Performance Evaluation 145 (2021) 102150

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Optimal multiserver scheduling with unknown job sizes in heavy traffic

Ziv Scully *, Isaac Grosof, Mor Harchol-Balter

Carnegie Mellon University, Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Available online 12 October 2020

Keywords:
Scheduling
Response time
Heavy traffic
M/G/k
Gittins policy
Shortest expected processing time (SERPT)

ABSTRACT

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

In this work we introduce *monotonic Gittins* (M-Gittins), a new variation of the Gittins policy, and show that it minimizes mean response time in the heavy-traffic M/G/k for a wide class of finite-variance job size distributions. We also show that the *monotonic shortest expected remaining processing time* (M-SERPT) policy, which is simpler than M-Gittins, is a 2-approximation for mean response time in the heavy traffic M/G/k under similar conditions. These results constitute the most general optimality results to date for the M/G/k with unknown job sizes. Our techniques build upon work by Grosof et al. (2018), who study simple policies, such as SRPT, in the M/G/k; Bansal et al. (2018), Kamphorst and Zwart (2020), and Lin et al. (2010), who analyze mean response time scaling of simple policies in the heavy-traffic M/G/1; and Aalto et al. (2009,2011) and Scully et al. (2018,2020), who characterize and analyze the Gittins policy in the M/G/1.

1. Introduction

Scheduling to minimize mean response time[1] of the M/G/k queue is an important problem in queueing theory. The single-server $k = 1$ case has been well studied. If the scheduler has access to each job's exact size, the *shortest remaining processing time* (SRPT) policy is easily shown to be optimal [1]. If the scheduler does not know job sizes, which is very often the case in practical systems, then a more complex policy called the *Gittins* policy is known to be optimal [2–4]. The Gittins policy tailors its priority scheme to the job size distribution, and it takes a simple form in certain special cases. For example, for distributions with *decreasing hazard rate* (DHR), Gittins becomes the *foreground-background* (FB) policy,[2] so FB is optimal in the M/G/1 for DHR job size distributions [2,3,5].

In contrast to the M/G/1, the M/G/k with $k \geq 2$ has resisted exact analysis, even for very simple scheduling policies. As such, much less is known about minimizing mean response time in the M/G/k, with the only nontrivial results holding

* Corresponding author.
E-mail addresses: zscully@cs.cmu.edu (Z. Scully), igrosof@cs.cmu.edu (I. Grosof), harchol@cs.cmu.edu (M. Harchol-Balter).

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

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

https://doi.org/10.1016/j.peva.2020.102150

under heavy traffic.[3] For known job sizes, recent work by Grosof et al. [6] shows that a multiserver analogue of SRPT is optimal in the heavy-traffic M/G/k. For unknown job sizes, Grosof et al. [6] address only the case of DHR job size distributions, showing that a multiserver analogue of FB is optimal in the heavy-traffic M/G/k.[4] But in general, optimal scheduling is an open problem for unknown job sizes, even in heavy traffic. We therefore ask:

> *What scheduling policy minimizes mean response time in the heavy-traffic M/G/k with unknown job sizes and general job size distribution?*

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

- As part of their heavy-traffic optimality proofs, Grosof et al. [6] use a tagged job method to stochastically bound M/G/k response time under each of SRPT and FB relative to M/G/1 response time (Fig. 2.1) under the same policy.
- Lin et al. [7] and Kamphorst and Zwart [8] characterize the heavy-traffic scaling of M/G/1 mean response time under SRPT and FB, respectively.
- Scully et al. [9] show that a policy called *monotonic shortest expected remaining processing time* (M-SERPT), which is considerably simpler than Gittins, has M/G/1 mean response time within a constant factor of that of Gittins.

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

When searching for a policy to minimize mean response time, a natural candidate is a multiserver analogue of Gittins. As a first step, one might hope to use the tagged job method of Grosof et al. [6] to stochastically bound M/G/k response time under Gittins relative to M/G/1 response time. Unfortunately, the tagged job method does not apply to multiserver Gittins, because it relies on both stochastic and worst-case properties of the scheduling policy, whereas Gittins has poor worst-case properties.

One of our key ideas is to introduce a new variant of Gittins, called *monotonic Gittins* (M-Gittins), that has better worst-case properties than Gittins while maintaining similar stochastic properties. This allows us to generalize the tagged job method [6] to M-Gittins, thus bounding its M/G/k response time relative to its M/G/1 response time.

Our M/G/k analysis of M-Gittins reduces the question of whether M-Gittins is optimal in the heavy-traffic M/G/k to analyzing the heavy-traffic scaling of M-Gittins's M/G/1 mean response time. However, there are no heavy-traffic scaling results for the M/G/1 under policies other than SRPT [7], FB [8], *first-come, first served* (FCFS) [10,11], and a small number of other simple policies [12,13]. To remedy this, we derive heavy-traffic scaling results for M-Gittins in the M/G/1. It turns out that analyzing M-Gittins directly is very difficult. Fortunately, M-Gittins has a simpler cousin, M-SERPT, which Scully et al. [9] introduce and analyze. We analyze M-SERPT in heavy traffic as a key stepping stone in our heavy-traffic analysis of M-Gittins.

This paper makes the following contributions:

- We introduce the M-Gittins policy and prove that it minimizes mean response time in the heavy-traffic M/G/k for a large class of finite-variance job size distributions (Theorem 3.1).
- We also prove that the simple and practical M-SERPT policy is a 2-approximation for mean response time in the heavy-traffic M/G/k for a large class of finite-variance job size distributions (Theorem 3.2).
- We characterize the heavy-traffic scaling of mean response time in the M/G/1 under Gittins, M-Gittins, and M-SERPT (Theorem 3.3).

Section 3 formally states these results and compares them to prior work. Their proofs rely on a large collection of intermediate results, which we outline in detail in Section 4 and prove in Sections 5–7.

2. Preliminaries

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

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

We write F for the cumulative distribution function of X and $\overline{F}(x) = 1 - F(x)$ for its tail. We assume that X has a continuous, piecewise-monotonic[5] hazard rate

$$h(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}x}F(x)}{\overline{F}(x)}.$$

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

[4] Both the SRPT and FB optimality results of Grosof et al. [6] hold under technical conditions similar to finite variance.

[5] A function is piecewise-monotonic if, roughly speaking, it switches between increasing and decreasing finitely many times in any compact interval.

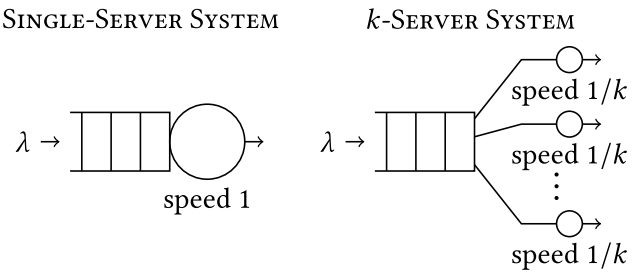


Fig. 2.1. Single-server and k-server systems.

We also frequently work with the expected remaining size of a job at age a, which is $\mathbf{E}[X - a \mid X > a]$. We assume it, too, is continuous and piecewise-monotonic as a function of a.

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

2.1. SOAP policies and rank functions

All of the scheduling policies considered in this work are in the class of *SOAP policies* [14], generalized to a multiserver setting. In a single-server setting, a SOAP policy π is specified by a *rank function*

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

which maps a job's *age*, namely the amount of service it has received so far, to its *rank*, or priority level. Single-server SOAP policies work by always serving the job of *minimal rank*, breaking ties in FCFS fashion.[6]

As an example, FB is a SOAP policy with $r^{\text{FB}}(a) = a$. Because lower age corresponds to lower rank, FB prioritizes the job of least age.[7]

A multiserver SOAP policy uses the same rank function as its single-server analogue. The only difference is that the system can serve up to k jobs, so a multiserver SOAP policy works as follows:

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

We often compare the k-server variant of a policy π to its single-server analogue. When it is necessary to distinguish between them, we write π-k for the k-server version of a policy, so π-1 is the single-server version. We write $T_x^{\pi\text{-}k}$ for the size-conditional response time distribution of jobs of size x under π-k, and we write $T^{\pi\text{-}k}$ for the overall response time distribution.

There are four main policies we consider in this work: SERPT, M-SERPT, Gittins, and M-Gittins. None of the policies need job size information, but each uses the job size distribution to tune its rank function. As an example, Fig. 2.2 shows the four rank functions for a bounded distribution with nonmonotonic hazard rate.

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

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

As a reminder, lower rank means better priority, so, as hinted by its name, SERPT prioritizes the job of least expected remaining size.

Definition 2.2. The *monotonic SERPT* (M-SERPT) policy is the SOAP policy with monotonic rank function

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

Definition 2.3. The *Gittins* policy is the SOAP policy with rank function

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

6 The full SOAP class allows a job's rank to depend on both its age and its "static" characteristics, such as its size or class, but we do not use this generality in this paper.

7 When multiple jobs are tied for least age, FB shares the server among all such jobs because the rank function is increasing. See Scully et al. [14, Appendix B] for details.

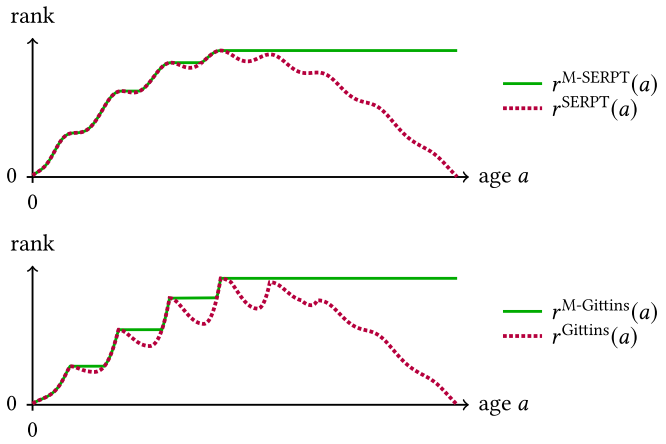


Fig. 2.2. Rank function examples.

Definition 2.4. The *monotonic Gittins* (M-Gittins) policy is the SOAP policy with monotonic rank function

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

The M-Gittins and M-SERPT policies, which both have monotonic rank functions, are the primary focus of this paper. Some of our intermediate results apply more broadly to any policy with a monotonic rank function.

Definition 2.5. A SOAP policy π is *monotonic* if its rank function is nondecreasing, meaning $r^{\pi}(a) \leq r^{\pi}(b)$ for all ages $a < b$.[8]

Figure 2.2 shows the SERPT, M-SERPT, Gittins, and M-Gittins rank functions for a bounded distribution with nonmonotonic hazard rate. Notice that SERPT and Gittins are not monotonic. This makes it hard to analyze their M/G/k response time (Appendix A). In contrast, the M-SERPT and M-Gittins are monotonic: their rank functions alternate between constant regions and strictly increasing regions.

While the rank functions of Gittins and SERPT may not be monotonic, they are still well behaved under our assumptions on the job size distribution.

Lemma 2.6. *Under the assumption that the job size distribution X has continuous and piecewise-monotonic hazard rate and expected remaining size functions, each of r^{SERPT}, $r^{\text{M-SERPT}}$, r^{Gittins}, and $r^{\text{M-Gittins}}$ is continuous and piecewise-monotonic.*

Proof. It suffices to prove the claims for r^{SERPT} and r^{Gittins}. The claim for r^{SERPT} is exactly our assumption on expected remaining size, and the claim for r^{Gittins} is a known result [3, Theorem 1]. □

2.2. Job size distribution classes

We consider several classes of job size distributions in this paper. We briefly describe each class before giving the formal definitions.

- The $\mathsf{OR}(-\infty, -1)$ class (Definition 2.7) contains, roughly speaking, distributions with Pareto-like tails.
  - We focus especially on the $\mathsf{OR}(-\infty, -2)$ subclass, all members of which have finite variance.
- The $\mathsf{MDA}(\Lambda)$ class (Definition 2.12) contains, roughly speaking, distributions with smooth tails that are lighter than Pareto tails. It includes, among others, exponential, normal, log-normal, Weibull, and Gamma distributions.
- The QDHR and QIMRL classes (Definitions 2.8 and 2.9) are relaxations of the well-known *decreasing hazard rate* (DHR) and *increasing mean residual lifetime* (IMRL) classes [2,3,5,15–19]. QDHR contains distributions whose hazard rate is roughly decreasing with age, even if it is not perfectly monotonic, and QIMRL contains distributions with roughly increasing expected remaining size.
  - We focus especially on the subclasses $\mathsf{MDA}(\Lambda) \cap$ QDHR and $\mathsf{MDA}(\Lambda) \cap$ QIMRL.

