Lemma 27, we first establish a Lyaponuv drift analysis for $\mathcal{E} = \mathcal{S}$ (the entire state space) in Lemma 31.

Lemma 31. *Consider Lyapunov function*

$$V(s) = s_{1,2} - \frac{p}{\mu_2}.$$

When $V(s) \geq \frac{\log N}{4\sqrt{N}}$, *we have*

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

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

$$\begin{aligned}\nabla V(s) =& p\mu_1 s_{1,1} - \mu_2 s_{1,2} && \text{(C.3)}\\ \leq& p\mu_1 - (p\mu_1 + \mu_2)s_{1,2} && \text{(C.4)}\\ =& \mu_1(p - \mu_2 s_{1,2}) \leq -\frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}} && \text{(C.5)}\end{aligned}$$

(C.3) to (C.4) holds because $s_{1,1} = s_1 - s_{1,2} \leq 1 - s_{1,2}$; (C.4) to (C.5) holds because $\frac{1}{\mu_1} + \frac{p}{\mu_2} = 1$ implies $p\mu_1 + \mu_2 = \mu_1\mu_2$. □

From Lemma 31, we know $B = \frac{\log N}{4\sqrt{N}}$ and $\gamma = \frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}}$. According to the definition of $q_{\max}$ and $\nu_{\max}$, we have $q_{\max} = N$ and $\nu_{\max} = \frac{1}{N}$. Since $\mathcal{E} = \mathcal{S}$ is the entire space, then $\Pr(s \notin \mathcal{E}) = 0$, we use Lemma 26 (or Theorem 1 in Bertsimas *et al.* (2001)) to obtain the following tail bound with $j = \frac{\sqrt{N}\log N}{8}$,

$$\begin{aligned}\Pr\left(V(S) \geq B + 2\nu_{max}j\right) =& \Pr\left(S_{1,2} - \frac{p}{\mu_2} \geq \frac{\log N}{2\sqrt{N}}\right) && \text{(C.6)}\\ \leq& \left(\frac{1}{1 + \frac{\mu_1\mu_2}{4}\frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{8}} && \text{(C.7)}\\ \leq& \left(1 - \frac{\mu_1\mu_2}{5}\frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{8}} && \text{(C.8)}\\ \leq& e^{-\frac{\mu_1\mu_2\log^2 N}{40}}\end{aligned}$$

- (C.6) holds by substituting $B = \frac{\log N}{4\sqrt{N}}$, $\nu_{max} = \frac{1}{N}$ and $j = \frac{\sqrt{N}\log N}{8}$;
- (C.6) to (C.7) holds based on Lemma 31;
- (C.7) to (C.8) holds because $\mu_1\mu_2 \leq \frac{\sqrt{N}}{\log N}$ for a large N.

C.2.2 Proof of Lemma 28: A Lower Bound on $S_{1,1}$.

To prove Lemma 28, we first establish a Lyaponuv drift analysis in Lemma 32.

Lemma 32. *Consider Lyapunov function*

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

we have

- $\nabla V(s) \leq -\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}}$, *when*
$$V(s) \geq \frac{\log N}{2\sqrt{N}} \quad \textit{and} \quad s_{1,2} \leq \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}};$$
- $\nabla V(s) \leq 1$, *when*
$$V(s) \geq \frac{\log N}{2\sqrt{N}} \quad \textit{and} \quad s_{1,2} \geq \frac{p}{\mu_2} + \frac{\log N}{2\sqrt{N}}.$$

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

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

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

$$\begin{aligned}
\nabla V(s) =& -\lambda 1_{\{s_1<1\}} + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.9)}\\
\leq& -\lambda + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.10)}\\
\leq& -\lambda + \mu_1 s_{1,1} && \text{(C.11)}\\
\leq& -\frac{\mu_1}{2}\frac{\log N}{\sqrt{N}} && \text{(C.12)}\\
\leq& -\frac{\mu_1}{3}\frac{\log N}{\sqrt{N}},
\end{aligned}$$

where

- (C.9) to (C.10) holds because $1_{\{s_1<1\}} = 1$ under JSQ;
- (C.11) to (C.12) holds because $s_{1,1} \leq \frac{\lambda}{\mu_1} - \frac{\log N}{2\sqrt{N}}$.

