Queueing Systems Vol. 100
Special Issue: 100 views on queues

Personalized Scheduling in Service Systems

Rouba Ibrahim

February 7, 2022

1 Introduction

Service systems where scarce resources need to be allocated to incoming customers are everywhere e.g., healthcare facilities, supermarkets, contact centers, online platforms, etc. In every such system, there is a scheduling question. Usually, service requests that cannot be processed immediately form a queue, either physical or virtual, and a scheduling policy is implemented to specify the order in which service requests from the queue should be processed. Given its immense practical relevance, the question of how to schedule effectively has been studied by queueing theorists, among others, for decades. Yet, given its complexity, it remains an important topic of study today.

[6] introduces a modeling paradigm for personalized queueing systems. In this note, we focus on *personalized scheduling* in service systems. Personalized scheduling exploits information about the individual realizations of underlying stochastic processes in the system, beyond the relevant probability distributions. Specifically, we focus here on scheduling policies which utilize information about the service time (the length of time needed to process the service request) of each individual customer.

When service times are perfectly known a priori, it is natural to give priority to customers with the shorter remaining processing times in order to minimize the mean sojourn time, i.e., the time from arrival until departure from the system. Indeed, a large body of literature shows that size-based scheduling policies such as the **s**hortest-**r**emaining-**p**rocessing-**t**ime (SRPT) and the **s**hortest-**j**ob-**f**irst (SJF) policies have, in general, superior performance. The SRPT policy allows preemptions, whereas SJF does not. At a high level, we are interested here in investigating whether the theoretical promise of personalized, or size-based, scheduling policies, such as SRPT and SJF, remains relevant for service systems in practice.

Rouba Ibrahim
School of Management, University College London, U.K. E-mail: rouba.ibrahim@ucl.ac.uk

2 Problem statement

Realistic queueing-model representations of service systems must account for several features. First, service systems typically have multiple agents that process service requests in parallel, thus it is important to consider multiserver queueing systems. Second, customers do not wait indefinitely for service, and they abandon the queue if their waiting time exceeds their patience time. Third, service times are typically not perfectly known before entry to service. Thus, they must be predicted based on previous customer data. For example, in the context of a call center, estimates of call durations are notoriously imprecise because service times are driven by randomly varying work contents of various customer questions and requests, which are difficult to predict ex-ante.

Size-based policies, such as SJF and SRPT, have been extensively studied for over 50 years, yet almost exclusively in single-server queues with infinitely patient jobs and perfectly known service times. For example, [8] demonstrates optimality properties of SRPT in the $M/G/1$ system. There is a notable stream of works, as in [3], that studies SRPT under heavy traffic. [10] develops a unified framework to analyze several age-based scheduling policies. Size-based scheduling with noisy service times is rarely considered in the literature. A notable exception is [11] which considers noisy service times in a single-server setting with infinite patience times. In contrast, there are very few relevant theoretical results in multiserver queues. With multiple servers, we know that SRPT is not necessarily optimal; e.g., see [5]. An important reference is [4], which studies the performance of the SRPT policy in the $M/G/k$ queue where jobs are infinitely patient; there, the SRPT policy is shown to achieve an asymptotically optimal mean sojourn time in the conventional heavy traffic regime. In the single-server queue, Gittins Index scheduling is known to be optimal in a wide array of settings, including where service times are known fully or partially. [9] shows that Gittins Index scheduling is optimal in heavy traffic, and near-optimal at all loads, in the $M/G/k$ queue.

Given those gaps in the literature, there is a need to investigate whether the superior performance of size-based policies, such as SJF and SRPT, continues to hold in multiserver queues where patience times are finite, and where service times may or may not be known with certainty. In precise terms, the problem that we would like to draw attention to in this note is: *Can we derive performance results about size-based scheduling in multiserver queues with abandonment, under noisy estimates of service times?*

3 Discussion

In general, analyzing size-based policies is complicated because it requires keeping track of all the processing times in the system, continuously over time. In single-server queues, where closed-form expressions for the performance of those policies are known, the “tagged job approach” is used to analyze busy periods and the steady-state workload. However, this type of analysis does not readily extend to multiserver queues (even without abandonment) because these systems are not work conserving. Moreover, scheduling decisions in systems with abandonment is notoriously difficult because the optimal scheduling policy can be complex and dependent on the patience-time distribution [7]. For example, when the system is critically loaded, the optimal diffusion control may no longer follow a simple fixed priority rule. Finally, very little is known about the performance of personalized policies when the information about service times is noisy. Indeed, analyzing the performance of personalized policies with noisy service times remains an open problem in general [2].

[1] makes partial progress towards studying the problem above. The authors consider the $M/GI/S+GI$ queue under SRPT, with *fully known service times*, and demonstrate a state-space collapse in the many-server overloaded limit. In particular, [1] proves that only customers with long service times (above a threshold) wait in the queue, and eventually abandon, whereas customers with short service times are immediately served. Importantly, [1] proves that, asymptotically, among all scheduling policies, SRPT maximizes the throughput in the system, minimizes the expected waiting time conditional on being served, and maximizes the expected waiting time conditional on abandoning. Performance in the SRPT queue is also shown to be, asymptotically, insensitive to the patience-time distribution beyond its mean. Those results all focus on steady-state performance measures. [1] circumvent the difficulty of doing direct analysis in the multiserver SRPT queue with abandonment by relying on a coupling proof with a loss queueing model instead. Turning now to noisy service times, coupling proofs, as above, are difficult to develop. This is because it is difficult to guarantee a strict ordering of sample paths when the scheduling is done based on noisy service times, since the noise may lead to violating the size-based scheduling rule in place. Thus, there remains ample opportunity for theoretical work in that vein.

References

1. J. Dong and R. Ibrahim. On the SRPT scheduling discipline in many-server queues with impatient customers. *arXiv preprint arXiv:2102.05789*, 2021.
2. D. G. Down. Open problem—size-based scheduling with estimation errors. *Stochastic Systems*, 9(3):295–296, 2019.
3. D. G. Down, H. C. Gromoll, and A. L. Puha. Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, 34(4):880–911, 2009.
4. I. Grosof, Z. Scully, and M. Harchol-Balter. SRPT for multiserver systems. *Performance Evaluation*, 127:154–175, 2018.
5. S. Leonardi and D. Raz. Approximating total flow time on parallel machines. *Journal of Computer and System Sciences*, 73(6):875–891, 2007.
6. A. Mandelbaum and P. Momčilović. Personalized queues: the customer view, via a fluid model of serving least-patient first. *Queueing Systems*, 87(1-2):23–53, 2017.
7. A. L. Puha and A. R. Ward. Scheduling an overloaded multiclass many-server queue with impatient customers. In *Operations Research & Management Science in the Age of Analytics*, pages 189–217. INFORMS, 2019.
8. L. Schrage. Letter to the editor-a proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16(3):687–690, 1968.
9. Z. Scully, I. Grosof, and M. Harchol-Balter. The Gittins policy is nearly optimal in the M/G/k under extremely general conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):1–29, 2020.
10. Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. Soap: One clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–30, 2018.
11. A. Wierman and M. Nuyens. Scheduling despite inexact job-size information. In *ACM SIGMETRICS Performance Evaluation Review*, volume 36, pages 25–36. ACM, 2008.