Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems

Xingyu Zhou*
zhou.2055@osu.edu
The Ohio State University
Department of ECE
Columbus, OH, USA

Fei Wu*
wu.1973@osu.edu
The Ohio State University
Department of CSE
Columbus, OH, USA

Jian Tan
tan.252@osu.edu
The Ohio State University
Department of ECE
Columbus, OH, USA

Kannan Srinivasan
kannan@cse.ohio-state.edu
The Ohio State University
Department of CSE
Columbus, OH, USA

Ness Shroff
shroff.11@osu.edu
The Ohio State University
Departments of ECE and CSE
Columbus, OH, USA

ABSTRACT

In this paper, we argue that heavy-traffic delay optimality is a coarse metric that does not necessarily imply good delay performance. Specifically, we show that any load balancing scheme is heavy-traffic delay optimal as long as it satisfies a fairly weak condition. This condition only requires that in the long-term the dispatcher favors, even slightly, shorter queues over longer queues. Hence, the empirical delay performance of heavy-traffic delay optimal schemes can range from very good (that of join-shortest-queue) to very bad (arbitrarily close to the performance of random routing). To overcome this limitation, we introduce a new metric called *degree of queue imbalance*, which measures the queue length difference between all the servers in steady-state. Given a heavy-traffic delay optimal load balancing scheme, we can characterize the resultant *degree of queue imbalance*. This, in turn, allows us to explicitly differentiate between good and poor load balancing schemes. Thus, this paper implies that good load balancing should not only be designed to be heavy-traffic delay optimal, but also have a low degree of queue imbalance.

KEYWORDS

Heavy-traffic delay optimality; Limitation; New metric

ACM Reference Format:

Xingyu Zhou, Fei Wu, Jian Tan, Kannan Srinivasan, and Ness Shroff. 2018. Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems Abstracts, June 18–22, 2018, Irvine, CA, USA*. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3219617.3219665>

*Co-primary author.

This work has been supported in part by ONR grant N00014-17-1-2417 and NSF CNS-1719371, CNS-1717060.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '18 Abstracts, June 18–22, 2018, Irvine, CA, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5846-0/18/06.

<https://doi.org/10.1145/3219617.3219665>

MAIN RESULTS

In this paper, we are particularly interested in the following two questions: (1) How large can the difference in the resultant delay be when using different heavy-traffic delay optimal load balancing schemes? (2) Can we characterize this difference, and explicitly differentiate between good and poor load balancing schemes that may all be heavy-traffic delay optimal? To answer these questions, we first introduce the following key notion. Given a policy $\eta(t)$ at time-slot t , the corresponding *dispatching distribution* is denoted by a vector $\mathbf{P}_{\eta(t)}(t)$, where its n -th component corresponds to the probability that the new arrivals are dispatched to the n -th shortest queue at time-slot t . For simplicity, we will omit the subscript $\eta(t)$ when there is no confusion.

Dispatching Preference: The dispatching preference at time-slot t is given by

$$\Delta(t) \triangleq \mathbf{P}(t) - \mathbf{P}_{\text{rand}},$$

where $\mathbf{P}_{\text{rand}}$ corresponds to the dispatching distribution of random routing, i.e., $\mathbf{P}_{\text{rand}} = (1/N, \dots, 1/N)$ for homogeneous servers.

Main Result 1: A load balancing scheme is heavy-traffic delay optimal if it satisfies the Long-term Dispatching Preference Condition (LDPC), which is defined as follows.

$$\tilde{\Delta}_1 \geq \tilde{\Delta}_2 \geq \dots \geq \tilde{\Delta}_N \quad \text{and} \quad \tilde{\Delta}_1 \neq \tilde{\Delta}_N,$$

where $\tilde{\Delta} \triangleq \mathbb{E}[\bar{\Delta}]$ and $\bar{\Delta}$ is a random vector which is equal in distribution to $\Delta(t)$ in the steady state.

Main Result 2: For any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E} \left[\left\| \bar{\mathbf{Q}}_{\perp}^{(\epsilon)} \right\|^2 \right] = \Theta \left(\frac{1}{\left\| \bar{\Delta} \right\|_1^2} \right).$$

The full version of this paper is available at [1].

REFERENCES

- [1] Xingyu Zhou, Fei Wu, Jian Tan, Kannan Srinivasan, and Ness Shroff. 2018. Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1, Article 21 (April 2018), 41 pages. <https://doi.org/10.1145/3179424>