J. Appl. Prob. **57,** 578–596 (2020)
*doi:*10.1017/jpr.2020.13

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

XIN LIU,* ** AND
LEI YING,* *** *Arizona State University*

Abstract

We study a class of load-balancing algorithms for many-server systems (N servers). Each server has a buffer of size $b-1$ with $b=O(\sqrt{\log N})$, i.e. a server can have at most one job in service and $b-1$ jobs queued. We focus on the steady-state performance of load-balancing algorithms in the heavy traffic regime such that the load of the system is $\lambda=1-\gamma N^{-\alpha}$ for $0<\alpha<0.5$ and $\gamma>0$, which we call the sub-Halfin–Whitt regime ($\alpha=0.5$ is the so-called Halfin–Whitt regime). We establish a sufficient condition under which the probability that an incoming job is routed to an idle server is 1 asymptotically (as $N\to\infty$) at steady state. The class of load-balancing algorithms that satisfy the condition includes join-the-shortest-queue, idle-one-first, join-the-idle-queue, and power-of-d-choices with $d\geq\frac{r}{\gamma}N^{\alpha}\log N$ (r a positive integer). The proof of the main result is based on the framework of Stein's method. A key contribution is to use a simple generator approximation based on state space collapse.

Keywords: Many-server systems; load balancing; heavy traffic; Stein's method; mean-field model; steady state; state space collapse; asymptotic zero delay

2010 Mathematics Subject Classification: Primary 90B15
Secondary 60K25; 68M20

1. Introduction

This paper 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 a buffer. We assume jobs arrive according to a Poisson process with rate λN, where $\lambda=1-\gamma N^{-\alpha}$ for $0<\alpha<0.5$ and $\gamma>0$, and have independent and identically distributed (i.i.d.) exponential service times with mean 1. 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) [9], join-the-idle-queue (JIQ) [11, 14], and power-of-d-choices (Pod) with $d\geq\frac{r}{\gamma}N^{\alpha}\log N$ [12, 17], 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 $d\geq\frac{r}{\gamma}N^{\alpha}\log N$, 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 $r\leq\frac{\log N}{72(b-1)^2}$.

Received 4 September 2018; revision received 16 December 2019.
* Postal address: School of Electrical, Computer and Energy Engineering, 436 Goldwater Center for Science and Engineering, 650 E Tyler Mall, Arizona State University, Tempe, AZ 85287, USA.
** Email address: xliu272@asu.edu
*** Email address: lei.ying.2@asu.edu

1.1. Related work and our contributions

Performance 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 (called the Erlang C model) was one of the earliest subjects studied. For systems with distributed queues where each server maintains a separate queue, it is well known that the JSQ algorithm is delay optimal [19, 20] under fairly general conditions. However, the exact stationary distribution of many-server systems under JSQ remains an open problem. A recent breakthrough in this area is [6], which shows that in the Halfin–Whitt regime ($\alpha = 0.5$) 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 $\Theta(\sqrt{N})$ servers have two jobs (one in service and one in a buffer). This seminal work has led to several significant developments: (i) [3] proved that the stationary distribution indeed converges to the stationary distribution of the two-dimensional diffusion limit based on Stein's method; (ii) via stochastic coupling, [13] 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 $\alpha < 1/6$, [10] 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. Interested readers can find a comprehensive survey of recent results in [16].

Let S_i denote the fraction of servers with at least i jobs, including the one in service, at steady state. In this paper 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 $\lambda + \frac{\bar{k}\log N}{\sqrt{N}}$, then the following bound holds for any positive integer $r \le \frac{\log N}{72(b-1)^2}$:

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

This result implies that

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

i.e. the expected queue length per server exceeds λ by at most $\frac{11\lambda b+11}{N^{r(0.5-\alpha)}}$, and that under JSQ, I1F, JIQ, and Pod ($d \ge \frac{r}{\gamma}N^{\alpha}\log N$), 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 2.1.

To 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 [1, 3, 10]. [1] and [3] analyze the steady-state distribution of JSQ in the Halfin–Whitt regime, and [10] studies Pod with $\alpha < 1/6$. This paper complements all three, as it applies to a class of load-balancing algorithms and to any sub-Halfin–Whitt regime. Table 1 summarizes the comparison between our results and existing ones in the literature. Since the existing works focused on steady-state queue length, we also present our results in terms of $\mathrm{E}[S_i]$ for comparison purposes.

Similar to [3] and [10], the result of this paper is proved using the mean-field approximation (fluid-limit approximation) based on Stein's method. The execution of Stein's method here, however, is quite different from [3, 10].

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

$$\dot{x} = -\frac{\log N}{\sqrt{N}}. \tag{1}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

TABLE 1: Our contributions and related work.

	Load balancing	Servers with one job	Servers with two or more jobs
Steady state ($\alpha < 0.5$) Xin & Ying	JSQ Pod ($d \geq \frac{r \log N}{1-\lambda}$) I1F JIQ	$N - \Theta\left(N^{1-\alpha}\right)$	$O\left(bN^{1-r(0.5-\alpha)}\right)$
Steady state ($\alpha = 0.5$) Braverman [3]	JSQ	$N - \Theta\left(\sqrt{N}\right)$	$\Theta\left(\sqrt{N}\right)$
Process level ($\alpha = 0.5$) Eschenfeldt & Gamarnik [6]	JSQ	$N - \Theta\left(\sqrt{N}\right)$	$\Theta\left(\sqrt{N}\right)$
Process level ($\alpha = 0.5$) Mukherjee et al. [13]	Pod ($d = \frac{\log N}{\gamma(1-\lambda)}$)	$N - \Theta\left(\sqrt{N}\right)$	$\Theta\left(\sqrt{N}\right)$

x can be viewed as a fluid approximation of the normalized total 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 (1) 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 total queue length is close to one.
- Regime 3: The state is in neither regime 1 nor regime 2. In this case we apply the tail bound in [2] 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 [15] (and called the drift-based fluid limits method) for fluid-limit analysis and in [4, 5] for steady-state diffusion approximation, the power of Stein's method for steady-state approximations has been recognized in a number of recent papers [3, 4, 5, 7, 8, 15, 21, 22].

