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. Our techniques build upon work by Grosof et al. [6], who study simple policies, such as SRPT, in the M/G/k; Bansal et al. [2], Kamphorst and Zwart [7], and Lin et al. [9], who analyze mean response time scaling of simple policies in the heavy-traffic M/G/1; and Aalto et al. [1] and Scully et al. [11, 13], who characterize and analyze the Gittins policy in the M/G/1.

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]. 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,[2] so FB is optimal in the M/G/1 for DHR job size distributions [1, 5].

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.[3] For known job sizes, recent work by Grosof et al. [6] shows that a multiserver analogue of SRPT is optimal in the heavy-traffic M/G/k. For unknown job sizes, Grosof et al. [6] 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. 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. [6] use a tagged job method to stochastically 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. [9] and Kamphorst and Zwart [7] characterize the heavy-traffic scaling of M/G/1 mean response time under SRPT and FB, respectively.
- Scully et al. [13] show that a policy called *monotonic shortest expected remaining processing time* (M-SERPT), which is considerably 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. [6] to stochastically 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, because it relies on both stochastic and worst-case properties of the scheduling policy, whereas 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

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

[2] FB is the policy that prioritizes the job of least age, meaning the job that has been served the least so far. It is also known as *least attained service* (LAS).

[3] Here "heavy traffic" refers to the limit as the system load approaches capacity for a fixed number of servers.

stochastic properties. This allows us to generalize the tagged job method [6] to M-Gittins, thus bounding its M/G/k response time relative to its M/G/1 response time.

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 [9], FB [7], *first-come, first served* (FCFS) [8], and a small number of other simple policies [2, 4]. 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, M-Gittins has a simpler cousin, M-SERPT, which Scully et al. [13] introduce and analyze. We analyze M-SERPT in heavy traffic as a key stepping stone in our heavy-traffic analysis of M-Gittins.

We make the following contributions:

- We introduce the M-Gittins policy 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 (Theorem 3.1).
- 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 for a large class of finite-variance job size distributions (Theorem 3.2).
- We characterize the heavy-traffic scaling of mean response time in the M/G/1 under Gittins, M-Gittins, and M-SERPT (Theorem 3.3).

Section 3 formally states these results and compares them to prior work. Their proofs can be found in the full version of this work [12].

2. PRELIMINARIES

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 system load is $\rho = \lambda\mathbf{E}[X]$. This allows us to easily compare the M/G/k system to a single-server M/G/1 system. We assume a preempt-resume model with no preemption overhead, so the single-server M/G/1 system can simulate any policy for the M/G/k system by time-sharing between k jobs.

Throughout this paper we consider the $\rho \to 1$ or *heavy-traffic* limit. This is the $\lambda \to 1/\mathbf{E}[X]$ limit with the job size distribution X and number of servers k held constant.

We write F for the cumulative distribution function of X and $\overline{F}(x) = 1-F(x)$ for its tail. We assume that X has a continuous, piecewise-monotonic hazard rate $h(x) = F'(x)/\overline{F}(x)$. We also frequently work with the expected remaining size of a job at age a, which is $\mathbf{E}[X - a \mid X > a]$. We assume it, too, is continuous and piecewise-monotonic as a function of a.

The above assumptions on hazard rate and expected remaining size are not restrictive and serve primarily to simplify presentation. It is very likely that our proofs can be generalized to relax them.

2.1 SOAP Policies and Rank Functions

All of the scheduling policies considered in this work are in the class of *SOAP policies* [11], generalized to a multiserver setting. In a single-server setting, a SOAP policy π is specified by a *rank function*

$$r^{\pi} : \mathbb{R}_+ \to \mathbb{R}$$

which maps a job's *age*, namely the amount of service it has received so far, to its *rank*, or priority level. Single-server SOAP policies work by always serving the job of *minimal rank*, breaking ties in FCFS fashion.

As an example, FB is a SOAP policy with $r^{\text{FB}}(a) = a$. Because lower age corresponds to lower rank, under FB, the server prioritizes the job of least age.[4]

We define multiserver SOAP policies in much the same way as the single-server case. The difference is that the system can serve up to k jobs.

- If there are at most k jobs, the policy serves all of them.
- If there are more than k jobs, the policy serves the k jobs of minimal rank, breaking ties in FCFS fashion.

