Power-of-d Choices Load Balancing in the Sub-Halfin Whitt Regime

Sushil Mahavir Varma*, Francisco Castro†, Siva Theja Maguluri*
* School of Industrial and Systems Engineering, Georgia Institute of Technology
† Anderson School of Management, University of California, Los Angeles

ACM Reference Format:
Sushil Mahavir Varma*, Francisco Castro†, Siva Theja Maguluri*. 2023. Power-of-d Choices Load Balancing in the Sub-Halfin Whitt Regime. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts), June 19–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3578338.3593564

1 Introduction

We study a load-balancing queuing system in which a single stream of jobs arrives governed by a Poisson process and is routed to one of n homogeneous servers, operating with a service rate equal to one. Each server has a queue of maximum buffer size b.

The job dispatcher uses a load balancing or routing algorithm to route arriving jobs to the queues. Many possible routing algorithms ranging from random routing to Joining the Shortest Queue (JSQ) are studied in the literature. Random routing routes a new job to a queue selected uniformly at random. On the other hand, the JSQ algorithm routes to the server with the lowest queue length. While random routing has no informational requirements—the dispatcher does not need to know any information about the system primitives and state—it does not provide optimal delay performance. In contrast, JSQ has more informational requirements—the dispatcher needs to know the system state to determine the shortest queue—but it has a proven near-optimal delay performance. In this paper, we consider an in-between policy known as Power-of-d choices, in which once a job arrives, d queues are sampled uniformly at random from the n queues. Then, the job joins the smallest among the d sampled queues. Note that $d = 1$ is the same as random routing, and $d = n$ is the same as JSQ.

In order to study the performance of different routing policies, the literature has considered different asymptotic regimes where the number of servers goes to infinity or the load of the system approaches its capacity, or both happen simultaneously. In fact, the performance of JSQ has been studied extensively under these regimes and combinations thereof. However, Power-of-d choices (for certain values of d) have been relatively less studied. We analyze the Power-of-d choices routing algorithm under the Sub-Halfin Whitt asymptotic regime. In this regime, the arrival rate of jobs increases with the number of servers at a rate $\lambda = n - n^{1-\gamma}$ with $\gamma \in (0, 0.5)$. Under this scaling, we characterize the system's asymptotic delay performance and steady-state behavior for growing choices, i.e., $d \to \infty$ as $n \to \infty$.

If d is sufficiently large ($d \geq n^{\gamma} \log n$), then [5] showed that Power-of-d behaves like JSQ, and the jobs experience zero asymptotic delays in steady-state. In particular, the asymptotic queue lengths at each server are either zero or one. However, for smaller values of d, one expects the delay to be higher. In particular, we show that the queue lengths can be finite but greater than one or even asymptotically infinite depending on how d scales with n. Thus the asymptotic queue lengths are qualitatively different from JSQ, i.e., they are not just zero-one but exhibit a rich steady-state distribution. Characterizing such a non-degenerate behavior under different scenarios warrants a new approach. We provide a unified framework to characterize the system performance for various scalings of d.

2 Model

A natural state descriptor for the system is the number of jobs in each queue. However, it is mathematically more convenient to consider $\mathbf{s} \in (\mathbb{Z}_+ \cup \{0\})^b$ as the state descriptor. Here, s_i is the number of queues with a length of at least i and b is the maximum buffer size.

Once a job arrives, d queues are sampled uniformly at random, with replacement from n queues. Then, the job joins the smallest among the d sampled queues. This algorithm is known as Power-of-d choices in the literature. Under this routing scheme, the process $\{\mathbf{s}(t) : t \geq 0\}$ is a finite state-space, irreducible, continuous time Markov chain. Thus, the CTMC $\{\mathbf{s}(t) : t \geq 0\}$ is positive recurrent and exhibits a unique stationary distribution. Denote by $\bar{\mathbf{s}}$ a random variable with the same distribution as the stationary distribution of the CTMC.

As the exact analysis is challenging, we consider a many-server-heavy-traffic asymptotic regime, wherein the number of servers scales to infinity ($n \to \infty$) and the arrival rate increases to the capacity ($\lambda/n \to 1$). In particular, consider a sequence of load-balancing systems parameterized by n. The arrival rate for the n^{th} system is given by $\lambda = n - n^{1-\gamma}$ for $\gamma \in (0, 0.5)$, known as the sub-Halfin Whitt regime. In addition, our focus is on growing choices in Power-of-d, i.e., $d \to \infty$ as $n \to \infty$. The goal is to characterize the limiting steady-state distribution $\bar{\mathbf{s}}^{(n)}$ as $n \to \infty$.

