Optimal Multiserver Scheduling with Unknown Job Sizes in Heavy Traffic

Ziv Scully
Computer Science Department
Carnegie Mellon University
zscully@cs.cmu.edu

Isaac Grosof
Computer Science Department
Carnegie Mellon University
igrosof@cs.cmu.edu

Mor Harchol-Balter
Computer Science Department
Carnegie Mellon University
harchol@cs.cmu.edu

ABSTRACT

We consider scheduling to minimize mean response time of the M/G/k queue with unknown job sizes. In the single-server $k = 1$ case, the optimal policy is the *Gittins* policy, but it is not known whether Gittins or any other policy is optimal in the multiserver case. Exactly analyzing the M/G/k under any scheduling policy is intractable, and Gittins is a particularly complicated policy that is hard to analyze even in the single-server case.

In this work we introduce *monotonic Gittins* (M-Gittins), a new variation of the Gittins policy, and show that it minimizes mean response time in the heavy-traffic M/G/k for a wide class of finite-variance job size distributions. We also show that the *monotonic shortest expected remaining processing time* (M-SERPT) policy, which is simpler than M-Gittins, is a 2-approximation for mean response time in the heavy traffic M/G/k under similar conditions. These results constitute the most general optimality results to date for the M/G/k with unknown job sizes.

1. INTRODUCTION

Scheduling to minimize mean response time[1] of the M/G/k queue is an important problem in queueing theory. The single-server $k = 1$ case has been well studied. If the scheduler has access to each job's exact size, the *shortest remaining processing time* (SRPT) policy is easily shown to be optimal. If the scheduler does not know job sizes, which is very often the case in practical systems, then a more complex policy called the *Gittins* policy is known to be optimal [1, 2]. The Gittins policy tailors its priority scheme to the job size distribution, and it takes a simple form in certain special cases. For example, for distributions with *decreasing hazard rate* (DHR), Gittins becomes the *foreground-background* (FB) policy, so FB is optimal in the M/G/1 for DHR job size distributions [1].

In contrast to the M/G/1, the M/G/k with $k \geq 2$ has resisted exact analysis, even for very simple scheduling policies. As such, much less is known about minimizing mean response time in the M/G/k, with the only nontrivial results holding under heavy traffic (Section 2). For known job sizes, recent work by Grosof et al. [3] shows that a multiserver analogue of SRPT is optimal in the heavy-traffic M/G/k. For unknown job sizes, Grosof et al. [3] address only the case of DHR job size distributions, showing that a multiserver analogue of FB is optimal in the heavy-traffic M/G/k.[2] But in general, optimal scheduling is an open problem for unknown job sizes, even in heavy traffic. We therefore ask: *What scheduling policy minimizes mean response time in the heavy-traffic M/G/k with unknown job sizes and general job size distribution?*

This is a very difficult question. In order to answer it, we draw upon several recent lines of work in scheduling theory.

- As part of their heavy-traffic optimality proofs, Grosof et al. [3] use a tagged job method to bound M/G/k response time under each of SRPT and FB relative to M/G/1 response time under the same policy.
- Lin et al. [6] and Kamphorst and Zwart [5] characterize the heavy-traffic scaling of M/G/1 mean response time under SRPT and FB, respectively.
- Scully et al. [8] show that the *monotonic shortest expected remaining processing time* (M-SERPT) policy, which is simpler than Gittins, has M/G/1 mean response time within a constant factor of that of Gittins.

While these prior results do not answer the question on their own, together they suggest a plan of attack for proving optimality in the heavy-traffic M/G/k.

When searching for a policy to minimize mean response time, a natural candidate is a multiserver analogue of Gittins. As a first step, one might hope to use the tagged job method of Grosof et al. [3] to bound M/G/k response time under Gittins relative to M/G/1 response time. Unfortunately, the tagged job method does not apply to multiserver Gittins: it relies on both stochastic and worst-case properties of the scheduling policy, and Gittins has poor worst-case properties.