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

$$\nabla V(s) = -\lambda 1_{\{s_1<1\}} + \mu_1 s_{1,1} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} \leq \mu_1 s_{1,1} < 1.$$

□

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

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

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

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

Based on Lemma 26 with $j = \frac{\sqrt{N} \log N}{4}$, we have

$$\begin{aligned}
\Pr\left(V(s) \geq B + 2\nu_{max} j\right) =& \Pr\left(\frac{\lambda}{\mu_1} - S_{1,1} \geq \frac{\log N}{\sqrt{N}}\right) && \text{(C.13)} \\
\leq& \left(\frac{1}{1 + \frac{\mu_1}{3} \frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N} \log N}{4}} + \beta \Pr\left(S_{1,2} \notin \mathcal{E}\right) && \text{(C.14)} \\
\leq& \left(1 - \frac{\mu_1}{4} \frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N} \log N}{4}} + \frac{4}{\mu_1} \frac{\sqrt{N}}{\log N} e^{-\frac{\mu_1 \mu_2 \log^2 N}{40}} && \text{(C.15)} \\
\leq& e^{-\frac{\mu_1 \log^2 N}{16}} + \frac{4}{\mu_1} \frac{\sqrt{N}}{\log N} e^{-\frac{\mu_1 \mu_2 \log^2 N}{40}} \\
\leq& \frac{5}{\mu_1} \frac{\sqrt{N}}{\log N} e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_1 \mu_2}{40}\right) \log^2 N},
\end{aligned}$$

where

- (C.13) holds by substituting $B = \frac{\log N}{2\sqrt{N}}$, $\nu_{max} = \frac{1}{N}$ and $j = \frac{\sqrt{N} \log N}{4}$;
- (C.13) to (C.14) holds based on Lemma 32;
- (C.14) to (C.15) holds because (i) in the first term in (C.15), $\mu_1 \leq \frac{\sqrt{N}}{\log N}$ for a large N, and (ii) the second term in (C.14) can be bounded by applying Lemma 27.

C.2.3 Proof of Lemma 29: A Lower Bound on $S_{1,2}$.

Lemma 33. *Consider Lyapunov function*

$$V(s) = \frac{p\lambda}{\mu_2} - s_{1,2},$$

we have

- $\nabla V(s) \leq -\frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}$, *when*
$$V(s) \geq \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}} \quad \textit{and} \quad s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}};$$
- $\nabla V(s) \leq 1$, *when*
$$V(s) \geq \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}} \quad \textit{and} \quad s_{1,1} \leq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}.$$

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

$$\begin{aligned}\nabla V(s) =& - \left(p\mu_1 s_{1,1} - \mu_2 s_{1,2}\right) & \text{(C.16)}\\ \leq& - \left(p\lambda - \frac{p\mu_1 \log N}{\sqrt{N}} - \mu_2 s_{1,2}\right) & \text{(C.17)}\\ \leq& - \frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}, & \text{(C.18)}\end{aligned}$$

where

- (C.16) to (C.17) holds because $s_{1,1} \geq \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$;
- (C.17) to (C.18) holds because $s_{1,2} \leq \frac{p\lambda}{\mu_2} - \left(\frac{p\mu_1}{\mu_2} + \frac{1}{2}\right)\frac{\log N}{\sqrt{N}}$.

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

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

□

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

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

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

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

Based on Lemma 26 with $j = \frac{\sqrt{N}\log N}{4}$, we have

$$\begin{aligned}
\Pr\left(V(s) \geq B + 2\nu_{max} j\right) =& \Pr\left(\frac{p\lambda}{\mu_2} - S_{1,2} \geq \left(\frac{p\mu_1}{\mu_2} + 1\right)\frac{\log N}{\sqrt{N}}\right) && \text{(C.19)}\\
\leq& \left(\frac{1}{1 + \frac{\mu_2}{2}\frac{\log N}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{4}} + \frac{2}{\mu_2}\frac{\sqrt{N}}{\log N}\Pr\left(S_{1,1} \notin \mathcal{E}\right) && \text{(C.20)}\\
\leq& \left(1 - \frac{\mu_2}{3}\frac{\log N}{\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{4}} + \frac{3}{\mu_2}\frac{\sqrt{N}}{\log N}\Pr\left(S_{1,1} \notin \mathcal{E}\right) && \text{(C.21)}\\
\leq& e^{-\frac{\mu_2 \log^2 N}{12}} + \frac{15}{\mu_1\mu_2}\frac{N}{\log^2 N}e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_1\mu_2}{40}\right)\log^2 N} && \text{(C.22)}\\
\leq& \frac{16}{\mu_1\mu_2}\frac{N}{\log^2 N}e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1\mu_2}{40}\right)\log^2 N},
\end{aligned}$$

where

- (C.19) holds by substituting B, ν_{max} and j;
- (C.19) to (C.20) holds due to Lemma 33;
- (C.20) to (C.21) holds because $\mu_2 \leq \frac{\sqrt{N}}{\log N}$ for a large N for the first term in (C.21);
- (C.21) to (C.22) holds by applying Lemma 28 to obtain the tail bound in the second term in (C.22).

