A Simple Steady-State Analysis of Load Balancing Algorithms in the Sub-Halfin-Whitt Regime

Xin Liu
Arizona State University
Tempe, Arizona
xliu272@asu.edu

Lei Ying
Arizona State University
Tempe, Arizona
lei.ying.2@asu.edu

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 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 exponential service times with mean one. We call the traffic regime *sub-Halfin-Whitt regime* because $\alpha = 0.5$ is the so-called the Halfin-Whitt regime [9]. 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) [8], join-the-idle-queue (JIQ) [11, 13] and power-of-d-choices (Pod) with $d = N^\alpha \log N$ [12, 15], and establish an upper bound on the mean queue length. From the queue-length bound, we further show that under JSQ, I1F, and Pod with $d = N^\alpha \log N$, the probability that a job is routed to a non-idle server and the expected waiting time per job are both $O\left(\frac{\log N}{\sqrt{N}}\right)$, which means only $O\left(\frac{\log N}{\sqrt{N}}\right)$ fraction of jobs experience non-zero waiting or are discarded. For JIQ, we show that the probability of waiting is $O\left(\frac{b}{N^{0.5-\alpha}\log N}\right)$.

Let S_i denote the fraction of servers with at least i jobs *at steady state*. In this paper, we prove that

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{29b}{\sqrt{N}\log N},$$

with $k = 1 + \frac{1}{2(b-1)}$, for a class of load balancing algorithms that route an incoming job to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$ when $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$. This result implies that (i)

$$E\left[\sum_{i=1}^{b} S_i\right] \leq \lambda + \frac{k\log N}{\sqrt{N}} + \frac{29b}{\sqrt{N}\log N},$$

i.e, the average queue length per server exceeds λ by at most $O\left(\frac{\log N}{\sqrt{N}}\right)$; and (ii) under JSQ, I1F, JIQ and Pod ($d = N^\alpha \log N$), the probability that an incoming job is routed to a non-idle server is asymptotically zero.

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 [3, 1, 10]. [3, 1] analyze the steady-state distribution of JSQ in the Halfin-Whitt regime and [10] studies the Pod with $\alpha < 1/6$. This paper complements [3, 1, 10], as it applies to a class of load balancing algorithms and to any sub-Halfin-Whitt regime.

Similar to [3, 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 in this paper, however, is quite different from [3, 10]. In our proof, a simple mean-field model (fluid-limit) model $\sum_{i=1}^{b} \dot{S}_i = -\frac{\log N}{\sqrt{N}}$ is used to partially approximate the evolution of the stochastic system when the system is away from the mean-field equilibrium. This is because in this paper, we are interested in bounding

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

i.e. when $\sum_{i=1}^{b} S_i \geq \lambda + \frac{k\log N}{\sqrt{N}} > \lambda$. Note that this simple mean-field model is not even accurate when $\sum_{i=1}^{b} S_i \geq \lambda + \frac{k\log N}{\sqrt{N}}$. However, using state-space collapse (SSC) approach based on the tail bound in [2], we show that the generator difference is small. In the literature, SSC has been used to show that the approximation error of using a low-dimensional system is order-wise smaller than the queue length (or some function of the queue length). Instead in this paper, we show that the error is a fraction of the term (1), but not negligible, with a high probability. We then deal with this error by subtracting it from the term (1) without bounding it explicitly. Furthermore, SSC is proved only in the regime $\sum_{i=1}^{b} S_i \geq \lambda + \frac{k\log N}{\sqrt{N}}$, which turns out to be sufficient and easy to prove. Pioneered in [14] (called drift-based-fluid-limits (DFL) method) for fluid-limit analysis and in [5, 4] for steady-state diffusion approximation, the power of Stein's method for steady-state approximations has been recognized in a number of recent papers [14, 5, 17, 4, 18, 6, 7, 3]. This paper is another an example that demonstrates the power of Stein's method for analyzing complex queueing systems.

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 one. We consider the sub-Halfin-Whitt regime such that $\lambda = 1 - N^{-\alpha}$ for some $0 < \alpha < 0.5$. As shown in Figure

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).