We denote the k-server version of policy π by π-k, so π-1 is the single-server version. We write $T_x^{\pi\text{-}k}$ for the size-conditional response time distribution of jobs of size x under π-k, and we write $T^{\pi\text{-}k}$ for the overall response time distribution.

There are three main policies we consider in this work: M-SERPT, Gittins, and M-Gittins. None of these policies require knowledge of job sizes, but each uses the job size distribution to tune its rank function.

Definition 2.1. The *monotonic shortest expected remaining processing time* (M-SERPT) policy is the SOAP policy with monotonic rank function

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

Definition 2.2. The *Gittins* policy is the SOAP policy with rank function

$$r^{\text{Gittins}}(a) = \inf_{b > a} \frac{\mathbf{E}[\min\{X, b\} - a \mid X > a]}{\mathbf{P}\{X \le b \mid X > a\}} = \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)}.$$

Definition 2.3. The *monotonic Gittins* (M-Gittins) policy is the SOAP policy with monotonic rank function

$$r^{\text{M-Gittins}}(a) = \max_{b \in [0,a]} r^{\text{Gittins}}(b).$$

2.2 Job Size Distribution Classes

There are several classes of job size distributions we consider in this paper. We first briefly describe each class, then give the formal definitions.

- The OR$(-\infty, -1)$ class (Definition 2.4) contains, roughly speaking, distributions with Pareto-like tails.
  - We focus especially on the OR$(-\infty, -2)$ subclass, all members of which have finite variance.
- The MDA(Λ) class from extreme value theory [10] contains distributions whose tails are lighter than Pareto tails. It includes, among others, exponential, normal, log-normal, Weibull, and Gamma distributions.
- The QDHR and QIMRL classes (Definition 2.5) are relaxations of the *decreasing hazard rate* (DHR) and *increasing mean residual lifetime* (IMRL) classes [1, 5]. 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 (Definition 2.6) is a relaxation of the *new better than used in expectation* (NBUE) class [1]. It contains distributions whose expected remaining size reaches a global maximum at some age.

[4] When multiple jobs are tied for least age, FB equally shares the server among all such jobs because the rank function is increasing. See Scully et al. [11, Appendix B] for details.

- We focus especially on the `Bounded` subclass, which contains all bounded distributions.

Definition 2.4. A job size distribution is *O-regularly varying* if there exist exponents $\beta \geq \alpha > 0$ along with constants $C_0, x_0 > 0$ such that for all $y \geq x > x_0$,

$$\frac{1}{C_0}\left(\frac{y}{x}\right)^{-\beta} \leq \frac{\overline{F}(y)}{\overline{F}(x)} \leq C_0\left(\frac{y}{x}\right)^{-\alpha}.$$

We write $\mathtt{OR}(-\beta_0, -\alpha_0)$ for the set of O-regularly varying distributions where the exponents α and β above may be chosen such that $\alpha_0 < \alpha \leq \beta < \beta_0$.[5]

Definition 2.5. A job size distribution is in the *quasi-decreasing hazard rate* (`QDHR`) class, if there exist a strictly increasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$, an exponent $\gamma \geq 1$, and constants $C_0, x_0 > 0$ such that for all $x > x_0$,

$$m(x) \leq \frac{1}{h(x)} \leq m(C_0 x^\gamma).$$

Similarly, a distribution is in the *quasi-increasing mean residual lifetime* (`QIMRL`) class if under the same conditions,

$$m(x) \leq \mathbf{E}[X - x \mid X > x] \leq m(C_0 x^\gamma).$$

Definition 2.6. A job size distribution is in the *eventually new better than used in expectation* (`ENBUE`) class, if there exists an age $a_* \geq 0$ at which a job's expected remaining size reaches a global maximum, meaning that for all $x \neq a_*$,

$$\mathbf{E}[X - a_* \mid X > a_*] \geq \mathbf{E}[X - x \mid X > x].$$

`ENBUE` contains `Bounded`, distributions with bounded support.

3. MAIN RESULTS

We now present our main results, beginning with our heavy-traffic M/G/k optimality result.

Theorem 3.1. *In an M/G/k, if*

$$X \in \mathtt{OR}(-\infty, -2) \cup (\mathtt{MDA}(\Lambda) \cap \mathtt{QDHR}) \cup \mathtt{Bounded},$$

then $\lim_{\rho \to 1} \mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] = 1$*. In such cases, M-Gittins-k minimizes mean response time in heavy traffic.*

The M-Gittins policy is based on the Gittins policy, which is somewhat complex to describe and compute. Fortunately, the M-SERPT policy, which can be much simpler to compute [13], also performs well in the heavy-traffic M/G/k.