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

C.2.4 Proof of Lemma 30: SSC on S_1 and $\sum_{i=2}^{b} S_i$.

Define $L_{1,1} = \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$ and $L_{1,2} = \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$. Recall

$$w_u = \max((1-p)\mu_1, \mu_2) \text{ and } w_l = \min((1-p)\mu_1, \mu_2),$$

$$k = \left(1 + \frac{w_u b}{w_l}\right)\left(\frac{1 + \mu_1 + \mu_2}{w_l} + 2\mu_1\right) \text{ and } c_1 = \frac{w_u b}{w_l}\left(\frac{1 + \mu_1 + \mu_2}{w_l} + 2\mu_1\right) + 2\mu_1.$$

Lemma 34. *Consider Lyapunov function*

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

we have

- $\nabla V(s) \leq -\frac{w_u \mu_1 \log N}{\sqrt{N}}$, *when* $V(s) \geq \frac{c_1 \log N}{\sqrt{N}}$ *with* $s_{1,1} \geq L_{1,1}$ *and* $s_{1,2} \geq L_{1,2}$;
- $\nabla V(s) \leq w_u$, *when* $V(s) \geq \frac{c_1 \log N}{\sqrt{N}}$ *with* $s_{1,1} \leq L_{1,1}$ *or* $s_{1,2} \leq L_{1,2}$.

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

$$s_1 \leq \lambda + \frac{(k-c_1)\log N}{\sqrt{N}} = \lambda + \frac{1+\mu_1+\mu_2}{w_l}\frac{\log N}{\sqrt{N}}, \tag{C.23}$$

$$\sum_{i=2}^{b} s_i \geq \frac{c_1 \log N}{\sqrt{N}}. \tag{C.24}$$

We have two observations based on (C.23) and (C.24):

- (C.23) implies $1_{\{s_1<1\}} = 1$ under JSQ;
- (C.24) implies $s_2 \geq \frac{c_1}{b}\frac{\log N}{\sqrt{N}}$ because $s_2 \geq s_3 \geq \cdots \geq s_b$, and we have

$$(1-p)\mu_1 s_{2,1} + \mu_2 s_{2,2} \geq w_l s_2 \geq \frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}} \tag{C.25}$$

We study the Lyapunov dirft and consider two cases:

- Supppose $\lambda + \frac{k\log N}{\sqrt{N}} - s_1 \geq \sum_{i=2}^{b} s_i \geq \frac{c_1 \log N}{\sqrt{N}}$. In this case, $V(s) = \sum_{i=2}^{b} s_i$, and

$$\begin{aligned}\nabla V(s) \leq& \lambda 1_{\{s_1=1\}} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} &\text{(C.26)}\\ \leq& -(1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} &\text{(C.27)}\\ \leq& -\frac{w_l c_1}{b}\frac{\log N}{\sqrt{N}} &\text{(C.28)}\\ \leq& -\frac{2 w_u \mu_1 \log N}{\sqrt{N}} &\text{(C.29)}\end{aligned}$$

  where

  - (C.26) to (C.27) holds because $1_{\{s_1=1\}} = 0$ under JSQ;
  - (C.27) to (C.28) holds because (C.25);
  - (C.28) to (C.29) holds because $c_1 \geq \frac{w_u b}{w_l} 2\mu_1$.

- Suppose $\sum_{i=2}^{b} s_i > \lambda + \frac{k \log N}{\sqrt{N}} - s_1 \geq \frac{c_1 \log N}{\sqrt{N}}$. In this case, $V(s) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1$, and

$$\begin{aligned}
\nabla V(s) \leq & -\lambda 1_{\{s_1<1\}} + (1-p)\mu_1 s_{1,1} + \mu_2 s_{1,2} - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.30)}\\
\leq & -\lambda + w_u s_1 - (w_u - (1-p)\mu_1)\, s_{1,1} - (w_u - \mu_2)\, s_{1,2} \\
& - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.31)}\\
\leq & -\lambda + w_u(s_1 - L_{1,1} - L_{1,2}) + ((1-p)\mu_1 L_{1,1} + \mu_2 L_{1,2}) \\
& - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.32)}\\
= & (w_u(k - c_1 + 1 + \mu_1) - (1-p)\mu_1 - \mu_1\mu_2) \frac{\log N}{\sqrt{N}} \\
& - (1-p)\mu_1 s_{2,1} - \mu_2 s_{2,2} && \text{(C.33)}\\
\leq & (w_u(k - c_1 + 1 + \mu_1) - (1-p)\mu_1 - \mu_1\mu_2) \frac{\log N}{\sqrt{N}} - \frac{w_l c_1}{b} \frac{\log N}{\sqrt{N}} && \text{(C.34)}\\
= & w_u \left(k - \left(1 + \frac{w_l}{w_u b}\right) c_1 + \mu_1\right) \frac{\log N}{\sqrt{N}} - ((1-p)\mu_1 + \mu_1\mu_2 - w_u) \frac{\log N}{\sqrt{N}} && \text{(C.35)}\\
\leq & w_u \left(k - \left(1 + \frac{w_l}{w_u b}\right) c_1 + \mu_1\right) \frac{\log N}{\sqrt{N}} && \text{(C.36)}\\
\leq & -\frac{w_u \mu_1 \log N}{\sqrt{N}}, && \text{(C.37)}
\end{aligned}$$

