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

ZIV SCULLY, Carnegie Mellon University, USA
LUCAS VAN KREVELD, University of Amsterdam, The Netherlands
ONNO BOXMA, Eindhoven University of Technology, The Netherlands
JAN-PIETER DORSMAN, University of Amsterdam, The Netherlands
ADAM WIERMAN, California Institute of Technology, USA

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*.

Additional Key Words and Phrases: 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. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 2, Article 30 (June 2020), 33 pages. https://doi.org/10.1145/3392148

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

Authors' addresses: Ziv Scully, Carnegie Mellon University, Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA, zscully@cs.cmu.edu; Lucas van Kreveld, University of Amsterdam, Korteweg-de Vries Institute for Mathematics, P.O.Box 94248, 1090 GE, Amsterdam, The Netherlands, l.r.vankreveld@uva.nl; Onno Boxma, Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands, o.j.boxma@tue.nl; Jan-Pieter Dorsman, University of Amsterdam, Korteweg-de Vries Institute for Mathematics, P.O.Box 94248, 1090 GE, Amsterdam, The Netherlands, j.l.dorsman@uva.nl; Adam Wierman, California Institute of Technology, Department of Computing and Mathematical Sciences, 1200 E California Blvd, Pasadena, CA, 91125, USA, adamw@caltech.edu.

2476-1249/2020/6-ART30 $15.00
https://doi.org/10.1145/3392148

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 [26] 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 [2]. If the job size distribution is also unknown, then the Randomized Multi-Level Feedback (RMLF) policy minimizes the competitive ratio for mean response time [16].

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, ranging from classical policies such as First Come First Served (FCFS) and SRPT, to newer ones such as Processor Sharing (PS) and its many generalizations [1, 3, 20]. 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, e.g., [10, 20, 30] and the references therein. 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.[1] 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 [12], in I/O patterns [23], the length of network sessions [22], 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 heavy-tailed traffic characteristics have a dramatic effect on the waiting times and response times experienced by users and 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 [30], but has a response time tail as bad as possible among work conserving policies under heavy-tailed job sizes. More precisely, assume that a job size X is regularly varying with index $-\alpha$ and denote this with $\mathrm{RV}(-\alpha)$.[2] Then, the response time in a $GI/GI/1$ FCFS queue is known to be $\mathrm{RV}(1-\alpha)$ [11]. A worse index is not possible under work-conserving policies since the residual busy period of such a queue is $\mathrm{RV}(1-\alpha)$. The response time in a $GI/GI/1$ SRPT queue, on the other hand, has the same index as the job size ($-\alpha$) in this setting. Since the response time of a job can never be smaller than its size, the response time index $-\alpha$ is optimal. So, SRPT is optimal in this heavy-tailed setting, whereas FCFS performs the worst in terms of response time tail index – the exact opposite of the light-tailed scenario.

Observations like this have led to significant research on the impact of the service discipline on delay asymptotics; cf. the surveys [7, 10]. 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, i.e., which policies

[1] A scheduling policy is *work conserving* if it always uses the server at full speed whenever a job is present in the system.
[2] A random variable X is regularly varying with index $-\alpha$ if $P(X > x) = L(x)x^{-\alpha}$ where the function $L(\cdot)$ is slowly varying, i.e., $L(ax)/L(x) \to 1$ for any $a > 0$.

have a response time tail that is of the same order as the tail of the job size distribution under regularly varying job sizes. This notion of "tail equivalence" (also referred to as tail optimality) has driven research for decades and at this point there is a variety of common policies that have been shown to be tail equivalent, including Processor Sharing (PS) [32], Foreground-Background (FB) [19], and Preemptive Shortest Job First (PSJF) [19].

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 [19], 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. [21], 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 [21] and a general characterization of tail-equivalent scheduling policies.

Contributions

In this paper, 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 [21], 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. Additionally, the sufficient conditions are 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 (Schedule Ordered by Age-based Priority) framework, recently introduced in [28]. In the SOAP framework, scheduling policies are expressed 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 Section 3, but intuitively the conditions ensure that old jobs do not receive priority over other jobs for too long. Specifically, for a job J, part one of the condition bounds the consecutive time that other jobs outrank J, and part two bounds the first age at which jobs will never in the future outrank J.

In general, there are three typical approaches for proving tail equivalence, see [7] for a survey. The first relies on a relationship between the tail behavior of a random variable Y and the behavior of its Laplace-Stieltjes transform (LST) $E[\mathrm{e}^{-sY}]$ for $s \downarrow 0$ ([6], p. 333). An expression for the response time LST of the single server queue under SOAP is indeed available (see [28]); however, it depends in such an intricate way on the rank function that this approach proved unsuitable for determining the tail behavior of the response time. Another common approach is to perform a sample path analysis of the policies, as was pursued by [21]. However, again, the form of dependence in the

rank function makes this difficult. Hence, in proving our main result, Theorem 3.3, we have adapted the probabilistic method developed by Núñez-Queija [19], which exploits a Markov-type inequality. While the method of [19] 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 (see Theorem 5.4).

To conclude, we summarize the contributions of the paper below:

- We provide a set of sufficient conditions for tail equivalence, i.e., optimality of the response time tail, 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 given their optimality among policies that do not use precise job size information.
- 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.

2 SYSTEM MODEL AND PRELIMINARIES

We consider an $M/G/1$ queue with arrival rate λ and job size X. We write $F(x) = \mathbf{P}\{X \le x\}$ and $\overline{F}(x) = 1 - F(x)$ for the distribution function and tail of X, respectively. The system load is denoted by $\rho = \lambda \mathbf{E}[X] < 1$. We write T for response time and T_x for size-conditional response time, that is, the response time for jobs of size x. Our focus is on the case where X is heavy-tailed. Specifically, we study the following class of heavy-tailed distributions.

Assumption 2.1.

(i) The tail $\overline{F}(\cdot)$ is intermediately regularly varying, meaning

$$\liminf_{\varepsilon \to 0^+} \liminf_{x \to \infty} \frac{\overline{F}((1+\varepsilon)x)}{\overline{F}(x)} = 1.$$

(ii) There exist $\beta > \alpha > 1$ such that the upper and lower Matuszewska indices of $\overline{F}(\cdot)$ are in $(-\beta, -\alpha)$.[3]

The class of intermediately regularly varying functions contains the class of regularly varying functions [13], which in turn contains Pareto distributions and power law distributions, among others. Roughly speaking, one can think of Assumption 2.1 as saying that $\overline{F}(\cdot)$ is bounded between two power law distributions, as the following Potter bound formalizes.

Lemma 2.2 ([6, Proposition 2.2.1]). *If Assumption 2.1 holds, then there exist constants $C, x_0 > 0$ such that for all $x_2 \ge x_1 \ge x_0$,*

$$\frac{1}{C}\left(\frac{x_2}{x_1}\right)^{-\beta} \le \frac{\overline{F}(x_2)}{\overline{F}(x_1)} \le C\left(\frac{x_2}{x_1}\right)^{-\alpha}.$$

[3] See Bingham et al. [6, Section 2.1].

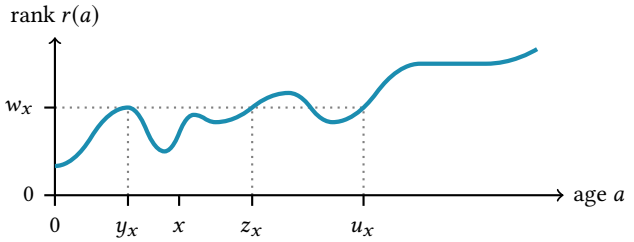

Fig. 3.1. Illustration of w_x, y_x, z_x, and u_x

The scheduling policies we study in this work are *SOAP* policies, a broad class of policies introduced by Scully et al. [28]. A SOAP policy is specified by a *rank function* $r : \mathbb{R}_+ \to \mathbb{R}$ mapping a job's *age*, which is the amount of time it has been served, to its *rank*, which is its priority level.[4] All SOAP policies have the same core scheduling rule: at every moment in time, *always serve the job of minimum rank*, breaking ties first-come, first-served. We assume a preemptive-resume model with no preemption overhead. We discuss SOAP policies in more detail, including how to analyze their response time, in Section 6.

3 OVERVIEW OF RESULTS

Our main result, Theorem 3.3, gives sufficient conditions for tail optimality in terms of properties of the rank function in SOAP. Thus, it characterizes properties of the prioritization of policies that guarantee optimal tail behavior. We first state a version of our main theorem in which the condition for tail optimality is slightly simplified. It states that a policy is tail-optimal if its rank function is bounded between two power functions in a specific way.

Theorem 3.1 (Simplified Result). *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1 using a SOAP scheduling policy whose rank function obeys*

$$r(a) \in \Omega(a^\gamma) \cap O(a^\delta)$$

for some $\delta > \gamma > 0$. If

$$\frac{\delta}{\gamma} - \frac{\gamma}{\delta} < \frac{\alpha - 1}{\beta},$$

then the policy is tail-optimal, i.e., $\lim_{x\to\infty} \mathbf{P}\left\{T > \frac{x}{1-\rho}\right\} / \overline{F}(x) = 1$.

The condition of Theorem 3.1 is easy to interpret and is suitable for tail-optimality proofs for many of the policies presented in Section 3. However, the result holds under more general conditions. To state these conditions formally, we need some notation. Let

- w_x be the worst rank attained by a job of size x,
- $y_x \le x$ be the earliest age with rank w_x,
- $z_x \ge x$ be the earliest age after x with rank $\ge w_x$, and
- $u_x \ge z_x$ be the latest age with rank $\le w_x$.

Figure 3.1 illustrates these quantities and they are defined formally in Definitions 6.2, 6.11, and 6.13.

Our sufficient conditions on the rank function are defined in terms of two quantities: $z_x - y_x$ and u_x.

[4]The full SOAP framework allows the rank function to be parametrized by additional information about a job, such as its size [28], but our results do not require using this feature.

Assumption 3.2.

(i) There exists $\zeta \in [0, \infty]$ such that $z_x - y_x = O(x^\zeta)$.
(ii) There exists $\eta \in [\max\{1, \zeta\}, \infty]$ such that $u_x = O(x^\eta)$.

Intuitively, ζ and η have the following interpretations. The smaller ζ is, the more quickly the system can preempt jobs. The smaller η is, the more each job is shielded from getting stuck behind larger jobs. Note that any rank function trivially satisfies Assumption 3.2 with $\zeta = \eta = \infty$, but, as suggested by the intuitive interpretations, we would like ζ and η to be small. Our main result states just how small ζ and η need to be to ensure tail optimality.

Theorem 3.3 (Main Theorem). *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1 and a SOAP scheduling policy whose rank function obeys Assumption 3.2. If*

$$\zeta - \frac{1}{\eta} < \frac{\alpha - 1}{\beta}, \tag{3.1}$$

then the policy is tail-optimal, i.e., $\lim_{x\to\infty} \mathbf{P}\left\{T > \frac{x}{1-\rho}\right\} / \overline{F}(x) = 1$.

As we prove now, Theorem 3.3 immediately implies its simplified version, Theorem 3.1.

Proof of Theorem 3.1. Precomposing any strictly increasing function with the rank function r yields an equivalent rank function that encodes the same SOAP policy, so we may assume without loss of generality that

$$r(a) \in \Omega(a^{\gamma/\delta}) \cap O(a).$$

This implies $w_x = O(x)$ and thus $u_x = \Omega(x^{\delta/\gamma})$. Therefore, Assumption 3.2 holds with $\zeta = \eta = \delta/\gamma$, so tail optimality follows from Theorem 3.3. □

The proof of Theorem 3.3 makes up the bulk of the remainder of the paper. However, a key component of our proof that we would like to highlight here is an analysis of the fractional moments of a busy period. The bounds we obtain are potentially of interest beyond the study of the tail of response time. In particular, let B_U be the length of a busy period with initial work U. Thus, a standard busy period would be B_X. We develop the following representation of the nth moment of a busy period for $n \in \mathbb{Z}_+$:

$$\mathbf{E}[B_U^n] = \sum_{i=1}^{I} d_i \frac{\mathbf{E}[U^{b_i}]}{(1-\rho)^{a_i}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X^{c_{ij}}],$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants that depend on n (see Lemma 5.2 and Corollary 5.3). Moreover, we show that this representation extends in a natural way to fractional moments of order $p = n - q > 0$, where $q \in (0, 1)$. Instead of equality, we obtain an upper bound in the case of fractional moments. We defer the full statement, which requires heavy notation, to Theorem 5.4.