3 Main Result

Fig. 1 (left) provides the performance of JSQ, and augmenting it with Power-of-d would correspond to adding a third dimension for d as a function of n. The case of JSQ depicted in Fig. 1 (left) corresponds to one slice of the three-dimensional figure with $d = n$.

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0074-3/23/06.
https://doi.org/10.1145/3578338.3593564

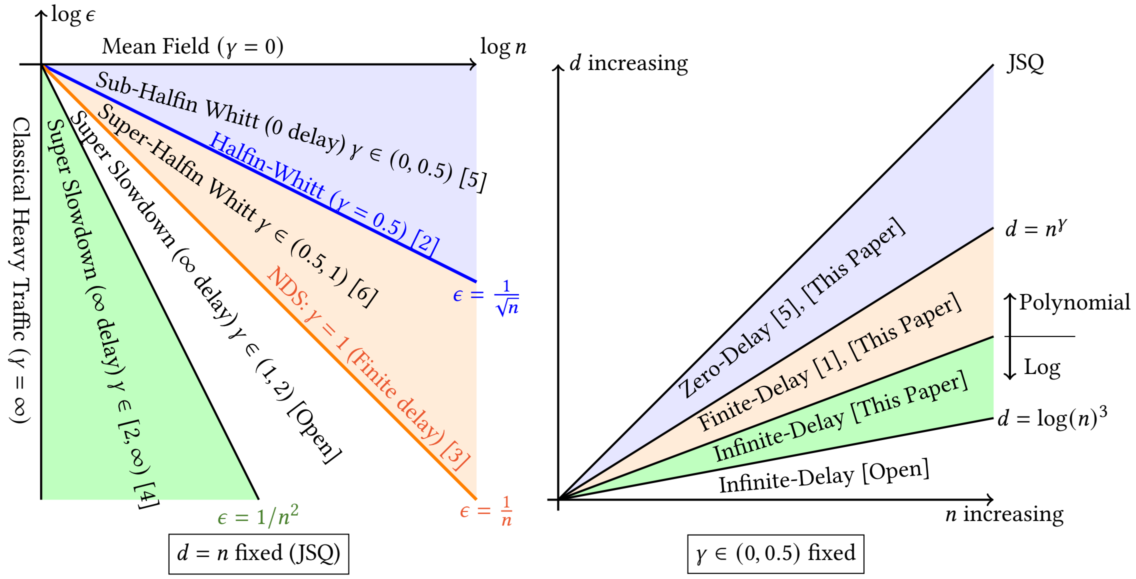
log ϵ
Mean Field ($\gamma = 0$)
log n
Sub-Halfin Whitt (0 delay) $\gamma \in (0, 0.5)$ [5]
Halfin-Whitt ($\gamma = 0.5$) [2]
Super-Halfin Whitt $\gamma \in (0.5, 1)$ [6]
NDS: $\gamma = 1$ (Finite delay) [3]
Super Slowdown (∞ delay) $\gamma \in (1, 2)$ [Open]
Super Slowdown (∞ delay) $\gamma \in [2, \infty)$ [4]
Classical Heavy Traffic ($\gamma = \infty$)
$\epsilon = \frac{1}{\sqrt{n}}$
$\epsilon = 1/n^2$
$\epsilon = \frac{1}{n}$
$d = n$ fixed (JSQ)
d increasing
JSQ
$d = n^\gamma$
Polynomial
Log
$d = \log(n)^3$
Zero-Delay [5], [This Paper]
Finite-Delay [1], [This Paper]
Infinite-Delay [This Paper]
Infinite-Delay [Open]
n increasing
$\gamma \in (0, 0.5)$ fixed

Figure 1: Performance of JSQ ($d = n$) under many-server-heavy-traffic regimes ($\gamma \in [0, \infty]$), where $\epsilon = n^{-\gamma}$ (left) and performance of Power-of-d for different choices of d under the sub-Halfin Whitt regime, i.e. $\gamma \in (0, 0.5)$ (right).

In this paper, we restrict ourselves to $\gamma \in (0, 0.5)$ and consider a broad range of d. We present the result below and illustrate it in Fig. 1 (right).

