The Gittins Policy is Nearly Optimal in the M/G/k under Extremely General Conditions

Ziv Scully
Carnegie Mellon University
Computer Science Department
Pittsburgh, PA, USA
zscully@cs.cmu.edu

Isaac Grosof
Carnegie Mellon University
Computer Science Department
Pittsburgh, PA, USA
igrosof@cs.cmu.edu

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

ABSTRACT

The Gittins scheduling policy minimizes mean response time in the preemptive M/G/1 queue in a wide variety of settings. Most famously, Gittins is optimal when service requirements are unknown but drawn from a known distribution. Gittins is also optimal much more generally, adapting to any amount of available information. However, scheduling to minimize mean response time in a multiserver setting, specifically the central-queue M/G/k, is a much more difficult problem.

In this work we give the first general analysis of Gittins in the M/G/k. Specifically, we show that under extremely general conditions, Gittins's mean response time in the M/G/k is at most its mean response time in the M/G/1 plus an $O(k \log(1/(1-\rho)))$ additive term, where ρ is the system load. A consequence of this result is that Gittins is heavy-traffic optimal in the M/G/k if the service requirement distribution S satisfies $\mathbf{E}[S^2(\log S)^+] < \infty$. This is the most general result on minimizing mean response time in the M/G/k to date.

To prove our results, we combine properties of the Gittins policy and Palm calculus in a novel way. Notably, our technique overcomes the limitations of tagged job methods used in prior scheduling analyses.

CCS CONCEPTS

• **General and reference** → **Performance**; • **Mathematics of computing** → **Queueing theory**; • **Networks** → **Network performance modeling**; • **Theory of computation** → *Routing and network design problems*; • **Computing methodologies** → *Model development and analysis*; • **Software and its engineering** → *Scheduling*.

KEYWORDS

M/G/k; response time; latency; sojourn time; Gittins policy; heavy traffic; Markov process

ACM Reference Format:
Ziv Scully, Isaac Grosof, and Mor Harchol-Balter. 2021. The Gittins Policy is Nearly Optimal in the M/G/k under Extremely General Conditions. In *Abstract Proceedings of the 2021 ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '21 Abstracts), June 14–18, 2021, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3410220.3456281

SIGMETRICS '21 Abstracts, June 14–18, 2021, Virtual Event, China

1 INTRODUCTION

The question of how to schedule jobs so as to minimize mean response time is a classic question in the queueing literature. Here the response time of a job is the time from when the job arrives until it completes service, and the goal is to minimize the average response time across all jobs. In the M/G/1 queue, the answer is the *Gittins* policy. Gittins minimizes mean response time in a huge variety of settings [1]. These include the case where one has perfect information about job service requirements (sizes); the case where one has zero information about job sizes, but one knows the distribution of the job sizes; and cases where one has partial knowledge about jobs' sizes.

Gittins is an index policy which assigns a "rank" to each job, where at every moment of time the job served is the one with lowest rank. Roughy speaking, a job has low rank if it is likely to complete soon. Note that in case where one has perfect information about job sizes, the Gittins policy is equivalent to the Shortest-Remaining-Processing-Time (SRPT) policy.

Unfortunately, minimizing mean response time in multiserver systems like the M/G/k is much harder. In the case where one has perfect information about job sizes, it was recently shown that SRPT, which is optimal for the M/G/1, is also optimal for the M/G/k in heavy traffic and near-optimal at all loads [3]. Here "near-optimal" specifically means that the response time differs from optimal by an $O(k \log(1/(1-\rho)))$ additive term, where $\rho < 1$ is the system load. One might therefore hope that in the case where job sizes are not known, or not fully known, that the Gittins policy, which is optimal for the M/G/1, is also optimal for the M/G/k.

In the full version of this work [4], we show that the Gittins policy in the M/G/k is optimal in heavy traffic, meaning the $\rho \to 1$ limit, and near-optimal at all loads. Here "near-optimal" again means within an $O(k \log(1/(1-\rho)))$ additive term of optimal. We do so by comparing the Gittins policy in the M/G/k to the Gittins policy in the M/G/1, where it is known to be optimal. Our approach involves applying a novel Palm calculus technique to bounding mean response time in the M/G/k, directly relating response time to certain quantities of steady-state work. Our Palm calculus technique leverages key properties of the Gittins rank function which are responsible for Gittins' optimality in the M/G/1. In contrast, recent prior work on analyzing scheduling in the M/G/k has focused on

tagged-job methods [3, 5], which are not well-suited to analyzing Gittins.

Our analysis allows us to prove Theorem 1.1, which upper-bounds the *gap* between the mean response time of the Gittins policy in the M/G/k and the M/G/1, where the server in the M/G/1 is k times faster than the servers in the M/G/k. This is important because the mean response time in the M/G/1 under Gittins is a lower bound on the mean response time in the M/G/k under any scheduling policy. In the common case where the variance of the service requirement distribution is finite, Theorem 1.2 gives a simpler bound on the mean response time gap. The tightness of our bound enables us to prove in Theorem 1.3 that Gittins is heavy-traffic optimal in the M/G/k.