The surprising part of our analysis is that the simple fluid system, which only 'partially' approximates the generator of the stochastic system, is sufficient for executing Stein's method when combining with the state-space-collapse approach. The advantage of using such a simple fluid system is that Stein's equation can be solved easily, which is often the key difficulty of applying Stein's method for queueing systems.

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

2. Model and main results

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

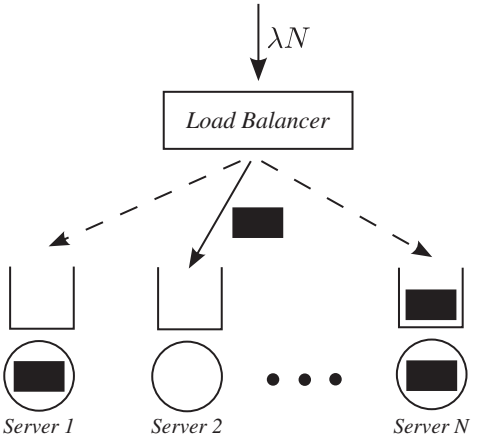

FIGURE 1. Load balancing in many-server systems.

1. We consider the sub-Halfin–Whitt regime such that $\lambda = 1 - \gamma N^{-\alpha}$ for some $0 < \alpha < 0.5$ and $\gamma > 0$. As shown in Figure 1, each server maintains a separate queue and we assume a buffer size of $b - 1$ (i.e. each server can have one job in service and $b - 1$ jobs in the queue).

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 $S_i(t) = 0$ for all $i \geq b + 1$ and all $t \geq 0$ for notational convenience. Furthermore, define the set $\mathbb{S} \subseteq \mathbb{R}^b$ such that

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

We then have $S(t) = [S_1(t), S_2(t), \ldots, S_b(t)]^\top \in \mathbb{S}$ for any $t \geq 0$. We consider a class of 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. Let $S \in \mathbb{S}$ be the random variables having the stationary distribution of $(S(t) : t \geq 0)$. Note that λ, $(S(t) : t \geq 0)$, and S all depend on N, the number of servers in the system. 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.

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

The main result of this paper is the following theorem.

Theorem 2.1. *Assume* $\lambda = 1 - \gamma N^{-\alpha}$ *for* $0 < \alpha < 0.5$ *and* $\gamma > 0$, *and* $b \leq 1 + \frac{\sqrt{\log N}}{9}$. *If a load-balancing algorithm guarantees* $A_1(s) \leq \frac{1}{\sqrt{N}}$ *for any* $s \in \mathbb{S}$ *such that* $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$, *then for any integers* N *and* r *such that* $N \geq \left(\frac{4\bar{k} \log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ *and* $1 \leq r \leq \frac{\log N}{72(b-1)^2}$, *the following bound holds at steady state:*

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

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

Note that the expectation in Theorem 2.1 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 the fraction $\frac{\gamma}{N^\alpha} - \frac{\bar{k} \log N}{\sqrt{N}}$ of servers are idle, an incoming job should be routed to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$. Note that $N \geq \left(\frac{4\bar{k} \log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{\gamma}{N^\alpha} \geq \frac{4\bar{k} \log N}{\sqrt{N}}$, which guarantee that $\lambda + \frac{\bar{k} \log N}{\sqrt{N}} < 1$ and $\frac{\gamma}{N^\alpha} > \frac{\bar{k} \log N}{\sqrt{N}}$. There are several well-known policies that satisfy this condition:

- JSQ routes an incoming job to the least-loaded server in the system, so $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$.
- I1F routes an incoming job to an idle server, if available, or to a server with one job, if available. Otherwise, the job is routed to a randomly selected server. Therefore, $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$.
- JIQ routes an incoming job to an idle server if possible; otherwise, it routes the job to a server chosen uniformly at random. Therefore, $A_1(s) = 0$ when $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$.
- 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 $d \geq \frac{r}{\gamma} N^\alpha \log N$, $A_1(s) \leq \frac{1}{\sqrt{N}}$ when $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$.

A direct consequence of Theorem 2.1 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 that $\mathcal{B} \subseteq \mathcal{W}$, 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 that are not blocked at steady state. We have the following results based on the main theorem.

Corollary 2.1. *Assume* $\lambda = 1 - \gamma N^{-\alpha}$ *for* $0 < \alpha < 0.5$ *and* $\gamma > 0$, *and* $b \leq 1 + \frac{\sqrt{\log N}}{9}$. *If a load-balancing algorithm guarantees* $A_1(s) \leq \frac{1}{N^{0.5r}}$ *for any* $s \in \mathbb{S}$ *such that* $s_1 \leq \lambda + \frac{\bar{k} \log N}{\sqrt{N}}$, *then the following results hold for any integers* N *and* r *such that* $N \geq \left(\frac{4\bar{k} \log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ *and* $1 \leq r \leq \frac{\log N}{72(b-1)^2}$ *at steady state:*

$$\textit{Waiting time per job}: \quad \mathbb{E}[W] \leq \frac{11b}{\gamma^r N^{r(0.5-\alpha)}},$$

$$\textit{Waiting probability}: \quad p_{\mathcal{W}} \leq \frac{11}{\gamma^r N^{r(0.5-\alpha)}},$$

$$\textit{Fraction of busy servers}: \quad \lambda - \frac{11}{\gamma^r N^{r(0.5-\alpha)}} \leq \mathbb{E}[S_1] \leq \lambda,$$