Theorem 3.1. *Let $\{m_n \in \mathbb{Z}_+ : n \in \mathbb{Z}_+\}$ be such that either $m_n \equiv m \in \mathbb{Z}_+$ or $m_n \longrightarrow \infty$ and consider $d = (2m_n n^\gamma)^{1/m_n} \log(d)^{1/m_n}$. If further $d = \Omega(\log(n)^3)$ and $b = O(\log(n)^3)$, then with probability at least $1 - \left(\frac{1}{n}\right)^{(m_n \log n)/9}$, for large enough n, we have*

$$\frac{1}{n}\bar{s}_i^{(n)} = \begin{cases} 1 - n^{-\gamma} d^{i-1}\left(1 + o(1)\right) & \forall i \in [m_n] \\ o\left(1\right) & \textit{for } i = m_n + 1 \\ o\left(\frac{1}{n}\right) & \textit{otherwise.} \end{cases}$$

Finite Delay: First, consider the case when $m_n \equiv m$, i.e. $d = (n^\gamma \log n)^{1/m}$ for some positive integer m. Theorem 3.1 shows that the queue lengths exhibits the following behavior with high probability: most of the queues are of length m and a vanishing fraction are either longer or shorter. In particular, we show that the fraction of queues with length less than i is equal to $n^{-\gamma} d^{i-1}(1 + o(1))$ for $i \leq m$ and the fraction of queues with length more than m is at most $o(n^{-\gamma} d^{m-1})$ which is $o(1)$. It is worth noting that these results are applicable for the pre-limit system as well, i.e. for all finite, large enough n, and we provide explicit expressions for all the $o(\cdot)$ terms in the technical report [7]. These results imply that when $m \geq 2$, the queue lengths are non-zero but finite, behaving qualitatively similar to that of JSQ in the non-degenerate slowdown (NDS) [3] regime. However, a fundamental difference in behavior is that while our results show that the queue lengths are essentially concentrated around m for Power-of-d in sub-Halfin Whitt regime, the limiting queue lengths of JSQ in NDS are spread over multiple values and the distribution has a nontrivial support. Also note that, when we pick $m = 1$, our result implies that the jobs experience zero asymptotic delay and the queue lengths are zero or one. The result in this special case was first established in [5].

Infinite Asymptotic Delay: Now, we consider the case when $m_n \longrightarrow \infty$, i.e. d is Poly-Log(n) (slower than any polynomial) but is at least $\Omega\left(\log(n)^3\right)$. We show that all the queue lengths are $\Theta(m_n) = \Theta(\log n / \log d)$ with high probability. This implies that the asymptotic queue lengths are infinite. Similar to the finite delay case, we characterize the fraction of queue lengths smaller or larger than m for the pre-limit system. Note that, such a behavior is qualitatively similar to that of JSQ in the super slowdown regime. However, there is again a fundamental difference in behavior because while we show that the queue lengths concentrate around $\Theta(\log n / \log d)$ for Power-of-d, JSQ in the super slowdown regime has a large support. The case when $d < \log(n)^3$ is an open future research direction.

References

[1] Shankar Bhamidi, Amarjit Budhiraja, and Miheer Dewaskar. 2022. Near equilibrium fluctuations for supermarket models with growing choices. *The Annals of Applied Probability* 32, 3 (2022), 2083–2138.

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

[3] Varun Gupta and Neil Walton. 2019. Load balancing in the nondegenerate slowdown regime. *Operations Research* 67, 1 (2019), 281–294.

[4] Daniela Hurtado-Lange and Siva Theja Maguluri. 2020. Load balancing system under Join the Shortest Queue: Many-Server-Heavy-Traffic Asymptotics. arXiv:arXiv:2004.04826

[5] Xin Liu and Lei Ying. 2020. Steady-state analysis of load-balancing algorithms in the sub-Halfin–Whitt regime. *Journal of Applied Probability* 57, 2 (2020), 578–596. https://doi.org/10.1017/jpr.2020.13

[6] Xin Liu and Lei Ying. 2022. Universal Scaling of Distributed Queues Under Load Balancing in the Super-Halfin-Whitt Regime. *IEEE/ACM Transactions on Networking* 30, 1 (2022), 190–201. https://doi.org/10.1109/TNET.2021.3105480

[7] Sushil Mahavir Varma, Francisco Castro, and Siva Theja Maguluri. 2022. Power-of-d Choices Load Balancing in the Sub-Halfin Whitt Regime. arXiv:arXiv:2208.07539