[8] The nonincreasing case is less interesting, because all nonincreasing rank functions encode FCFS.

- The ENBUE class (Definition 2.10) is a relaxation of the well-known *new better than used in expectation* (NBUE) class [2,3,15].[9] It contains distributions whose expected remaining size reaches a global maximum at some age.
  - We focus especially on the Bounded subclass, which contains all bounded distributions.

These classes play two different roles in our analysis.

- Some of the classes broadly characterize the asymptotic behavior of the tail $\overline{F}$. These include $\mathsf{OR}(-\infty, -1)$, $\mathsf{MDA}(\Lambda)$, and ENBUE. Virtually all job size distributions of interest are in one of these classes, so requiring membership in one of them, as in Theorem 3.3, should not be viewed as a major restriction.
- Some of the classes impose additional conditions on the job size distribution that help us bound the M-Gittins and M-SERPT rank functions (Section 6). These include QDHR, QDHR, and Bounded. While these classes are much broader than those previously studied (Section 3.1), they do not cover all distributions of interest. Requiring membership in one of them, as in Theorems 3.1 and 3.2, represents a genuine restriction.

Definition 2.7. A function f is *O-regularly varying* if there exist exponents $\beta \geq \alpha > 0$ along with constants $C_0, x_0 > 0$ such that for all $y \geq x \geq x_0$,

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

We write $\mathsf{OR}(-\beta_0, -\alpha_0)$ for the set of *O*-regularly varying functions where the exponents α and β above may be chosen such that $\alpha_0 < \alpha \leq \beta < \beta_0$.[10] We use the same $\mathsf{OR}(-\beta_0, -\alpha_0)$ notation to represent the class of distributions whose tails are in $\mathsf{OR}(-\beta_0, -\alpha_0)$.

Definition 2.8. A job size distribution is in the *quasi-decreasing hazard rate* class, denoted as QDHR, if there exist a strictly increasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$, an exponent $\gamma \geq 1$, and constants $C_0, x_0 > 0$ such that for all $x \geq x_0$,

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

Definition 2.9. A job size distribution is in the *quasi-increasing mean residual lifetime* class, denoted as QIMRL, if there exist a strictly increasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$, an exponent $\gamma \geq 1$, and constants $C_0, x_0 > 0$ such that for all $x \geq x_0$,

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

Definition 2.10. A job size distribution is in the *eventually new better than used in expectation* class, denoted as ENBUE, if there exists an age $a_* \geq 0$ at which a job's expected remaining size reaches a global maximum, meaning that for all $x \neq a_*$,

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

Definition 2.11. A job size distribution is in the *bounded* class, denoted as Bounded, if there exists $x_{\max} < \infty$ such that $\overline{F}(x_{\max}) = 0$.

Definition 2.12. A job size distribution is said to be in the *Gumbel domain of attraction*, denoted as $\mathsf{MDA}(\Lambda)$, under certain conditions specified in extreme value theory [21].

The exact characterization of $\mathsf{MDA}(\Lambda)$ is outside the scope of this paper. The most important property is that distributions in $\mathsf{MDA}(\Lambda)$ are lighter-tailed than all Pareto distributions.

Lemma 2.13. *If $X \in \mathsf{MDA}(\Lambda)$, then $\overline{F}(x) = o(x^{-\alpha})$ for all $\alpha > 0$.*

Proof. The result follows from a known characterization of $\mathsf{MDA}(\Lambda)$ [21, Proposition 1.4]. □

3. Main results

We now present our main results, explaining how they relate to prior work in Section 3.1. We begin with our heavy-traffic M/G/k optimality result.

[9] Because the NBUE terminology originates in reliability analysis, the word "better" here means "longer".

[10] This is not the standard definition of *O*-regular variation, but it is equivalent to it [20, Section 2.2.1]. Specifically, our $\mathsf{OR}(-\beta_0, -\alpha_0)$ contains the *O*-regularly varying functions whose Matuszewska indices are in the interval $(-\beta_0, -\alpha_0)$.

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

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

then

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T^{\text{M-Gittins-}k}]}{\mathbf{E}[T^{\text{Gittins-1}}]} = 1.$$

In such cases, M-Gittins-k is optimal for mean response time in heavy traffic.

The M-Gittins policy is based on the Gittins policy, which is somewhat complex to describe and compute. Fortunately, the M-SERPT policy, which can be much simpler to compute [9], also performs well in the heavy-traffic M/G/k.

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

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

then

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T^{\text{M-SERPT-}k}]}{\mathbf{E}[T^{\text{Gittins-1}}]} \leq 2.$$

In such cases, M-SERPT-k is a 2-approximation for mean response time in heavy traffic.

Theorems 3.1 and 3.2 apply to a broad class of finite-variance job size distributions. Roughly speaking, $\mathsf{OR}(-\infty, -2)$ covers heavy-tailed distributions, and $\mathsf{MDA}(\Lambda)$ covers non-heavy-tailed distributions that are unbounded (Section 2.2). Assuming membership in these sets is standard for heavy-traffic analysis [8]. The main restriction the results impose is on $\mathsf{MDA}(\Lambda)$ distributions, for which we additionally require membership in QDHR or QIMRL. While slightly relaxing this restriction is possible,[11] removing it entirely appears to be very difficult (Section 8).

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

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

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

and if $X \in \mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda) \cup \mathsf{ENBUE}$, then

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

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

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

3.1. Relationship to prior work

Theorem 3.1 is the first result proving optimality of a scheduling policy in the heavy-traffic M/G/k with unknown job sizes and general job size distribution. As mentioned in Section 1, the only prior results of this type were shown by Grosof et al. [6], who prove similar results for SRPT and FB, that latter for *decreasing hazard rate* (DHR) job size distributions.

- SRPT was shown to be optimal in the heavy-traffic M/G/k for job size distributions whose tail has upper Matuszewska index less than -2 [6, Theorem 6.1], which corresponds to satisfying the upper bound in Definition 2.7 for some $\alpha > 2$. This is somewhat broader than the precondition of Theorem 3.1, though it is still limited to finite-variance distributions.
  - Given that SRPT is designed for known job sizes while M-Gittins is designed for unknown job sizes, Theorem 3.1 complements the prior SRPT results.

[11] For example, we only need the QDHR and QIMRL assumptions to prove Theorems 6.3 and 6.5, so we could instead assume the results of those theorems.

- FB was shown to be optimal in the heavy-traffic M/G/k for job size distributions in the class $\mathsf{DHR} \cap (\mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda))$ [6, Theorem 7.13].[12] The DHR class is much more restrictive than QDHR, so this is much narrower than the precondition of Theorem 3.1.
  - Given that FB is equivalent to M-Gittins in the DHR case [2,3], Theorem 3.1 subsumes the prior FB results.

There is another result that follows from two prior works that complements Theorem 3.1, although to the best of our knowledge it has never been explicitly stated. Köllerström [10,11] shows that under FCFS, the mean response times in the M/G/1 and M/G/k converge. This means that if Gittins and M-Gittins happen to be equivalent to FCFS for a given job size distribution, then FCFS minimizes mean response time in the heavy-traffic M/G/k. Aalto et al. [2,3] show this occurs exactly for job size distributions in the *new better than used in expectation* (NBUE) class, which includes some distributions that Theorem 3.1 does not cover.

Finally, versions of the Gittins policy have been shown to be heavy-traffic optimal for two discrete-state versions of the M/G/k queue [22,23]. These models support some features our model does not, such as multiple job classes, but discretizing the state space imposes some limitations. Specifically, Glazebrook and Niño-Mora [22] require each job to be composed of phases where each phase has exponentially distributed size; and Glazebrook [23] allows nonexponential job size distributions but discretizes time and additionally requires ENBUE job size distributions (Definition 2.10). In contrast, Theorem 3.1 applies to heavy-tailed and other non-ENBUE job size distributions that are of practical importance in computer systems [24–27].

Theorem 3.2 shows that a simple scheduling policy, namely M-SERPT, has mean response time within a constant factor of optimal in the heavy-traffic M/G/k with unknown job sizes and general job size distribution. Specifically, we show M-SERPT is a 2-approximation. This complements the result of Scully et al. [9], who show that in the M/G/1, M-SERPT is a 5-approximation for M/G/1 mean response time at all loads. Our result is tighter and applies to multiserver systems, not just single-server systems, but it applies only in heavy traffic. The techniques we introduce could be useful for tightening the upper bound on M-SERPT's M/G/1 approximation ratio, which is conjectured to be 2 [9].

Theorem 3.3 characterizes the heavy-traffic scaling of M/G/1 mean response time under Gittins, M-Gittins, and M-SERPT. There are three other policies whose heavy-traffic scaling has been characterized: FB, SRPT, and a policy called *randomized multilevel feedback* (RMLF) [28,29]. We now compare Theorem 3.3 to each of these prior results.

Kamphorst and Zwart [8] study FB in heavy traffic. They show that if $X \in \mathsf{OR}(-2, -1)$, then

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

matching the first expression in Theorem 3.3. They also show that if $X \in \mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda)$, then

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

This is similar to the second expression in Theorem 3.3, except it replaces the monotonic $r^{\text{M-SERPT}}$ with the nonmonotonic r^{SERPT}, which pinpoints the suboptimality of FB's heavy-traffic scaling.

Lin et al. [7] study SRPT in heavy traffic. They show that if $X \in \mathsf{OR}(-2, -1)$, then

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

and if $\overline{F}$ has upper Matuszewska index less than -2, which covers $X \in \mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda)$, then

$$\mathbf{E}[T^{\text{SRPT-1}}] = \Theta\left(\frac{1}{(1-\rho)\cdot \overline{G}^{-1}(1-\rho)}\right),$$

where

$$\overline{G}(x) = 1 - \frac{\mathbf{E}[X\mathbb{1}(X \le x)]}{\mathbf{E}[X]} = \overline{F}_e(x) + \frac{x\overline{F}(x)}{\mathbf{E}[X]}.$$

Recall that SRPT minimizes mean response time in the presence of job size information, whereas Gittins does not use job size information, so the heavy-traffic scaling of SRPT is a lower bound on that of Gittins. By comparing the above result for SRPT with our result for Gittins (Theorem 3.3), we learn when knowledge of job sizes yields an asymptotic improvement in mean response time.

- For $X \in \mathsf{OR}(-2, -1)$, meaning X is heavy-tailed with infinite variance, the heavy-traffic scaling of Gittins matches that of SRPT.

[12] While Grosof et al. [6, Theorem 7.13] claim that this result applies to all distributions in DHR with upper Matuszewska index less than -2, their proof incorrectly cites the preconditions of results of Kamphorst and Zwart [8]. Correcting the precondition narrows the result to what we state here.

- For $X \in \mathsf{OR}(-\infty, -2)$, meaning X is heavy-tailed with finite variance, the heavy-traffic scaling of Gittins still matches that of SRPT. Specifically, we later show $r^{\text{M-SERPT}}(a) = \Theta(a)$ (Theorem 6.2), and one can also show $\bar{G}^{-1}(1-\rho) = \Theta(\bar{F}_e^{-1}(1-\rho))$.
- For $X \in \mathsf{MDA}(\Lambda)$, meaning X is not heavy-tailed, one can show $r^{\text{M-SERPT}}(a) = o(a)$ [21], implying Gittins has worse heavy-traffic scaling than SRPT in those cases.

We see that, roughly speaking, Gittins matches the heavy-traffic scaling of SRPT if and only if the job size distribution is heavy-tailed. We conclude that knowledge of job sizes yields an asymptotic improvement in mean response time for non-heavy-tailed job size distributions.

Bansal et al. [12] study RMLF in heavy traffic. They show that if $\mathbf{E}[X^\alpha] < \infty$ for some $\alpha > 2$, then

$$\mathbf{E}[T^{\text{RMLF-1}}] = O\left(\mathbf{E}[T^{\text{SRPT-1}}] \cdot \log \frac{1}{1-\rho}\right). \tag{3.1}$$

Because Gittins minimizes M/G/1 mean response time, this serves as an upper bound on the heavy-traffic scaling of Gittins. However, as previously discussed when comparing Theorem 3.3 to prior results on SRPT, there are cases where Gittins matches the heavy-traffic scaling of SRPT, so our result is a tighter bound. With that said, requiring $\mathbf{E}[X^\alpha] < \infty$ for some $\alpha > 2$ is more lenient than the precondition of Theorem 3.3, so there are still instances where (3.1) is the best known bound on Gittins's heavy-traffic scaling.