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

The proof of this corollary is an application of the Markov inequality and Little's law, and can be found in Section 4. We note that the corollary above requires $A_1(s) \leq \frac{1}{N^{0.5r}}$ for any $s \in \mathbb{S}$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

such that $s_1 \le \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$, which is more restrictive than the assumption in the theorem that only requires $A_1(s) \le \frac{1}{\sqrt{N}}$ for the same s. However, it is easy to verify that JSQ, I1F, JIQ, and Pod with $d \ge \frac{r}{\gamma} N^{\alpha} \log N$ also satisfy this condition. We further remark that the probability of waiting and the expected waiting time are both $O\big(\frac{b}{N^{r(0.5-\alpha)}}\big)$. Under the assumption that $b = O\big(\sqrt{\log N}\big)$ 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 (2), we have

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

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 is the expected number of buffered jobs per server almost zero, but so also is the total number of buffered jobs in all N servers.

3. Proof of Theorem 2.1

In this section, we present the proof, based on Stein's method, of our main theorem. As modularized in [4], this approach includes three key ingredients: generator approximation, gradient bounds, and state space collapse (SSC).

3.1. 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.

$$A_i(s) = \mathbb{P}(\text{an incoming job is routed to a server with at least } i \text{ jobs} \mid S(t) = s).$$

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 $\lambda N\,(A_{i-1}(s) - A_i(s))$, which occurs when an arrival comes and is routed to a 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 the CTMC $(S(t) : t \ge 0)$. Given a function $f : \mathbb{S} \to \mathbb{R}$, we have

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

For any bounded function $f : \mathbb{S} \to \mathbb{R}$ we have $\mathbb{E}[Gf(S)] = 0$, 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 moment bounds on the following function:

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

where $\bar{k} - \frac{r}{\sqrt{N}\log N} \le k \le \bar{k}$. The moment bounds are used to bound the probability that the total number of jobs in the system $\left(N\sum_{i=1}^{b} S_i\right)$ exceeds $N\lambda + k\sqrt{N}\log N$ at steady state, and can also be used to bound the probability that an incoming job is routed to an idle server in Corollary 2.1.

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

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

and a function $g(x)$ which is the solution of the following Stein's equation [21]:

$$g'(x)\left(-\frac{\log N}{\sqrt{N}}\right) = \left(\max\left\{x - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right)^r \quad \text{for all } x, \tag{3}$$

where $g'(x) = \frac{\mathrm{d}g(x)}{\mathrm{d}x}$ and r is a positive integer. The left-hand side of (3) is applying the generator of the simple fluid system to the function $g(x)$, i.e.

$$\frac{\mathrm{d}g(x)}{\mathrm{d}t} = g'(x)\dot{x} = g'(x)\left(-\frac{\log N}{\sqrt{N}}\right).$$

It is easy to verify that the solution to (3) is

$$g(x) = -\frac{\sqrt{N}}{(r+1)\log N}\left(x - \lambda - \frac{k\log N}{\sqrt{N}}\right)^{r+1}\mathbf{1}_{x\ge\lambda+\frac{k\log N}{\sqrt{N}}}, \tag{4}$$

and

$$g'(x) = -\frac{\sqrt{N}}{\log N}\left(x - \lambda - \frac{k\log N}{\sqrt{N}}\right)^{r}\mathbf{1}_{x\ge\lambda+\frac{k\log N}{\sqrt{N}}}. \tag{5}$$

We 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

$$f(s) = g\left(\sum_{i=1}^{b} s_i\right), \tag{6}$$

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

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

$$\mathbb{E}[Gf(S)] = E\left[Gg\left(\sum_{i=1}^{b} S_i\right)\right] = 0. \tag{7}$$

Now define

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

Based on (3) and (7), we obtain

$$\mathbb{E}\left[h_k^r\left(\sum_{i=1}^{b} S_i\right)\right] = \mathbb{E}\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - Gg\left(\sum_{i=1}^{b} S_i\right)\right]. \tag{8}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

Note that, according to the definitions of $f(s)$ in (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 $1\le j\le b$. 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 into (8), we have

$$\begin{aligned}&\mathbb{E}\left[h_k^r\left(\sum_{i=1}^{b}S_i\right)\right]\\&=\mathbb{E}\left[g'\left(\sum_{i=1}^{b}S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right)-N\lambda(1-A_b(S))\left(g\left(\sum_{i=1}^{b}S_i+\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}S_i\right)\right)\right.\\&\qquad\left.-NS_1\left(g\left(\sum_{i=1}^{b}S_i-\frac{1}{N}\right)-g\left(\sum_{i=1}^{b}S_i\right)\right)\right].\end{aligned}\tag{9}$$

Let $\eta=\lambda+\frac{k\log N}{\sqrt{N}}$ to simplify the notation. From the closed forms of g and g' in (4) and (5), note that, for any $x<\eta$, $g(x)=g'(x)=0$. Also note that when $x>\eta+\frac{1}{N}$,

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

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

$$g''(x)=-\frac{r\sqrt{N}}{\log N}\left(x-\lambda-\frac{k\log N}{\sqrt{N}}\right)^{r-1}.$$