One of our key ideas is to introduce a new variant of Gittins, called *monotonic Gittins* (M-Gittins), that has better worst-case properties than Gittins while maintaining similar stochastic properties. This allows us to generalize the tagged job method [3] to M-Gittins.

Our M/G/k analysis of M-Gittins reduces the question of whether M-Gittins is optimal in the heavy-traffic M/G/k to analyzing the heavy-traffic scaling of M-Gittins's M/G/1 mean response time. However, there are no heavy-traffic scaling results for the M/G/1 under policies other than SRPT [6], FB [5], and a small number of other simple policies. To remedy this, we derive heavy-traffic scaling results for M-Gittins in the M/G/1. It turns out that analyzing M-Gittins directly is very difficult. Fortunately, Scully et al. [8] introduced a simpler cousin of M-Gittins, namely M-SERPT. We analyze M-SERPT in heavy traffic as a key stepping stone in our

[1]A job's *response time*, also called *sojourn time* or *latency*, is the amount of time between its arrival and its completion.

[2]Both the SRPT and FB optimality results of Grosof et al. [3] hold under technical conditions similar to finite variance.

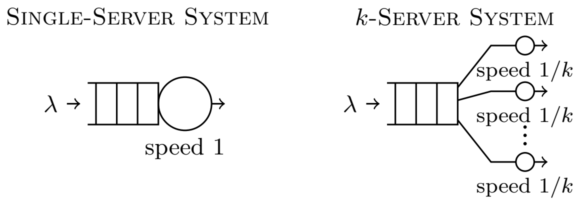

Figure 2.1: Single-Server and k-Server Systems

analysis of M-Gittins.

Our paper [4] makes the following contributions:

- We introduce M-Gittins and prove that it minimizes mean response time in the heavy-traffic M/G/k for a large class of finite-variance job size distributions.
- We also prove that the simple and practical M-SERPT policy is a 2-approximation for mean response time in the heavy-traffic M/G/k under similar conditions.
- We characterize the heavy-traffic scaling of mean response time in the M/G/1 under Gittins, M-Gittins, and M-SERPT.

We now state our main results using the notation of Section 2.

Theorem 1.1. *If X in $\mathtt{OR}(-\infty, -2)$, $\mathtt{MDA}(\Lambda) \cap \mathtt{QDHR}$, or $\mathtt{Bounded}$, then $\lim_{\rho \to 1} \mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] = 1$, in which case M-Gittins-k minimizes mean response time in the heavy-traffic M/G/k.*

Theorem 1.2. *If X in $\mathtt{OR}(-\infty, -2)$, $\mathtt{MDA}(\Lambda) \cap (\mathtt{QDHR} \cup \mathtt{QIMRL})$, or $\mathtt{Bounded}$, then $\lim_{\rho \to 1} \mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] \le 2$, in which case M-SERPT-k is a 2-approximation for mean response time in the heavy-traffic M/G/k.*

Theorem 1.3. *Let π-1 be one of Gittins-1, M-Gittins-1, or M-SERPT-1. In the $\rho \to 1$ limit, if $X \in \mathtt{OR}(-2, -1)$, then $\mathbf{E}[T^{\pi\text{-}1}] = \Theta(-\log(1-\rho))$; and if X is in $\mathtt{OR}(-\infty, -2)$, $\mathtt{MDA}(\Lambda)$, or $\mathtt{ENBUE}$, then*

$$\mathbf{E}[T^{\pi\text{-}1}] = \Theta\left(\frac{1}{(1-\rho) \cdot r^{\text{M-SERPT}}(\overline{F}_e^{-1}(1-\rho))} \right),$$

where $\overline{F}_e^{-1}$ is the inverse of the tail of the excess of X, namely $\overline{F}_e(x) = \int_x^\infty \mathbf{P}\{X > t\}\,\mathrm{d}t/\mathbf{E}[X]$.

2. NOTATION AND TERMINOLOGY

We consider an M/G/k queue with arrival rate λ and job size distribution X. Each of the k servers has speed $1/k$, so regardless of the number of servers, the total service rate is 1 and the system load is $\rho = \lambda \mathbf{E}[X]$. This allows us to easily compare the M/G/k to an M/G/1, as shown in Figure 2.1 We assume a preempt-resume model with no preemption overhead, so a single-server M/G/1 system can simulate any M/G/k policy by time-sharing between k jobs.

