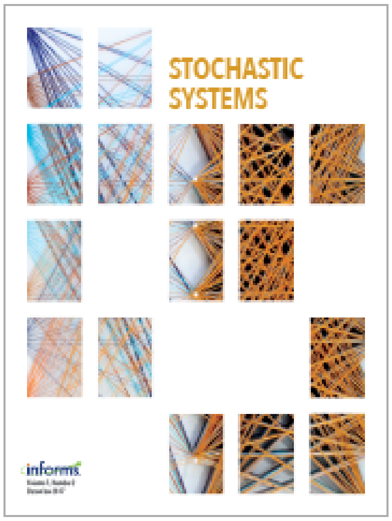

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Open Problem—Load Balancing Using Delayed Information

David Lipshutz

To cite this article:
David Lipshutz (2019) Open Problem—Load Balancing Using Delayed Information. Stochastic Systems 9(3):305-306.
https://doi.org/10.1287/stsy.2019.0045

informs
http://pubsonline.informs.org/journal/stsy
STOCHASTIC SYSTEMS
Vol. 9, No. 3, September 2019, pp. 305–306
ISSN 1946-5238 (online)

Special Section: Open Problems in Applied Probability

Open Problem—Load Balancing Using Delayed Information

David Lipshutz[a]

[a] Faculty of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel
Contact: lipshutz@technion.ac.il, https://orcid.org/0000-0001-9347-8326 (DL)

Published Online in Articles in Advance: September 19, 2019

https://doi.org/10.1287/stsy.2019.0045

History: This paper was accepted for the *Stochastic Systems* Special Section on Open Problems in Applied Probability, presented at the 2018 INFORMS Annual Meeting in Phoenix, Arizona, November 4–7, 2018.

Funding: D. Lipshutz was supported in part at the Technion by a Zuckerman fellowship.

Keywords: stochastic networks • control • optimization

Introduction

Load-balancing algorithms are important for efficiently routing jobs in systems of parallel queues; however, there has been relatively little attention devoted to developing algorithms in the presence of messaging and/or routing delays. Given a system of parallel queues with infinite capacity buffers; first-come, first-serve service discipline; and a single stream of incoming jobs that are routed by a *dispatcher* upon arrival, it has been shown that join the shortest queue (JSQ) satisfies certain optimality properties, including minimizing mean wait time when the job sizes are exponentially distributed (Winston 1977, Ephremides et al. 1980) and state space collapse of the queue lengths under heavy traffic scaling for general service distributions (Reiman 1984). However, implementation of JSQ uses up-to-date information about the state of the buffers, which requires instantaneous exchange of multiple messages between the dispatcher and the queues. This challenge has led to the development of efficient algorithms that require fewer messages, including join the shortest of d queues (JSQ(d); Vvedenskaya et al. 1996, Mitzenmacher 2001), join the idle queue (JIQ; Badonnel and Burgess 2008, Lu et al. 2011), and persistent idle load distribution (PILD; Atar et al. 2019a, b); see van der Boor et al. (2018) for a survey of some recent results in the many-server limit. Although these algorithms do not use full information about the system, they still use up-to-date information about the state of some of the queues. Because of the physical separation between the dispatcher and the queues, processing effects, or periodic updates from the queues, the dispatcher may have access only to information about the delayed states of the queues. (There may also be routing delays because of the time it takes a job to travel between the dispatcher and its assigned server, for which these algorithms also do not account.) In such settings, it has been shown that JSQ can perform quite poorly and lead to sustained oscillations in queue lengths (Mitzenmacher 2000).

Prior Work

There are relatively few load-balancing algorithms that account for the presence of delays, and mathematical analysis of systems with delays is challenging because of the need to keep track of the history of the system over the delay interval. Mitzenmacher (2000) studied the performance of JSQ(d) with delays using numerical simulations and analysis of fluid equations. He demonstrated that incorporating increased randomness into routing decisions as the delay increases reduces the size of oscillations in queue lengths. Pender et al. (2017, 2018) and Novitzky et al. (2019a, b) studied fluid models for decentralized systems in which customers (i.e., the jobs) probabilistically choose to join a queue based on posted wait times and there are information delays because the posted status may not be current and/or routing delays resulting from the travel time for the customer to join the queue. Such decentralized systems are relevant in applications ranging from customers choosing which ride to join at Disneyland to which hospital to route to in an emergency. They analyze conditions under which there are sustained oscillations and estimate the size of the oscillations. Atar and Lipshutz (2018) considered join the shortest estimated queue algorithms in which the dispatcher maintains an estimate of the current queue lengths based on the delayed queue lengths and past routing decisions. They prove state space collapse for the estimated queue lengths under heavy traffic scaling.