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

$$\begin{aligned}g\left(x+\frac{1}{N}\right)-g(x)&=\left(g\left(x+\frac{1}{N}\right)-g(x)\right)\left(\mathbf{1}_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}}+\mathbf{1}_{x>\eta+\frac{1}{N}}\right)\\&=\frac{g'(\xi)}{N}\mathbf{1}_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}}+\left(\frac{g'(x)}{N}+\frac{g''(\zeta)}{2N^2}\right)\mathbf{1}_{x>\eta+\frac{1}{N}},\end{aligned}\tag{10}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

$$g\Big(x-\frac{1}{N}\Big)-g(x)=\Big(g\Big(x-\frac{1}{N}\Big)-g(x)\Big)\Big(\mathbf{1}_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}}+\mathbf{1}_{x>\eta+\frac{1}{N}}\Big)$$
$$=-\frac{g'(\tilde{\xi})}{N}\mathbf{1}_{\eta-\frac{1}{N}\le x\le\eta+\frac{1}{N}}+\Big(-\frac{g'(x)}{N}+\frac{g''(\tilde{\zeta})}{2N^2}\Big)\mathbf{1}_{x>\eta+\frac{1}{N}},\tag{11}$$

where $\xi,\zeta\in\big(x,x+\frac{1}{N}\big)$ and $\tilde{\xi},\tilde{\zeta}\in\big(x-\frac{1}{N},x\big)$. Substituting (10) and (11) into the generator difference in (9), we have

$$\mathbb{E}\Big[h_k^r\Big(\sum_{i=1}^b S_i\Big)\Big]$$
$$=\mathbb{E}\Big[g'\Big(\sum_{i=1}^b S_i\Big)\Big(\lambda A_b(S)-\lambda-\frac{\log N}{\sqrt{N}}+S_1\Big)\mathbf{1}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\Big]\tag{12}$$
$$+\mathbb{E}\Big[\Big(g'\Big(\sum_{i=1}^b S_i\Big)\Big(-\frac{\log N}{\sqrt{N}}\Big)-\lambda(1-A_b(S))g'(\xi)+S_1g'(\tilde{\xi})\Big)\mathbf{1}_{\eta-\frac{1}{N}\le\sum_{i=1}^b S_i\le\eta+\frac{1}{N}}\Big]\tag{13}$$
$$-\mathbb{E}\Big[\frac{1}{2N}(\lambda(1-A_b(S))g''(\zeta)+S_1g''(\tilde{\zeta}))\mathbf{1}_{\sum_{i=1}^b S_i>\eta+\frac{1}{N}}\Big].\tag{14}$$

Note that in (12) and (14), $\xi,\zeta\in\Big(\sum_{i=1}^b S_i,\sum_{i=1}^b S_i+\frac{1}{N}\Big)$ and $\tilde{\xi},\tilde{\zeta}\in\Big(\sum_{i=1}^b S_i-\frac{1}{N},\sum_{i=1}^b S_i\Big)$ are random variables whose values depend on $\sum_{i=1}^b S_i$. We do not include $\sum_{i=1}^b S_i$ in the notation for simplicity.

Next, we study g' and g'' to bound the terms in (13) and (14), and SSC to bound the term in (12).

3.2. Gradient bounds

We summarize the bounds on g' and g'' in the following two lemmas.

Lemma 3.1. *For any* $x\in\Big[\lambda+\frac{k\log N}{\sqrt{N}}-\frac{2}{N},\lambda+\frac{k\log N}{\sqrt{N}}+\frac{2}{N}\Big]$, *we have*

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

Proof. For any $x\in\Big[\lambda+\frac{k\log N}{\sqrt{N}}-\frac{2}{N},\lambda+\frac{k\log N}{\sqrt{N}}+\frac{2}{N}\Big]$, from the closed-form expression of g' in (5) we have

$$|g'(x)|\le\frac{\sqrt{N}}{\log N}\Big|x-\lambda-\frac{k\log N}{\sqrt{N}}\Big|^r\le\frac{\sqrt{N}}{\log N}\Big(\frac{2}{N}\Big)^r=\frac{2^r}{N^{r-0.5}\log N}.$$ □

Lemma 3.2. *For* $x>\lambda+\frac{k\log N}{\sqrt{N}}$, *we have*

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

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

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

which implies

$$g''(x) = \frac{r\left(x - \lambda - \frac{k \log N}{\sqrt{N}}\right)^{r-1}}{-\frac{\log N}{\sqrt{N}}}$$

and

$$|g''(x)| = \left|\frac{r\left(x - \lambda - \frac{k \log N}{\sqrt{N}}\right)^{r-1}}{-\frac{\log N}{\sqrt{N}}}\right| \le \frac{r\sqrt{N}}{\log N}\left(\max\left\{x - \lambda - \frac{k \log N}{\sqrt{N}}, 0\right\}\right)^{r-1}. \qquad \square$$

Based on Lemma 3.1, we bound the term (13):

$$\mathbb{E}\left[\left(g'\left(\sum_{i=1}^{b} S_i\right)\left(-\frac{\log N}{\sqrt{N}}\right) - \lambda(1 - A_b(S))g'(\xi) + S_1 g'(\tilde{\xi})\right)\mathbf{1}_{\eta - \frac{1}{N} \le \sum_{i=1}^{b} S_i \le \eta + \frac{1}{N}}\right]$$
$$\le \left(\lambda + \frac{\log N}{\sqrt{N}} + 1\right)\frac{2^r}{N^{r-0.5}\log N} \le \frac{2^{r+1}}{N^{r-0.5}\log N}, \tag{15}$$

where $\lambda + \frac{\log N}{\sqrt{N}} \le 1$ in the first inequality according to the assumption $N \ge \left(\frac{4\bar{k}\log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ in Theorem 2.1.

