SRPT for Multiserver Systems

Isaac Grosof*
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA, US
igrosof@cs.cmu.edu

Ziv Scully*
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA, US
zscully@cs.cmu.edu

Mor Harchol-Balter*
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA, US
harchol@cs.cmu.edu

ABSTRACT

The Shortest Remaining Processing Time (SRPT) scheduling policy and its variants have been extensively studied in both theoretical and practical settings. While beautiful results are known for single-server SRPT, much less is known for *multiserver SRPT*. In particular, stochastic analysis of the M/G/k under SRPT is entirely open. Intuition suggests that multiserver SRPT should be optimal or near-optimal for minimizing mean response time. However, the only known analysis of multiserver SRPT is in the worst-case adversarial setting, where SRPT can be far from optimal. In this paper, we give the *first stochastic analysis* bounding mean response time of the M/G/k under SRPT. Using our response time bound, we show that multiserver SRPT has *asymptotically optimal* mean response time in the heavy-traffic limit. The key to our bounds is a strategic combination of stochastic and worst-case techniques. Beyond SRPT, we prove similar response time bounds and optimality results for several other multiserver scheduling policies.

This article is an introduction to our longer paper, [1].

Categories and Subject Descriptors

A.0 [**General Literature**]: General—*Performance*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Queueing Theory*; D.2 [**Software**]: Software Engineering—*Scheduling*; F.2 [**Theory of Computation**]: Analysis of algorithms and problem complexity—*Scheduling algorithms*

Keywords

M/G/k, Shortest remaining processing time (SRPT), Preemptive shortest job first (PSJF), Foreground background (FB), heavy traffic, response time bound

1. INTRODUCTION

The Shortest Remaining Processing Time (SRPT) scheduling policy and variants thereof have been deployed in many computer systems, including web servers, networks, databases, operating systems. SRPT has also long been a topic of fascination for queueing theorists due to its optimality properties. In 1966, the mean response time for SRPT was first derived [7], and in 1968 SRPT was shown to minimize mean response time both in a stochastic sense and in a worst-case sense [6]. However, these beautiful optimality results and the analysis of SRPT are only known for *single-server* systems. Almost nothing is known for *multiserver* systems, such as the M/G/k, even for the case of just $k = 2$ servers.

The SRPT policy for the M/G/k is defined as follows: at all times, the k jobs with smallest remaining processing time receive service, preempting jobs in service if necessary.

We assume a central queue, where any job can be dispatched or migrated to any server at any time, and a preempt-resume model, so preemption incurs no cost or loss of work.

It seems believable that SRPT should minimize mean response time in multiserver systems because it gives priority to the jobs which will finish soonest, which seems like it should minimize the number of jobs in the system. However, it was shown in 1997 that SRPT is not optimal for multiserver systems in the worst case [2]. There exist adversarial arrival sequences for which the mean response time under SRPT is larger that the optimal mean response time. In fact, the ratio by which SRPT's mean response time exceeds the optimal mean response time can be arbitrarily large [2].

The fact that multiserver SRPT is not optimal in the worst case provokes a natural question about the *stochastic* case.

> *Is SRPT optimal or near-optimal for minimizing mean response time in the M/G/k?*

Unfortunately, this question is entirely open. Not only is it not known whether SRPT is optimal, but multiserver SRPT has also eluded stochastic analysis.

> *What is the mean response time for the M/G/k under SRPT?*

The purpose of this paper is to answer both of these questions in the high-load setting. Under low load, response time is dominated by service time, which is not affected by the scheduling policy. In contrast, under high load, response time is dominated by queueing time, which can vary wildly under different scheduling policies. We thus focus on the high-load setting, and specifically on the *heavy-traffic limit* as load approaches capacity.

Our main result is that, under mild assumptions on the service requirement distribution,

> *SRPT is an optimal multiserver policy for minimizing mean response time in the M/G/k in the heavy-traffic limit.*

*This research was supported by NSF-XPS-1629444, NSF-CMMI-1538204, NSF-CSR-180341, and a 2018 Faculty Award from Microsoft. Additionally, Ziv Scully was supported by an ARCS Foundation scholarship and the NSF Graduate Research Fellowship Grant No. DGE-1745016.

IFIP WG 7.3 Performance 2018 Toulouse, France

We also give the *first mean response time bound for the M/G/k under SRPT*. The bound is valid for all loads and is tight for load near capacity.

In addition to SRPT, we give the *first mean response time bounds for the M/G/k with three other scheduling policies*, specifically Preemptive Shortest Job First (PSJF) [8], Remaining Size Times Original Size (RS) [9], and Foreground-Background (FB) [4]. Our bounds imply that in the heavy-traffic limit, under the same mild assumptions as for SRPT above,

- multiserver PSJF and RS are also optimal multiserver scheduling policies; and
- multiserver FB is optimal in the same setting where single-server FB is optimal [5], which is when the service requirement distribution has decreasing hazard rate and the scheduler does not have access to job sizes.