Open Problems

Although there has been some progress developing load-balancing algorithms in the presence of delays, there are a number of open problems remaining, some of which are listed here.

i. Analysis of load-balancing algorithms that rely on queues sending messages to the dispatcher upon idling (e.g., JIQ or PILD) in the presence of delays. Here, messaging delays can result in the queue idling during the length of the messaging and routing delays, so it may be advantageous for a queue to message the dispatcher prior to idling.

ii. Development of load-balancing algorithms at the heavy-traffic scale for fixed numbers of queues that do not account for past routing decisions (in contrast to the algorithms developed in Atar and Lipshutz (2018), which utilize past routing decisions) and analysis of ergodic properties of the scaling limits. The work (Kinnally and Williams 2010) on stationary distributions of delay differential equations with nonnegativity constraints may be relevant.

iii. Development of load-balancing algorithms that account for delays in the many-server limit. The works to date have focused on systems with a fixed number of parallel queues. See the recent work (van der Boor et al. 2019) in this direction that is applicable to the setting of periodic or random information updates on the queues.

iv. Development of load-balancing algorithms for systems with heterogeneous, stochastic, or state-dependent delays, which can, respectively, arise because of the physical locations of the queues, stochastic fluctuations, or effects from congestion within the network.

References

Atar, R Keslassy I, Mendelson G, Orda A, Vargaftik S (2019a) On the persistent-idle load distribution policy under batch arrivals and random service capacity. Technion Technical Report.

Atar, R Keslassy I, Mendelson G, Orda A, Vargaftik S (2019b) Persistent-idle load-distribution. Technion Technical Report.

Atar R, Lipshutz D (2018) Heavy traffic limits for join the shortest estimated queue policy using delayed information. Technion Technical Report.

Badonnel R, Burgess M (2008) Dynamic pull-based load balancing for autonomic servers. Brunner M, Westphall CB, Granville LZ, eds. *NOMS 2008–2008 IEEE Network Oper. Management Sympos.* (IEEE, Piscataway, NJ), 751–754.

Ephremides A, Varaiya P, Walrand J (1980) A simple dynamic routing problem. *IEEE Trans. Automatic Control* 25(4):690–693.

Kinnally M, Williams RJ (2010) On existence and uniqueness of stationary distributions for stochastic delay differential equations with positivity constraints. *Electronic J. Probab.* 15(15):409–451.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

Mitzenmacher M (2000) How useful is old information? *IEEE Trans. Parallel Distributed Systems* 11(1):6–20.

Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distributed Systems* 12(10):1094–1104.

Novitzky S, Pender J, Rand R, Wesson E (2019a) Limiting the oscillations in queues with delayed information through a novel type of delay announcement. Preprint, submitted February 20, https://arxiv.org/abs/1902.07617.

Novitzky S, Pender J, Rand RH, Wesson E (2019b) Nonlinear dynamics in queueing theory: Determining the size of oscillations in queues with delay. *SIAM J. Appl. Dynamical Systems* 18(1):279–311.

Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *Internat. J. Bifurcation Chaos* 27(4), https://doi.org/10.1142/S0218127417300166.

Pender J, Rand RH, Wesson E (2018) An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics* 91(4): 2411–2427.

Reiman MI (1984) Some diffusion approximations with state space collapse. Baccelli F, Fayolle G, eds. *Modelling and Performance Evaluation Methodology,* Lecture Notes in Control and Information Sciences, vol. 60 (Springer, Berlin, Heidelberg), 207–240.

van der Boor M, Borst SC, van Leeuwaarden JSH (2019) Hyper-scalable JSQ with sparse feedback. *2019 SIGMETRICS/Performance Joint Internat. Conf. Measurement Modeling Comp. Systems* (ACM, New York), 61–62.

van der Boor M, Borst SC, van Leeuwaarden JSH, Mukherjee D (2018) Scalable load balancing in networked systems: A survey of recent advances. Preprint, submitted June 14, https://arxiv.org/abs/1806.05444.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems Inform Transmission* 32(1):20–34.

Winston W (1977) Optimality of the shortest line discipline. *J. Appl. Probab.* 14(1):181–189.