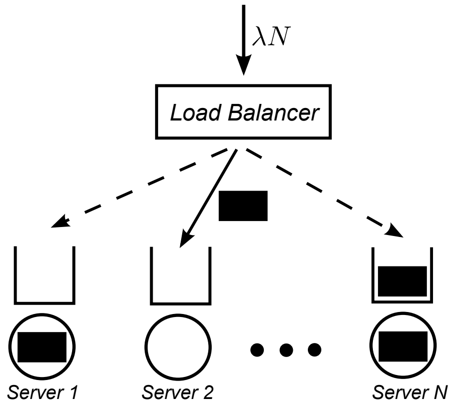

Figure 1: Load Balancing in Many-Server Systems.

We study a class of load balancing algorithms which route each incoming job to a server upon its arrival. Denote by $S_i(t)$ the fraction of servers with queue length at least i at time t. Under the finite buffer assumption with buffer size b, $S_i = 0, \forall i \geq b+1$. Define $\mathcal{S}$ to be

$$\mathcal{S} = \{s \mid 1 \geq s_1 \geq \cdots \geq s_b \geq 0\},$$

and $S(t) = [S_1(t), S_2(t), \cdots, S_b(t)]$. We consider load balancing algorithms such that $S(t) \in \mathcal{S}$ is a continuous-time Markov chain (CTMC) and has a unique stationary distribution, denoted by S, for any λ. Note λ, $S(t)$ 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 state of the system is S. Our main result of this paper is the following theorem.

THEOREM 1. *Assume* $\lambda = 1 - N^{-\alpha}$, $0 < \alpha < 0.5$, *and* $b = o(\sqrt{\log N})$. *Under any load balancing algorithm such that* $A_1(S) \leq \frac{1}{\sqrt{N}}$ *when* $S_1 \leq \lambda + \frac{k \log N}{\sqrt{N}}$ *with* $k = 1 + \frac{1}{2(b-1)}$, *the following bound holds when* N *is sufficiently large:*

$$E\left[\max\left\{\sum_{i=1}^{b} S_i - \lambda - \frac{k\log N}{\sqrt{N}}, 0\right\}\right] \leq \frac{29b}{\sqrt{N}\log N}.$$

Note that the condition $A_1(S) \leq \frac{1}{\sqrt{N}}$ when $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$ implies that an incoming job should be routed to an idle server with probability at least $1 - \frac{1}{\sqrt{N}}$ when at least $\frac{1}{N^\alpha} - \frac{k\log N}{\sqrt{N}}$ fraction of servers are idle. There are several well-known policies that satisfy this condition.

- **Join-the-Shortest-Queue (JSQ)**: JSQ routes an incoming job to the least loaded server in the system, so $A_1(S) = 0$ when $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$.
- **Idle-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 $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$.
- **Join-the-Idle-Queue (JIQ)**: JIQ routes an incoming job to an idle server if possible and otherwise, routes the job to server chosen uniformly at random. Therefore, $A_1(S) = 0$ when $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$.
- **Power-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. When $d = N^\alpha \log N$, $A_1(S) \leq \frac{1}{\sqrt{N}}$ when $S_1 \leq \lambda + \frac{k\log N}{\sqrt{N}}$.

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

COROLLARY 1. *Assume* $\lambda = 1 - N^{-\alpha}$, $0 < \alpha < 0.5$, *and* $b = o\left(\sqrt{\log N}\right)$. *For sufficiently large* N, *we have*

- *Under JSQ, IF1, and Pod with* $d = N^\alpha \log N$,
$$E[W_N] \leq \frac{3\log N}{\sqrt{N}}, \quad \text{and} \quad p_{\mathcal{W}_N} \leq \frac{4\log N}{\sqrt{N}}.$$
- *Under JIQ,*
$$p_{\mathcal{W}_N} \leq \frac{30b}{N^{0.5-\alpha}\log N}.$$

The proof of this lemma is a simple application of the Markov inequality, which can be found in [10].

We next provide an overview of the proof of our main theorem. The details are presented in [10]. The proof is based on Stein's method. As modularized in [4], this approach includes three key ingredients: generator approximation, gradient bounds and state space collapse (SSC).