Our approach to analyzing SRPT on k servers is to compare its performance to that of SRPT on a single server which is k times as fast, where both systems have the same arrival rate λ and service requirement distribution S. Specifically, let *SRPT-k* be the policy which uses multiserver SRPT on k servers of speed $1/k$. Ordinary SRPT on a single server is simply SRPT-1. The *system load* $\rho = \lambda \mathbf{E}[S]$ is the average rate at which work enters the system. The maximal total rate at which the k servers can do work is 1, so the system is stable for $\rho < 1$, which we assume throughout.

Our main result is that in the $\rho \to 1$ limit, the mean response time under SRPT-k, $\mathbf{E}\left[T^{\text{SRPT-}k}\right]$, approaches the mean response time under SRPT-1, $\mathbf{E}\left[T^{\text{SRPT-1}}\right]$. Because SRPT-1 minimizes response time among all scheduling policies, this means that SRPT-k is asymptotically optimal among k-server policies.

Specifically, we prove the following sequence of theorems.

Our first theorem is an upper bound on the mean response time of a job of size x under SRPT-k, written $\mathbf{E}\left[T^{\text{SRPT-}k}(x)\right]$. As in the classic SRPT-1 analysis [7], the response time of a job of size x depends on the system load contributed by jobs of size at most x, written $\rho_{\le x}$.

Theorem 1.1. *In an M/G/k, the mean response time of a job of size x under SRPT-k is bounded by*

$$\mathbf{E}\left[T^{\text{SRPT-}k}(x)\right] \le \frac{\int_0^x \lambda t^2 f_S(t)\,dt}{2(1-\rho_{\le x})^2} + \frac{k\rho_{\le x} x}{1-\rho_{\le x}} + \int_0^x \frac{k}{1-\rho_{\le t}}\,dt,$$

where $f_S(\cdot)$ is the probability density function of the service requirement distribution S.

The bound given in Theorem 1.1 holds for any load ρ and any service requirement distribution S. We use this bound to prove that, under mild conditions on S, the performance of SRPT-k approaches that of SRPT-1 in the $\rho \to 1$ limit, which implies asymptotic optimality of SRPT-k.

Theorem 1.2. *In an M/G/k with any service requirement distribution S which is either (i) bounded or (ii) unbounded with a tail function which has upper Matuszewska index*[1] *less than −2,*

$$\lim_{\rho \to 1} \frac{\mathbf{E}\left[T^{\text{SRPT-}k}\right]}{\mathbf{E}\left[T^{\text{SRPT-1}}\right]} = 1.$$

To prove Theorem 1.2, we make use of results from [3], which characterizes $\mathbf{E}\left[T^{\text{SRPT-1}}\right]$ based on the Matuszewska index of the tail fuction of S.

[1]This technical condition is roughly equivalent to finite variance. See [1].

The technique by which we bound response time under SRPT-k is widely generalizable. We also use it to give mean response time bounds and optimality results for PSJF-k, RS-k, and FB-k (See [1]).

Our approach is inspired by two very different worlds: the stochastic world and the adversarial worst-case world. Purely stochastic approaches are difficult to generalize to the M/G/k for many reasons, including the fact that multiserver systems are not work-conserving. Purely adversarial worst-case analysis is easier but leads to weak bounds when directly applied to the stochastic setting. For instance, Leonardi and Raz [2] show that for an adversarial arrival sequence, SRPT-k has worse mean response time than the optimal offline k-server policy by a factor of $\Omega(\log(\min(n/k, P))$, where n is the total number of jobs in the arrival sequence and P is the ratio of the largest job size to the smallest job size. This factor can be arbitrarily large in the context of the M/G/k, because $n \to \infty$ if the arrival sequence is an infinite Poisson process, and $P \to \infty$ if the service requirement distribution is unbounded or allows for arbitrarily small jobs.

What makes our analysis work is a strategic combination of the stochastic and worst-case techniques. We use the more powerful stochastic tools where possible and use worst-case techniques to bound variables for which exact stochastic analysis is intractable.

2. REFERENCES

[1] Grosof, I., Scully, Z., and Harchol-Balter, M. SRPT for multiserver systems. In *IFIP Performance Conference 2018* (Tolouse, France, December 2018).

[2] Leonardi, S., and Raz, D. Approximating total flow time on parallel machines. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing* (1997), ACM, pp. 110–119.

[3] Lin, M., Wierman, A., and Zwart, B. Heavy-traffic analysis of mean response time under shortest remaining processing time. *Performance Evaluation* (2011).

[4] Nuyens, M., and Wierman, A. The foreground–background queue: a survey. *Performance evaluation 65*, 3-4 (2008), 286–307.

[5] Righter, R., and Shanthikumar, J. G. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences 3*, 3 (1989), 323–333.

[6] Schrage, L. Letter to the editor-a proof of the optimality of the shortest remaining processing time discipline. *Operations Research 16*, 3 (1968), 687–690.

[7] Schrage, L. E., and Miller, L. W. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research 14*, 4 (1966), 670–684.

[8] Wierman, A., and Harchol-Balter, M. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *ACM SIGMETRICS Performance Evaluation Review* (2003), ACM, pp. 238–249.

[9] Wierman, A., Harchol-Balter, M., and Osogami, T. Nearly insensitive bounds on smart scheduling. In *ACM SIGMETRICS Performance Evaluation Review* (2005), ACM, pp. 205–216.