Theorem 3.2. *In an M/G/k, if*

$$X \in \mathtt{OR}(-\infty, -2) \cup (\mathtt{MDA}(\Lambda) \cap (\mathtt{QDHR} \cup \mathtt{QIMRL})) \cup \mathtt{Bounded},$$

then $\lim_{\rho \to 1} \mathbf{E}[T^{\text{M-SERPT-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] \leq 2$*. In such cases, M-SERPT-k is a 2-approximation for mean response time in heavy traffic.*

Theorems 3.1 and 3.2 apply to a broad class of finite-variance job size distributions. Roughly speaking, $\mathtt{OR}(-\infty, -2)$ covers heavy-tailed distributions, and $\mathtt{MDA}(\Lambda)$ covers non-heavy-tailed distributions that are unbounded (Section 2.2). Assuming membership in these sets is standard for heavy-traffic analysis [7]. The main restriction the results impose is on $\mathtt{MDA}(\Lambda)$ distributions, for which we additionally require membership in `QDHR` or `QIMRL`. While slightly relaxing this restriction is possible, removing it entirely appears to be very difficult [12, Section 8].

A key step in the proofs of Theorems 3.1 and 3.2 is analyzing M-Gittins and M-SERPT in the heavy-traffic M/G/1. This analysis is itself a new result of independent interest. Notably, it extends to ordinary Gittins in addition to M-Gittins, thus characterizing the optimal heavy-traffic scaling attainable by any scheduling policy.

Theorem 3.3. *Let π-1 be one of Gittins-1, M-Gittins-1, or M-SERPT-1. If* $X \in \mathtt{OR}(-2, -1)$*, then in the* $\rho \to 1$ *limit,*

$$\mathbf{E}[T^{\pi\text{-}1}] = \Theta\left(\log \frac{1}{1-\rho}\right)$$

and if $X \in \mathtt{OR}(-\infty, -2) \cup \mathtt{MDA}(\Lambda) \cup \mathtt{ENBUE}$*, then in the* $\rho \to 1$ *limit,*

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

where $\overline{F}_e^{-1}$ *is the inverse of the tail of the excess of* X*, namely*

$$\overline{F}_e(x) = \frac{1}{\mathbf{E}[X]} \int_x^\infty \overline{F}(t)\,\mathrm{d}t.$$

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] N. Bansal, B. Kamphorst, and B. Zwart. Achievable performance of blind policies in heavy traffic. *Mathematics of Operations Research*, 43(3):949–964, 2018.

[3] N. Bingham, C. Goldie, and J. Teugels. *Regular Variation*. Cambridge University Press, 1987.

[4] Y. Chen and J. Dong. Scheduling with service-time information: The power of two priority classes. Preprint, 2020.

[5] H. Feng and V. Misra. Mixed scheduling disciplines for network flows. In *ACM SIGMETRICS Performance Evaluation Review*, volume 31, pages 36–39. ACM, 2003.

[6] I. Grosof, Z. Scully, and M. Harchol-Balter. SRPT for multiserver systems. *Performance Evaluation*, 127–128:154–175, 2018. ISSN 0166-5316. doi: 10.1016/j.peva.2018.10.001.

[7] B. Kamphorst and B. Zwart. Heavy-traffic analysis of sojourn time under the foreground–background scheduling policy. *Stochastic Systems*, 10(1):1–28, 2020. doi: 10.1287/stsy.2019.0036.

[8] J. Köllerström. Heavy traffic theory for queues with several servers. ii. *Journal of Applied Probability*, 16(2):393–401, 1979. ISSN 00219002.

[9] 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.

[10] S. I. Resnick. *Extreme values, regular variation and point processes*. Springer, 2013.

[11] 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. ISSN 2476-1249. doi: 10.1145/3179419.

[12] Z. Scully, I. Grosof, and M. Harchol-Balter. Optimal multiserver scheduling with unknown job sizes in heavy traffic. *arXiv*, 2020.

[13] 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. doi: 10.1145/3379477.

[5] This is not the standard definition of O-regular variation, but it is equivalent to it [3, Section 2.2.1].