4. Technical overview

Our main goal is to show that M-Gittins minimizes M/G/k mean response time in the $\rho \to 1$ limit. Specifically, we show

$$\mathbf{E}[T^{\text{M-Gittins-}k}] \le \mathbf{E}[T^{\text{Gittins-1}}] + o(\mathbf{E}[T^{\text{Gittins-1}}]). \tag{4.1}$$

The only existing technique for proving a bound like (4.1) is the M/G/k tagged job method of Grosof et al. [6]. In general, tagged job methods work as follows [6,14,30–35]: one focuses on a "tagged" job J throughout its time in the system, tracking how much each other job delays J. The amount of time for which another job can delay J is called the *relevant work* due to that other job. The specific M/G/k tagged job method [6] relates the amount of relevant work in an M/G/k under π-k to the amount of relevant work in an M/G/1 under π-1.

As a first approach, we might try to prove a result like (4.1) for Gittins-k using the M/G/k tagged job method. Unfortunately, the method turns out not to work for Gittins, because Gittins can have a nonmonotonic rank function. It turns out that under nonmonotonic rank functions, jobs can contribute more relevant work in an M/G/k than in an M/G/1 (Appendix A), resulting in a much looser response time bound.

Our key insight is that we can generalize the M/G/k tagged job method of Grosof et al. [6] to any SOAP policy, provided it has a monotonic rank function. In Theorem 5.1 we show that for any monotonic SOAP policy π,

$$\mathbf{E}[T^{\pi\text{-}k}] \le \mathbf{E}[Q^{\pi\text{-}1}] + k\mathbf{E}[R^{\pi\text{-}1}] + (k-1)\mathbf{E}[S^{\pi\text{-}1}], \tag{4.2}$$

where the quantities on the right hand side, defined formally in Section 5, can be thought of as follows:

- $Q^{\pi\text{-}1}$ and $R^{\pi\text{-}1}$ are distributions called *waiting time* and *residence time*, respectively [14]. Response time in the M/G/1 is the sum of waiting time and residence time.
- $S^{\pi\text{-}1}$ is a new distribution we call *inflated residence time*, which is similar to residence time but longer.

Proving (4.2) is the first stepping stone to proving Theorem 3.1 because it reduces an M/G/k analysis to an M/G/1 analysis. Only the $\mathbf{E}[R^{\pi\text{-}1}]$ and $\mathbf{E}[S^{\pi\text{-}1}]$ coefficients depend on k, so to prove Theorem 3.1, we show the $\mathbf{E}[Q^{\pi\text{-}1}]$ term dominates in the $\rho \to 1$ limit when π is M-Gittins. Figure 4.1 gives an overview of the main proof steps.

In the remainder of this section, our goal is to bound $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$, where π is either M-Gittins or M-SERPT. We begin in Section 4.1 by explaining in more detail the concepts of relevant work and of waiting, residence, and inflated residence time. In doing so, we introduce *age cutoffs*, quantities which characterize the relevant work due to each job. It turns out that to bound $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$, we first need to bound the age cutoffs. Section 4.2 presents our age cutoff bounds, deferring proofs to Section 6, and Section 4.3 presents our bounds on $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$, deferring proofs to Section 7. Finally, in Section 4.4, we formally prove Theorems 3.1–3.3 by combining the intermediate results discussed throughout this section.

4.1. Understanding the tagged job method and relevant work

In this section we give intuition for the tagged job method, deferring some formalities to Section 5.

Recall that the tagged job method works by focusing on the journey of a "tagged" job J through the system. Roughly speaking, the relevant work due to any other job is the amount of time by which that job delays J's departure. A key insight from the M/G/1 SOAP analysis [14] is that to figure out how much another job delays J, we need to look not at

Key Definitions

- (Section 2.2) *Job size distribution classes:* QDHR, OR$(-\infty, -2)$, MDA(Λ), etc.
- (Sections 4 and 5) *Single-server quantities:* $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$.
- (Section 4.1) *Age cutoffs:* y_x^π and z_x^π.

Proof Steps

- (Section 5) *Compare M/G/k to M/G/1:* $\mathbf{E}[T^{\pi\text{-}k}] \leq \mathbf{E}[Q^{\pi\text{-}1}] + k\mathbf{E}[R^{\pi\text{-}1}] + (k-1)\mathbf{E}[S^{\pi\text{-}1}]$, whereas $\mathbf{E}[T^{\pi\text{-}1}] = \mathbf{E}[Q^{\pi\text{-}1}] + \mathbf{E}[R^{\pi\text{-}1}]$.
- *Show* $\mathbf{E}[Q^{\pi\text{-}1}]$ *dominates* $\mathbf{E}[R^{\pi\text{-}1}]$ *and* $\mathbf{E}[S^{\pi\text{-}1}]$ *in* $\rho \to 1$ *limit.*
  - (Section 6) *Job size distribution classes imply bounds on age cutoffs:* for example, if $X \in$ QDHR, then $z_x^\pi = O(x^\gamma)$ for some $\gamma \geq 1$.
  - (Section 7) *Job size distribution classes and bounds on age cutoffs imply* $\mathbf{E}[Q^{\pi\text{-}1}]$ *dominates:* for example, if $X \in$ MDA(Λ) and $z_x^\pi = O(x^\gamma)$ for some $\gamma \geq 1$, then $\mathbf{E}[S^{\pi\text{-}1}] = o(Q^{\pi\text{-}1})$.
- (Section 4.4) *Compare M-Gittins-k and M-SERPT-k to Gittins-1.*
  - *M-Gittins-k vs. Gittins-1:* prior work shows $\mathbf{E}[Q^{\text{M-Gittins-1}}] \leq \mathbf{E}[T^{\text{Gittins-1}}]$, implying $\lim_{\rho\to 1} \mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] = 1$.
  - *M-SERPT-k vs. Gittins-1:* prior work shows $\mathbf{E}[Q^{\text{M-SERPT-1}}] \leq 2\mathbf{E}[T^{\text{Gittins-1}}]$, implying $\lim_{\rho\to 1} \mathbf{E}[T^{\text{M-SERPT-}k}]/\mathbf{E}[T^{\text{Gittins-1}}] \leq 2$.

Throughout, π stands for either M-Gittins or M-SERPT.

Fig. 4.1. Proof overview.

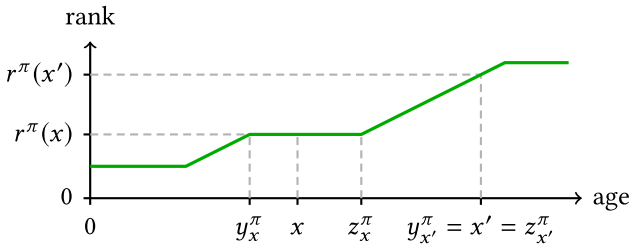


Fig. 4.2. New job and old job age cutoffs.

J's current rank but at its *worst future rank*. This is because even if J has priority over another job at first, if J's rank later increases, the other job can get priority.

Suppose that J has size x. Under a monotonic SOAP policy π, such as M-Gittins or M-SERPT, the worst future rank J will have is always the rank it will have just before completion, namely $r^\pi(x)$. The amount of relevant work due to another job J' is the amount of time J' is served while J is in the system until J' either completes or reaches rank $r^\pi(x)$. Due to the FCFS tiebreaking rule (Section 2.1), exactly what "reaches" means depends on when J' arrives.

- *New jobs*, those that arrive *after* J, contribute relevant work until they first have rank greater than *or equal* to $r^\pi(x)$. This occurs at a specific age called the *new job age cutoff*, denoted as y_x^π.
- *Old jobs*, those that arrive *before* J, contribute relevant work until they first have rank *strictly* greater than $r^\pi(x)$. This occurs at a specific age called the *old job age cutoff*, denoted as z_x^π.

Table 4.1
New job and old job age cutoff bounds.

Size distribution	Quantity	Bound	Reference
OR$(-\infty, -1)$	$y_x^{\text{M-Gittins-1}}$	$\Theta(x)$	Theorem 6.4
	$z_x^{\text{M-Gittins-1}}$	$\Theta(x)$	
	$y_x^{\text{M-SERPT-1}}$	$\Theta(x)$	Theorem 6.2
	$z_x^{\text{M-SERPT-1}}$	$\Theta(x)$	
QDHR	$y_x^{\text{M-Gittins-1}}$	$\Omega(x^{1/\gamma})$ for some $\gamma \geq 1$	Theorem 6.5
	$z_x^{\text{M-Gittins-1}}$	$O(x^{\gamma})$ for some $\gamma \geq 1$	
QDHR∪QIMRL	$y_x^{\text{M-SERPT-1}}$	$\Omega(x^{1/\gamma})$ for some $\gamma \geq 1$	Theorem 6.3
∪QIMRL	$z_x^{\text{M-SERPT-1}}$	$O(x^{\gamma})$ for some $\gamma \geq 1$	

These bounds on y_x^π and z_x^π are critical for characterizing heavy-traffic scaling of $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$.

Figure 4.2 illustrates the new job and old job age cutoffs y_x^π and z_x^π, which are formally defined below.[13] Roughly speaking,

- if r^π is increasing at x, then $y_x^\pi = x = z_x^\pi$; and
- if r^π is constant at x, then y_x^π and z_x^π are the endpoints of the constant region containing x.

As Fig. 4.2 illustrates, we always have

$$y_x^\pi \leq x \leq z_x^\pi. \tag{4.3}$$

Definition 4.1. Let π be a monotonic SOAP policy. The *new job age cutoff* and *old job age cutoff* of size x are, respectively,

$$y_x^\pi = \sup\{a \geq 0 \mid r^\pi(a) < r^\pi(x)\},$$
$$z_x^\pi = \sup\{a \geq 0 \mid r^\pi(a) \leq r^\pi(x)\}.$$

When the policy in question is clear, we drop the superscript π.

One can use new job and old job age cutoffs to write M/G/1 mean response time under a monotonic SOAP policy [9]. As a first step, we write M/G/1 response time $T^{\pi\text{-}1}$ as a sum of two parts, called *waiting time* $Q^{\pi\text{-}1}$ and *residence time* $R^{\pi\text{-}1}$ [14]:

$$\mathbf{E}[T^{\pi\text{-}1}] = \mathbf{E}[Q^{\pi\text{-}1}] + \mathbf{E}[R^{\pi\text{-}1}].$$

We define waiting and residence times formally in Section 5. For now, we just need to know that their means can be written in terms of y_x^π and z_x^π. Specifically, Scully et al. [9, Propositions 4.7 and 4.8] show

$$\begin{aligned}\mathbf{E}[Q^{\pi\text{-}1}] &= \int_0^\infty \frac{\tau(z_x^\pi)}{\overline{\rho}(y_x^\pi)\overline{\rho}(z_x^\pi)}\,\mathrm{d}F(x),\\ \mathbf{E}[R^{\pi\text{-}1}] &= \int_0^\infty \frac{x}{\overline{\rho}(y_x^\pi)}\,\mathrm{d}F(x),\end{aligned} \tag{4.4}$$

where $\overline{\rho}$ and τ are defined as

$$\begin{aligned}\overline{\rho}(a) &= 1 - \lambda\mathbf{E}[\min\{X, a\}] = 1 - \int_0^a \lambda\overline{F}(t)\,\mathrm{d}t,\\ \tau(a) &= \frac{\lambda}{2}\mathbf{E}[\min\{X, a\}^2] = \int_0^a \lambda t\overline{F}(t)\,\mathrm{d}t.\end{aligned} \tag{4.5}$$

The proof of Theorem 5.1 explains the intuition behind (4.4).

The significance of (4.2) is that it expresses M/G/k response time in terms of waiting and residence times, which are M/G/1 quantities. It also features a third quantity called *inflated residence time* $S^{\pi\text{-}1}$. We define inflated residence time formally in Section 5. For now, we just need to know that its mean,

$$\mathbf{E}[S^{\pi\text{-}1}] = \int_0^\infty \frac{z_x^\pi}{\overline{\rho}(y_x^\pi)}\,\mathrm{d}F(x), \tag{4.6}$$

can be written in terms of y_x^π and z_x^π. Note that $\mathbf{E}[R^{\pi\text{-}1}] \leq \mathbf{E}[S^{\pi\text{-}1}]$.