Applications of Theorems 3.1 and 3.3

To illustrate the generality of the sufficient conditions in Theorems 3.1 and 3.3 it is interesting to consider how they can be applied to understand the response time tail of common policies. In this section, we illustrate the application of the theorems to understand tail optimality of policies for which no analysis is known. We focus on four examples: FB with limited preemption, Gittins, SERPT, and RMLF.

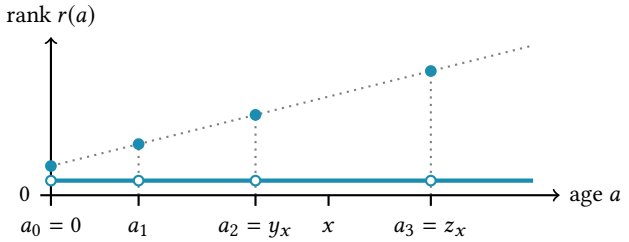

Fig. 3.2. Rank Function of FB with Limited Preemption

3.1 FB with Limited Preemption

Our first example focuses on a policy that is known to be tail optimal—FB—but limits the amount of preemption it may use. We consider FB here, but the same analysis can be performed for other policies that satisfy the conditions of Theorem 3.3. FB is particularly interesting because it is the optimal policy for job size distributions with a decreasing failure rate when no job size information is known. FB works by always serving the job of least age, sharing the processor equally in the case of ties. That is, FB is the SOAP policy with rank function $r(a) = a$ [28, Example 3.1].[5] As a result, it preempts jobs frequently and rarely works on a single job without interruption. In situations where there is a cost to preemption this is a signficant drawback. Thus, it is important to understand the performance of FB when preemption is limited.

To this end, we study a variation of FB with limited preemption (FB-LP) where preempting a job is only allowed when its age is one of a limited set of *checkpoints* $A \subseteq \mathbb{R}_+$. Specifically, FB-LP is the SOAP policy with rank function[6]

$$r(a) = \begin{cases} a + 2 & \text{if } a \in A \\ 1 & \text{otherwise.} \end{cases}$$

Figure 3.2 illustrates an example of FB-LP where A is a sequence $a_0 = 0, a_1, a_2, \ldots$.

The design of the FB-LP policy amounts to choosing the set of checkpoints A. In the extreme where $A = \mathbb{R}_+$, FB-LP is the same as using ordinary FB, which is tail-optimal but has frequent preemption and processor sharing. In the other extreme, setting $A = \varnothing$ is the same as using FCFS, which never preempts jobs but has pessimal response time tail behavior [11]. We therefore ask:

How frequently must checkpoints occur in order to ensure tail optimality?

We can answer this question using Theorem 3.3.

Consider a sequence of checkpoints $A = \{0, a_1, a_2, \ldots\}$. When $x \in (a_i, a_{i+1}]$, we have $y_x = a_i$ and $z_x = a_{i+1}$, as shown in Fig. 3.2. This means if $a_{i+1} - a_i = O\big(a_i^{\zeta}\big)$, then Assumption 3.2 holds with the same value of ζ and $\eta = \infty$. By Theorem 3.3, tail optimality holds if $\zeta < (\alpha - 1)/\beta$, implying the following result.

Corollary 3.4. *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1. The FB-LP policy with checkpoints $a_0 = 0, a_1, a_2, \ldots$ is tail-optimal if*

$$a_{i+1} - a_i = O\big(a_i^{\zeta}\big)$$

for some $\zeta < (\alpha - 1)/\beta$.

[5] FB is consistent with SOAP's FCFS tie-breaking convention [28, Algorithm B.1]: when two jobs have the same age, FB serves whichever came first for an instant, but this increases that job's age, causing FB to switch to the other.
[6] FB-LP generalizes the *discretized FB* policy introduced by Scully et al. [28, Example 3.7].

3.2 The Gittins Policy

Our next example application of Theorem 3.3 is the Gittins policy, which is the policy that minimizes mean response time of the $M/G/1$ queue when the job size distribution is known but individual job sizes are unknown. Gittins can be viewed as a SOAP policy whose rank function depends on the job size distribution [28, Example 3.6]:

$$r(a) = \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)}. \tag{3.2}$$

While Gittins is known to be optimal for the mean response time, the response time tail behavior of Gittins has resisted analysis. In this section we show that, under Assumption 2.1 and an additional technical condition, the Gittins policy is tail-optimal.

Given Theorem 3.1, it suffices to bound the Gittins rank function. Because $b = \infty$ is a possibility for the infimum in (3.2), by Lemma 2.2,

$$r(a) \le \int_a^\infty \frac{\overline{F}(t)}{\overline{F}(a)}\,\mathrm{d}t \le O(1) \int_a^\infty \left(\frac{t}{a}\right)^{-\alpha} \mathrm{d}t = O(a).$$

By Theorem 3.1, Gittins is tail-optimal if its rank function satisfies $r(a) = \Omega\left(a^\gamma\right)$, where $\frac{1}{\gamma} - \gamma < \frac{\alpha-1}{\beta}$. However, this is not the case for all job size distributions satisfying Assumption 2.1. For example, if the job size distribution has positive mass at some value x, then Gittins has $r(x-) = 0$.

Fortunately, under a mild additional condition, we can prove a lower bound on the Gittins rank function. Suppose that for sufficiently large x, the job size distribution has a well defined density $f(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\overline{F}(x)$ and hazard rate $h(x) = f(x)/\overline{F}(x)$. Then for sufficiently large ages a,

$$\begin{aligned}
r(a) &= \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)} \\
&= \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\int_a^b h(t)\overline{F}(t)\,\mathrm{d}t} \\
&\ge \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\left(\sup_{c\in(a,b)} h(c)\right)\int_a^b \overline{F}(t)\,\mathrm{d}t} \\
&= \inf_{b>a} \frac{1}{h(b)}.
\end{aligned}$$

This means that if $h(a) = O(a^{-\gamma})$ for $\gamma > 0$, then $r(a) = \Omega(a^\gamma)$, so Theorem 3.1 yields the following result.

Corollary 3.5. *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1. The Gittins policy is tail-optimal if the job size distribution has hazard rate $h(x) = O(x^{-\gamma})$ for some $\gamma > 0$ satisfying*

$$\frac{1}{\gamma} - \gamma < \frac{\alpha - 1}{\beta}.$$

In particular, $h(x) = O(x^{-\min\{1,\beta/(\alpha-1)\}})$ is sufficient for the Gittins policy to be tail-optimal.

3.3 Shortest Expected Remaining Processing Time

Shortest Expected Remaining Processing Time (SERPT) is a variation of SRPT for settings when the precise remaining sizes of jobs are not known, but the expected remaining size can be computed

given knowledge of the job size distribution. As the name implies, SERPT always serves whichever job has the least expected remaining size. Like Gittins, SERPT is a SOAP policy whose rank function depends on the job size distribution [28, Example 3.5]:

$$r(a) = \mathbf{E}[X - a \mid X > a] = \frac{\int_a^\infty \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a)}.$$

We show that SERPT is always tail-optimal. By Lemma 2.2, the rank function is bounded by

$$\Omega(a) = O(1)\int_a^\infty \left(\frac{t}{a}\right)^{-\beta} \mathrm{d}t \le r(a) \le O(1)\int_a^\infty \left(\frac{t}{a}\right)^{-\alpha} \mathrm{d}t = O(a),$$

so Theorem 3.1 implies tail optimality.

The *Monotonic SERPT* (M-SERPT) policy is a variant of SERPT introduced by Scully et al. [29]. Its rank function is the increasing envelope of SERPT's:

$$r(a) = \max_{0 \le b \le a} \mathbf{E}[X - b \mid X > b].$$

As with SERPT, Lemma 2.2 implies $r(a) \in \Omega(a) \cap O(a)$ for M-SERPT, so M-SERPT is also tail-optimal.

Corollary 3.6. *In an $M/G/1$ queue whose job size distribution obeys Assumption 2.1, SERPT and M-SERPT are both tail-optimal.*

The tail optimality of M-SERPT is particularly significant because M-SERPT has mean response time within a factor of 5 of Gittins's [29], and is simpler to understand and implement. Moreover, unlike our result for Gittins in Corollary 3.5, we require no additional assumptions on the job size distribution to ensure M-SERPT's tail optimality. Thus, M-SERPT is a policy that, for all distributions satisfying Assumption 2.1, is within a constant factor of optimal for both the mean response time and the tail of the response time.

Corollary 3.7. *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1. There exists a policy that is blind to job size information, namely M-SERPT, that is both tail-optimal and a constant-factor approximation for mean response time.*

3.4 Randomized Multi-Level Feedback

The Randomized Multi-Level Feedback (RMLF) policy is designed to have low mean response time when neither individual job sizes nor the job size distribution is known. Originally introduced in the worst-case scheduling literature [5, 16], RMLF was studied in the stochastic $GI/GI/1$ setting by Bansal et al. [4], who showed that RMLF is $O\left(\log \frac{1}{1-\rho}\right)$-competitive with SRPT for mean response time. However, no previous results exist for the tail of the response time under RMLF.

Here, we seek to apply our sufficient condition for tail optimality to RMLF. Unfortunately, RMLF does not fit into the SOAP framework as stated so far because not every job follows the same rank function: each job chooses a random parameter $v \in [0, 1]$ and then follows rank function[7]

$$r_v(a) = \min\{2^n \mid n \in \mathbb{N}, 2^{n+v} > a\}.$$

Nevertheless, we still have $a/2 \le r_v(a) \le 2a$ for all ages $a \ge 1$ and parameters $v \in [0, 1]$, so it seems that some adaptation of Theorem 3.1 ought to imply tail optimality of RMLF. This is indeed the case, but stating the adaptation requires some new terminology.

While RMLF is not a SOAP policy, it is what [27] calls a *SOAP Bubble* policy. The SOAP Bubble class of policies is a superset of the SOAP class. Much like SOAP, under a SOAP Bubble policy,

[7]The full SOAP framework is general enough to handle some policies with parametrized rank functions, but it requires the parameter to be chosen i.i.d. for each job [28], whereas RMLF draws $v \in [0, 1]$ using a different distribution for each job in the busy period.

every job's rank is a function of its age, and the system always serves the job of minimal rank, but *different jobs can have different rank functions*. Specifically, a SOAP Bubble policy is characterized by *lower and upper rank functions* $r^-, r^+ : \mathbb{R}_+ \to \mathbb{R}$, and the rank function r_j of each job j can be any function obeying $r^-(a) \le r_j(a) \le r^+(a)$. Therefore, RMLF is a SOAP Bubble policy with lower and upper rank functions

$$r^-(a) = \min\{2^n \mid n \in \mathbb{N}, 2^{n+1} > a\}$$
$$r^+(a) = \min\{2^n \mid n \in \mathbb{N}, 2^n > a\}.$$

In Appendix D, we formulate adaptations of Theorems 3.1 and 3.3 that apply to SOAP Bubble policies. For example, Theorem D.3 is the same as Theorem 3.1, except its precondition is $r^-(a) = \Omega(a^\gamma)$ and $r^+(a) = O(a^\delta)$. Applying Theorem D.3 to RMLF with $\gamma = \delta = 1$ implies that RMLF is tail-optimal.

Corollary 3.8. *In an $M/G/1$ queue whose job size distribution obeys Assumption 2.1, RMLF is tail-optimal.*

4 TAIL OPTIMALITY OF SOAP POLICIES

In the remainder of the paper we present a proof of our main result Theorem 3.3, namely that a SOAP policy is tail-optimal under certain conditions on the rank function. The foundation of the proof is an adaptation of a result by Núñez-Queija [19], which gives sufficient conditions for tail optimality. Hence, proving Theorem 3.3 amounts to verifying these conditions when Assumption (3.1) holds. The six steps of which this verification consists are presented in this section. Some of these steps rely on technical lemmas that require more background information. These lemmas are extensively introduced in Sections 5–7 and proven in Appendices A–C.

We now present a slight reformulation of the conditions in Núñez-Queija [19], relating to the conditional response time of a policy.

