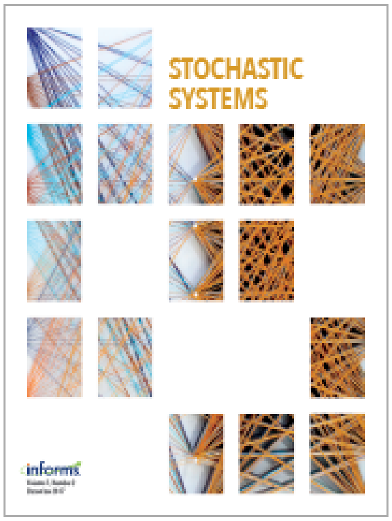

Stochastic Systems

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

Open Problem—Size-Based Scheduling with Estimation Errors

Douglas G. Down

To cite this article:
Douglas G. Down (2019) Open Problem—Size-Based Scheduling with Estimation Errors. Stochastic Systems 9(3):295-296. https://doi.org/10.1287/stsy.2019.0041

STOCHASTIC SYSTEMS
Vol. 9, No. 3, September 2019, pp. 295–296
ISSN 1946-5238 (online)
http://pubsonline.informs.org/journal/stsy

Special Section: Open Problems in Applied Probability

Open Problem—Size-Based Scheduling with Estimation Errors

Douglas G. Down[a]

[a] McMaster University, Hamilton, Ontario, L8S 4L8 Canada
Contact: downd@mcmaster.ca, https://orcid.org/0000-0003-0881-831X (DGD)

Published Online in Articles in Advance: September 18, 2019

https://doi.org/10.1287/stsy.2019.0041

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: The research was supported by the Discovery Grant program of the Natural Sciences and Engineering Council of Canada.

Keywords: shortest remaining processing time • estimation errors • server scheduling

For queueing systems, leveraging knowledge of job sizes to perform size-based scheduling leads to policies with attractive performance characteristics. Although there is a body of literature in this area, in the interest of space, we highlight one classical result: for a single-server system, the shortest remaining processing time (SRPT) policy (priority is given to the job closest to completion) is known to minimize the mean response time (Schrage and Miller 1966). Although this and related performance results have been known for some time, such size-based scheduling policies have not been deployed to any great extent in practice. One objection to their deployment is that the assumption that one knows job sizes exactly is problematic; the typical scenario would be that estimates of job sizes are available to make scheduling decisions. There is not a large literature on the study of queueing systems in which there are estimation errors for job sizes. In the next paragraph, we discuss some typical approaches and then conclude the section with two open problems of interest in this area.

In Dell'Amico et al. (2014), performance issues for SRPT with estimated job sizes are discussed in some detail. In particular, they demonstrate how underestimation of job sizes could lead to severe performance degradation. Suppose that the estimation error is proportional to the job size. If a large job has its size underestimated, then a naive implementation of SRPT eventually leads to an estimated size of zero (when the amount of time processed equals the estimated job size), at which point this job blocks all other jobs (as they all have positive estimated sizes). This could lead to long periods of time in which a large number of (small) jobs are blocked, in particular when the job size distribution is skewed to have few very large jobs and many small jobs. To address such situations, two possible directions have been undertaken. The first is to modify the scheduling policy such that jobs that have reached zero estimated errors (so-called *late* jobs) are treated differently; this is the focus of Dell'Amico et al. (2014). The other possibility is to characterize how large estimation errors can become before performance exhibits significant degradation. This is the focus of Lu et al. (2004) and Wierman and Nuyens (2008). To gain a greater fundamental understanding of the problem, it would be useful to understand the structure of the optimal policy in the face of estimation errors. At this point, we state a concrete open problem.

- Consider an M/G/1 queue in which the service time of arriving job i, S_i, follows a given distribution, B. The service times form an independent and identically distributed (i.i.d.) sequence. The estimated service time of job i,$\hat{S}_i$, is given by S_iX_i, where $\{X_i\}$ is another i.i.d. sequence, and X_i follows a given distribution, G. Determine the optimal scheduling policy (for example, with mean waiting time as the performance measure of interest) given B and G and the estimated service times (upon arrival).

This particular error model is discussed in greater detail in Dell'Amico et al. (2014). It has the desirable features that the estimation errors are proportional to the actual job size and that such a model has been seen to be appropriate in practical settings. This is not to say that other estimation error models would not be of interest. Even if the optimal policy is sufficiently complicated that it is difficult to implement, the hope is that the structure can give insight into how appropriate (near-optimal) heuristics could be constructed. Finding the optimal policy would also be useful for benchmarking purposes. Finally, in terms of a solution methodology, a Gittins index approach is one possibility.

Finally, the multiserver model is also of interest. There appears to be no work done in this area in terms of service times with estimation errors. An obvious reason for this is that, as opposed to the single server case, the optimal policy for the multiserver case without estimation errors is an open problem. Recent work by Grosof et al. (2018) has shown that multiserver SRPT is asymptotically optimal in the heavy traffic limit for an $M/G/k$ system. With estimation errors, numerical examples in Mailach and Down (2017) suggest that, as the number of servers grows, the performance of a system using multiserver SRPT is insensitive to estimation errors. So the second open problem is the resolution of the following conjecture.

- Consider an $M/G/k$ queue in which the service time of arriving job i, S_i, follows a given distribution, B. The service times form an i.i.d. sequence. The estimated service time of job $i, \hat{S}_i$, is given by $S_i X_i$, where $\{X_i\}$ is another i.i.d. sequence, and X_i follows a given distribution, G. Asymptotically, as $k \to \infty$ and the overall system load approaches one, the performance of multiserver SRPT approaches that of a system with perfect estimates, that is, when $\hat{S}_i = S_i$.

If the system load is fixed (and is less than one), then one would expect that nearly every arrival finds an idle server, and thus, estimation errors do not come into play. In heavy traffic, one could then reasonably expect that the level of blocking will not have a catastrophic effect.

References

Dell'Amico M, Carra D, Pastorelli M, Michiardi P (2014) Revisiting size-based scheduling with estimated job sizes. *Proc. IEEE 22nd Internat. Sympos. Model., Anal. Simulation Comput. Telecommunication Systems* (IEEE, Piscataway, NJ), 411–420.

Grosof I, Scully Z, Harchol-Balter M (2018) SRPT for multiserver systems. *Performance Evaluation* 127–128:154–175.

Lu D, Sheng H, Dinda P (2004) Size-based scheduling policies with inaccurate scheduling information. *Proc. IEEE 12th Internat. Sympos. Model., Anal. Simulation Comput. Telecommunication Systems* (IEEE, Piscataway, NJ), 31–38.

Mailach R, Down DG (2017) Scheduling jobs with estimation errors for multi-server systems. *Proc. 29th Internat. Teletraffic Congress* (IEEE, Piscataway, NJ), 10–18.

Schrage LE, Miller LW (1966) The queue M/G/1 with the shortest remaining processing time discipline. *Oper. Res.* 14(4):670–684.

Wierman A, Nuyens M (2008) Scheduling despite inexact job-size information. *ACM SIGMETRICS Performance Evaluation Rev.* 36(1):25–36.