4.2. Bounding new and old age cutoffs

Recall that proving our main results rests on characterizing the heavy-traffic scaling of $\mathbf{E}[Q^\pi]$, $\mathbf{E}[R^\pi]$, and $\mathbf{E}[S^\pi]$, where π is either M-Gittins or M-SERPT. As we see in (4.4) and (4.6), both y_x^π and z_x^π feature prominently in the formulas of

[13] The new job and old job age cutoffs of x are equivalent to what Scully et al. [9] call the *previous and next hill ages* of x.

Table 4.2
Heavy-traffic scaling of waiting, residence, and inflated residence times.

Size distribution	Quantity	Heavy-traffic scaling	Reference
OR$(-2,-1)$	$\mathbf{E}[Q^{\pi\text{-}1}]$	$O(-\log(1-\rho))$	Theorems 7.4 and 7.11
	$\mathbf{E}[R^{\pi\text{-}1}]$	$O(-\log(1-\rho))$	
OR$(-\infty,-2)$	$\mathbf{E}[Q^{\pi\text{-}1}]$	$\Omega((1-\rho)^{-\delta})$ for some $\delta>0$	Theorems 7.9 and 7.11
	$\mathbf{E}[R^{\pi\text{-}1}]$	$O(-\log(1-\rho))$	
	$\mathbf{E}[S^{\pi\text{-}1}]$	$O(-\log(1-\rho))$	Theorems 7.9 and 7.12
MDA(Λ)	$\mathbf{E}[Q^{\pi\text{-}1}]$	$\Omega((1-\rho)^{-(1-\epsilon)})$ for all $\epsilon>0$	Theorems 7.10 and 7.11
	$\mathbf{E}[R^{\pi\text{-}1}]$	$O((1-\rho)^{-\epsilon})$ for all $\epsilon>0$	
MDA$(\Lambda)\cap$ QDHR	$\mathbf{E}[S^{\pi\text{-}1}]$	$O((1-\rho)^{-\epsilon})$ for all $\epsilon>0$	Theorems 7.10 and 7.12
MDA$(\Lambda)\cap$ QIMRL	$\mathbf{E}[S^{\text{M-SERPT-}1}]$	$O((1-\rho)^{-\epsilon})$ for all $\epsilon>0$	Theorem 7.10
ENBUE	$\mathbf{E}[Q^{\pi\text{-}1}]$	$\Theta((1-\rho)^{-1})$	Theorems 7.5 and 7.11
	$\mathbf{E}[R^{\pi\text{-}1}]$	$\Theta(1)$	
Bounded	$\mathbf{E}[S^{\pi\text{-}1}]$	$\Theta(1)$	Theorems 7.5 and 7.12

These bounds hold when π is either M-Gittins or M-SERPT, except for the MDA$(\Lambda)\cap$ QIMRL case, in which the bound holds only for M-SERPT.

$\mathbf{E}[Q^{\pi}]$, $\mathbf{E}[R^{\pi}]$, and $\mathbf{E}[S^{\pi}]$. This means the first step of characterizing the heavy-traffic scaling of $\mathbf{E}[Q^{\pi}]$, $\mathbf{E}[R^{\pi}]$, and $\mathbf{E}[S^{\pi}]$ is understanding y_x^{π} and z_x^{π}. This is the subject of Section 6, in which we prove bounds on y_x^{π} and z_x^{π} for a wide class of job size distributions. Table 4.1 summarizes these results. The main takeaway is that y_x^{π} and z_x^{π} are always polynomially bounded relative to x.

4.3. Characterizing heavy traffic scaling

Armed with bounds on age cutoffs, we are ready to characterize heavy-traffic scaling of mean waiting, residence, and inflated residence times. This is the subject of Section 7, in which

- Theorems 7.4, 7.5, 7.9 and 7.10 characterize M-SERPT's heavy-traffic scaling; and
- Theorems 7.11 and 7.12 characterize M-Gittins's heavy-traffic scaling in terms of M-SERPT's.

Table 4.2 summarizes these results. The main takeaway of the table is that for all of the finite-variance job size distribution classes considered,[14] if π is either M-Gittins or M-SERPT, $\mathbf{E}[Q^{\pi\text{-}1}]$ dominates $\mathbf{E}[R^{\pi\text{-}1}]$ and $\mathbf{E}[S^{\pi\text{-}1}]$, with the latter sometimes requiring an additional condition. Specifically,

- $\mathbf{E}[Q^{\pi\text{-}1}]$ grows polynomially in $1/(1-\rho)$, whereas
- $\mathbf{E}[R^{\pi\text{-}1}]$ and $\mathbf{E}[S^{\pi\text{-}1}]$ grow subpolynomially in $1/(1-\rho)$.

4.4. From intermediate results to main results

We now prove our main results. The proofs of Theorems 3.1 and 3.2 both follow the same three main steps, where π is M-Gittins or M-SERPT, respectively:

- Theorem 5.1 bounds $\mathbf{E}[T^{\pi\text{-}k}]$ in terms of M/G/1 quantities.
- The results in Table 4.2 show $\lim_{\rho\to 1}\mathbf{E}[T^{\pi\text{-}k}]/\mathbf{E}[Q^{\pi\text{-}1}]=1$.
- Prior work relates $\mathbf{E}[Q^{\pi\text{-}1}]$ to $\mathbf{E}[T^{\text{Gittins-}1}]$.

Proof of Theorem 3.1. An M/G/1 can simulate any M/G/k policy by sharing the server, so the fact that Gittins minimizes M/G/1 mean response time means $\mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-}1}] \geq 1$. It therefore suffices to show $\lim_{\rho\to 1}\mathbf{E}[T^{\text{M-Gittins-}k}]/\mathbf{E}[T^{\text{Gittins-}1}]\leq 1$.

Theorem 5.1 implies

$$\frac{\mathbf{E}[T^{\text{M-Gittins-}k}]}{\mathbf{E}[Q^{\text{M-Gittins-}1}]} \leq 1+\frac{k\mathbf{E}[R^{\text{M-Gittins-}1}]+(k-1)\mathbf{E}[S^{\text{M-Gittins-}1}]}{\mathbf{E}[Q^{\text{M-Gittins-}1}]}.$$

Theorems 7.5 and 7.9–7.12 imply that the second term vanishes in the $\rho\to 1$ limit. A result of Scully et al. [9, Proposition 4.7] implies

$$\mathbf{E}[Q^{\text{M-Gittins-}1}]\leq \mathbf{E}[Q^{\text{Gittins-}1}]\leq \mathbf{E}[T^{\text{Gittins-}1}], \tag{4.7}$$

implying the desired result. □

[14] That is, for all the classes in Table 4.2 except OR$(-2,-1)$.

Table 5.1
Summary of notation.

Notation	Description	Reference
π-k	k-server version of SOAP policy π	Section 2.1
$\overline{\rho}(a)$, $\tau(a)$	functions of moments of $\min\{X, a\}$	(4.5)
y_x^π, z_x^π	new job and old job age cutoffs	Definition 4.1
$T^{\pi\text{-}k}$	response time under π-k	Section 2.1
$Q^{\pi\text{-}1}$	waiting time under π-1	(4.4)
$R^{\pi\text{-}1}$	residence time under π-1	(4.4)
$S^{\pi\text{-}1}$	inflated residence time under π-1	(4.6)

Additionally, $T_x^{\pi\text{-}k}$ is size-conditional response time for size x, and similarly for $Q_x^{\pi\text{-}1}$, $R_x^{\pi\text{-}1}$, and $S_x^{\pi\text{-}1}$.

Proof of Theorem 3.2. Theorem 5.1 implies

$$\frac{\mathbf{E}[T^{\text{M-SERPT-}k}]}{\mathbf{E}[Q^{\text{M-SERPT-1}}]} \le 1 + \frac{k\mathbf{E}[R^{\text{M-SERPT-1}}] + (k-1)\mathbf{E}[S^{\text{M-SERPT-1}}]}{\mathbf{E}[Q^{\text{M-SERPT-1}}]}.$$

Theorems 7.5, 7.9 and 7.10 imply that the second term vanishes in the $\rho \to 1$ limit. Scully et al. [9, Lemma 5.6] show[15]

$$\mathbf{E}[Q^{\text{M-SERPT-1}}] \le 2\mathbf{E}[Q^{\text{M-Gittins-1}}],$$

which combines with (4.7) to imply the desired result. □

To prove Theorem 3.3, we simply combine the results in Table 4.2.

Proof of Theorem 3.3. We examine each case in turn.

- For $X \in \mathsf{OR}(-2, -1)$, we use Theorems 7.4 and 7.11.
- For $X \in \mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda)$, we use Theorems 7.9–7.11.
- For $X \in \mathsf{ENBUE}$, we have $r^{\text{M-SERPT}}(a) = \Theta(1)$ by Definition 2.10, so we use Theorems 7.5 and 7.11. □

5. M/G/k Response time bound

This section bounds M/G/k mean response time under any monotonic SOAP policy π. The notation used in Theorem 5.1 below is summarized in Table 5.1.

Theorem 5.1. *For any monotonic SOAP policy π,*

$$\mathbf{E}[T_x^{\pi\text{-}k}] \le \frac{1}{\overline{\rho}(y_x^\pi)}\left(\frac{\tau(z_x^\pi)}{\overline{\rho}(z_x^\pi)} + kx + (k-1)z_x^\pi\right), \tag{5.1}$$

and therefore

$$\mathbf{E}[T^{\pi\text{-}k}] \le \mathbf{E}[Q^{\pi\text{-}1}] + k\mathbf{E}[R^{\pi\text{-}1}] + (k-1)\mathbf{E}[S^{\pi\text{-}1}].$$

Proof. In order to bound M/G/k mean response time, we use a tagged job method in the style of Grosof et al. [6], but we generalize it to allow an arbitrary monotonic SOAP policy π. We consider an arbitrary "tagged" job J of size x arriving to a steady-state system. Our goal is to analyze the distribution of J's response time.

The first step is a shift in perspective: instead of thinking about *time passing*, we reason in terms of *work completed*. Since each of the k servers works at rate $1/k$, the system can complete work at rate 1. While J is in the system, servers sometimes complete work and are sometimes left idle. This means J's response time is the sum of

- the amount of work completed while J is in the system and
- the amount of work "wasted", meaning service capacity left idle, while J is in the system.

We bound J's response time by bounding the total amount of work above. We do so by dividing it into several pieces:

- *Tagged work*: the work of J itself.
- *Virtual work*: work on jobs prioritized behind J, plus wasted work due to servers left idle.
- *Relevant work*: work on jobs prioritized ahead of J. We divide this into two subcategories:
  - *Old* relevant work: relevant work on *old jobs*, namely those present when J arrives.

[15] While Scully et al. [9, Lemma 5.6] mention Gittins instead of M-Gittins, they prove the desired statement for M-Gittins as an intermediate step of their proof.

– *New* relevant work: relevant work on *new jobs*, namely those that arrive after J.

For the first two categories, we have the same simple bound as Grosof et al. [6]: tagged work and virtual work add up to at most kx. This is because tagged work is J's size x, and the scheduling policy ensures that a server only completes virtual work while J is in service at another server. However, bounding the two relevant work categories is more complicated than in Grosof et al. [6].

We begin by asking: what rank must a job have to contribute to relevant work? Note that the job J will never have rank greater than its rank upon completion, $r^\pi(x)$, since π is a monotonic policy. As a result, all new relevant work is from jobs with rank *strictly* less than $r^\pi(x)$, and all old relevant work is from jobs with rank less than *or equal* to $r^\pi(x)$. We can put this in terms of the age cutoffs defined in Definition 4.1:

- jobs contribute new relevant work up to at most age y_x^π, and
- jobs contribute old relevant work up to at most age z_x^π.

In the rest of this proof, y_x and z_x refer to y_x^π and z_x^π, respectively.

To help us bound the amount of old relevant work completed while J is in the system, we define a new concept: the amount of relevant work in the M/G/k system under π.

Definition 5.2. Let $\text{RelWork}_x^{\pi\text{-}k}(t)$ denote the amount of work in the M/G/k at time t which is relevant to a job J of size x:

$$\text{RelWork}_x^{\pi\text{-}k}(t) = \sum_{\text{jobs } J'} \left(\min\{z_x, x_{J'}\} - a_{J'}(t)\right)^+,$$

where $x_{J'}$ is the size of job J' and $a_{J'}(t)$ is its age at time t. We write $\text{RelWork}_x^{\pi\text{-}k}$ for the steady state distribution of the amount of relevant work in the M/G/k system.