Condition 4.1. T_x is stochastically increasing in x.[8]

Condition 4.2. We have $\lim_{x\to\infty} \mathbf{E}[T_x]/x = 1/(1-\rho)$.

Condition 4.3.

(i) For all $\varepsilon > 0$,

$$\lim_{x\to\infty} \mathbf{P}\left\{T_X < \frac{(1-\varepsilon)x}{1-\rho} \;\middle|\; X > x\right\} = 0.$$

(ii) For all $\varepsilon > 0$,

$$\lim_{x\to\infty} \frac{1}{\bar{F}(x)} \mathbf{P}\left\{T_X > \frac{(1+\varepsilon)x}{1-\rho} \;\middle|\; X \le x\right\} = 0.$$

Based on these conditions, Núñez-Queija [19] deduces the following tail-optimality result.

Proposition 4.4. *If X obeys Assumption 2.1 and Conditions 4.1–4.3 hold, then*

$$\lim_{x\to\infty} \frac{1}{\bar{F}(x)} \mathbf{P}\left\{T > \frac{x}{1-\rho}\right\} = 1.$$

Remark. Proposition 4.4 differs from Núñez-Queija [19, Theorem 2.3] in that, instead of assuming Condition 4.3 directly, Núñez-Queija [19, Lemmas 2.1 and 2.2] proves it starting from a stronger condition. This adapted version is more appropriate for our analysis.

[8] That is, $\mathbf{P}\{T_x > t\} \le \mathbf{P}\{T_y > t\}$ for all $t \ge 0$ and all $0 \le x \le y$.

Since Proposition 4.4 immediately implies Theorem 3.3, what remains is to prove that Conditions 4.1–4.3 hold if the rank function parameters ζ, η satisfy $\zeta - \frac{1}{\eta} < \frac{\alpha-1}{\beta}$. We break down this proof in the following six steps.

Step 1: Express the tails of Condition 4.3 in terms of moments of T_x.
Step 2: Bound the moments obtained in Step 1. These bounds are used in Steps 4-6.
Step 3: Verify Condition 4.1.
Step 4: Verify Condition 4.2.
Step 5: Verify Condition 4.3(i).
Step 6: Verify Condition 4.3(ii).

In the remainder of this section we go through each step individually and refer to later sections for more technical details.

Step 1: *From tails to moments.*

To relate the tails of Condition 4.3 to moments of T_x, we need the following lemma, which does not rely on any specifics of the $M/G/1$ model or SOAP.

Let

$$g_x^p(t) = t^p - \mathbf{E}[T_x]^p - p\mathbf{E}[T_x]^{p-1}(t - \mathbf{E}[T_x]). \tag{4.1}$$

We can think of $g_x^p(t)$ as t^p minus the first two terms of the Taylor series of t^p about $t = \mathbf{E}[T_x]$.

Lemma 4.5. *For all $p, t > 0$,*

$$\mathbf{P}\{T_x > t\} \le \frac{\mathbf{E}[T_x^p] - \mathbf{E}[T_x]^p}{g_x^p(t)} \quad \text{if } t > \mathbf{E}[T_x],$$
$$\mathbf{P}\{T_x < t\} \le \frac{\mathbf{E}[T_x^p] - \mathbf{E}[T_x]^p}{g_x^p(t)} \quad \text{if } t < \mathbf{E}[T_x].$$

Proof. Note that $g_x^p(t)$ is decreasing in t for $t < \mathbf{E}[T_x]$ and increasing for $t > \mathbf{E}[T_x]$. Therefore, if $t < \mathbf{E}[T_x]$, then

$$\mathbf{P}\{T_x < t\} = \mathbf{P}\{T_x < t \text{ and } g_x^p(T_x) > g_x^p(t)\} \le \mathbf{P}\{g_x^p(T_x) > g_x^p(t)\},$$

and if $t > \mathbf{E}[T_x]$, then

$$\mathbf{P}\{T_x > t\} = \mathbf{P}\{T_x > t \text{ and } g_x^p(T_x) > g_x^p(t)\} \le \mathbf{P}\{g_x^p(T_x) > g_x^p(t)\}.$$

In both cases, Markov's inequality implies the desired bound. □

Step 2: *Moment bounds.*

The bounds presented in this step will be used to verify Conditions 4.2 and 4.3 (Steps 4-6). First, we split a job's response time T_x into two independent non-negative components, *waiting time* $Q[w_x]$ and *residence time* R_x [28]:

$$T_x =_{\text{st}} Q[w_x] + R_x. \tag{4.2}$$

We bound $\mathbf{E}[Q[w_x]]$ and $\mathbf{E}[R_x]$ in Lemmas 4.6 and 4.7 below, subject to the precondition of Theorem 3.3. For more details we refer to Section 7.

In the sequel we use the following notation. We write $f(x) = \check{o}(g(x))$ if there exists $\delta > 0$ such that $f(x) = o(x^{-\delta} g(x))$.

Lemma 4.6. *If (3.1) holds, then there exists $\beta' > \beta$ such that for all $p \in (0, \beta')$,*

$$\mathbf{E}[Q[w_x]^p] \le \check{o}(x^p),$$
$$\mathbf{E}[R_x^p] \le \left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p).$$

Lemma 4.7. *If* (3.1) *holds, then*

$$\mathbf{E}[R_x] \geq \frac{x}{1-\rho} - \breve{o}(x).$$

The proofs of Lemmas 4.6 and 4.7, presented in Section 7, require additional analysis of $M/G/1$ busy periods and SOAP policies, which is the purpose of Sections 5 and 6. Specifically, Section 5 derives a general bound for fractional busy period moments, and Section 6 describes how to express waiting and residence times in terms of busy periods.

We now use the moment bounds of Lemmas 4.6 and 4.7 for $Q[w_x]$ and R_x to obtain moment bounds for T_x.

Lemma 4.8. *If* (3.1) *holds, then*

$$\mathbf{E}[T_x^p] \leq \left(\frac{x}{1-\rho}\right)^p + \breve{o}(x^p) \qquad \text{for all } p \in [1, \beta')$$

$$\mathbf{E}[T_x]^p \geq \mathbf{E}[R_x]^p \geq \left(\frac{x}{1-\rho}\right)^p - \breve{o}(x^p) \quad \text{for all } p > 0.$$

Proof. Note that

$$(x \pm \breve{o}(x))^p = x^p \pm \breve{o}(x^p). \tag{4.3}$$

The lower bound on $\mathbf{E}[T_x]^p$ follows directly from (4.2), (4.3), and Lemma 4.7. For the upper bound on $\mathbf{E}[T_x^p]$, we use Minkowski's inequality to compute

$$\begin{aligned}
\mathbf{E}[T_x^p] &\leq \left(\mathbf{E}[Q[w_x]^p]^{1/p} + \mathbf{E}[R^p]^{1/p}\right)^p && \text{[by (4.2), Minkowski]} \\
&\leq \left(\breve{o}(x) + \left(\left(\frac{x}{1-\rho}\right)^p + \breve{o}(x^p)\right)^{1/p}\right)^p && \text{[by Lemma 4.6]} \\
&= \left(\frac{x}{1-\rho}\right)^p + \breve{o}(x^p). && \text{[by (4.3)]}
\end{aligned}$$

□

A direct consequence of Lemma 4.8 is the following.

Lemma 4.9. *If* (3.1) *holds, then there exists $p > \beta$ such that $\mathbf{E}[T_x^p] - \mathbf{E}[T_x]^p = \breve{o}(x^p)$ in the $x \to \infty$ limit.*

Proof. Choose $p \in (\beta, \beta')$ in Lemma 4.8. □

Remark. Núñez-Queija [19] uses a slightly different version of Lemma 4.9, showing that Condition 4.3 holds if there exists $p > \beta$ such that $\mathbf{E}[|T_x - \mathbf{E}[T_x]|^p] = \breve{o}(x^p)$. Unfortunately, working with the absolute central moment is difficult unless p is an even integer, which suffices for the simple policies considered by Núñez-Queija [19] but not for the broad class of SOAP policies we consider. Our Lemma 4.9 is easier to work with for odd and fractional p and, as shown below, still allows us to verify Condition 4.3.

Step 3: *Verification of Condition 4.1.*

Condition 4.1 is immediate for all SOAP policies (see Lemma 6.15).

Step 4: *Verification of Condition 4.2.*

Condition 4.2 follows from choosing $p = 1$ in Lemma 4.8.

Step 5: *Verification of Condition 4.3(i).*

Let $p > \beta$ be as in Lemma 4.9. We compute

$$\begin{aligned}
&\lim_{x\to\infty} \frac{1}{x^p} g_x^p\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right) \\
&\quad= \lim_{x\to\infty} \frac{1}{x^p}\left(\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right)^p - \mathbf{E}[T_x]^p - p\mathbf{E}[T_x]^{p-1}\left(\frac{(1\pm\varepsilon)x}{1-\rho} - \mathbf{E}[T_x]\right)\right) && \text{[by (4.1)]} \\
&\quad= \frac{(1\pm\varepsilon)^p - 1 - p(1\pm\varepsilon - 1)}{(1-\rho)^p} && \text{[by Condition 4.2]} \\
&\quad= \frac{(1\pm\varepsilon)^p - (1\pm\varepsilon p)}{(1-\rho)^p} \\
&\quad> 0,
\end{aligned}$$

and therefore

$$g_x^p\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right) = \Omega(x^p). \tag{4.4}$$

Combining this with Condition 4.1 and Lemma 4.5 implies Condition 4.3(i):

$$\begin{aligned}
\lim_{x\to\infty} \mathbf{P}\left\{T_X < \frac{(1-\varepsilon)x}{1-\rho} \;\middle|\; X > x\right\} &\le \lim_{x\to\infty} \mathbf{P}\left\{T_x < \frac{(1-\varepsilon)x}{1-\rho}\right\} && \text{[by Condition 4.1]} \\
&\le \lim_{x\to\infty} \frac{\mathbf{E}[T_x^\beta] - \mathbf{E}[T_x]^\beta}{g_{T_x}^\beta\left(\frac{(1-\varepsilon)x}{1-\rho}\right)} && \text{[by Lemma 4.5]} \\
&= \lim_{x\to\infty} \frac{\check{o}(x^\beta)}{\Omega(x^\beta)} && \text{[by (4.4) and Lemma 4.9]} \\
&= 0.
\end{aligned}$$

Step 6: *Verification of Condition 4.3(ii).*

We begin by applying Lemma 4.5:

$$\begin{aligned}
\mathbf{P}\left\{T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \le x\right\} &= \int_0^x \mathbf{P}\left\{T_t > \frac{(1+\varepsilon)x}{1-\rho}\right\} \mathrm{d}F(t) \\
&\le \int_0^x \frac{\mathbf{E}[T_t^p] - \mathbf{E}[T_t]^p}{g_t^p\left(\frac{(1+\varepsilon)x}{1-\rho}\right)} \mathrm{d}F(t).
\end{aligned} \tag{4.5}$$

We would like to apply (4.4) to the denominator, but the variables in the subscript and function argument do not match. To make them match, observe in (4.1) that $g_x^p(t)$ is decreasing in $\mathbf{E}[T_x]$ as long as $\mathbf{E}[T_x] < t$. By Conditions 4.1 and 4.2, for all sufficiently large x and all $t \in (0, x)$,

$$\mathbf{E}[T_t] \le \mathbf{E}[T_x] < \frac{(1+\varepsilon)x}{1-\rho},$$

so for sufficiently large x, we may replace t with x in the subscript in the denominator in (4.5). Using this along with Lemma 4.9 and (4.4) gives us

$$\begin{aligned}
\mathbf{P}\left\{T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \le x\right\} &\le \int_0^x \frac{\mathbf{E}[T_t^p] - \mathbf{E}[T_t]^p}{g_x^p\left(\frac{(1+\varepsilon)x}{1-\rho}\right)} \,\mathrm{d}F(t) \\
&\le \frac{1}{\Omega(x^p)} \int_0^x \left(\left(\frac{t}{1-\rho}\right)^p + \breve{o}(t^p) - \left(\frac{t}{1-\rho} - \breve{o}(t)\right)^p\right) \mathrm{d}F(t) \\
&\le O(x^{-p}) \int_0^x \breve{o}(t^p) \,\mathrm{d}F(t).
\end{aligned}$$

We continue the computation by integrating by parts and applying Lemma 2.2:

$$\begin{aligned}
\mathbf{P}\left\{T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \le x\right\} &\le O(x^{-p}) - O(x^{-p}) \int_{x_0}^x \breve{o}(t^p) \,\mathrm{d}\overline{F}(t) \\
&= O(x^{-p}) - O(x^{-p})\left(\breve{o}(x^p)\overline{F}(x) - O(1) - \int_{x_0}^x \breve{o}(t^{p-1})\overline{F}(t) \,\mathrm{d}t\right) \\
&\le O(x^{-p}) + O(x^{-p}) \int_{x_0}^x \breve{o}(t^{p-1}) \cdot C\overline{F}(x)\left(\frac{t}{x}\right)^{-\beta} \mathrm{d}t \\
&= O(x^{-p}) + O(x^{-(p-\beta)})\overline{F}(x) \int_{x_0}^x \breve{o}(t^{(p-\beta)-1}) \,\mathrm{d}t \\
&= O(x^{-p}) + \breve{o}(1)\overline{F}(x).
\end{aligned}$$

Lemma 2.2 and the fact that $p > \beta > \alpha$ imply that this is $\breve{o}(\overline{F}(x))$, so Condition 4.3(ii) holds.