Based on Lemma 3.2, we bound the term (14):

$$-\mathbb{E}\left[\frac{1}{2N}(\lambda(1 - A_b(S))g''(\zeta) + S_1 g''(\tilde{\zeta}))\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]$$
$$\le \mathbb{E}\left[\frac{1}{2N}(\lambda|g''(\zeta)| + S_1|g''(\tilde{\zeta})|)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \le \frac{r\mathbb{E}\left[h_k^{r-1}\left(\sum_{i=1}^{b} S_i + \frac{1}{N}\right)\right]}{\sqrt{N}\log N}. \tag{16}$$

3.3. State space collapse

In this section we consider the term in (12):

$$\mathbb{E}\left[g'\left(\sum_{i=1}^{b} S_i\right)\left(\lambda A_b(S) - \lambda - \frac{\log N}{\sqrt{N}} + S_1\right)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]$$
$$= \mathbb{E}\left[-\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k \log N}{\sqrt{N}}\right)^r\left(\lambda A_b(S) - \lambda - \frac{\log N}{\sqrt{N}} + S_1\right)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

$$= \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \left(\lambda + \frac{\log N}{\sqrt{N}} - S_1 - \lambda A_b(S)\right)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]$$

$$\le \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right], \tag{17}$$

where the last inequality holds because

$$\left(\sum_{i=1}^{b} s_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \mathbf{1}_{\sum_{i=1}^{b} s_i > \eta + \frac{1}{N}} \ge 0$$

for any $s \in \mathbb{S}$.

We define a Lyapunov function $V : \mathbb{S} \to \mathbb{R}$ to be

$$V(s) = \min\left\{\sum_{i=2}^{b} s_i,\ \lambda + \frac{k\log N}{\sqrt{N}} - s_1\right\}. \tag{18}$$

Lemma 3.3. *Under any load-balancing algorithm such that* $A_1(s) \le \frac{1}{\sqrt{N}}$ *when* $s_1 \le \lambda + \frac{\bar{k}\log N}{\sqrt{N}}$, *we have, for* $N \ge \left(\frac{4\bar{k}\log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$, *that*

$$\nabla V(s) \le -\frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} + \frac{2}{\sqrt{N}}$$

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

Proof. For the Lyapunov function defined in (18), the Lyapunov drift is

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

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

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

$$V(s+e_1) \le \sum_{i=2}^{b} s_i, \qquad V(s-e_1) = \sum_{i=2}^{b} s_i,$$
$$V(s+e_j) \le \sum_{i=2}^{b} s_i + \frac{1}{N}, \qquad V(s-e_j) = \sum_{i=2}^{b} s_i - \frac{1}{N}, \quad \text{for all } 2 \le j \le b.$$

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

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

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

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

$$V(s+e_1) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1 - \frac{1}{N}, \; V(s-e_1) \le \lambda + \frac{k \log N}{\sqrt{N}} - s_1 + \frac{1}{N},$$

$$V(s+e_j) = \lambda + \frac{k \log N}{\sqrt{N}} - s_1, \qquad V(s-e_j) \le \lambda + \frac{k \log N}{\sqrt{N}} - s_1, \text{ for all } 2 \le j \le b.$$

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

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

where the second inequality holds because $s_1 \le \lambda + (k-1)\frac{\log N}{\sqrt{N}}$ and it implies $A_1(s) \le \frac{1}{\sqrt{N}}$, the last inequality holds because $s_2 \ge \frac{1}{b-1} \frac{\log N}{\sqrt{N}}$, and the last equality holds because $1 + \frac{1}{2(b-1)} - \frac{r}{\sqrt{N} \log N} \le k \le 1 + \frac{1}{2(b-1)}$. □

Before moving forward, we present the following result from [2]. The following version of the lemma is from [18], but the result was proved in [2].

Lemma 3.4. *Let $(X(t) : t \ge 0)$ 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 $(X(t) : t \ge 0)$. Consider a Lyapunov function $V : \mathbb{X} \to \mathbb{R}^+$ and define the drift of V at a state $i \in \mathbb{X}$ as*

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

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

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

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

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

Then, for any non-negative integer j, we have

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

where

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

Based on the drift analysis in Lemmas 3.3 and 3.4 (Lemma B.1 in [18]), we have the following tail bound on $V(S)$.

Lemma 3.5. *Given the Lyapunov function defined in (18) and denoting* $\tilde{k} = 1 + \frac{1}{4(b-1)}$, *we have*

$$\mathbb{P}\left(V(S) \ge \frac{\tilde{k}\log N}{\sqrt{N}}\right) \le \exp\left\{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{8(b-1)}\right\}.$$

Proof. From Lemma 3.3, we have

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

and it is easy to verify that $q_{\max} \le N$ and $\nu_{\max} \le \frac{1}{N}$.

Based on Lemma 3.4 with $j = \frac{\sqrt{N}\log N}{8(b-1)}$, we have

$$\begin{aligned}
\mathbb{P}\left(V(S) \ge \frac{\tilde{k}\log N}{\sqrt{N}}\right) &\le \left(\frac{1}{1 + \frac{1}{2(b-1)}\frac{\log N}{\sqrt{N}} - \frac{1}{\sqrt{N}}}\right)^{\frac{\sqrt{N}\log N}{8(b-1)}+1} \\
&\le \left(1 - \frac{1}{4(b-1)}\frac{\log N}{\sqrt{N}} + \frac{1}{2\sqrt{N}}\right)^{\frac{\sqrt{N}\log N}{8(b-1)}} \\
&\le \exp\left\{-\frac{\log^2 N}{32(b-1)^2} + \frac{\log N}{16(b-1)}\right\},
\end{aligned}$$