Since J is a Poisson arrival, $\text{RelWork}_x^{\pi\text{-}k}$ is the distribution of the amount of relevant work in the system when J arrives. That amount is an upper bound on the amount of old relevant work that will be completed while J is in the system.

To bound new relevant work, note that if a job J' of size x' arrives while J is in the system, then J' contributes at most $\min\{x', y_x\}$ new relevant work. As a result, new relevant work can be upper bounded by considering a transformed M/G/1 system in which the job size distribution is

$$X_{\overline{y_x}} =_{\text{st}} \min\{X, y_x\}.$$

The amount of new relevant work that arrives to our real system is upper bounded by the total amount of work that arrives to the transformed system. Let $B_{\overline{y_x}}(w)$ be the length of a busy period in the transformed M/G/1 system started by an initial amount of work w. If w is the total amount of tagged, virtual, and old relevant work, then the amount of new relevant work is at most $B_{\overline{y_x}}(w) - w$.

Combining our bounds, we obtain

$$T_x^{\pi\text{-}k} \leq_{\text{st}} B_{\overline{y_x}}\left(kx + \text{RelWork}_x^{\pi\text{-}k}\right).$$

Applying Lemma 5.3, stated and proven later in this section, yields

$$T_x^{\pi\text{-}k} \leq_{\text{st}} B_{\overline{y_x}}\left(kx + \text{RelWork}_x^{\pi\text{-}1} + (k-1)z_x\right). \tag{5.2}$$

Taking expectations gives us

$$\mathbf{E}[T_x^{\pi\text{-}k}] \leq \frac{\mathbf{E}[\text{RelWork}_x^{\pi\text{-}1}] + kx + (k-1)z_x}{\overline{\rho}(y_x)}.$$

Because π-1 is work conserving with respect to relevant work, the Pollaczek-Khinchine formula tells us

$$\mathbf{E}[\text{RelWork}_x^{\pi\text{-}1}] = \frac{\tau(z_x)}{\overline{\rho}(z_x)},$$

which completes the proof of (5.1).

To connect (5.1) to the quantities $\mathbf{E}[Q^\pi]$, $\mathbf{E}[R^\pi]$, and $\mathbf{E}[S^\pi]$, we rewrite (5.2) as

$$T_x^{\pi\text{-}k} \leq_{\text{st}} B_{\overline{y_x}}(\text{RelWork}_x^{\pi\text{-}1}) + \sum_1^k B_{\overline{y_x}}(x) + \sum_1^{k-1} B_{\overline{y_x}}(z_x), \tag{5.3}$$

where all of the relevant busy periods are independent. Prior work on SOAP policies [9,14] gives names to some of the distributions on the right-hand side.[16]

16 We define waiting, residence, and inflated residence times in terms of relevant busy periods. Waiting and residence times also have natural definitions as components of M/G/1 response time [9,14], but we do not need them in this paper.

- The *size-conditional waiting time* for size x is the random variable $Q_x^{\pi\text{-}1} =_{\text{st}} B_{\overline{y_x}}(\text{RelWork}_x^{\pi\text{-}1})$, and *waiting time* is $Q^{\pi\text{-}1} =_{\text{st}} Q_X^{\pi\text{-}1}$.
- The *size-conditional residence time* for size x is the random variable $R_x^{\pi\text{-}1} =_{\text{st}} B_{\overline{y_x}}(x)$, and *residence time* is $R^{\pi\text{-}1} =_{\text{st}} R_X^{\pi\text{-}1}$.
- As there is no concise name for $B_{\overline{y_x}}(z_x)$ in prior work, we define *size-conditional inflated residence time* for size x to be the random variable $S_x^{\pi\text{-}1} =_{\text{st}} B_{\overline{y_x}}(z_x)$, and we define *inflated residence time* to be $S^{\pi\text{-}1} =_{\text{st}} S_X^{\pi\text{-}1}$.

With these definitions in place, (5.3) gives us

$$T_x^{\pi\text{-}k} \leq_{\text{st}} Q_x^{\pi\text{-}1} + \sum_1^k R_x^{\pi\text{-}1} + \sum_1^{k-1} S_x^{\pi\text{-}1},$$

so the result follows by taking the expectation of $T^{\pi\text{-}k} =_{\text{st}} T_X^{\pi\text{-}k}$. □

Theorem 5.1 applies only to monotonic SOAP policies. It is tempting to try to apply the same technique to SOAP policies with nonmonotonic rank functions, but as we discuss in Appendix A, the argument does not readily generalize.

The proof of Theorem 5.1 assumes a bound on $\text{RelWork}_x^{\pi\text{-}k}$. We prove the bound in the following lemma, which generalizes a similar lemma of Grosof et al. [6, Lemma 7.10].

Lemma 5.3. *Let*

$$\Delta_x(t) = \text{RelWork}_x^{\pi\text{-}k}(t) - \text{RelWork}_x^{\pi\text{-}1}(t).$$

Then $\Delta_x(t) \leq (k-1)z_x^\pi$ *for all times* t*, and therefore*

$$\text{RelWork}_x^{\pi\text{-}k} \leq_{\text{st}} \text{RelWork}_x^{\pi\text{-}1} + (k-1)z_x^\pi.$$

Proof. Throughout this proof, z_x refers to z_x^π. We consider a pair of coupled systems with the same arrival sequence:

- *System 1*, an M/G/1 using π-1; and
- *System k*, an M/G/k using π-k.

Our approach is to bound the difference in relevant work between Systems 1 and k at any time t.

Call a job *relevant* if it has age less than z_x. These are the only jobs that contribute relevant work. To bound $\Delta_x(t)$, we divide times t into two types of intervals:

- *few-jobs intervals*, during which there are fewer than k relevant jobs in System k; and
- *many-jobs intervals*, during which there are at least k relevant jobs in System k.

Note that both types of intervals are defined based on System k alone, so System 1 may or may not have relevant jobs during either type of interval.

Any time t is in either a few-jobs interval or a many-jobs interval. If t is in a few-jobs interval, the argument is simple: there are at most $k-1$ relevant jobs in System k at time t, so

$$\Delta_x(t) \leq \text{RelWork}_x^{\pi\text{-}k}(t) \leq (k-1)z_x.$$

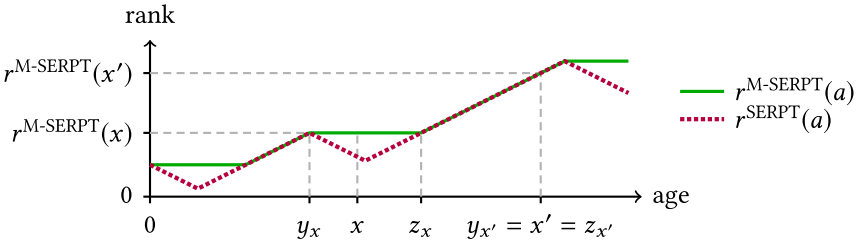
Suppose instead that t is in a many-jobs interval. Let $s \leq t$ be the start of the many-jobs interval containing t. We will show

$$\Delta_x(t) \leq \Delta_x(s) \leq (k-1)z_x.$$

We begin by showing $\Delta_x(t) \leq \Delta_x(s)$. Note that arrivals do not affect Δ_x, because the two systems experience the same arrivals and have the same definition of relevant work. Next, note that service to irrelevant jobs does not affect Δ_x, because irrelevant jobs never become relevant under π, since π is a monotonic policy. In fact, the only way that Δ_x changes over a many-jobs period is due to service to relevant jobs. System k serves relevant jobs on all k servers throughout a many-jobs period, completing relevant work at rate 1. System 1 may or may not serve relevant jobs during a many-jobs period, so it completes relevant work at rate at most 1. This means $\Delta_x(t) \leq \Delta_x(s)$, as desired.

All that remains is to show that $\Delta_x(s) \leq (k-1)z_x$. Recall that s is the start of a many-jobs interval. Many-jobs intervals cannot start due to irrelevant jobs becoming relevant, because π is a monotonic policy. This means each many-jobs interval starts due to a relevant job arriving while System k has $k-1$ relevant jobs. Relevant jobs arriving do not change Δ_x, as discussed above. This means $\Delta_x(s) = \Delta_x(s^-)$, where s^- is the instant before the arrival that starts the many-jobs interval. But s^- is in a few-jobs interval, so

$$\Delta_x(s) = \Delta_x(s^-) \leq (k-1)z_x. \quad \square$$



Here y_x stands for $y_x^{\text{M-SERPT}}$, and similarly for z_x, $y_{x'}$, and $z_{x'}$.

Fig. 6.1. Relationship between SERPT and M-SERPT rank functions.

6. Rank function bounds

We now have a bound on M/G/k mean response time under monotonic SOAP policies π, including M-Gittins and M-SERPT. The bound (Theorem 5.1) is expressed in terms of $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$, quantities which in turn are expressed in terms of the new job and old job age cutoffs y_x^π and z_x^π. In order to prove optimality of M-Gittins in the heavy-traffic M/G/k, we need to understand the heavy-traffic behavior of $\mathbf{E}[Q^{\pi\text{-}1}]$, $\mathbf{E}[R^{\pi\text{-}1}]$, and $\mathbf{E}[S^{\pi\text{-}1}]$, which, as we will see in Section 7, boils down to understanding the behavior of y_x^π and z_x^π in the $x \to \infty$ limit. This section is thus devoted to asymptotically bounding the new job and old job age cutoffs, and more generally the rank functions, of M-Gittins and M-SERPT.

Recall from Definition 2.2 that SERPT's rank function is used to define M-SERPT's. The following lemma shows that the two rank functions are equal at the new job and old job age cutoffs, and similarly for Gittins and M-Gittins. Figure 6.1 gives an intuitive picture of the result.

Lemma 6.1. *The SERPT and M-SERPT rank functions are related by*

$$r^{\text{SERPT}}(y_x^{\text{M-SERPT}}) = r^{\text{M-SERPT}}(y_x^{\text{M-SERPT}}) = r^{\text{M-SERPT}}(x) = r^{\text{M-SERPT}}(z_x^{\text{M-SERPT}}) = r^{\text{SERPT}}(z_x^{\text{M-SERPT}}),$$

and analogously for Gittins and M-Gittins.

Proof. We prove the statement for SERPT and M-SERPT, as the proof for Gittins and M-Gittins is analogous. Throughout this proof, y_x and z_x refer to $y_x^{\text{M-SERPT}}$ and $z_x^{\text{M-SERPT}}$, respectively. The illustration in Fig. 6.1 may provide helpful intuition for the following argument.

We first show the outer equalities. Definition 4.1 implies that $r^{\text{M-SERPT}}$ is increasing in the intervals $(y_x - \delta, y_x)$ and $(z_x, z_x + \delta)$ for some $\delta > 0$. By Definition 2.2, for $r^{\text{M-SERPT}}$ to be increasing at age a, we must have $r^{\text{M-SERPT}}(a) = r^{\text{SERPT}}(a)$, so continuity of $r^{\text{M-SERPT}}$ (Lemma 2.6) implies the outer equalities.

By (4.3) and the monotonicity of $r^{\text{M-SERPT}}$, it remains only to show $r^{\text{M-SERPT}}(y_x) = r^{\text{M-SERPT}}(z_x)$. This is immediate if $y_x = z_x$, and if $y_x < z_x$, then $r^{\text{M-SERPT}}$ is constant over the interval $[y_x, z_x)$, so the result follows by the continuity of $r^{\text{M-SERPT}}$ (Lemma 2.6). □

6.1. Bounds on the M-SERPT rank function

In this section we show two bounds on $y_x^{\text{M-SERPT}}$ and $z_x^{\text{M-SERPT}}$, each subject to a different assumption on the job size distribution.