Finally, the proof of our main result combines all of the pieces outlined in this section.

Proof of Theorem 3.3. Conditions 4.1–4.3 have been proven above and the theorem now follows from Proposition 4.4. □

The bulk of the remainder of the paper is devoted to proving Lemmas 4.6 and 4.7, which give bounds on moments of size-conditional waiting and residence times. Our proofs of these lemmas rely on detailed analysis of fractional moments of busy periods (Section 5) and on new general results about SOAP policies (Section 6).

5 FRACTIONAL MOMENTS OF BUSY PERIODS

A key component of our proof of Theorem 3.3 is an analysis of the fractional moments of an $M/G/1$ queue. We write B for the length of a busy period and B_U for the length of a busy period with initial work U.

We denote the Laplace-Stieltjes transform (LST) of a random variable V by

$$\widetilde{V}(s) = \mathbf{E}[\exp(-sV)].$$

We shall also encounter the *excess* $\mathcal{E}V$ of a random variable V. It has distribution

$$\mathbf{P}\{\mathcal{E}V > x\} = \int_0^x \frac{\mathbf{P}\{V > t\}}{\mathbf{E}[V]} \,\mathrm{d}t$$

and has LST

$$\widetilde{\mathcal{E}V}(s) = \frac{1 - \widetilde{V}(s)}{s\mathbf{E}[V]}. \tag{5.1}$$

Letting

$$\sigma(s) = s + \lambda(1 - \widetilde{B}(s)), \tag{5.2}$$

we can write the LSTs of B and B_U as

$$\begin{aligned} \widetilde{B}(s) &= \widetilde{X}(\sigma(s)), \\ \widetilde{B_U}(s) &= \widetilde{U}(\sigma(s))). \end{aligned} \tag{5.3}$$

Although the expression for $\widetilde{B}(s)$ is recursive, it suffices for extracting moments.

Let $\mathcal{D}$ be the derivative operator.

Lemma 5.1. *The derivative of $\sigma(s)$ satisfies*

$$\mathcal{D}\sigma(s) = \frac{1}{1 - \lambda(-\mathcal{D})\widetilde{X}(\sigma(s))}. \tag{5.4}$$

Proof. Differentiating (5.2) yields

$$\mathcal{D}\sigma(s) = 1 - \lambda\mathcal{D}\sigma(s) \cdot \mathcal{D}\widetilde{X}(\sigma(s)),$$

which rearranges to the desired equation. □

Lemma 5.2. *For all $n \in \mathbb{Z}_+$,*

$$(-\mathcal{D})^n \widetilde{B_U}(s) = \sum_{i=1}^{I} d_i (\mathcal{D}\sigma(s))^{a_i} \cdot (-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda(-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants, independent of the system parameters λ and X, satisfying

$$\begin{aligned} a_i, b_i &\geq 1 && \text{for all } i, \\ c_{ij} &\geq 2 && \text{for all } i, j, \\ b_i + \sum_{j=1}^{J_i} (c_{ij} - 1) &= n && \text{for all } i, \\ b_1 &> \ldots > b_n, \\ a_1 &= b_1 = n, \\ d_1 &= 1, \\ J_1 &= 0. \end{aligned}$$

Proof. See Appendix A.

As an immediate consequence, we have the following.

Corollary 5.3. *For all $n \in \mathbb{Z}_+$,*

$$\mathbf{E}[B_U^n] = \sum_{i=1}^{I} d_i \frac{\mathbf{E}[U^{b_i}]}{(1-\rho)^{a_i}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X^{c_{ij}}],$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are as in Lemma 5.2.

The main result of this subsection is that nearly the same formula works for fractional moments, though it gives an upper bound instead of an exact result. To bound $\mathbf{E}[B_U^p]$ for $p = n - q$, $n \in \mathbb{Z}_+$, we start with the formula for $\mathbf{E}[B_U^n]$, then decrease some of the exponents by q. Specifically, for each i, we decrease a_i and one more exponent of our choice, either b_i or one of the c_{ij}.

Theorem 5.4. *Let $p = n - q > 0$ for $n \in \mathbb{Z}_+$ and $q \in (0, 1)$. Then for all choices of $\chi_{ij} \in \{0, 1\}$ such that $\sum_{j=0}^{J_i} \chi_{ij} = 1$ for all i, we have*

$$\mathbf{E}[B_U^p] \le \sum_{i=1}^{I} d_i \frac{\mathbf{E}[U^{b_i - q\chi_{i0}}]}{(1-\rho)^{a_i - q}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X^{c_{ij} - q\chi_{ij}}],$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are as in Lemma 5.2.

Proof. See Appendix A.

Remark. Remerova et al. [25, Lemma 3] discuss the finiteness of $\mathbf{E}[f(B)]$ for the $M/G/1$ busy period B for a quite general class of functions $f(\cdot)$. They obtain bounds on moments of B that are more general but less sharp than Theorem 5.4. Bansal et al. [4] formulate a bound on fractional moments of $GI/GI/1$ busy periods, but their bound only characterizes the growth rate in the $\rho \to 1$ limit. Focusing on the $M/G/1$ setting, in which a recursive LST is known for busy periods, enables us to obtain a much sharper bound in Theorem 5.4 that characterizes the coefficients of each $1/(1-\rho)^b$ term.

6 SOAP BACKGROUND

Recall from Section 2 that a SOAP policy is one defined by a rank function $r : \mathbb{R}_+ \to \mathbb{R}$ mapping each job's age, or attained service, to its rank. The scheduler always serves the job of minimum rank, so lower rank means higher priority.

In this section we give background on how to analyze the mean response time of SOAP policies. Sections 6.1 and 6.2 review the response time analysis in [28], adapting the notation slightly to suit our needs. These expressions are hard to work with directly, and the complexity grows when considering higher moments. As such, we introduce new concepts and results in Sections 6.3 and 6.4 which help simplify the analysis.

6.1 Core SOAP Concepts

All of the definitions in the remainder of this section are given in terms of a generic SOAP policy with rank function r. We make the following standard assumption on rank functions [28, Appendix B].

Assumption 6.1. The rank function r is piecewise monotonic and piecewise differentiable.

The way [28] analyzes response time of SOAP policies is with the "tagged job" approach, following the journey of a specific job from arrival to departure. Suppose the tagged job has size x. One of their key insights is that to determine the tagged job's response time, its current rank is less important than the worst rank it will attain in its remaining time in the system.

Definition 6.2. The *worst future rank* of a job of size x at age a, written $w_x(a)$, is

$$w_x(a) = \sup_{a \le b < x} r(b).$$

The *worst ever rank* of a job of size x is $w_x = w_x(0)$.

When the tagged job initially enters the system, there may be a number of other jobs already present. Any other job with rank w_x or less is "relevant" to the tagged job, meaning it will receive some amount of service during the tagged job's time in the system.

Definition 6.3. The amount of *w-relevant work* a job has is the amount of service it needs to either finish or attain rank greater than w. Similarly, the amount of w-relevant work in a system is the total amount of w-relevant work of all jobs in the system.

To find the response time of the tagged job, we need to know the amount of w_x-relevant work it encounters upon arrival. Because the arrival process is Poisson, this means finding the steady-state distribution of the amount of w_x-relevant work in the system, for which we need the following definition.

Definition 6.4.

(i) The *kth w-relevant age interval* is $(b_k[w], c_k[w])$, where

$$\begin{aligned} b_0[w] &= 0 \\ c_0[w] &= \inf\{a \geq 0 \mid r(a) > w\} \\ b_k[w] &= \inf\{a > c_{k-1}[w] \mid r(a) \leq w\} \quad \text{for all } k \geq 1 \\ c_k[w] &= \inf\{a > b_k[w] \mid r(a) > w\} \quad \text{for all } k \geq 1. \end{aligned}$$

Additionally, let $K[w]$ be the number of w-relevant age intervals, namely the maximum k such that $b_k[w] < \infty$. It may be that $K[w] = \infty$.

(ii) The *kth w-relevant job segment* is

$$X_k[w] =_{\text{st}} \max\{0, \min\{X, c_k[w]\} - b_k[w]\}.$$

(iii) The *kth w-relevant load* is

$$\rho_k[w] = \lambda \mathbf{E}[X_k[w]].$$

For convenience, we also define

$$\rho_\Sigma[w] = \sum_{k=0}^{K[w]} \rho_k[w].$$

The tagged job can also be delayed by jobs arriving after it. The following definition helps us quantify this delay.

Definition 6.5. The *w-relevant busy period*, written $B[w]$, is the length of an $M/G/1$ busy period with arrival rate λ and job size $X_0[w]$. Similarly, the *w-relevant busy period with initial work U*, written $B_U[w]$, is the length of such a busy period with initial work U.

6.2 SOAP Response Time Analysis

To study the response time of SOAP policies, we introduce the following random variables.

Definition 6.6. The *residence time* of a job of size x, written R_x, is a random variable with transform

$$\widetilde{R_x}(s) = \exp\left(-\int_0^x (s + \lambda(1 - B[w_x(a)-])) \, \mathrm{d}a\right).$$

Abusing notation slightly, we can write the residence time as an integral of busy periods:

$$R_x =_{\text{st}} \int_0^x B_{\mathrm{d}a}[w_x(a)-].$$

Definition 6.7. The *rank-w waiting time*, written $Q[w]$, is a random variable that has the same distribution as a particular busy period:

$$Q[w] =_{\text{st}} B_{U[w]}[w-],$$

where $U[w]$ is the steady-state amount of w-relevant work, which is a random variable with transform [28, Lemma 5.2]

$$\widetilde{U[w]}(s) = \frac{1 - \rho_\Sigma[w] + \sum_{i=1}^{K[w]} \rho_i[w] \widetilde{\mathcal{E}X_i[w]}(s)}{1 - \rho_0[w] \widetilde{\mathcal{E}X_0[w]}(s)}.$$

Scully et al. [28, Theorem 5.4] show that for any SOAP policy, T_x is the independent sum of waiting and residence times, namely $T_x =_{\text{st}} Q[w_x] + R_x$, which implies the following formula for mean response time.

Corollary 6.8 ([28, Theorem 5.5]). *Under any SOAP policy,*

$$\begin{aligned}\mathbf{E}[Q[w]] &= \frac{\frac{\lambda}{2}\sum_{i=0}^{K[w]} \mathbf{E}[X_i[w]^2]}{(1-\rho_0[w])(1-\rho_0[w-])},\\ \mathbf{E}[R_x] &= \int_0^x \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a,\\ \mathbf{E}[T_x] &= \mathbf{E}[Q[w_x]] + \mathbf{E}[R_x].\end{aligned}$$

6.3 Stochastic Response Time Bounds

The next two lemmas, proven in Appendix B, bound the residence and waiting time in terms of busy periods. The main concept in their proofs is the observation that jobs with rank larger than w will not be served before the w-relevant busy period ends.