Our results hold under an extremely broad job model and under a very permissive condition on the service requirement distribution. Both the job model and the condition on the service requirement distribution are far broader than typically assumed [2, 3, 5].

Our extremely broad job model allows the information that is revealed as a job runs to be very complex and general. As a basic example, the job could reveal partial information about its service requirement upon arrival, such as a noisy estimate of the service requirement. Beyond that, each job could be broken into a series of stages, with information revealed whenever a stage completes. We even support scenarios where service requirement information is revealed continuously over time as a job runs. In general, we allow a job to consist of an arbitrary absorbing Markov process, revealing information according to the evolution of the state.

The condition we require on the service requirement distribution are also extremely general. For example, Theorem 1.1, which proves the near-optimality of Gittins in the M/G/k, requires only that the service requirement distribution have a finite αth moment for some $\alpha > 1$. Moreover, this condition is independent of the underlying job Markov process.

1.1 Main Results

Our main results concern the M/G/k queue with arrival rate λ and service requirement distribution S. All three theorems compare a k-server system with a single-server system with the total service capacity 1, meaning each server has speed $1/k$. We denote the load by $\rho = \lambda \mathbf{E}[S]$ and the mean response time in the M/G/k by $\mathbf{E}[T_k]$. See the full version of this work [4] for a full description of the system model.

Theorem 1.1. *For all $\alpha > 1$, if $\mathbf{E}[S^\alpha] < \infty$, then the mean response time gap between the M/G/k and M/G/1 under Gittins is at most*

$$\mathbf{E}[T_k] - \mathbf{E}[T_1] \le (k-1)\,\mathbf{E}[S]\left(\frac{1}{\alpha-1}\left(\log\frac{1}{1-\rho} + \log\frac{\mathbf{E}[S^\alpha]}{\alpha\,\mathbf{E}[S]^\alpha} + 1\right) + 4.547\right).$$

Theorem 1.2. *If $C_S^2 = \mathbf{Var}[S^2]/\mathbf{E}[S]^2 < \infty$, then the mean response time gap between the M/G/k and M/G/1 under Gittins is at most*

$$\mathbf{E}[T_k] - \mathbf{E}[T_1] \le (k-1)\,\mathbf{E}[S]\left(\log\frac{1}{1-\rho} + \log(1+C_S^2) + 4.811\right).$$

Theorem 1.3. *If $\mathbf{E}[S^2(\log S)^+] < \infty$, then under Gittins,*

$$\lim_{\rho\to 1}\frac{\mathbf{E}[T_k]}{\mathbf{E}[T_1]} = 1,$$

so Gittins minimizes mean response time in the M/G/k in the heavy-traffic limit.

1.2 Contributions and Outline

Our main result is Theorem 1.1, where we prove that Gittins is near-optimal in the M/G/k. As a corollary of this result, in Theorem 1.3, we prove that Gittins is heavy-traffic optimal for the M/G/k. Along the way to proving these results, we make several contributions of independent interest. The first part of our paper [4] introduces our model.

- We lay out an extremely general job model, allowing a highly flexible handling of the information that is revealed as a job runs.
- We give a treatment of the Gittins policy under our highly general job model.

The second part of our paper [4] proves our main results.

- We introduce the "Gittins game", a framework for understanding the Gittins policy which we use throughout the paper to prove crucial properties of the policy.
- We derive a new formula for mean response time in the M/G/k in terms of the mean amount of different "relevant" subsets of work in the system.
- We present a new decomposition law that bounds the gap between the mean relevant work in the M/G/k and the M/G/1.
  - In addition to being useful in the M/G/k, our techniques also yield an elegant new proof of the optimality of Gittins in the M/G/1.

ACKNOWLEDGMENTS

This work was supported by NSF grants CMMI-1938909, XPS-1629444, and CSR-1763701; and a Google 2020 Faculty Research Award.

REFERENCES

[1] John C. Gittins, Kevin D. Glazebrook, and Richard Weber. 2011. *Multi-armed Bandit Allocation Indices.* John Wiley & Sons.
[2] Kevin D Glazebrook and José Niño-Mora. 2001. Parallel scheduling of multi-class M/M/m queues: Approximate and heavy-traffic optimization of achievable performance. *Operations Research* 49, 4 (2001), 609–623.
[3] Isaac Grosof, Ziv Scully, and Mor Harchol-Balter. 2018. SRPT for Multiserver Systems. *Performance Evaluation* 127–128 (2018), 154–175. https://doi.org/10.1016/j.peva.2018.10.001
[4] Ziv Scully, Isaac Grosof, and Mor Harchol-Balter. 2020. The Gittins Policy is Nearly Optimal in the M/G/k under Extremely General Conditions. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 3, Article 43 (Nov. 2020), 29 pages. https://doi.org/10.1145/3428328
[5] Ziv Scully, Isaac Grosof, and Mor Harchol-Balter. 2021. Optimal Multiserver Scheduling with Unknown Job Sizes in Heavy Traffic. *Performance Evaluation* 145, Article 102150 (Jan. 2021), 31 pages.