Theorem 6.2. *If $X \in \mathsf{OR}(-\infty, -1)$, then*

$$\begin{aligned} r^{\text{SERPT}}(a) &= \Theta(a), \\ r^{\text{M-SERPT}}(a) &= \Theta(a), \\ y_x^{\text{M-SERPT}} &= \Theta(x), \\ z_x^{\text{M-SERPT}} &= \Theta(x). \end{aligned}$$

Proof. By Definition 2.7, there exists $\alpha > 1$ such that

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

and $r^{\text{SERPT}}(a) = \Omega(a)$ follows similarly. This implies

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

so the result follows from Lemma 6.1. □

Theorem 6.3. *If $X \in \mathsf{QDHR} \cup \mathsf{QIMRL}$ with exponent γ, then*

$$y_x^{\text{M-SERPT}} = \Omega(x^{1/\gamma}),$$
$$z_x^{\text{M-SERPT}} = O(x^{\gamma}).$$

Proof. The QDHR case follows from Theorem 6.5 (Section 6.2) and a result of Scully et al. [9, Eq. (3.8)] stating

$$y_x^{\text{M-Gittins}} \le y_x^{\text{M-SERPT}} \le z_x^{\text{M-SERPT}} \le z_x^{\text{M-Gittins}},$$

so only the QIMRL case remains.

In the rest of this proof, y_x and z_x refer to $y_x^{\text{M-SERPT}}$ and $z_x^{\text{M-SERPT}}$, respectively. By (4.3), it suffices to show $z_x = O(y_x{}^{\gamma})$. Because $X \in \mathsf{QIMRL}$ with exponent γ, there exists strictly increasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all ages a,

$$a \le m^{-1}\big(r^{\text{SERPT}}(a)\big) \le O(a^{\gamma}).$$

The result follows by plugging in $a = y_x$ and $a = z_x$ and applying Lemma 6.1. □

6.2. Bounds on the M-Gittins rank function

In this section we show two bounds on $y_x^{\text{M-Gittins}}$ and $z_x^{\text{M-Gittins}}$, each subject to a different assumption on the job size distribution.

Theorem 6.4. *If $X \in \mathsf{OR}(-\infty, -1)$, then*

$$y_x^{\text{M-Gittins}} = \Theta(x),$$
$$z_x^{\text{M-Gittins}} = \Theta(x).$$

Theorem 6.5. *If $X \in \mathsf{QDHR}$ with exponent γ, then*

$$y_x^{\text{M-Gittins}} = \Omega(x^{1/\gamma}),$$
$$z_x^{\text{M-Gittins}} = O(x^{\gamma}).$$

These bounds are harder to prove than their M-SERPT counterparts from Section 6.1. The most important component is the following definition, which helps us better understand the M-Gittins rank function and relate it to the simpler M-SERPT rank function.

Definition 6.6. The *time per completion* over an age interval $(a, b]$ is[17]

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

We extend this definition to the $b \to a$ and $b \to \infty$ limits:

$$\eta(a, a) = \frac{1}{h(a)},$$
$$\eta(a, \infty) = \mathbf{E}[X - a \mid X > a].$$

We can write the rank functions of SERPT, M-SERPT, Gittins, and M-Gittins in terms of η as

$$\begin{aligned} r^{\text{SERPT}}(a) &= \eta(a, \infty), \\ r^{\text{M-SERPT}}(a) &= \max_{b \in [0,a]} \eta(b, \infty), \\ r^{\text{Gittins}}(a) &= \min_{b \in [a,\infty]} \eta(a, b), \\ r^{\text{M-Gittins}}(a) &= \max_{b \in [0,a]} \min_{c \in [b,\infty]} \eta(b, c). \end{aligned} \tag{6.1}$$

Armed with Definition 6.6 and (6.1), we are ready to prove Theorems 6.4 and 6.5. The former proof relies on some technical lemmas that we defer to Section 6.3.

[17] Our time per completion function is the reciprocal of what Aalto et al. [2,3] call the *efficiency function*.

Proof of Theorem 6.4. Throughout this proof, y_x and z_x refer to $y_x^{\text{M-Gittins}}$ and $z_x^{\text{M-Gittins}}$, respectively. By (4.3), it suffices to show there exist $C_0, x_0 > 0$ such that for all $x \geq x_0$,

$$z_x \leq C_0 y_x.$$

We will set $C_0 \geq 2$, which covers the $z_x \leq 2y_x$ case. The rest of the proof is thus devoted to the $z_x > 2y_x$ case. Our approach is to show there exist C_1, C_2 such that for all $x \geq x_0$,

$$C_1 y_x \geq r^{\text{Gittins}}(y_x) \geq C_2 z_x. \tag{6.2}$$

We begin with the upper bound on $r^{\text{Gittins}}(y_x)$. By Lemma 6.1, we have $r^{\text{Gittins}}(y_x) = r^{\text{M-Gittins}}(y_x)$ for all sizes x, and by (6.1), we have $r^{\text{M-Gittins}}(a) \leq r^{\text{M-SERPT}}(a)$ for all ages a. Combining these observations with Theorem 6.2 implies $r^{\text{Gittins}}(y_x) = O(y_x)$ and thereby implies the desired upper bound from (6.2).[18]

We now turn to the lower bound on $r^{\text{Gittins}}(y_x)$. This requires Lemmas 6.7 and 6.8, which are facts about η that we prove in Section 6.3. Combining Lemma 6.7 with (6.1) and the fact that we are in the $z_x > 2y_x$ case gives us

$$r^{\text{Gittins}}(y_x) = \eta(y_x, z_x) \geq \eta\left(\frac{z_x}{2}, z_x\right).$$

By Lemma 6.8, there exist C_2, x_2 such that for all x with $z_x/2 > x_2$,

$$\eta\left(\frac{z_x}{2}, z_x\right) \geq C_2 z_x,$$

implying the desired lower bound from (6.2). □

Proof of Theorem 6.5. Throughout this proof, y_x and z_x refer to $y_x^{\text{M-Gittins}}$ and $z_x^{\text{M-Gittins}}$, respectively. By (4.3), it suffices to show $z_x = O({y_x}^\gamma)$. Because $X \in \mathsf{QDHR}$ with exponent γ, there exists a strictly increasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all sizes x,

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

We have $r^{\text{Gittins}}(y_x) \leq 1/h(y_x)$ by (6.1), and Lemma 6.1 implies $r^{\text{Gittins}}(z_x) = r^{\text{Gittins}}(y_x)$, so

$$r^{\text{Gittins}}(z_x) \leq m(O({y_x}^\gamma)).$$

It remains only to lower bound $r^{\text{Gittins}}(z_x)$. We do so using the observation that for any age a,

$$\begin{aligned} r^{\text{Gittins}}(a) &= \min_{b \in [a,\infty]} \eta(a,b) \\ &= \left(\max_{b \in [a,\infty]} \frac{\int_a^b \overline{F}(t) h(t)\, \mathrm{d}t}{\int_a^b \overline{F}(t)\, \mathrm{d}t} \right)^{-1} \\ &\geq \left(\sup_{b>a} h(b)\right)^{-1} \\ &= \inf_{b>a} \frac{1}{h(b)} \\ &\geq m(a), \end{aligned}$$

where the first inequality follows from viewing the ratio of integrals as a weighted average. Plugging in $a = z_x$ implies $m(z_x) \leq m(O({y_x}^\gamma))$, so the result follows because m is strictly increasing. □

6.3. Time per completion lemmas

Lemma 6.7. *For all sizes x and ages a, if $y_x < a < z_x$, then*

$$r^{\text{Gittins}}(y_x) = \eta(y_x, z_x) \geq \eta(a, z_x).$$

Proof. A property of the Gittins index [4, Lemma 2.2] implies[19]

$$r^{\text{Gittins}}(y_x) = \eta(y_x, z_x).$$

In particular, for any $a \neq z_x$,

$$\eta(y_x, a) \geq \eta(y_x, z_x). \tag{6.3}$$

18 This would be more subtle if $\lim_{x\to\infty} y_x$ were finite, but Theorem 6.2 and a result of Aalto et al. [3, Proposition 9] imply $\lim_{x\to\infty} y_x = \infty$.

19 The proof given by Gittins et al. [4] is in a discrete setting, but essentially the same proof carries over to our continuous setting.

A basic property of the η function [9, Eq. (D.3)] is that for any $d < e < f$,

$$\eta(d,e) \geq \eta(d,f) \;\Leftrightarrow\; \eta(d,f) \geq \eta(e,f).$$

Plugging in $d = y_x$, $e = a$, and $f = z_x$ and applying (6.3) yields $\eta(y_x, z_x) \geq \eta(a, z_x)$, as desired. □

Lemma 6.8. *If $X \in \mathsf{OR}(-\infty, -1)$, then there exist constants $C_0, x_0 > 0$ such that for all $b > a > x_0$,*

$$\eta(a,b) \geq C_0 a\left(1 - \frac{a}{b}\right).$$

Proof. We can write $\eta(a,b)$ as

$$\eta(a,b) = \frac{\int_a^b \overline{F}(t)/\overline{F}(a)\,\mathrm{d}t}{1 - \overline{F}(b)/\overline{F}(a)} \geq \int_a^b \frac{\overline{F}(t)}{\overline{F}(a)}\,\mathrm{d}t.$$

Because $X \in \mathsf{OR}(-\infty, -1)$, there exist $\beta > 1$ and $C_1, x_1 > 0$ such that for all $t > a > x_1$,

$$\frac{\overline{F}(t)}{\overline{F}(a)} \geq C_1\left(\frac{t}{a}\right)^{-\beta}.$$

For all $b > a > x_1$, we have

$$\eta(a,b) \geq C_1 \int_a^b \left(\frac{t}{a}\right)^{-\beta} \mathrm{d}t = \frac{C_1 a}{\beta - 1}\left(1 - \left(\frac{b}{a}\right)^{-(\beta-1)}\right).$$

We now consider two cases: $\beta \geq 2$ or $1 < \beta < 2$. If $\beta \geq 2$, then $(b/a)^{-(\beta-1)} \leq a/b$ and therefore

$$\eta(a,b) \geq \frac{C_1 a}{\beta - 1}\left(1 - \frac{a}{b}\right), \tag{6.4}$$

so setting $C_0 = C_1/(\beta - 1)$ and $x_0 = x_1$ suffices. If $1 < \beta < 2$, we use the fact that for all $u > 0$,

$$u^{\beta-1} \leq 1 + (\beta - 1)(u - 1).$$

Substituting $u = a/b$ and combining this with (6.4) yields

$$\eta(a,b) \geq C_1 a\left(1 - \frac{a}{b}\right),$$

so setting $C_0 = C_1$ and $x_0 = x_1$ suffices. □

7. Heavy-traffic scaling of M/G/1 waiting and residence times

In this section we characterize the heavy-traffic scaling of mean waiting, residence, and inflated residence times, which are the M/G/1 quantities that appear Theorem 5.1. Because M-SERPT is a simpler policy than M-Gittins, our approach is to first study M-SERPT's heavy-traffic scaling (Sections 7.2 and 7.3) then show that the results extend to M-Gittins (Section 7.4).

7.1. Key parts of waiting and residence time

Before starting the heavy-traffic analyses of M-Gittins and M-SERPT, we introduce some new notation. Let

$$H_\rho(x) = \frac{\overline{F}(x)}{\overline{\rho}(x)}.$$

Definition 7.1. The *key M/G/1 response time quantities*, or simply "key quantities", of a monotonic SOAP policy π are the following:

$$\mathrm{I}_\mathrm{Q}^\pi = \int_0^\infty \left(H_\rho(y_x^\pi) + H_\rho(z_x^\pi)\right) \frac{\lambda \tau(z_x^\pi)\overline{F}(x)}{\overline{\rho}(x)^2}\,\mathrm{d}x,$$

$$\mathrm{II}_\mathrm{Q}^\pi = \int_0^\infty \lambda x H_\rho(y_x^\pi)^2 \cdot \frac{\overline{F}(x)}{\overline{F}(y_x^\pi)}\,\mathrm{d}x,$$

$$\mathrm{II}_\mathrm{R}^\pi = \int_0^\infty \lambda z_x^\pi H_\rho(y_x^\pi) H_\rho(z_x^\pi) \cdot \frac{\overline{F}(x)}{\overline{F}(y_x^\pi)}\,\mathrm{d}x,$$

$$\mathrm{III}_\mathrm{R}^\pi = \int_0^\infty H_\rho(y_x^\pi) \cdot \frac{\overline{F}(x)}{\overline{F}(y_x^\pi)}\,\mathrm{d}x,$$

$$\mathrm{II}_S^\pi = \mathrm{II}_R^\pi,$$
$$\mathrm{III}_S^\pi = \int_0^\infty H_\rho(y_x^\pi)\,\mathrm{d}x.$$

When the policy in question is clear, we drop the superscript π.