2.1 SOAP Policies and Rank Functions

All of the scheduling policies considered in this work are in the class of *SOAP policies* [7], generalized to a multiserver setting. In a single-server setting, a SOAP policy π is specified by a *rank function* $r^\pi : \mathbb{R}_+ \to \mathbb{R}$ mapping a job's *age*, the amount of service it has received so far, to its *rank*, or priority level. Single-server SOAP policies always serve the job of *minimal rank*, breaking ties first-come, first-served (FCFS).

A multiserver SOAP policy uses the same rank function as its single-server analogue, but it serves the k jobs of minimal rank, breaking ties FCFS. We write π-k for the k-server version of a policy, so π-1 is the single-server version. We write $T^{\pi\text{-}k}$ for the response time distribution under π-k.

We primarily consider four policies: shortest expected remaining processing time (SERPT), monotonic SERPT (M-SERPT), Gittins, and monotonic Gittins (M-Gittins). Each uses the job size distribution to tune its rank function:

$$r^{\text{SERPT}}(a) = \mathbf{E}[X - a \mid X > a],$$
$$r^{\text{M-SERPT}}(a) = \max_{b \in [0,a]} r^{\text{SERPT}}(b),$$
$$r^{\text{Gittins}}(a) = \inf_{b > a} \frac{\mathbf{E}[\min\{X, b\} - a \mid X > a]}{\mathbf{P}\{X \le b \mid X > a\}},$$
$$r^{\text{M-Gittins}}(a) = \max_{b \in [0,a]} r^{\text{Gittins}}(b).$$

2.2 Job Size Distribution Classes

We consider several classes of job size distributions, briefly described below. See our paper [4] for the full definitions.

- For any $\beta > \alpha > 0$, the $\mathtt{OR}(-\beta, -\alpha)$ class contains, roughly speaking, distributions with Pareto-like tails asymptotically between $x^{-\beta}$ and $x^{-\alpha}$. For example, all distributions in $\mathtt{OR}(-\infty, -2)$ have finite variance.
- The $\mathtt{MDA}(\Lambda)$ class contains, roughly speaking, distributions with lighter-than-Pareto tails, such as exponential, normal, log-normal, Weibull, and Gamma distributions.
- The `QDHR` and `QIMRL` classes are relaxations of the well-known *decreasing hazard rate* (`DHR`) and *increasing mean residual lifetime* (`IMRL`) classes. `QDHR` contains distributions whose hazard rate is roughly decreasing with age, even if it is not perfectly monotonic, and `QIMRL` contains distributions with roughly increasing expected remaining size.
- The `ENBUE` class contains distributions whose expected remaining size reaches a global maximum at some age. The `Bounded` subclass contains distributions with bounded support.

Acknowledgments. This work was supported by NSF-CMMI-1938909, NSF-XPS-1629444, and NSF-CSR-1763701.

References

[1] S. Aalto, U. Ayesta, and R. Righter. On the Gittins index in the M/G/1 queue. *Queueing Systems*, 63(1):437–458, 2009.

[2] S. Aalto, U. Ayesta, and R. Righter. Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25(03):269–288, 2011.

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

[4] I. Grosof, Z. Scully, and M. Harchol-Balter. Optimal multiserver scheduling with unknown job sizes in heavy traffic. *Performance Evaluation*, 2020. To appear.

[5] B. Kamphorst and B. Zwart. Heavy-traffic analysis of sojourn time under the foreground–background scheduling policy. *Stochastic Systems*, 10(1):1–28, 2020.

[6] M. Lin, A. Wierman, and B. Zwart. The average response time in a heavy-traffic SRPT queue. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 12–14. ACM, 2010.

[7] Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. Soap: One clean analysis of all age-based scheduling policies. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1):16:1–16:30, Apr. 2018.

[8] Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. Simple near-optimal scheduling for the M/G/1. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(1):11:1–11:29, Mar. 2020.