Lemma 6.9. *For any SOAP policy, the residence time of a job of size x is stochastically bounded by*

$$R_x \le_{\text{st}} B_x[w_x].$$

Lemma 6.10. *For any SOAP policy, the rank-w waiting time is stochastically bounded by*

$$Q[w] \le_{\text{st}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w],\\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w],\\ \vdots & \\ 0 & \text{w.p. } 1-\rho_\Sigma[w], \end{cases}$$

where

$$\begin{aligned}\pi_0[w] &= \frac{\rho_0[w](1-\rho_\Sigma[w])}{1-\rho_0[w]},\\ \pi_k[w] &= \frac{\rho_k[w]}{1-\rho_0[w]} \qquad \text{for all } k \ge 1.\end{aligned}$$

These lemmas allow us to express response time moments in terms of busy period moments, which can be further analyzed using Theorem 5.4.

6.4 Additional SOAP Bounds

The purpose of the next few definitions is to formalize Assumption 3.2. All of them relate to the worst rank for a job of size x.

Definition 6.11. The *maximum relevant age* of a job of size x is, roughly speaking, the latest age at which another job can possibly outrank it:

$$u_x = c_{K[w_x]}[w_x] = \sup\{a > 0 \mid r(a) \le w_x\}.$$

The next two definitions are due to Scully et al. [29].

Definition 6.12. A *hill age* is an age b such that $r(a) < r(b)$ for all ages $a < b$. An age that is not a hill age is called a *valley age*.[9]

[9]The full definition is more subtle [29, Definition 4.1], such as including ages z_x (see Definition 6.13) as hill ages, but we do not need the subtleties in this paper.

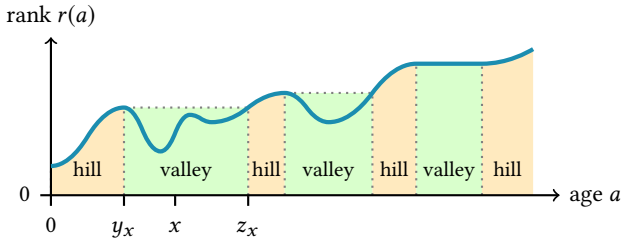

Fig. 6.1. Hills and Valleys

Definition 6.13. The *previous and next hill ages* of x are, respectively,

$$y_x = c_0[w_x-],$$
$$z_x = c_0[w_x].$$

For any x such that $y_x < z_x$, we call the interval (y_x, z_x) a *valley*, and any interval that does not overlap with a valley is called a *hill*. Figure 6.1 illustrates hills and valleys, including previous and next hill ages.

The next definition is not specific to SOAP but, as we will soon see, can be helpful when analyzing the moments of a SOAP policy's response time.

Definition 6.14.

(i) The *a-cutoff job segment* is

$$X\langle a\rangle =_{\text{st}} \min\{X, a\}.$$

(ii) The *a-cutoff load* is

$$\rho\langle a\rangle = \lambda \mathbf{E}[X\langle a\rangle].$$

The y_x- and z_x-cutoff job segments give us another way to write $X_0[w_x-]$ and $X_0[w_x]$:

$$\begin{aligned} X_0[w_x-] &=_{\text{st}} X\langle y_x\rangle, \\ X_0[w_x] &=_{\text{st}} X\langle z_x\rangle. \end{aligned} \tag{6.1}$$

The following lemma follows immediately from Definitions 6.6 and 6.7, observing that w-relevant busy periods are stochastically increasing in w.

Lemma 6.15.

(i) $Q[w]$ is stochastically increasing in w.
(ii) R_x is stochastically increasing in x.
(iii) T_x is stochastically increasing in x.

Note that Lemma 6.15 completes Step 3 of the proof described in Section 4. In fact, the only part of the proof outlined in Section 4 that remains to prove is Step 2, specifically Lemmas 4.6 and 4.7. We prove these lemmas in Section 7 with the help of the useful results given in the remainder of this section.

The next lemmas follow from integration by parts.

Lemma 6.16. *For any $p > 0$,*

$$\mathbf{E}[X_k[w]^p] = \int_{b_k[w]}^{c_k[w]} p(t - b_k[w])^{p-1} \overline{F}(t)\, \mathrm{d}t.$$

Lemma 6.17. *For any $p > 0$,*

$$\mathbf{E}[X\langle a\rangle^p] = \int_0^a pt^{p-1}\overline{F}(t)\,\mathrm{d}t.$$

Previous and next hill ages are also useful for bounding moments of $X_k[w_x]$ for $k \geq 1$, specifically by combining Lemma 6.16 with the following lemma.

Lemma 6.18. *For all ranks w and $k \geq 1$, if $x \in (b_k[w], c_k[w])$, then*

$$y_x \leq b_k[w] < x < c_k[w] \leq z_x.$$

Proof. By Definition 6.13, we have $x \in (y_x, z_x)$, where y_x is the first age at which the rank function reaches rank w_x, and z_x is the first age at which the rank function strictly exceeds w_x. Because $k \geq 1$, there must be some age $a \leq b_k[w]$ at which $r(a) > b_k[w]$, so $w_x > w$. But by Definition 6.4, a job's rank is at most w during $(b_k[w], c_k[w])$, so $y_x, z_x \notin (b_k[w], c_k[w])$. We therefore must have $y_x \leq b_k[w]$ and $z_x \geq c_k[w]$. □

7 PROVING TAIL OPTIMALITY

Recall from Section 4 that the proof of our main result, Theorem 3.3, is complete once we prove Lemmas 4.6 and 4.7. This section is devoted to proving these last two lemmas.

Many of the lemma statements in this section use similar preconditions on a parameter p. For convenience, we name these conditions $\Phi(p)$ and $\Psi(p)$:

$$\Phi(p) \iff \zeta - \frac{1}{\eta} < \frac{\alpha - 1}{p} \text{ or } p \leq 0,$$
$$\Psi(p) \iff 1 - \frac{1}{\zeta} < \frac{\alpha - 1}{p} \text{ or } p \leq 0.$$

For all $p \geq q$, we have[10]

$$\begin{aligned} &\Phi(p) \implies \Psi(p), \\ &\Phi(p) \implies \Phi(q), \\ &\Psi(p) \implies \Psi(q). \end{aligned} \tag{7.1}$$

We prove Lemmas 4.6 and 4.7 by way of the following more general statements.

Lemma 7.1. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\mathbf{E}[Q[w_x]^p] \leq \breve{o}(x^p).$$

Lemma 7.2. *For all $p > 0$ satisfying $\Psi(p-1)$, in the $x \to \infty$ limit,*

$$\mathbf{E}[R_x^p] \leq \left(\frac{x}{1-\rho}\right)^p + \breve{o}(x^p).$$

Lemma 7.3. *If $\zeta < 1$ or $\eta < \infty$, then in the $x \to \infty$ limit,*

$$\mathbf{E}[R_x] \geq \frac{x}{1-\rho} - \breve{o}(x).$$

Lemmas 7.1 and 7.2 immediately imply Lemma 4.6, and Lemma 7.3 immediately implies Lemma 4.7, so it remains only to prove Lemmas 7.1–7.3. In the remainder of this section we prove Lemma 7.2. We also give the main ideas of the proofs of Lemmas 7.1 and 7.3, deferring their full proofs to Appendix C.

[10]The $\Phi(p) \implies \Psi(p)$ implication follows from $\zeta \leq \eta$ and the fact that $\Psi(p)$ is vacuously true for $\zeta \leq 1$.

To prove Lemma 7.2, we use Lemma 6.9 to bound residence times using a busy period, namely $R_x \leq_{\text{st}} B_x[w_x]$. We can apply Theorem 5.4 to bound moments of the busy period $B_x[w_x]$ in terms of moments of its initial work, which is simply x, and its job size, which is $X_0[w_x]$. Thus, to bound moments of R_x, it suffices to bound moments of $X_0[w_x]$, which is the purpose of the following lemma.

Lemma 7.4. *For all $p > 0$ satisfying $\Psi(p)$, in the $x \to \infty$ limit,*

$$\mathbf{E}[X_0[w_x]^{p+1}] = \breve{o}(x^p).$$

Proof. By (6.1) and Lemma 6.17,

$$\mathbf{E}[X_0[w_x]^{p+1}] = \mathbf{E}[X\langle z_x\rangle^{p+1}] = \int_0^{z_x} (p+1)t^p \bar{F}(t)\,\mathrm{d}t.$$

Hence for $x \to \infty$, Assumption 3.2 implies

$$\begin{aligned}\mathbf{E}[X_0[w_x]^{p+1}] &= \int_0^{O(x^{\max\{1,\zeta\}})} O(t^{p-\alpha})\,\mathrm{d}t \\ &= O(x^{\max\{0,(p-(\alpha-1))\max\{1,\zeta\}\}}). \end{aligned} \tag{7.2}$$

If $\zeta \leq 1$, then (7.2) is $O(x^{\max\{0,p-(\alpha-1)\}}) = \breve{o}(x^p)$. If instead $\zeta > 1$, then $\Psi(p)$ implies (7.2) is $\breve{o}(x^p)$. □

Armed with bounds on moments of $X_0[w_x]$, we are now ready to prove Lemma 7.2.

Proof of Lemma 7.2. Let $p = n - q$ for $n \in \mathbb{Z}_+$ and $q \in (0,1)$. We again apply Theorem 5.4, choosing $\chi_{i0} = 1$ for all i. Using that and Lemma 6.9, we obtain

$$\begin{aligned}\mathbf{E}[R_x^p] &\leq \mathbf{E}[B_x^p[w_x]] && \text{[by Lemma 6.9]} \\ &\leq \left(\frac{x}{1-\rho_0[w_x]}\right)^p + \sum_{i=2}^{I} d_i \frac{x^{b_i-q}}{(1-\rho_0[w_x])^{a_i-q}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X_0[w_x]^{c_{ij}}]. && \text{[by Theorem 5.4]}\end{aligned}$$

Recall from Lemma 5.2 that

$$b_i - q + \sum_{j=1}^{J_i}(c_{ij}-1) = n - q = p.$$

This means for all i and j, we have $c_{ij} - 1 \leq p - b_i \leq p - 1$, so $\Psi(c_{ij}-1)$ holds by (7.1). We can therefore apply Lemma 7.4, which yields

$$\begin{aligned}\sum_{i=2}^{I} d_i \frac{x^{b_i-q}}{(1-\rho_0[w_x])^{a_i-q}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X_0[w_x]^{c_{ij}}] &= \sum_{i=2}^{I} O(x^{b_i-q}) \prod_{j=1}^{J_i} \breve{o}(x^{c_{ij}-1}) \\ &= \sum_{i=2}^{I} \breve{o}\big(x^{b_i-q+\sum_{j=1}^{J_i}(c_{ij}-1)}\big) \\ &= \breve{o}(x^p). \end{aligned}$$

□

It remains only to prove Lemma 7.1 and Lemma 7.3. The proof of Lemma 7.1, an upper bound on moments of waiting time, follows essentially the same outline as the proof of Lemma 7.2: we use Lemma 6.10 to bound waiting time in terms of busy periods, use Theorem 5.4 to bound the moments of those busy periods in terms of moments of $X_i[w_x]$, then use Lemma 7.5 below to bound those moments. Finally, we prove Lemma 7.3 by combining several results from Section 6.

Lemma 7.5. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\sum_{k=1}^{K[w_x]} \mathbf{E}[X_k[w_x]^{p+1}] = \breve{o}(x^p).$$

Proofs of Lemmas 7.1, 7.3, and 7.5. See Appendix C.

8 DISCUSSION

Over the past decades, much effort has been given to the task of designing policies that maintain the optimal response time tail, i.e., a response time tail that is equally heavy as the tail of the job size distribution. While the analysis of individual policies has been successful in many cases, e.g., SRPT and FB, there are many important policies that have resisted analysis and, further, little is known about which scheduling mechanisms provably lead to tail optimality. In this paper, we provide general sufficient conditions on the type of prioritization that ensures tail optimality in policies that do not have access to job sizes. Our sufficient conditions enable the first results on tail-optimality for Gittins, RMLF, SERPT, and FB with limited preemption.