In Theorems B.1–B.3 (Appendix B) we show that for any monotonic SOAP policy π,

$$\mathbf{E}[Q^\pi] = \mathrm{I}_Q^\pi + \mathrm{II}_Q^\pi,$$
$$\mathbf{E}[R^\pi] = \mathrm{II}_R^\pi + \mathrm{III}_R^\pi,$$
$$\mathbf{E}[S^\pi] = \mathrm{II}_S^\pi + \mathrm{III}_S^\pi.$$

Bounding mean waiting, residence, and inflated residence times thus amounts to bounding the key quantities.

For the most of the rest of this section we focus on the case where π is M-SERPT, deferring the M-Gittins case to Section 7.4. Until then, y_x, z_x, and the key quantities are understood to have an implicit superscript M-SERPT.

The most important step of bounding the key quantities is bounding $H_\rho(y_x)$ and $H_\rho(z_x)$. As a first step, we bound $H_\rho(x)$. Let

$$\overline{F}_e(x) = \frac{1}{\mathbf{E}[X]}\int_x^\infty \overline{F}(t)\,\mathrm{d}t \tag{7.1}$$

be the tail of the excess of X. We can write $\overline{\rho}(x)$ as

$$\overline{\rho}(x) = (1-\rho) + \rho\overline{F}_e(x). \tag{7.2}$$

This means that for all $\epsilon \in [0, 1]$, we have

$$H_\rho(x) \le \frac{\overline{F}(x)}{\max\{1-\rho,\, \rho\overline{F}_e(x)\}} \le \frac{\overline{F}(x)}{(1-\rho)^\epsilon(\rho\overline{F}_e(x))^{1-\epsilon}} = \frac{\overline{F}(x)^\epsilon H_1(x)^{1-\epsilon}}{(1-\rho)^\epsilon\rho^{1-\epsilon}}, \tag{7.3}$$

where $H_1(x) = \overline{F}(x)/\overline{F}_e(x) = \lim_{\rho\to 1} H_\rho(x)$. This bound is useful because it separates $H_\rho(x)$'s dependence on x and ρ: the numerator depends only on x, and the denominator depends only on ρ. We will typically choose ϵ to be either 0 or arbitrarily small.

Having bounded $H_\rho(x)$ in (7.3), we now turn to bounding $H_\rho(y_x)$ and $H_\rho(z_x)$. Recalling the definition of r^{SERPT} (Definition 2.1),

$$H_1(x) = \frac{\overline{F}(x)}{\overline{F}_e(x)} = \frac{\mathbf{E}[X]}{r^{\text{SERPT}}(x)},$$

so Lemma 6.1 and the monotonicity of $r^{\text{M-SERPT}}$ imply

$$H_1(y_x) = H_1(z_x) = \frac{\mathbf{E}[X]}{r^{\text{M-SERPT}}(x)} = O(1). \tag{7.4}$$

Combining this with (7.3) yields bounds on $H_\rho(y_x)$ and $H_\rho(z_x)$, though the bounds still have $\overline{F}(y_x)$ and $\overline{F}(z_x)$ terms. To better understand $H_\rho(y_x)$ and $H_\rho(z_x)$, we need to use our results from Section 6 in arguments that depend on what class of distributions contains X. We do this over the course of Sections 7.2 and 7.3.

7.2. Infinite-variance job size distributions

In this section we study the heavy-traffic scaling of M-SERPT's waiting, residence, and inflated residence times for infinite-variance job size distributions, specifically those in $\mathsf{OR}(-2,-1)$. With that said, many of the intermediate results we prove will also be useful for the finite-variance $\mathsf{OR}(-\infty,-2)$ case (Section 7.3).

Suppose that $X \in \mathsf{OR}(-\infty,-1)$. Combining Theorem 6.2 and (7.4) gives us

$$\begin{aligned} y_x, z_x &= \Theta(x), \\ H_1(y_x), H_1(z_x) &= \Theta\left(\frac{1}{x}\right). \end{aligned} \tag{7.5}$$

This alone is enough to bound all of the key quantities except I_Q.

Lemma 7.2. *Under M-SERPT, if $X \in \mathsf{OR}(-\infty,-1)$, then*

$$\mathrm{II}_Q, \mathrm{II}_R, \mathrm{III}_R, \mathrm{II}_S, \mathrm{III}_S = O\left(\log\frac{1}{1-\rho}\right).$$

Proof. Our approach is to use the fact that, by (4.5),

$$\int_0^\infty H_\rho(x)\,\mathrm{d}x = \int_0^\infty \frac{\overline{F}(x)}{\overline{\rho}(x)}\,\mathrm{d}x = \frac{\mathbf{E}[X]}{\rho}\log\frac{1}{1-\rho}. \tag{7.6}$$

Because $\mathrm{II}_R = \mathrm{II}_S$ and $\mathrm{III}_R \le \mathrm{III}_S$, it suffices to show that the integrands of II_Q, II_S, and III_S are all $O(H_\rho(x))$.

We begin by showing that III_S's integrand is $O(H_\rho(x))$. By (7.5) and the fact that $X \in \mathsf{OR}(-\infty, -1)$, we have

$$\overline{F}(y_x) = \overline{F}(\Theta(x)) = \Theta(\overline{F}(x)),$$

which yields

$$H_\rho(y_x) = \frac{\overline{F}(y_x)}{\overline{\rho}(y_x)} \le \frac{\overline{F}(y_x)}{\overline{\rho}(x)} = \frac{O(\overline{F}(x))}{\overline{\rho}(x)} = O(H_\rho(x)). \tag{7.7}$$

This implies the desired bound for III_S and III_R.

We show II_S's integrand is $O(H_\rho(x))$ by applying (7.3) with $\epsilon = 0$, (7.5), and (7.7):

$$\lambda z_x H_\rho(y_x) H_\rho(z_x) \le \lambda z_x H_\rho(y_x) H_1(z_x) = O(H_\rho(x)).$$

This implies the desired bound for II_S and II_R. Similarly,

$$\lambda x H_\rho(y_x)^2 \cdot \frac{\overline{F}(x)}{\overline{F}(y_x)} \le \lambda x H_\rho(y_x) H_1(y_x) = O(H_\rho(x)),$$

implying the bound for II_Q. □

It remains only to characterize the heavy-traffic scaling of I_Q. Treating the $\mathsf{OR}(-\infty, -2)$ case requires some additional care, so we defer it to Section 7.3, focusing on the $\mathsf{OR}(-2, -1)$ case for now. The first step is to bound $\tau(x)$.

Lemma 7.3. *If $X \in \mathsf{OR}(-2, -1)$, then*

$$\tau(x) = \Theta(x^2\overline{F}(x)).$$

Proof. By Definition 2.7, there exists $\beta \in (1, 2)$ such that

$$\frac{\tau(x)}{\overline{F}(x)} = \int_0^x \frac{\lambda t\overline{F}(t)}{\overline{F}(x)}\,\mathrm{d}t \le O(1)\int_0^x t\left(\frac{t}{x}\right)^{-\beta}\mathrm{d}t = O(x^2),$$

and similarly for the lower bound. □

We now have bounds on every term in I_Q's integrand, allowing us to bound I_Q and thereby mean response time.

Theorem 7.4. *If $X \in \mathsf{OR}(-2, -1)$, then in the $\rho \to 1$ limit,*

$$\mathbf{E}[Q^{\text{M-SERPT-1}}] = O\left(\log\frac{1}{1-\rho}\right),$$

$$\mathbf{E}[R^{\text{M-SERPT-1}}] = O\left(\log\frac{1}{1-\rho}\right),$$

and therefore

$$\mathbf{E}[T^{\text{M-SERPT-1}}] = O\left(\log\frac{1}{1-\rho}\right).$$

Proof. By Lemma 7.2, it suffices to upper bound I_Q. We compute

$$\begin{aligned}
\big(H_\rho(y_x) + H_\rho(z_x)\big)\frac{\lambda\tau(z_x)\overline{F}(x)}{\overline{\rho}(x)^2} &\le \big(H_1(y_x) + H_1(z_x)\big)\frac{\lambda\tau(z_x)H_1(x)}{\overline{\rho}(x)} && \text{[by (7.3)]}\\
&= \big(H_1(y_x) + H_1(z_x)\big)\frac{O(z_x{}^2\overline{F}(z_x))\cdot H_1(x)}{\overline{\rho}(x)} && \text{[by Lemma 7.3]}\\
&= \frac{O(\overline{F}(x))}{\overline{\rho}(x)} && \text{[by (7.5)]}\\
&= O(H_\rho(x)),
\end{aligned}$$

so (7.6) implies the desired bound. □

7.3. Finite-variance job size distributions

We now turn to finite-variance job size distributions, specifically those in $\mathsf{OR}(-\infty, -2)$, $\mathsf{MDA}(\Lambda)$, and ENBUE. We begin with the simplest case, which is ENBUE.

Theorem 7.5. *If $X \in \mathsf{ENBUE}$, then in the $\rho \to 1$ limit,*

$$\mathbf{E}[Q^{\text{M-SERPT-1}}] = \Theta\left(\frac{1}{1-\rho}\right),$$
$$\mathbf{E}[R^{\text{M-SERPT-1}}] = \Theta(1),$$

and therefore

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

If additionally $X \in \mathsf{Bounded}$, then in the $\rho \to 1$ limit,

$$\mathbf{E}[S^{\text{M-SERPT-1}}] = \Theta(1).$$

Proof. Let $x_{\max}$ be the supremum of X's support, so we may have $x_{\max} = \infty$. Because $X \in \mathsf{ENBUE}$, there exists age $a_* < x_{\max}$ such that

- $r^{\text{M-SERPT}}(a) < r^{\text{M-SERPT}}(a_*)$ for all $a < a_*$, and
- $r^{\text{M-SERPT}}(a) = r^{\text{M-SERPT}}(a_*)$ for all $a \geq a_*$.

This means

- $y_x \leq a_*$ for all sizes x,
- $z_x \leq a_*$ for all sizes $x \leq a_*$, and
- $z_x = x_{\max}$ for all sizes $x > a_*$.

Because

$$\overline{\rho}(a_*) < \overline{\rho}(x_{\max}) = 1 - \rho,$$

applying (4.4) yields

$$\mathbf{E}[Q^{\text{M-SERPT-1}}] = \Theta(1) + \int_{a_*}^{\infty} \frac{\tau(x_{\max})}{\overline{\rho}(a_*) \cdot (1-\rho)}\, \mathrm{d}F(x) = \Theta\left(\frac{1}{1-\rho}\right),$$
$$\mathbf{E}[R^{\text{M-SERPT-1}}] = \Theta(1) + \int_{a_*}^{\infty} \frac{x}{\overline{\rho}(a_*)}\, \mathrm{d}F(x) = \Theta(1).$$

If additionally $X \in \mathsf{Bounded}$, then $x_{\max} < \infty$, so

$$\mathbf{E}[S^{\text{M-SERPT-1}}] = \Theta(1) + \int_{a_*}^{\infty} \frac{x_{\max}}{\overline{\rho}(a_*)}\, \mathrm{d}F(x) = \Theta(1). \quad \square$$

We now turn to the $\mathsf{OR}(-\infty, -2)$ and $\mathsf{MDA}(\Lambda)$ cases, which require the following technical lemma.

Lemma 7.6. *Let*

$$L^{\pi}(u) = \frac{1}{r^{\pi}\left(\overline{F}_e^{-1}(1/u)\right)},$$

where π is SERPT or M-SERPT. If $X \in \mathsf{OR}(-\infty, -2)$, then

$$L^{\text{SERPT}}, L^{\text{M-SERPT}} \in \mathsf{OR}(-1, 0),$$

and if $X \in \mathsf{MDA}(\Lambda)$, then

$$L^{\text{SERPT}}, L^{\text{M-SERPT}} \in \mathsf{OR}(-\epsilon, \epsilon) \text{ for all } \epsilon > 0.$$

Proof. Because $L^{\text{M-SERPT}}$ is the nonincreasing envelope of L^{SERPT}, it suffices to prove the result for L^{SERPT}. The $\mathsf{OR}(-\infty, -2)$ case follows from closure properties of Matuszewska indices [8, Lemmas 4.5 and 4.6]. The $\mathsf{MDA}(\Lambda)$ case follows from a result of Kamphorst and Zwart [8, Section 4.2.2] which states that if $X \in \mathsf{MDA}(\Lambda)$, then L^{SERPT} is *slowly varying*, a property implying $L^{\text{SERPT}} \in \mathsf{OR}(-\epsilon, \epsilon)$ for all $\epsilon > 0$ [20]. $\square$

One implication of Lemma 7.6 is that if $X \in \mathsf{MDA}(\Lambda)$, then

$$H_1(x) = O(\overline{F}(x)^{-\epsilon}) \text{ for all } \epsilon > 0. \tag{7.8}$$

We are now ready to tackle the $\mathsf{OR}(-\infty, -2)$ and $\mathsf{MDA}(\Lambda)$ cases. As in Section 7.2, we begin by bounding the five key quantities other than I_Q. Lemma 7.2 does so for $\mathsf{OR}(-\infty, -2)$, and the following lemma does so for $\mathsf{MDA}(\Lambda)$.