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

Given the SSC result in Lemma 3.5, we now bound (12) (continuing from (17)) by considering two regimes, $V(s) \le \frac{\tilde{k}\log N}{\sqrt{N}}$ and $V(s) > \frac{\tilde{k}\log N}{\sqrt{N}}$, as follows:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \\
&= \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbf{1}_{V(S)\le\frac{\tilde{k}\log N}{\sqrt{N}}}\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right] \quad (19) \\
&\quad + \mathbb{E}\left[\frac{\sqrt{N}}{\log N}\left(\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}\right)^r \left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right)\mathbf{1}_{V(S)>\frac{\tilde{k}\log N}{\sqrt{N}}}\mathbf{1}_{\sum_{i=1}^{b} S_i > \eta + \frac{1}{N}}\right]. \quad (20)
\end{aligned}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

To bound (19), we consider state s such that $\mathbf{1}_{V(s)\le \frac{\tilde{k}\log N}{\sqrt{N}}}=1$ and $\mathbf{1}_{\sum_{i=1}^{b} s_i>\eta+\frac{1}{N}}=1$, because otherwise (19) $=0$. 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

$$V(s)=\lambda+\frac{k\log N}{\sqrt{N}}-s_1. \tag{21}$$

Given (21), $V(s)\le \frac{\tilde{k}\log N}{\sqrt{N}}$ means

$$\lambda+\frac{\log N}{\sqrt{N}}-s_1\le(\tilde{k}-k+1)\frac{\log N}{\sqrt{N}}\le\left(1-\frac{1}{5(b-1)}\right)\frac{\log N}{\sqrt{N}}.$$

Therefore, we have

$$(19)\le\left(1-\frac{1}{5(b-1)}\right)\mathbb{E}\left[\left(\max\left\{\sum_{i=1}^{b}S_i-\lambda-\frac{k\log N}{\sqrt{N}},0\right\}\right)^r\right]. \tag{22}$$

To bound (20), we have

$$\begin{aligned}(20)&\le\frac{b^r\sqrt{N}}{\log N}\mathbb{E}\left[\mathbf{1}_{V(S)>\frac{\tilde{k}\log N}{\sqrt{N}}}\right]\\&\le\frac{b^r\sqrt{N}}{\log N}\exp\left\{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}\right\},\end{aligned} \tag{23}$$

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 large N, and the second inequality holds due to Lemma 3.5.

Based on (22) and (23), we obtain the following upper bound on (12):

$$\begin{aligned}&\mathbb{E}\left[g'\left(\sum_{i=1}^{b}S_i\right)\left(\lambda B(S)-\lambda-\frac{\log N}{\sqrt{N}}+S_1\right)\mathbf{1}_{\sum_{i=1}^{b}S_i>\eta}\right]\\&\quad\le\left(1-\frac{1}{5(b-1)}\right)\mathbb{E}\left[\left(\max\left\{\sum_{i=1}^{b}S_i-\lambda-\frac{k\log N}{\sqrt{N}},0\right\}\right)^r\right]\\&\qquad+\frac{b^r\sqrt{N}}{\log N}\exp\left\{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}\right\}.\end{aligned} \tag{24}$$

3.4. Higher moment bounds

Given the results (24), (15), and (16), which bound (12), (13), and (14), respectively, we have

$$\begin{aligned}\mathbb{E}\left[h_k^r\left(\sum_{i=1}^{b}S_i\right)\right]&\le\left(1-\frac{1}{5(b-1)}\right)\mathbb{E}\left[h_k^r\left(\sum_{i=1}^{b}S_i\right)\right]\\&\quad+\frac{b^r\sqrt{N}}{\log N}\exp\left\{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}\right\}\end{aligned}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

$$+\frac{2^{r+1}}{N^{r-0.5}\log N}+\frac{r}{\sqrt{N}\log N}\mathbb{E}\left[h_k^{r-1}\left(\sum_{i=1}^{b}S_i\right)\right]$$

$$\le\left(1-\frac{1}{5(b-1)}\right)\mathbb{E}\left[h_k^{r}\left(\sum_{i=1}^{b}S_i\right)\right]+\frac{1+2^{r+1}}{N^{r-0.5}\log N} \tag{25}$$

$$+\frac{r}{\sqrt{N}\log N}\mathbb{E}\left[h_k^{r-1}\left(\sum_{i=1}^{b}S_i+\frac{1}{N}\right)\right], \tag{26}$$

where the second inequality holds because, given $0\le b\le 1+\frac{\sqrt{\log N}}{9}$ and $1\le r\le\frac{\log N}{72(b-1)^2}$, we have

$$\frac{b^r\sqrt{N}}{\log N}\exp\left\{-\frac{\log^2 N}{32(b-1)^2}+\frac{\log N}{16(b-1)}\right\}\le\frac{b^r\sqrt{N}}{\log N}\exp\left\{-\frac{\log^2 N}{36(b-1)^2}\right\}$$

$$\le\frac{b^r\sqrt{N}}{\log N}\mathrm{e}^{-2r\log N}=\frac{b^r}{N^{2r-0.5}\log N}\le\frac{1}{N^{r-0.5}\log N}.$$

Finally, by moving the first term in (25) to the left-hand side of the inequality and then multiplying by $4(b-1)$ on both sides of the inequality, we obtain

$$\mathbb{E}\left[h_k^{r}\left(\sum_{i=1}^{b}S_i\right)\right]\le\frac{4(1+2^{r+1})(b-1)}{N^{r-0.5}\log N}+\frac{4r(b-1)}{\sqrt{N}\log N}\mathbb{E}\left[h_k^{r-1}\left(\sum_{i=1}^{b}S_i+\frac{1}{N}\right)\right],$$

where $\bar{k}-\frac{r}{\sqrt{N}\log N}\le k\le\bar{k}$.