where

  - (C.30) to (C.31) holds by adding and substructing $w_u s_1 = w_u(s_{1,1} + s_{1,2})$;
  - (C.31) to (C.32) holds because $s_{1,1}$ and $s_{1,2}$ taking the lower bounds at $L_{1,1}$ and $L_{1,2}$ gives an upper bound;
  - (C.32) to (C.33) holds by substituting $L_{1,1} = \frac{\lambda}{\mu_1} - \frac{\log N}{\sqrt{N}}$, $L_{1,2} = \frac{p\lambda}{\mu_2} - \frac{\mu_1 \log N}{\sqrt{N}}$ and $s_1 \leq \lambda + \frac{(k-c_1)\log N}{\sqrt{N}}$. We have $s_1 - L_{1,1} - L_{1,2} = (k - c_1 + 1 + \mu_1)\frac{\log N}{\sqrt{N}}$ and $(1-p)\mu_1 L_{1,1} + \mu_2 L_{1,2} = \lambda - ((1-p)\mu_1 + \mu_1\mu_2)\frac{\log N}{\sqrt{N}}$.
  - (C.33) to (C.34) holds by substituting the lower bound of $(1-p)\mu_1 s_{2,1} + \mu_2 s_{2,2}$ in (C.25);
  - (C.34) to (C.35) holds by combining the terms with c_1;
  - (C.35) to (C.36) holds because $(1-p)\mu_1 + \mu_1\mu_2 - w_u = \mu_1 + \mu_2 - w_u \geq 0$;
  - (C.36) to (C.37) holds because $k - \left(1 + \frac{w_l}{w_u b}\right) c_1 \leq -2\mu_1$.

□

Let $\mathcal{E} = \{s \mid s_{1,1} \geq L_{1,1},\ s_{1,2} \geq L_{1,2}\}$ and $V(s) = \min\left\{\lambda + \frac{k \log N}{\sqrt{N}} - s_1, \sum_{i=2}^{b} s_i\right\}$ satisfying the following two conditions based on Lemma 34:

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

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

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

Based on Lemma 26 with $j = \frac{\mu_1 \sqrt{N} \log N}{2}$, we have

$$\begin{aligned}
&\Pr\left(V(S) \geq B + 2\nu_{max} j\right) \\
=&\Pr\left(V(S) \geq \frac{c_1 \log N}{\sqrt{N}} + \frac{\mu_1 \log N}{\sqrt{N}}\right) && \text{(C.38)} \\
\leq&\left(\frac{1}{1 + \frac{w_u \mu_1 \log N}{\sqrt{N}}}\right)^{\frac{\mu_1 \sqrt{N} \log N}{2}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right) \Pr\left(s \notin \mathcal{E}\right) && \text{(C.39)} \\
\leq&\left(1 - \frac{w_u \mu_1}{2} \frac{\log N}{\sqrt{N}}\right)^{\frac{\mu_1 \sqrt{N} \log N}{2}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right) \Pr\left(s \notin \mathcal{E}\right) && \text{(C.40)} \\
\leq&e^{-\frac{w_u \mu_1^2 \log^2 N}{4}} + \left(\frac{\sqrt{N}}{\mu_1 \log N} + 1\right) \frac{32}{\mu_1 \mu_2} \frac{N}{\log^2 N} e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1 \mu_2}{40}\right) \log^2 N} && \text{(C.41)} \\
\leq&\frac{34}{\mu_1^2 \mu_2} \frac{N^{1.5}}{\log^3 N} e^{-\min\left(\frac{\mu_1}{16}, \frac{\mu_2}{12}, \frac{\mu_1 \mu_2}{40}\right) \log^2 N},
\end{aligned}$$

where

- (C.38) holds holds by substituting B, ν_{max} and j;
- (C.38) to (C.39) holds based on Lemma 34;
- (C.39) to (C.40) holds $w_u \mu_1 \leq \frac{\sqrt{N}}{\log N}$ for a large N for the first term in (C.40);
- (C.40) to (C.41) holds by applying the union bound on $\Pr\left(s \notin \mathcal{E}\right)$ such that

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