Characterizing Policies with Optimal Response Time Tails under Heavy-Tailed Job Sizes

Ziv Scully
Carnegie Mellon University
zscully@cs.cmu.edu

Lucas van Kreveld
University of Amsterdam
l.r.vankreveld@uva.nl

Onno Boxma
Eindhoven University of Technology
o.j.boxma@tue.nl

Jan-Pieter Dorsman
University of Amsterdam
j.l.dorsman@uva.nl

Adam Wierman
California Institute of Technology
adamw@caltech.edu

ABSTRACT

We consider the tail behavior of the response time distribution in an $M/G/1$ queue with heavy-tailed job sizes, specifically those with intermediately regularly varying tails. In this setting, the response time tail of many individual policies has been characterized, and it is known that policies such as Shortest Remaining Processing Time (SRPT) and Foreground-Background (FB) have response time tails of the same order as the job size tail, and thus such policies are tail-optimal. Our goal in this work is to move beyond individual policies and characterize the set of policies that are tail-optimal. Toward that end, we use the recently introduced SOAP framework to derive sufficient conditions on the form of prioritization used by a scheduling policy that ensure the policy is tail-optimal. These conditions are general and lead to new results for important policies that have previously resisted analysis, including the Gittins policy, which minimizes mean response time among policies that do not have access to job size information. As a by-product of our analysis, we derive a general upper bound for fractional moments of $M/G/1$ busy periods, which is of independent interest.

CCS CONCEPTS

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

KEYWORDS

response time; sojourn time; tail latency; tail optimality; Gittins policy; shortest expected processing time (SERPT); randomized multi-level feedback (RMLF); M/G/1

ACM Reference Format:
Ziv Scully, Lucas van Kreveld, Onno Boxma, Jan-Pieter Dorsman, and Adam Wierman. 2020. Characterizing Policies with Optimal Response Time Tails under Heavy-Tailed Job Sizes. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '20 Abstracts), June 8–12, 2020, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3393691.3394179

SIGMETRICS '20 Abstracts, June 8–12, 2020, Boston, MA, USA

ACM ISBN 978-1-4503-7985-4/20/06.
https://doi.org/10.1145/3393691.3394179

1 INTRODUCTION

The scheduling of jobs in single server queues has been an important topic of study over the past decades. On one hand, much attention has been devoted to identifying scheduling policies that minimize the mean response time in a variety of settings. For example, in preemptive settings it is widely known that Shortest Remaining Processing Time (SRPT) minimizes the mean response time regardless of the job size distribution when job sizes are known to the scheduler. When sizes are unknown to the scheduler but the job size distribution is known, the optimal scheduling policy is the Gittins policy, which serves the job with maximum Gittins index. If the job size distribution is also unknown, then the Randomized Multi-Level Feedback (RMLF) policy minimizes the competitive ratio for mean response time.

On the other hand, in many applications it is important to avoid large response times, not just minimize the mean response time. Thus, significant research has been devoted to analyzing the distribution of response times under a large variety of scheduling policies, including SRPT, First Come First Served (FCFS), and Processor Sharing (PS) and its many generalizations. In some simple settings it is possible to precisely characterize the response time distribution, but in general research focuses on characterizing the *tail* of the response time distribution.

The task of characterizing the response time tail is more complex than that of optimizing the mean response time. Initially, response time tail asymptotics were studied in the case of light-tailed job size distributions. In this context, it has been shown that FCFS maintains the optimal (lightest) tail of the response time distribution, whereas under SRPT the response time tail is the heaviest possible under any work conserving scheduling policy. This is a stark contrast to the optimality of SRPT for the mean response time.

While the focus of response time tail asymptotics was initially on light-tailed settings, a shift occurred in the late 1990s when it was observed that heavy-tailed distributions occur frequently in computer and communications systems, e.g., in file sizes in the web, in I/O patterns, the length of network sessions, and more. These observations triggered significant research into the impact of heavy-tailed phenomena on the design and performance of computer and communications systems. The resulting literature has demonstrated that scheduling and priority mechanisms need to be designed with heavy-tailed phenomena in mind.

A key observation from the research that followed is that scheduling policies that perform well under light-tailed settings may not perform well under heavy-tailed settings, and vice versa. A prime example is FCFS, which has the optimal response time tail under light-tailed job sizes, but has a response time tail as bad as possible among work conserving policies under heavy-tailed job sizes. In contrast, SRPT, which has the worst possible tail behavior in light-tailed settings, is optimal in heavy-tailed settings.

Observations like this have led to significant research on the impact of the service discipline on delay asymptotics. Given the prominence of heavy-tailed phenomena in computer and communications systems, a driving question for the community has been to characterize which policies have the optimal response time tail asymptotics under heavy tailed job sizes. This notion of "tail equivalence" (also known as tail optimality) has driven research for decades and at this point there is a variety of common policies known to be tail equivalent, including PS, Foreground-Background (FB), and Preemptive Shortest Job First (PSJF).

