Designing Low-Complexity Heavy-Traffic Delay-Optimal Load Balancing Schemes: Theory to Algorithms

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

Yin Sun
yys0078@auburn.edu
Auburn University
Department of ECE
Auburn, AL, USA

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

ABSTRACT

We establish a unified analytical framework for designing load balancing algorithms that can simultaneously achieve low latency, low complexity, and low communication overhead. We first propose a general class Π of load balancing policies and prove that they are both throughput optimal and heavy-traffic delay optimal. This class Π includes popular policies such as join-shortest-queue (JSQ) and power-of- d as special cases, but not the recently proposed join-idle-queue (JIQ) policy. In fact, we show that JIQ is not heavy-traffic delay optimal even for homogeneous servers. By exploiting the flexibility offered by the class Π , we design a new load balancing policy called join-below-threshold (JBT- d), in which the arrival jobs are preferentially assigned to queues that are no greater than a threshold, and the threshold is updated infrequently. JBT- d has several benefits: (i) JBT- d belongs to the class Π and hence is throughput optimal and heavy-traffic delay optimal. (ii) JBT- d has zero dispatching delay, like JIQ and other pull-based policies, and low message overhead due to infrequent threshold updates. (iii) Extensive simulations show that JBT- d has good delay performance, comparable to the JSQ policy in various system settings.

KEYWORDS

Load balancing; Throughput optimality; Heavy-traffic delay optimality; Sufficient conditions

ACM Reference Format:

Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. 2018. Designing Low-Complexity Heavy-Traffic Delay-Optimal Load Balancing Schemes: Theory to Algorithms. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*

This work has been supported in part by NSF grants CNS-1446582, CNS-1719371, CNS-1717060, ONR grant N00014-17-1-2417, ARO grant W911NF-14-1-0368, and a gift from Huawei.

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/10.1145/3219617.3219670>

Abstracts, June 18–22, 2018, Irvine, CA, USA. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3219617.3219670>

MAIN RESULTS

In this paper, we are particularly interested in the following question: how to design load balancing policies that are heavy-traffic delay optimal and also possess the nice features of pull-based policies such as zero dispatching delay, low message overhead, and low computational complexity? We answer this question by providing the following step-by-step contributions: (1) We derive inner-product based sufficient conditions for load-balancing policies to be both throughput and heavy-traffic delay optimal. We design a general class of policies Π that satisfy these sufficient conditions. This class of load balancing policies contains the famous (push-based) JSQ and the power-of- d policies as special cases, but not the (pull-based) JIQ policy. (2) Further, we show that JIQ, which is *not in* Π , is *not heavy-traffic delay optimal* even for *homogeneous servers*. While it has been empirically shown in the past that the delay using JIQ is poor at high loads, the question of whether it is heavy-traffic delay optimal for homogeneous servers had been previously left unresolved. As a by-product of our approach, we develop a novel Lyapunov-drift approach that offers a new avenue to show that a policy is not heavy-traffic delay optimal. (3) By exploiting the flexibility offered by class Π , we design a new policy called Join-Below-Threshold (JBT- d). To the best of our knowledge, this is the first load balancing policy that guarantees heavy-traffic delay optimality while enjoying the nice features of pull-based policies. Through extensive simulations, we demonstrate that JBT- d has excellent delay performance for different system sizes and various arrival and service processes over a large range of traffic loads. The full version is available at [1].

REFERENCES

- [1] Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. 2017. Designing Low-Complexity Heavy-Traffic Delay-Optimal Load Balancing Schemes: Theory to Algorithms. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 39 (Dec. 2017), 30 pages. <https://doi.org/10.1145/3154498>