Although our sufficient conditions define a broad class of tail-optimal policies, it must be stressed that they are not necessary. For instance, Processor Sharing (PS), which is known to be tail-optimal, does not use job sizes and does not satisfy our sufficient conditions since it is not a SOAP policy. Thus, it is important to continue to develop both necessary and sufficient conditions for tail optimality. An interesting open question is to identify sufficient conditions that unify the results in [21] for size-based policies with the results in this paper on policies that do not have access to job size information. Additionally, the only necessary condition known for tail optimality is given by [31], which proves that all tail-optimal policies must "remain stable when faced with the arrival of a job with infinite size." It is not known if this condition is also sufficient. Toward this end, Guillemin et al. [14] have developed an interesting probabilistic method to prove tail equivalence of the response time and service time for a large class of $M/G/1$ processor-sharing queues (with and without impatience, and with finite or infinite capacity). It would be interesting to investigate whether their approach is applicable to a class of SOAP policies. See also Sections 2.4 and 3 of [8], where the approaches of [14] and [19] are compared and unified.

Another interesting research topic is to weaken the goal and, instead of looking to characterize policies that are tail-optimal, characterize classes of policies with near-optimal response time tails. It is known that the orders of the job size and response time tails can differ by any number $\gamma \in (0, 1]$ [9, 18], and so a natural question is: what forms of prioritization achieve these intermediate response time tails?

It is also worth considering tail optimality among light-tailed job size distributions. Are there sufficient conditions on prioritization that ensure tail optimality in the light-tailed setting? The results of [31] highlight that if a policy is tail-optimal under heavy-tailed job sizes it cannot be tail-optimal under light-tailed job sizes, and thus it is clear that the sufficient conditions must change. However, little is known about the general class of policies that are (nearly) tail-optimal under light-tailed job sizes. More broadly, an important open question first posed in [18] is: what are sufficient conditions that ensure a policy is near-optimal for the response time tail under both heavy- and light-tailed job size distributions?

Finally, it remains to be seen what tail optimality results extend to more complicated queueing models. This includes single-server systems with variable service rate, such as systems using computational sprinting [17, 24], as well as systems with multiple servers.

ACKNOWLEDGMENTS

The authors are grateful to Bert Zwart for providing some useful references. 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] S. Aalto, U. Ayesta, S. Borst, V. Misra, and R. Núñez-Queija. 2007. Beyond processor sharing. *ACM SIGMETRICS Performance Evaluation Review* 34 (2007), 36–43.
[2] S. Aalto, U. Ayesta, and R. Righter. 2009. On the Gittins index in the M/G/1 queue. *Queueing Systems* 63 (2009), 437–458.
[3] E. Altman, K. Avrachenkov, and U. Ayesta. 2006. A survey on discriminatory processor sharing. *Queueing Systems* 53 (2006), 53–63.
[4] N. Bansal, B. Kamphorst, and B. Zwart. 2018. Achievable performance of blind policies in heavy traffic. *Mathematics of Operations Research* 43, 3 (2018), 949–964.
[5] L. Becchetti and S. Leonardi. 2004. Nonclairvoyant scheduling to minimize the total flow time on single and parallel machines. *Journal of the ACM (JACM)* 51, 4 (2004), 517–539.
[6] N. Bingham, C. Goldie, and J. Teugels. 1987. *Regular Variation.* Cambridge University Press.
[7] S. Borst, O. Boxma, R. Núñez Queija, and B. Zwart. 2003. The impact of the service discipline on delay asymptotics. *Performance Evaluation* 54 (2003), 175–206.
[8] S.C. Borst, R. Núñez-Queija, and B. Zwart. 2006. Sojourn-time asymptotics in processor-sharing queues. *Queueing Systems* 53 (2006), 31–51.
[9] O. Boxma and D. Denisov. 2011. Sojourn time tails in the single server queue with heavy-tailed service times. *Queueing Systems* 69, 2 (2011), 101–119.
[10] O. Boxma and B. Zwart. 2007. Tails in scheduling. *ACM SIGMETRICS Performance Evaluation Review* 34, 4 (2007), 13–20.
[11] J. Cohen. 1973. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability* 10, 2 (1973), 343–353.
[12] M. Crovella and A. Bestavros. 1996. Self-similarity in World Wide Web traffic: evidence and possible causes. *Proceedings of ACM Sigmetrics '96* (1996), 160–169.
[13] S. Foss, D. Korshunov, and S. Zachary. 2013. *An Introduction to Heavy-tailed and Subexponential Distributions.* Springer, New York, NY.
[14] F. Guillemin, Ph. Robert, and B. Zwart. 2004. Tail asymptotics for processor-sharing queues. *Advances in Applied Probability* 36 (2004), 525–543.
[15] W. Johnson. 2002. The curious history of Faà di Bruno's formula. *The American Mathematical Monthly* 109, 3 (2002), 217–234.
[16] B. Kalyanasundaram and K. R. Pruhs. 1997. Minimizing flow time nonclairvoyantly. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*. IEEE, 345–352.
[17] N. Morris, C. Stewart, L. Chen, R. Birke, and J. Kelley. 2018. Model-driven computational sprinting. In *Proceedings of the Thirteenth EuroSys Conference*. 1–13.
[18] J. Nair, A. Wierman, and B. Zwart. 2010. Tail-robust scheduling via limited processor sharing. *Performance Evaluation* 67, 11 (2010), 978–995.
[19] 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
[20] M. Nuyens and A. Wierman. 2008. The foreground-background queue: a survey. *Performance Evaluation* 65 (2008), 286–307.
[21] M. Nuyens, A. Wierman, and B. Zwart. 2008. Preventing large sojourn times using SMART scheduling. *Operations Research* 56 (2008), 88–101.
[22] K. Park and W. Willinger. 2000. *Self-similar Network Traffic and Performance Evaluation.* Wiley.
[23] D. Peterson. 1996. Data center I/O patterns and power laws. *CMG Proceedings* (1996).
[24] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. Pipe, T. Wenisch, and M. Martin. 2012. Computational sprinting. In *Proceedings of the 18th Symposium on High Performance Computer Architecture*. IEEE, 1–12.
[25] M. Remerova, S. Foss, and B. Zwart. 2014. Random fluid limits of an overloaded polling model. *Advances in Applied Probability* 46 (2014), 76–101.

[26] L. Schrage and L. Miller. 1966. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research* 14 (1966), 670–684.
[27] Z. Scully and M. Harchol-Balter. 2018. SOAP Bubbles: Robust Scheduling Under Adversarial Noise. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 144–154. https://doi.org/10.1109/ALLERTON.2018.8635963
[28] 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
[29] Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. 2020. Simple Near-Optimal Scheduling for the M/G/1. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4, 1, Article 11 (March 2020), 29 pages. https://doi.org/10.1145/3379477
[30] A. Stolyar and K. Ramanan. 2001. Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability* 11 (2001), 1–48.
[31] A. Wierman and B. Zwart. 2012. Is tail-optimal scheduling possible? *Operations Research* 60, 5 (2012), 1249–1257.
[32] B. Zwart and O. Boxma. 2000. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems* 35, 1-4 (2000), 141–166.

A PROOFS FOR SECTION 5

Given a random variable V, it is well known how to obtain positive integer moments of V from its LST: for all $n \in \mathbb{Z}_+$,

$$(-\mathcal{D})^n \widetilde{V}(s) = \mathbf{E}[V^n \exp(-sV)], \tag{A.1}$$

so in particular $\mathbf{E}[V^n] = (-\mathcal{D})^n \widetilde{V}(0)$. Obtaining fractional moments of V from its LST is trickier but also possible: for all $p \geq 0$, letting $p = n - q > 0$ for $n \in \mathbb{Z}_+$ and $q \in (0, 1)$, we have

$$\begin{aligned}
\mathbf{E}[V^p] &= \int_0^\infty t^{n-q} \, \mathrm{d}\mathbf{P}\{V \leq t\} \\
&= \int_{t=0}^\infty \frac{t^{n-q}}{\Gamma(q)} \int_{s=0}^\infty (st)^{q-1} \exp(-st) \cdot t \, \mathrm{d}s \, \mathrm{d}\mathbf{P}\{V \leq t\} \\
&= \int_{s=0}^\infty \frac{1}{s^{1-q}\Gamma(q)} \int_{t=0}^\infty t^n \exp(-st) \, \mathrm{d}\mathbf{P}\{V \leq t\} \, \mathrm{d}s \\
&= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} (-\mathcal{D})^n \widetilde{V}(s) \, \mathrm{d}s.
\end{aligned} \tag{A.2}$$

Lemma 5.2. *For all $n \in \mathbb{Z}_+$,*

$$(-\mathcal{D})^n \widetilde{B_U}(s) = \sum_{i=1}^I d_i (\mathcal{D}\sigma(s))^{a_i} \cdot (-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda (-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants, independent of the system parameters λ and X, satisfying

$$\begin{aligned}
a_i, b_i &\geq 1 && \text{for all } i, \\
c_{ij} &\geq 2 && \text{for all } i, j, \\
b_i + \sum_{j=1}^{J_i} (c_{ij} - 1) &= n && \text{for all } i, \\
b_1 &> \ldots > b_n, \\
a_1 &= b_1 = n, \\
d_1 &= 1, \\
J_1 &= 0.
\end{aligned}$$

Proof. We proceed by induction on n. The base case of $n = 0$ is immediate by (5.3), so we turn to the inductive step. By relabeling, we can have $a_1 > \ldots > a_n$ without loss of generality. We address the constraint on the $i = 1$ constants at the end of the proof.

For $a, b, c_j \in \mathbb{Z}_+$, let

$$\tau_{a,b,\langle c_1,\ldots,c_J\rangle}(s) = (\mathcal{D}\sigma(s))^a \cdot (-\mathcal{D})^b \widetilde{U}(\sigma(s)) \prod_{j=1}^{J} \lambda(-\mathcal{D})^{c_j} \widetilde{X}(\sigma(s)).$$

We abbreviate $c = \langle c_1, \ldots, c_J\rangle$. Call $b + \sum_{j=1}^{J}(c_j - 1)$ the *degree* of $\tau_{a,b,c}(s)$. For the inductive step, it suffices to show that the derivative of a term with degree n is a sum of terms with degree $n + 1$. Using Lemma 5.1, we compute[11]

$$\begin{aligned} -\mathcal{D}\tau_{a,b,c}(s) = {} & \tau_{a+1,b+1,c}(s) + a\tau_{a+2,b,\langle c_1,\ldots,c_J,2\rangle}(s) \\ & + \sum_{j=1}^{J} \tau_{a+1,b,\langle c_1,\ldots,c_j+1,\ldots,c_J\rangle}(s), \end{aligned} \tag{A.3}$$

in which each term has degree $n + 1$, as desired.

We now address the constraint on the $i = 1$ term, again by induction on n. The base case of $n = 0$ is immediate by (5.3), and the inductive step follows from plugging $a = b = n$ into (A.3). □

Lemma A.1. *Let $p = m - q > 0$ for $m \in \mathbb{Z}_+$ and $q \in (0, 1)$. Then for any nonnegative random variable V, we have*

$$\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} (-\mathcal{D})^m \widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \le \frac{\mathbf{E}[V^p]}{(1-\rho)^{1-q}}.$$

Proof. We first show that for all $s > 0$,

$$\frac{\sigma(s)}{s} \le \frac{1}{1-\rho}. \tag{A.4}$$

By (5.2),

$$\begin{aligned} \sigma(s) &= s + \lambda(1 - \tilde{X}(\sigma(s))) \\ &= s + \lambda \int_0^\infty (1 - \exp(-x\sigma(s)))\,\mathrm{d}\mathbf{P}\{X \le x\} \\ &\le s + \lambda \mathbf{E}[X]\sigma(s), \end{aligned}$$

which implies (A.4). Making a change of variable $\sigma = \sigma(s)$, we compute

$$\begin{aligned} & \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} (-\mathcal{D})^m \widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \\ & \quad \le \frac{1}{(1-\rho)^{1-q}} \int_0^\infty \frac{1}{(\sigma(s))^{1-q}\Gamma(q)} (-\mathcal{D})^m \widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \qquad \text{[by (A.4)]} \\ & \quad = \frac{1}{(1-\rho)^{1-q}} \int_0^\infty \frac{1}{\sigma^{1-q}\Gamma(q)} (-\mathcal{D})^m \widetilde{V}(\sigma)\,\mathrm{d}\sigma \\ & \quad = \frac{\mathbf{E}[V^p]}{(1-\rho)^{1-q}}. \qquad \text{[by (A.2)]} \end{aligned}$$

□

[11] Two clarifications about the list notation below. First, in the second term on the right-hand side, we append c with an extra 2. Second, for each j, in the jth summand of the third term on the right-hand side, we increase the jth element of c by 1.