However, despite significant progress, there are still many important policies for which we do not know if they are tail equivalent or not. Examples are the Gittins policy and RMLF. Further, no precise characterization of which properties a scheduling policy must have in order to be tail equivalent is known.

The first attempt at a general set of conditions that ensure tail equivalence was by Núñez-Queija [1], who provided analytic conditions that can be used to simplify the analysis of scheduling policies when studying the response time tail. It was these analytic conditions that enabled the first analysis of policies such as SRPT, PSJF, and FB. However, the conditions are defined in terms of the analysis of the policy rather than the prioritization rules used by the policy, and so they do not provide insight into which policies are tail equivalent. For that, the most general result to this point is by Nuyens et al. [2], who introduce a set of properties based on job sizes that are sufficient conditions for tail equivalence. These properties ensure that the scheduler always prioritizes jobs with small sizes and are satisfied by both SRPT and PSJF, but not by policies that do not make use of job sizes, such as Gittins, RMLF, FB, etc. Thus, there is a considerable gap between the sufficient conditions outlined by [2] and a general characterization of tail-equivalent scheduling policies.

In our paper [4], we provide sufficient conditions that ensure optimality of the tail of the response time distribution (a.k.a. tail equivalence) for scheduling policies in M/GI/1 queues with job size distributions that are intermediately regularly varying. Our results provide guidelines on how scheduling policies can perform prioritization in order to ensure tail equivalence without having access to job sizes, and are thus complementary to the conditions in [2], which focus on prioritization based on job size. The conditions are general and are satisfied by important policies such as the Gittins and RMLF policies, for which no previous analysis of the response time tail is known. The conditions are also satisfied by policies that use limited preemption, for the first time highlighting the preemption frequency needed to achieve tail-optimality.

The key building block underlying the sufficient conditions we develop is the SOAP framework, recently introduced in [3]. SOAP expresses scheduling policies as rank functions which assign a rank to each job and serve the job with lowest rank. A job's rank is determined by a function of its age (the amount of time the job has already received service), and a job-specific descriptor, e.g., the size or priority level of the job. Using this framework, our sufficient conditions for tail equivalence are defined in terms of the rank function of a policy. The formal conditions can be found in the full version of this work [4, Section 3]. Intuitively, the conditions ensure that old jobs do not receive priority over other jobs for too long.

Our proof adapts the probabilistic method developed by Núñez-Queija [1], which exploits a Markov-type inequality. While the method of [1] does not apply directly off-the-shelf, we are able to extend it to apply to the analysis of our sufficient conditions. This extension requires technical effort and, in particular, relies on a new analysis of the fractional moments of busy periods that is of independent interest.

To conclude, we summarize the contributions of our paper [4]:

- We provide a set of sufficient conditions for tail equivalence when job sizes are intermediately regularly varying for policies that do not have access to job size information. These conditions highlight that tail equivalence depends on imposing a bound on the amount of consecutive time that a job has priority over others.
- Our sufficient conditions provide the first proof of tail equivalence for a number of well-known scheduling policies, including the Gittins policy, RMLF, and the Shortest Expected Remaining Processing Time first (SERPT) policy. Tail equivalence of these policies is a long-standing open question.
- Our sufficient conditions provide the first insight into how much preemption is needed in order to maintain tail equivalence. We specifically state which preemption frequencies guarantee tail optimality.
- Our proof of sufficiency includes an interesting foundational result for $M/G/1$ queues: a bound on the fractional moments of the $M/G/1$ busy period. Previously, only expressions for its integer moments were known.

ACKNOWLEDGMENTS

Ziv Scully was supported by an ARCS Foundation scholarship and the NSF Graduate Research Fellowship Program under Grant Nos. DGE-1745016 and DGE-125222. Lucas van Kreveld, Onno Boxma, and Jan-Pieter Dorsman were supported by the Netherlands Organisation for Scientific Research (NWO) through the Gravitation project NETWORKS, grant number 024.002.003. Adam Wierman was supported by NSF grant CNS-1518941.

REFERENCES

[1] R. Núñez-Queija. 2002. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research* 113, 1 (01 Jul 2002), 101–117. https://doi.org/10.1023/A:1020905810996

[2] M. Nuyens, A. Wierman, and B. Zwart. 2008. Preventing large sojourn times using SMART scheduling. *Operations Research* 56 (2008), 88–101.

[3] Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. 2018. SOAP: one clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 2, 1, Article 16 (April 2018), 30 pages. https://doi.org/10.1145/3179419

[4] Z. Scully, L. van Kreveld, O. Boxma, J. Dorsman, and A. Wierman. 2020. Characterizing Policies with Optimal Response Time Tails under Heavy-Tailed Job Sizes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4, 2, Article 30 (June 2020), 32 pages. https://doi.org/10.1145/3392148