Define $w_r=\frac{5r(b-1)}{\sqrt{N}\log N}$ and $z_r=\frac{5(1+2^{r+1})(b-1)}{N^{r-0.5}\log N}$. The inequality above can be written as

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

By recursively applying this inequality, we obtain

$$\mathbb{E}\left[\left(\max\left\{\sum_{i=1}^{b}S_i-\lambda-\frac{\bar{k}\log N}{\sqrt{N}},0\right\}\right)^r\right]\le\prod_{j=1}^{r}w_j+\sum_{i=1}^{r}z_i\left(\prod_{j=i+1}^{r}w_j\right),$$

where we define $\prod_{j=r+1}^{r}w_j=1$ for notational convenience. We note that $z_i\prod_{j=i+1}^{r}w_j$ is a decreasing sequence in i for $2\le i\le r$, because

$$\frac{z_i\prod_{j=i+1}^{r}w_j}{z_{i-1}\prod_{j=i}^{r}w_j}=\frac{z_i}{z_{i-1}w_i}=\frac{1+2^{i+1}}{1+2^i}\frac{1}{4i(b-1)}\frac{\log N}{\sqrt{N}}\le\frac{1}{2i(b-1)}\frac{\log N}{\sqrt{N}}\le 1,$$

and

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

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

because $w_1 = \frac{5(b-1)}{\sqrt{N}\log N} \le z_1 = \frac{25(b-1)}{\sqrt{N}\log N}$. Therefore,

$$\mathbb{E}\left[h_k^r\left(\sum_{i=1}^{b} S_i\right)\right] \le (r+1)z_1 \prod_{j=2}^{r} w_j$$

$$\le (r+1)z_1 (w_r)^{r-1} \tag{27}$$

$$\le 10\left(\frac{5r(b-1)}{\sqrt{N}\log N}\right)^r, \tag{28}$$

where the inequality (27) holds because w_i is increasing in i, and (28) holds by substituting z_1 and w_r. This concludes the proof.

4. Proof of Corollary 2.1

Based on the moment bound in Theorem 2.1, we study the waiting probability $p_{\mathcal{W}}$ and the waiting time $\mathbb{E}[W]$, $\mathbb{E}[S_1]$, and $\mathbb{E}[\sum_{i=2}^{b} S_i]$ for JSQ, I1F, and JIQ. The analysis for Pod is similar and will be provided later. We begin with the waiting probability $p_{\mathcal{W}}$:

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

$$\le \mathbb{P}\left(h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right) \ge \left(\frac{\gamma}{N^\alpha} - \frac{\bar{k}\log N}{\sqrt{N}}\right)^r\right)$$

$$\le \frac{\mathbb{E}\left[h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{\gamma}{N^\alpha} - \frac{\bar{k}\log N}{\sqrt{N}}\right)^r} \tag{29}$$

$$\le \frac{\mathbb{E}\left[h_{\bar{k}}^r\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{\gamma}{2N^\alpha}\right)^r} \tag{30}$$

$$\le 10\left(\frac{10r(b-1)}{\gamma N^{0.5-\alpha}\log N}\right)^r \tag{31}$$

$$\le \frac{10}{\gamma^r N^{r(0.5-\alpha)}}, \tag{32}$$

where (29) is a result of Markov's inequality, (30) holds because $N \ge \left(\frac{4\bar{k}\log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{\bar{k}\log N}{\sqrt{N}} \le \frac{\gamma}{2N^\alpha}$, (31) holds by substituting (28), and (32) holds because $r \le \frac{\log N}{72(b-1)^2}$ implies $\log N \ge 10r(b-1)$.

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

$$\mathbb{E}[W] = \mathbb{E}[W \mid \text{a job routed to busy servers}] \times p_{\mathcal{W}} \le b p_{\mathcal{W}},$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

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

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

Therefore, we have

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

Further, according to the work conservation law, we have the following lower bound on $\mathbb{E}[S_1]$:

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

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

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

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

$$\begin{aligned}p_{\mathcal{W}} &= \mathbb{P}\left(\mathcal{W}\,\middle|\, S_1 \le 1-\frac{\gamma}{2N^{\alpha}}\right)\mathbb{P}\left(S_1 \le 1-\frac{\gamma}{2N^{\alpha}}\right)\\ &\quad+\mathbb{P}\left(\mathcal{W}\,\middle|\, S_1 > 1-\frac{\gamma}{2N^{\alpha}}\right)\mathbb{P}\left(S_1 > 1-\frac{\gamma}{2N^{\alpha}}\right)\\ &\le \mathbb{P}\left(\mathcal{W}\,\middle|\, S_1 \le 1-\frac{\gamma}{2N^{\alpha}}\right)+\mathbb{P}\left(S_1 > 1-\frac{\gamma}{2N^{\alpha}}\right)\\ &\le \left(1-\frac{\gamma}{2N^{\alpha}}\right)^{\frac{r}{\gamma}N^{\alpha}\log N}+\mathbb{P}\left(\sum_{i=1}^{b} S_i > 1-\frac{\gamma}{2N^{\alpha}}\right)\\ &\le N^{-\frac{r}{2}}+\mathbb{P}\left(h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right) \ge \left(\frac{\gamma}{2N^{\alpha}}-\frac{\bar{k}\log N}{\sqrt{N}}\right)^{r}\right) \qquad (33)\\ &\le \frac{1}{N^{0.5r}}+\frac{\mathbb{E}\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{\gamma}{2N^{\alpha}}-\frac{\bar{k}\log N}{\sqrt{N}}\right)^{r}} \qquad (34)\end{aligned}$$