Proof of Theorem 5.4. Let V be a nonnegative random variable. By (A.1), for all $m \in \mathbb{Z}_+$ and $s > 0$,

$$(-\mathcal{D})^m \widetilde{V}(\sigma(s)) \le (-\mathcal{D})^m \widetilde{V}(0) = \mathbf{E}[V^m], \tag{A.5}$$

which when applied to (5.4) implies

$$\mathcal{D}\sigma(s) \le \frac{1}{1-\rho}. \tag{A.6}$$

Combining (A.6) and Lemma 5.2 yields

$$\begin{aligned}
\mathbf{E}[B_U^p] &= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} (-\mathcal{D})^n \widetilde{B_U}(s)\,\mathrm{d}s \\
&= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} \left(\sum_{i=1}^I d_i (\mathcal{D}\sigma(s))^{a_i} \cdot (-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda (-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)) \right) \mathrm{d}s \\
&\le \sum_{i=1}^I \frac{d_i}{(1-\rho)^{a_i - 1}} \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} \left((-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda (-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)) \right) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s. \qquad \text{(A.7)}
\end{aligned}$$

It remains only to bound the integral in (A.7), which we do separately for each value of i. If $\chi_{i0} = 1$, then applying Lemma A.1 with $V = U$ and $m = b_i$ along with (A.5) yields

$$\begin{aligned}
&\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} \left((-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda (-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)) \right) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \\
&\quad \le \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} \left((-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \right) \prod_{j=1}^{J_i} \lambda \mathbf{E}[X^{c_{ij}}] \qquad \text{[by (A.5)]} \\
&\quad \le \frac{\mathbf{E}[U^{b_i - q}]}{(1-\rho)^{1-q}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X^{c_{ij}}], \qquad \text{[by Lemma A.1]}
\end{aligned}$$

which gives the desired bound for the ith summand in (A.7). The case where $\chi_{ij} = 1$ for some $j \ge 1$ is very similar, except we apply Lemma A.1 with $V = X$ and $m = c_{ij}$. □

Remark. Note that one might hope to get a simpler expression for $(-\mathcal{D})^n \widetilde{B_U}(s) = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}s^n} \widetilde{U}(\sigma(s))$ *by applying the following compact form of Faà di Bruno's formula for derivatives of composite functions [15]:*

$$\frac{\mathrm{d}^n \widetilde{U}(\sigma(s))}{\mathrm{d}s^n} = \sum_{k=1}^n \left. \frac{\mathrm{d}^k \widetilde{U}(y)}{\mathrm{d}y^k} \right|_{y=\sigma(s)} B_{n,k}(\sigma'(s), \sigma^{(2)}(s), \ldots, \sigma^{(n-k+1)}(s)), \tag{A.8}$$

where the partial or incomplete exponential Bell polynomials $B_{n,k}$ *are given by*

$$B_{n,k}(x_1, \ldots, x_{n-k+1}) = \sum \frac{n!}{j_1! \ldots j_{n-k+1}!} \left(\frac{x_1}{1!} \right)^{j_1} \cdots \left(\frac{x_{n-k+1}}{(n-k+1)!} \right)^{j_{n-k+1}},$$

where the sum is taken over all nonnegative integers $(j_1, \ldots, j_{n-k+1})$ *with* $j_1 + \ldots + j_{n-k+1} = k$ *and* $j_1 + 2j_2 + \ldots + (n-k+1)j_{n-k+1} = n$. *In particular,* $B_{n,n}(x) = x^n$, *so that the leading* $(k = n)$ *term of* $(-\mathcal{D})^n \widetilde{B_U}(s)$ *is bounded by* $\mathbf{E}[U^n](\sigma'(s))^n$. *In the proof of Lemma 7.2, where we have a busy period with initial work* x *and job size* $X_0[w_x]$, *that would very quickly lead to the leading term* $\left(\frac{x}{1-\rho_0[w_x]}\right)^p$. *However, to show that the remaining* $n-1$ *terms in* (A.8) *are* $\check{o}(x^p)$ *requires a detailed study of the higher derivatives of* $\sigma(s)$.

B PROOFS FOR SECTION 6

Lemma 6.9. *For any SOAP policy, the residence time of a job of size x is stochastically bounded by*

$$R_x \leq_{\text{st}} B_x[w_x].$$

Proof. By Definition 6.6,

$$R_x =_{\text{st}} \int_0^x B_{\text{d}a}[w_x(a)-].$$

It is clear from Definition 6.5 that $B_U[w]$ is stochastically increasing in w for any U. Definition 6.2 implies $w_x \geq w_x(a)$ for all $a \geq 0$, so

$$R_x \leq_{\text{st}} \int_0^x B_{\text{d}a}[w_x] =_{\text{st}} B_x[w_x]. \qquad \square$$

Lemma 6.10. *For any SOAP policy, the rank-w waiting time is stochastically bounded by*

$$Q[w] \leq_{\text{st}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w], \\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w], \\ \vdots & \\ 0 & \text{w.p. } 1-\rho_\Sigma[w], \end{cases}$$

where

$$\pi_0[w] = \frac{\rho_0[w](1-\rho_\Sigma[w])}{1-\rho_0[w]},$$
$$\pi_k[w] = \frac{\rho_k[w]}{1-\rho_0[w]} \qquad \text{for all } k \geq 1.$$

Proof. Following the approach of [28, Section 5], one can think of $Q[w]$ as defined by the following process. A job J with initial rank r arrives at a random time. Because the system uses FCFS tiebreaking between jobs of the same rank, job J is first served when

- all jobs that arrived *before* J either complete or have rank strictly greater than r, and
- all jobs that arrived *after* J either complete or have rank greater than or equal to r.

Then $Q[w]$ is the amount of time from J's arrival to its first service.

Define $Q'[w]$ in the same way as $Q[w]$ but in a system that breaks rank ties by prioritizing all other jobs over J. Clearly, $Q[w] \leq_{\text{st}} Q'[w]$. But we can succinctly describe $Q'[w]$: it is either 0 or the excess of a w-relevant busy period with some amount of initial work. Specifically, the initial work is a kth w-relevant job segment for some $k \geq 0$. Thus, letting $\pi_k[w]$ be the steady-state probability that the system is in a w-relevant busy period started by a kth w-relevant segment, we have

$$Q'[w] =_{\text{st}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w], \\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w], \\ \vdots & \\ 0 & \text{w.p. } 1-\rho_\Sigma[w]. \end{cases}$$

All that remains is to compute the probabilities $\pi_k[w]$. For $k \geq 1$, each job's kth w-relevant segment starts a w-relevant busy period with expected length $\mathbf{E}[X_k[w]]/(1-\rho_0[w])$,[12] and jobs

[12]The possibility of a job completing before reaching its kth w-relevant segment is not a problem: this corresponds to the outcome $X_k[w] = 0$, in which case we think of the segment as starting a w-relevant busy period of length 0.

arrive at rate λ, so for $k \geq 1$,

$$\pi_k = \frac{\rho_k[w]}{1 - \rho_0[w]}.$$

The $k = 0$ case is similar, except that a job's 0th w-relevant segment only starts a w-relevant busy period if the system has no w-relevant work. Thus, the arrival rate of jobs whose 0th w-relevant segment starts a w-relevant busy period is $\lambda(1 - \rho_\Sigma[w])$, so

$$\pi_0 = \frac{\rho_0[w](1 - \rho_\Sigma[w])}{1 - \rho_0[w]}.$$ □

C PROOFS FOR SECTION 7

Lemma 7.1. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\mathbf{E}[Q[w_x]^p] \leq \check{o}(x^p).$$

Proof. By Lemma 6.10,

$$\mathbf{E}[Q[w_x]^p] \leq \sum_{k=0}^{K[w_x]} \pi_k[w_x] \cdot \mathbf{E}[\mathcal{E}B_{X_k[w_x]}[w_x]^p], \tag{C.1}$$

where

$$\pi_0[w_x] = \frac{\rho_0[w_x](1 - \rho_\Sigma[w_x])}{1 - \rho_0[w_x]} = O(1) \cdot \rho_0[w_x],$$
$$\pi_k[w_x] = \frac{\rho_k[w_x]}{1 - \rho_0[w_x]} \qquad = O(1) \cdot \rho_k[w_x] \quad \text{for all } k \geq 1.$$

We start by bounding each term of the sum in (C.1). Observe first that for any random variable V and any $p \geq 0$,

$$\mathbf{E}[\mathcal{E}V^p] = \frac{\mathbf{E}[V^{p+1}]}{(p+1)\mathbf{E}[V]}.$$

Then for all $k \geq 0$, we compute

$$\begin{aligned}
\pi_k[w_x] \cdot \mathbf{E}[\mathcal{E}B_{X_k[w_x]}[w_x]^p] &= O(1) \cdot \rho_k[w_x] \cdot \mathbf{E}[\mathcal{E}B_{X_k[w_x]}[w_x]^p] \\
&= O(1) \cdot \rho_k[w_x] \cdot \frac{\mathbf{E}[B_{X_k[w_x]}[w_x]^{p+1}]}{\mathbf{E}[X_k[w_x]]} \\
&= O(1) \cdot \mathbf{E}[B_{X_k[w_x]}[w_x]^{p+1}].
\end{aligned} \tag{C.2}$$

Bounding the right hand side of (C.2) requires bounding fractional busy period moments. We therefore apply Theorem 5.4 to the $(p+1)$th moment above, letting $p + 1 = n - q$ for $n \in \mathbb{Z}_+$ and $q \in (0, 1)$. We choose $\chi_{i0} = 1$ for all i such that $b_i \geq 2$ and $\chi_{i1} = 1$ for all other i.[13] This choice ensures that

$$b_i - q\chi_{i0} \geq 1$$
$$c_{ij} - q\chi_{ij} > 1,$$

which will allow the use of Lemmas 7.4 and 7.5 later in the proof.

[13]This choice requires checking that $J_i \geq 1$ for all i such that $b_i = 1$, which holds by (C.4) and the fact that $n \geq 2$.

Applying Theorem 5.4 to (C.2) yields, for $x \to \infty$,

$$\pi_k[w_x] \cdot \mathbf{E}[\mathcal{E}B_{X_k[w_x]}[w_x]^p] \le O(1) \cdot \sum_{i=1}^{I} d_i \frac{\mathbf{E}[X_k[w_x]^{b_i - q\chi_{i0}}]}{(1-\rho_0[w_x])^{a_i-q}} \prod_{j=1}^{J_i} \lambda \mathbf{E}[X_0[w_x]^{c_{ij}-q\chi_{ij}}]$$

$$= O(1) \cdot \sum_{i=1}^{I} \mathbf{E}[X_k[w_x]^{b_i - q\chi_{i0}}] \prod_{j=1}^{J_i} \mathbf{E}[X_0[w_x]^{c_{ij}-q\chi_{ij}}]. \tag{C.3}$$

Recall from Lemma 5.2 that[14]

$$b_i - q\chi_{i0} + \sum_{j=1}^{J_i} (c_{ij} - q\chi_{ij} - 1) = n - q = p + 1. \tag{C.4}$$

This means for all i and j, we have $c_{ij} - q\chi_{ij} - 1 \le p$, so $\Psi(c_{ij} - q\chi_{ij} - 1)$ holds by (7.1). Returning to (C.3), applying Lemma 7.4 and (C.4) gives us

$$\begin{aligned}
&\pi_k[w_x] \cdot \mathbf{E}[\mathcal{E}B_{X_k[w_x]}[w_x]^p] \\
&\le \sum_{i=1}^{I} \mathbf{E}[X_k[w_x]^{b_i - q\chi_{i0}}] \prod_{j=1}^{J_i} \check{o}(x^{c_{ij}-q\chi_{ij}-1}) && \text{[by Lemma 7.4]} \\
&= \sum_{i=1}^{I} \mathbf{E}[X_k[w_x]^{b_i - q\chi_{i0}}] \max\left\{O(1), \check{o}\left(x^{\sum_{j=1}^{J_i}(c_{ij}-q\chi_{ij}-1)}\right)\right\} \\
&= \sum_{i=1}^{I} \mathbf{E}[X_k[w_x]^{b_i - q\chi_{i0}}] \max\{O(1), \check{o}(x^{p+1-(b_i-q\chi_{i0})})\}, && \text{[by (C.4)]}
\end{aligned} \tag{C.5}$$

where the $O(1)$ covers the $J_i = 0$ case, in which the product is empty.