Define e_i to be a b-dimensional vector such that the ith entry is $1/N$ and all other 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. For convenience, define $A_0(S) = 1$ and $A_{b+1}(S) = B(S)$, where $B(S)$ is the probability that an incoming job is discarded. Let G be the generator of CTMC $S(t)$. Given function $g : \mathcal{S} \to R$, we have

$$\begin{aligned} Gg(S) = &\sum_{i=1}^{b} \lambda N(A_{i-1}(S) - A_i(S))(g(S+e_i) - g(S)) \\ &+ N(S_i - S_{i+1})(g(S-e_i) - g(S)) \end{aligned}$$

For a bounded function $g : \mathcal{S} \to R$,

$$E[Gg(S)] = 0.$$

Following the framework of Stein's method, the first step of our proof is generator approximation. We propose a simple, almost trivial, generator L such that

$$Lg(s) = g'(s)\left(-\frac{\log N}{\sqrt{N}}\right),$$

and assume $g(s)$ is the solution of the following Stein's equation (also called Poisson equation):

$$Lg(s) = g'(s)\left(-\frac{\log N}{\sqrt{N}}\right) = h(s).$$

Following Stein's method, we bound $E[h(s)]$ by studying generator difference between L and G :

$$E[h(S)] = E[Lg(S) - Gg(S)] = E[g'(S)\left(-\frac{\log N}{\sqrt{N}}\right) - Gg(S)]$$
$$= E\left[g'(S)\left(\lambda B(S) - \lambda - \frac{\log N}{\sqrt{N}} + S_1\right) + \frac{c}{N}g''(S)\right]$$

for some constant $c > 0$. The second term can be bounded by using the gradient bound on $g''(s)$, which has a very simple form and is almost trivial to calculate. The first term is bounded based on SSC in the regime $\sum_{i=1}^{b} S_i \geq \lambda + \frac{k \log N}{\sqrt{N}}$, where a key step is to show that

$$\left(\lambda + \frac{\log N}{\sqrt{N}} - S_1\right) \mathbb{I}_{\sum_{i=1}^{b} S_i > \lambda + \frac{k \log N}{\sqrt{N}} + \frac{1}{N}} \tag{2}$$

is $O\left(\frac{\log N}{\sqrt{N}}\right)$. The intuition is that when the average number of jobs per server ($\sum_i S_i$) exceeds λ by $\frac{k \log N}{\sqrt{N}} + \frac{1}{N}$, the fraction of busy servers should be close to or exceed λ under a good load balancing algorithm. We prove this result by using the following Lyapunov function

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

and establishing the following Lemma

Lemma 1. *For sufficient large N, we have*

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

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

Based on the lemma above, we can obtain a tail bound on $V(S)$ by applying the result in [2, 16], which results in an upper bound on (2) and further prove the main theorem. Readers can find the details in [10].

3. CONCLUSION

In this paper, we studied the steady-state performance of a class of load balancing algorithms for many-server (N servers) systems in the sub-Halfin-Whitt regime. We established an upper bound on the expected 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.

Acknowledgment

This work was supported in part by NSF ECCS-1547294, ECCS-1609202, ECCS-1739344 and the U.S. Office of Naval Research (ONR Grant No. N00014-15-1-2169).

4. REFERENCES

[1] S. Banerjee and D. Mukherjee. Join-the-shortest queue diffusion limit in halfin-whitt regime: tail asymptotics and scaling of extrema. *arXiv preprint arXiv:1803.03306*, 2018.

[2] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Adv. in Appl. Probab.*, 2001.

[3] A. Braverman. Steady-state analysis of the join the shortest queue model in the halfin-whitt regime. *arXiv preprint arXiv:1801.05121*, 2018.

[4] A. Braverman and J. G. Dai. Steins method for steady-state diffusion approximations of $m/Ph/n + m$ systems. *Ann. Appl. Probab.*, 27(1):550–581, 02 2017.

[5] A. Braverman, J. G. Dai, and J. Feng. Steins method for steady-state diffusion approximations: An introduction through the erlang-a and erlang-c models. *Stoch. Syst.*, 6(2):301–366, 2016.

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

[7] N. Gast and B. Van Houdt. A refined mean field approximation. In *Proc. Ann. ACM SIGMETRICS Conf.*, Irvien, CA, 2018.

[8] V. Gupta and N. Walton. Load Balancing in the Non-Degenerate Slowdown Regime. *arXiv preprint arXiv:1707.01969*, July 2017.

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

[10] X. Liu and L. Ying. A simple steady-state analysis of load balancing algorithms in the sub-halfin-whitt regime. *arXiv preprint arXiv:1804.02622*, 2018.

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

[12] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California at Berkeley, 1996.

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

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

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

[16] W. Wang, 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, Nov. 2017.

[17] L. Ying. On the approximation error of mean-field models. In *Proc. Ann. ACM SIGMETRICS Conf.*, Antibes Juan-les-Pins, France, 2016.

[18] L. Ying. 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, June 2017.