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

$$\leq \frac{1}{N^{0.5r}} + \frac{\mathbb{E}\left[h_{\bar{k}}^{r}\left(\sum_{i=1}^{b} S_i\right)\right]}{\left(\frac{\gamma}{4N^{\alpha}}\right)^{r}} \tag{35}$$

$$\leq \frac{1}{N^{0.5r}} + 10\left(\frac{20r(b-1)}{\gamma N^{0.5-\alpha}\log N}\right)^{r} \tag{36}$$

$$\leq \frac{1}{N^{0.5r}} + \frac{10}{\gamma^{r} N^{r(0.5-\alpha)}}$$

$$\leq \frac{11}{\gamma^{r} N^{r(0.5-\alpha)}}, \tag{37}$$

where (33) holds because $\left(1-\frac{1}{x}\right)^{x} \leq \frac{1}{\mathrm{e}}$ for $x \geq 1$, (34) is a result of the Markov inequality, (35) holds because $N \geq \left(\frac{4\bar{k}\log N}{\gamma}\right)^{\frac{1}{0.5-\alpha}}$ implies $\frac{\gamma}{4N^{\alpha}} \geq \frac{\bar{k}\log N}{\sqrt{N}}$, (36) holds by substituting (28), and (37) holds because $r \leq \frac{\log N}{72(b-1)^2}$ implies $\log N \geq 20r(b-1)$. The remaining analysis to obtain $\mathbb{E}[W]$, $\mathbb{E}[S_1]$, and $\mathbb{E}[\sum_{i=1}^{b} S_i]$ is the same as the analysis for JSQ.

5. Conclusion and discussion

In this paper we have studied the steady-state performance of a class of load-balancing algorithms for many-server (N servers) systems in the sub-Halfin–Whitt regime ($\alpha < 0.5$). We established an upper bound on the higher moment on a distance function of the queue length with Stein's method, and studied the probability that an incoming job is routed to a busy server under JSQ, I1F, JIQ, and Pod.

We 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 above state space collapse result does not hold. It would require a different fluid model and a different state space collapse analysis, which is an interesting problem to be studied in the future.

Acknowledgements

The authors would like to thank Anton Braverman and Weina Wang for stimulating discussions that led to this result. This work was supported in part by NSF CNS-1824393, ECCS-1547294, ECCS-1609202, ECCS-1739344, and the U.S. Office of Naval Research (ONR Grant No. N00014-15-1-2169).

References

[1] Banerjee, S. and Mukherjee, D. (2018). Join-the-shortest queue diffusion limit in Halfin–Whitt regime: Tail asymptotics and scaling of extrema. arXiv:1803.03306.

[2] Bertsimas, D., Gamarnik, D. and Tsitsiklis, J. N. (2001). Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Prob.* **11**, 1384–1428.

[3] Braverman, A. (2018). Steady-state analysis of the join the shortest queue model in the Halfin–Whitt regime. arXiv:1801.05121.

[4] Braverman, A. and Dai, J. G. (2017). Stein's method for steady-state diffusion approximations of $m/Ph/n+m$ systems. *Ann. Appl. Prob.* **27**, 550–581.

[5] Braverman, A., Dai, J. G. and Feng, J. (2016). Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* **6**, 301–366.

[6] Eschenfeldt, P. and Gamarnik, D. (2018). Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Operat. Res.* **43**, 867–886.

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press

[7] GAST, N. (2017). Expected values estimated via mean-field approximation are $1/n$-accurate. *Proc. ACM Meas. Anal. Comput. Syst.* **1**, 17:1–17:26.
[8] GAST, N. AND VAN HOUDT, B. (2018). A refined mean field approximation. In *Proc. Ann. ACM SIGMETRICS Conf.*, Irvine, CA.
[9] GUPTA, V. AND WALTON, N. (2017). Load balancing in the non-degenerate slowdown regime. arXiv:1707.01969.
[10] LIU, X. AND YING, L. (2018). On achieving zero delay with power-of-*d*-choices load balancing. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, Honolulu, Hawaii.
[11] LU, Y., XIE, Q., KLIOT, G., GELLER, A., LARUS, J. R. AND GREENBERG, A. (2011). Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* **68**, 1056–1071.
[12] MITZENMACHER, M. (1996). The power of two choices in randomized load balancing. Ph.D. thesis, University of California at Berkeley.
[13] MUKHERJEE, D., BORST, S. C., VAN LEEUWAARDEN, J. S. AND WHITING, P. A. (2016). Universality of power-of-*d* load balancing in many-server systems. arXiv:1612.00723.
[14] STOLYAR, A. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Syst.* **80**, 341–361.
[15] STOLYAR, A. (2015). Tightness of stationary distributions of a flexible-server system in the Halfin–Whitt asymptotic regime. *Stoch. Syst.* **5**, 239–267.
[16] VAN DER BOOR, M., BORST, S. C., VAN LEEUWAARDEN, J. S. AND MUKHERJEE, D. (2017). Scalable load balancing in networked systems: Universality properties and stochastic coupling methods. arXiv:1712.08555.
[17] VVEDENSKAYA, N. D., DOBRUSHIN, R. L. AND KARPELEVICH, F. I. (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* **32**, 20–34.
[18] WANG, W., MAGULURI, S. T., SRIKANT, R. AND YING, L. (2017). Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *ACM SIGMETRICS Performance* **45**, 232–245.
[19] WEBER, R. R. (1978). On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15**, 406–413.
[20] WINSTON, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14**, 181–189.
[21] YING, L. (2016). On the approximation error of mean-field models. In *Proc. Ann. ACM SIGMETRICS Conf.*, Antibes Juan-les-Pins, France, pp. 285–297.
[22] YING, L. (2017). Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.* **1**, 12:1–12:27.

https://doi.org/10.1017/jpr.2020.13 Published online by Cambridge University Press