We now return to bounding the right-hand side of (C.1), substituting in (C.5) and interchanging the order of summation:

$$\mathbf{E}[Q[w_x]^p] \le \sum_{i=1}^{I} \max\{O(1), \check{o}(x^{p+1-(b_i-q\chi_{i0})})\} \sum_{k=0}^{K[w_x]} \mathbf{E}[X_k[w_x]^{b_i-q\chi_{i0}}].$$

It suffices to show that each term of the outer sum is $\check{o}(x^p)$. We would like to use Lemmas 7.4 and 7.5. We know $\Phi(b_i - q\chi_{i0} - 1)$ holds by (7.1) and (C.4). However, the lemmas also require $b_i - q\chi_{i0} > 1$, yet it may be the case that $b_i - q\chi_{i0} = 1$. To handle this case, we use the fact that by Definition 6.4,

$$\sum_{k=0}^{K[w_x]} \mathbf{E}[X_k[w_x]] = \mathbf{E}[\sum_{k=0}^{K[w_x]} X_k[w_x]] \le \mathbf{E}[X] = O(1).$$

Combining this with Lemmas 7.4 and 7.5 gives us

$$\begin{aligned}
&\max\{O(1), \check{o}(x^{p+1-(b_i-q\chi_{i0})})\} \sum_{k=0}^{K[w_x]} \mathbf{E}[X_k[w_x]^{b_i-q\chi_{i0}}] \\
&\le \max\{O(1), \check{o}(x^{p+1-(b_i-q\chi_{i0})})\} \cdot \max\{O(1), \check{o}(x^{b_i-q\chi_{i0}-1})\} \\
&= \check{o}(x^p).
\end{aligned}$$

□

[14]Note that we are applying Theorem 5.4 to a $(p+1)$th moment, not a pth moment.

Lemma 7.3. *If $\zeta < 1$ or $\eta < \infty$, then in the $x \to \infty$ limit,*

$$\mathbf{E}[R_x] \geq \frac{x}{1-\rho} - \breve{o}(x).$$

Proof. We consider the $\zeta < 1$ and $\eta < \infty$ cases separately.

Case 1: $\zeta < 1$. Definitions 6.4 and 6.13 imply (see also Fig. 3.1)

$$w_x(a) = w_x \quad \text{for all } a \in [0, y_x). \tag{C.6}$$

From this we compute

$$\begin{aligned}
\mathbf{E}[R_x] &= \int_0^x \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a && \text{[by Corollary 6.8]}\\
&\geq \int_0^{y_x} \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a\\
&= \frac{y_x}{1-\rho_0[w_x-]} && \text{[by (C.6)]}\\
&= \frac{y_x}{1-\rho\langle y_x\rangle} && \text{[by (6.1)]}\\
&\geq \frac{x - O(x^\zeta)}{1-\rho\langle x - O(x^\zeta)\rangle} && \text{[by Assumption 3.2 and Lemma 6.18]}\\
&= \frac{x}{1-\rho\langle \Omega(x)\rangle} - \breve{o}(x).
\end{aligned}$$

For any $\rho' \in (0, \rho)$, we have

$$\frac{1}{1-\rho'} = \frac{1}{1-\rho} \cdot \frac{1}{1+\frac{\rho-\rho'}{1-\rho}} \geq \frac{1}{1-\rho} - \frac{\rho-\rho'}{(1-\rho)^2}. \tag{C.7}$$

By (C.7) with $\rho' = \rho\langle\Omega(x)\rangle$, it suffices to show that $\rho - \rho\langle\Omega(x)\rangle = \breve{o}(1)$. This indeed holds by Lemma 6.17 and Assumption 2.1:

$$\begin{aligned}
\rho - \rho\langle\Omega(x)\rangle &= \lambda \int_0^\infty \bar{F}(t)\,\mathrm{d}t - \lambda \int_0^{\Omega(x)} \bar{F}(t)\,\mathrm{d}t && \text{[by Lemma 6.17]}\\
&= \int_{\Omega(x)}^\infty O(t^{-\alpha})\,\mathrm{d}t && \text{[by Assumption 2.1]}\\
&= O(x^{-(\alpha-1)})\\
&= \breve{o}(1).
\end{aligned}$$

Case 2: $\eta < \infty$. A job's worst future rank $w_x(a)$ is decreasing in a by Definition 6.2, so for all $a \in [0, x)$,

$$w_x(a) \geq w_x(x-) = r(x-).$$

Applying this to Corollary 6.8 yields

$$\mathbf{E}[R_x] = \int_0^x \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a \geq \frac{x}{1-\rho_0[r(x-)]}.$$

By (C.7) with $\rho' = \rho_0[r(x-)]$, it suffices to show $\rho - \rho_0[r(x-)] = \breve{o}(1)$.

Let $f(\cdot)$ be a strictly increasing function such that for sufficiently large t,

$$u_{t+} \leq f(t) \leq 2u_{t+}. \tag{C.8}$$

Definition 6.4 tells us that for all ages $a > f(t)$, we have $r(a) > w_{t+}$. But by Definitions 6.2 and 6.4, we have $r(x-) \le r(c_0[r(x-)]) = w_{c_0[r(x-)]+}$, so it must be that $x \le f(c_0[r(x-)])$. Because $f(\cdot)$ is strictly increasing, it is invertible, so Assumption 3.2 and (C.8) imply

$$c_0[r(x-)] \ge f^{-1}(x) = \Omega(x^{1/\eta}).$$

Combining this with (6.1) and Lemma 6.17, we compute, similarly to the previous case,

$$\begin{aligned}
\rho - \rho_0[r(x-)] &= \rho - \rho\langle\Omega(x^{1/\eta})\rangle && \text{[by (6.1)]}\\
&= \int_{\Omega(x^{1/\eta})}^{\infty} O(t^{-\alpha})\,\mathrm{d}t && \text{[by Lemma 6.17]}\\
&= O(x^{-(\alpha-1)/\eta})\\
&= \breve{o}(1).
\end{aligned}$$

□

Lemma 7.5. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\sum_{k=1}^{K[w_x]} \mathbf{E}[X_k[w_x]^{p+1}] = \breve{o}(x^p).$$

Proof. We compute

$$\begin{aligned}
\sum_{k=1}^{K[w_x]} \mathbf{E}[X_k[w_x]^{p+1}] &= \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(t - b_k[w_t])^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 6.16]}\\
&\le \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(z_t - y_t)^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 6.18]}\\
&\le \sum_{k=1}^{K[w_x]} \int_{c_{k-1}[w_x]}^{c_k[w_x]} (p+1)(z_t - y_t)^p \overline{F}(t)\,\mathrm{d}t && \text{[by Definition 6.4]}\\
&\le \int_0^{u_x} (p+1)(z_t - y_t)^p \overline{F}(t)\,\mathrm{d}t. && \text{[by Definition 6.11]}
\end{aligned}$$

Hence for $x \to \infty$, Assumption 3.2 implies

$$\begin{aligned}
\sum_{k=1}^{K[w_x]} \mathbf{E}[X_k[w_x]^{p+1}] &\le \int_0^{O(x^\eta)} O(t^{\zeta p - \alpha})\,\mathrm{d}t\\
&= O(x^{\max\{0,\,\eta(\zeta p - (\alpha-1))\}}),
\end{aligned}$$

which $\Phi(p)$ implies is $\breve{o}(x^p)$. □

D GENERALIZATION TO SOAP BUBBLE POLICIES

In this appendix we generalize our main results, Theorems 3.1 and 3.3, to SOAP Bubble policies, which are a superset of SOAP policies that is introduced in [27]. To review (see Section 3.4), a SOAP Bubble policy has *lower and upper rank functions*

$$r^-, r^+ : \mathbb{R}_+ \to \mathbb{R}.$$

A SOAP Bubble policy works like a SOAP policy, except each job j can have a different rank function r_j. Each job's rank function may be set arbitrarily, provided it remains within the "bubble" between the lower and upper rank functions, meaning for all jobs j and ages a,

$$r^-(a) \le r_j(a) \le r^+(a).$$

One can view ordinary SOAP policies as the special case with $r^-(a) = r(a) = r^+(a)$.

To upper bound the response time of a SOAP bubble policy, one can essentially replicate the analysis of SOAP policies, but replacing each use of r with either r^- or r^+ as appropriate. The intuition is that a tagged job has maximal response time if it follows r^+ while every other job follows r^-. Specifically,

- when defining worst future rank (Definition 6.2), replace r with r^+; and
- when defining w-relevant work, intervals, segments, and load (Definitions 6.3 and 6.4), replace r with r^-.

For details, see [27].

To generalize Theorems 3.1 and 3.3, we begin by defining some new notation. Let[15]

$$\begin{aligned} w_x^- &= \sup_{0 \le b < x} r^-(b), \\ w_x^+ &= \sup_{0 \le b < x} r^+(b), \\ y_x^- &= c_0[w_x^-], \\ z_x^- &= c_0[w_x^-], \\ y_x^+ &= c_0[w_x^+], \\ z_x^+ &= c_0[w_x^+], \\ u_x^+ &= c_{K[w_x^+]}[w_x^+]. \end{aligned}$$

Throughout our proofs, we can simply replace u_x with u_x^+, but y_x and z_x are more subtle.

- The main use of y_x and z_x is through Lemma 6.18, which is used in Lemma 7.5. The lemma statement now holds with y_x^- and z_x^-.
- There is one more use of y_x in Lemma 7.3, and this one needs to be replaced with y_x^+.
- There is one more use of z_x in Lemma 7.4, and this one needs to be replaced with z_x^+.

This implies the following generalizations of Assumption 3.2 and Theorem 3.3. The only substantial change is that we need two versions of ζ because there are two version of y_x and z_x. This ends up breaking the $\Phi(p) \Rightarrow \Psi(p)$ implication in (7.1), so we add some extra preconditions to our result.

Assumption D.1.

(i) There exists $\zeta^- \in [0, \infty]$ such that $z_x^- - y_x^- = O(x^{\zeta^-})$.
(ii) There exists $\zeta^+ \in [\zeta^-, \infty]$ such that $z_x^+ - y_x^+ = O(x^{\zeta^+})$.
(iii) There exists $\eta^+ \in [\max\{1, \zeta^+\}, \infty]$ such that $u_x = O(x^{\eta^+})$.

Theorem D.2. *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1 and a SOAP Bubble scheduling policy whose lower and upper rank functions obey Assumption D.1. If*

$$\begin{gathered} \zeta^- - \frac{1}{\eta^+} < \frac{\alpha - 1}{\beta}, \\ 1 - \frac{1}{\zeta^+} < \frac{\alpha - 1}{\beta}, \\ \text{and either } \zeta^+ < 1 \text{ or } \eta^+ < \infty, \end{gathered}$$

then the policy is tail-optimal, i.e., $\lim_{x \to \infty} \frac{1}{\overline{F}(x)} \mathbf{P}\left\{T > \frac{x}{1-\rho}\right\} = 1$.

One can obtain the following simplified condition in much the same way as done in Theorem 3.1.

[15]Recall that $c_i[w]$ is defined using r^-.

Theorem D.3. *Consider an $M/G/1$ queue whose job size distribution obeys Assumption 2.1 using a SOAP Bubble scheduling policy whose lower and upper rank functions obey*

$$r^-(a) = \Omega(a^\gamma),$$
$$r^+(a) = O(a^\delta)$$

for some $\delta > \gamma > 0$. If

$$\frac{\delta}{\gamma} - \frac{\gamma}{\delta} < \frac{\alpha - 1}{\beta},$$

then the policy is tail-optimal, i.e., $\lim_{x\to\infty} \frac{1}{\bar{F}(x)} \mathbf{P}\left\{T > \frac{x}{1-\rho}\right\} = 1$.

Received January 2020; revised February 2020; accepted March 2020