Lemma 7.7. *Under M-SERPT, if $X \in \mathsf{MDA}(\Lambda)$, then*

$$\mathrm{II}_Q, \mathrm{II}_R, \mathrm{III}_R, \mathrm{II}_S = O\left(\frac{1}{(1-\rho)^\epsilon}\right) \text{ for all } \epsilon > 0.$$

If additionally $X \in \mathsf{MDA}(\Lambda) \cap (\mathsf{QDHR} \cup \mathsf{QIMRL})$, then

$$\mathrm{III}_S = O\left(\frac{1}{(1-\rho)^\epsilon}\right) \text{ for all } \epsilon > 0.$$

Proof. Our overall approach is to use (7.3) on each key quantity to bound it by an expression of the form $(1-\rho)^{-\epsilon} \cdot \int_0^\infty \Phi(\epsilon, x)\,\mathrm{d}x$, where $\Phi(\epsilon, x)$ does not depend on ρ. The challenge is then to show that the integral converges for arbitrarily small $\epsilon > 0$.

We begin with two bounds on $H_\rho(y_x) \cdot \overline{F}(x)/\overline{F}(y_x)$, a term which appears in the integrands of several key quantities. By (4.3),

$$H_\rho(y_x) \cdot \frac{\overline{F}(x)}{\overline{F}(y_x)} \le H_\rho(y_x), \tag{7.9}$$

$$H_\rho(y_x) \cdot \frac{\overline{F}(x)}{\overline{F}(y_x)} = \frac{\overline{F}(x)}{\overline{\rho}(y_x)} \le \frac{\overline{F}(x)}{\overline{\rho}(x)} = H_\rho(x). \tag{7.10}$$

Combining (7.10) with (7.6) implies the desired bound for III_R.

We now bound II_Q. To do so, we apply (7.3) twice, choosing $\epsilon = 0$ for $H_\rho(y_x)$ and arbitrarily small $\epsilon > 0$ for $H_\rho(x)$:

$$\begin{aligned}
\mathrm{II}_Q &\le \int_0^\infty \lambda x H_\rho(y_x) H_\rho(x)\,\mathrm{d}x && \text{[by (7.10)]}\\
&\le \frac{1}{(1-\rho)^\epsilon} \int_0^\infty \lambda x \overline{F}(x)^\epsilon H_1(y_x) H_1(x)^{1-\epsilon}\,\mathrm{d}x && \text{[by (7.3)]}\\
&\le \frac{O(1)}{(1-\rho)^\epsilon} \int_0^\infty x \overline{F}(x)^\epsilon \overline{F}(x)^{-\epsilon(1-\epsilon)}\,\mathrm{d}x && \text{[by (7.4), (7.8)]}\\
&\le \frac{O(1)}{(1-\rho)^\epsilon} \int_0^\infty x^{1-\alpha\epsilon^2}\,\mathrm{d}x, && \text{[by Lemma 2.13]}
\end{aligned}$$

where we may choose $\alpha > 0$ arbitrarily large. Choosing $\alpha > 2/\epsilon^2$ makes the integral converge, so $\mathrm{II}_Q = O((1-\rho)^{-\epsilon})$. The computation for II_S is similar:

$$\begin{aligned}
\mathrm{II}_S &\le \frac{1}{(1-\rho)^\epsilon} \int_0^\infty \lambda z_x \overline{F}(z_x)^\epsilon H_1(y_x) H_1(z_x)^{1-\epsilon}\,\mathrm{d}x && \text{[by (7.3), (7.9)]}\\
&\le \frac{O(1)}{(1-\rho)^\epsilon} \int_0^\infty {z_x}^{1-\alpha\epsilon}\,\mathrm{d}x. && \text{[by (7.4), Lemma 2.13]}
\end{aligned}$$

Because $z_x \ge x$, the integral converges if we choose $\alpha > 2/\epsilon$, so $\mathrm{II}_S = O((1-\rho)^{-\epsilon})$. This also covers II_R because $\mathrm{II}_R = \mathrm{II}_S$.

If additionally $X \in \mathsf{MDA}(\Lambda) \cap (\mathsf{QDHR} \cup \mathsf{QIMRL})$ with exponent γ, then we can similarly bound III_S:

$$\begin{aligned}
\mathrm{III}_S &\le \frac{1}{(1-\rho)^\epsilon} \int_0^\infty \overline{F}(y_x)^\epsilon H_1(y_x)^{1-\epsilon}\,\mathrm{d}x && \text{[by (7.3)]}\\
&\le \frac{O(1)}{(1-\rho)^\epsilon} \int_0^\infty {y_x}^{-\alpha\epsilon}\,\mathrm{d}x && \text{[by (7.4), Lemma 2.13]}\\
&\le \frac{O(1)}{(1-\rho)^\epsilon} \int_0^\infty x^{-\alpha\epsilon/\gamma}\,\mathrm{d}x, && \text{[by Theorem 6.3]}
\end{aligned}$$

so choosing $\alpha > \gamma/\epsilon$ shows that $\mathrm{III}_S = O((1-\rho)^{-\epsilon})$. □

It remains only to characterize the heavy-traffic scaling of I_Q.

Lemma 7.8. *Under M-SERPT, if* $X \in \mathsf{OR}(-\infty, -2) \cup \mathsf{MDA}(\Lambda)$*, then*

$$\mathrm{I}_Q = \left(\frac{1}{(1-\rho) \cdot r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \right).$$

Proof. Because $\mathbf{E}[X^2] < \infty$, we have $\tau(x) = \Theta(1)$, so by (7.3) and (7.4),

$$\mathrm{I}_Q = \int_0^\infty \frac{\Theta(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x.$$

For the lower bound, we integrate up to $\overline{F}_e^{-1}(1-\rho)$ instead of ∞. For $x \le \overline{F}_e^{-1}(1-\rho)$, we have $\overline{F}_e(x) \ge 1-\rho$, so (7.2) implies

$$\rho \overline{F}_e(x) \le \overline{\rho}(x) \le (1+\rho)\overline{F}_e(x).$$

Using this fact along with the monotonicity of $r^{\text{M-SERPT}}$ yields

$$\begin{aligned}
\mathrm{I}_Q &\ge \frac{\Omega(1)}{r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \int_0^{\overline{F}_e^{-1}(1-\rho)} \frac{\overline{F}(x)}{\overline{F}_e(x)^2}\, \mathrm{d}x \\
&= \frac{\Omega(1)}{r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \left(\frac{1}{\overline{F}_e\left(\overline{F}_e^{-1}(1-\rho)\right)} - 1 \right) && \text{[by (7.1)]} \\
&= \Omega\left(\frac{1}{(1-\rho) \cdot r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \right).
\end{aligned}$$

For the upper bound, we split the integration region at $\overline{F}_e^{-1}(1-\rho)$:

$$\mathrm{I}_Q = \int_0^{\overline{F}_e^{-1}(1-\rho)} \frac{O(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x + \int_{\overline{F}_e^{-1}(1-\rho)}^\infty \frac{O(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x. \tag{7.11}$$

The second integral in (7.11) is simple to bound using the monotonicity of $r^{\text{M-SERPT}}$:

$$\begin{aligned}
&\int_{\overline{F}_e^{-1}(1-\rho)}^\infty \frac{O(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x \\
&\quad \le \frac{O(1)}{r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \int_{\overline{F}_e^{-1}(1-\rho)}^\infty \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x \\
&\quad \le \frac{O(1)}{r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \left(\frac{1}{1-\rho} - \frac{1}{1-\rho+\rho\overline{F}_e^{-1}(1-\rho)} \right) && \text{[by (4.5), (7.2)]} \\
&\quad = O\left(\frac{1}{(1-\rho) \cdot r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)} \right).
\end{aligned}$$

To bound the first integral in (7.11), we change variables to $u = 1/\overline{F}_e(x)$:

$$\begin{aligned}
\int_0^{\overline{F}_e^{-1}(1-\rho)} \frac{O(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{\rho}(x)^2}\, \mathrm{d}x &\le \int_0^{\overline{F}_e^{-1}(1-\rho)} \frac{O(1)}{r^{\text{M-SERPT}}(x)} \cdot \frac{\overline{F}(x)}{\overline{F}_e(x)^2}\, \mathrm{d}x && \text{[by (7.2)]} \\
&= \int_1^{1/(1-\rho)} \frac{O(1)}{r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1/u)\right)}\, \mathrm{d}u \\
&= O(1) \int_1^{1/(1-\rho)} L^{\text{M-SERPT}}(u)\, \mathrm{d}u,
\end{aligned}$$

where $L^{\text{M-SERPT}}$ is as in Lemma 7.6. By Lemma 7.6, we have $L^{\text{M-SERPT}} \in \mathsf{OR}(-1, \infty)$, so a result in Karamata theory [20, Theorem 2.6.1] implies

$$\int_1^v L^{\text{M-SERPT}}(u)\, \mathrm{d}u = O(v L^{\text{M-SERPT}}(v))$$

in the $v \to \infty$ limit. Letting $v = 1/(1-\rho)$ yields the desired bound. □

Having characterized the heavy-traffic scaling of all the key quantities, the main heavy-traffic results for $\mathsf{OR}(-\infty, -2)$ and $\mathsf{MDA}(\Lambda)$ follow easily.

Theorem 7.9. *If $X \in \mathsf{OR}(-\infty, -2)$, then in the $\rho \to 1$ limit,*

$$\begin{aligned}\mathbf{E}[Q^{\text{M-SERPT-1}}] &= \Theta\left(\frac{1}{(1-\rho)\cdot r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)}\right)\\ &= \Omega\left(\frac{1}{(1-\rho)^{\delta}}\right) \text{ for some } \delta > 0,\\ \mathbf{E}[R^{\text{M-SERPT-1}}] &\le \mathbf{E}[S^{\text{M-SERPT-1}}]\\ &= \Theta\left(\log\frac{1}{1-\rho}\right),\end{aligned}$$

and therefore

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

Proof. After applying Lemmas 7.2 and 7.8, it remains only to show $\mathrm{I}_Q = \Omega((1-\rho)^{-\delta})$. Using $L^{\text{M-SERPT}}$ from Lemma 7.6, we can rewrite Lemma 7.8 as

$$\mathrm{I}_Q = \Theta\left(\frac{1}{1-\rho}L^{\text{M-SERPT}}\left(\frac{1}{1-\rho}\right)\right). \tag{7.12}$$

By Lemma 7.6, we have $L \in \mathsf{OR}(-1, 0)$, which means there exists $\beta \in (0, 1)$ such that $L(u) = \Omega(u^{-\beta})$ in the $u \to \infty$ limit. Letting $\delta = 1 - \beta$ and $u = 1/(1-\rho)$ yields the desired bound. □

Theorem 7.10. *If $X \in \mathsf{MDA}(\Lambda)$, then in the $\rho \to 1$ limit,*

$$\begin{aligned}\mathbf{E}[Q^{\text{M-SERPT-1}}] &= \Theta\left(\frac{1}{(1-\rho)\cdot r^{\text{M-SERPT}}\left(\overline{F}_e^{-1}(1-\rho)\right)}\right)\\ &= \Omega\left(\frac{1}{(1-\rho)^{1-\epsilon}}\right) \text{ for all } \epsilon > 0,\\ \mathbf{E}[R^{\text{M-SERPT-1}}] &= O\left(\frac{1}{(1-\rho)^{\epsilon}}\right) \text{ for all } \epsilon > 0,\end{aligned}$$

and therefore

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

If additionally $X \in \mathsf{MDA}(\Lambda) \cap (\mathsf{QDHR} \cup \mathsf{QIMRL})$, then

$$\mathbf{E}[S^{\text{M-SERPT-1}}] = O\left(\frac{1}{(1-\rho)^{\epsilon}}\right) \text{ for all } \epsilon > 0.$$

Proof. After applying Lemmas 7.7 and 7.8, it remains only to show $\mathrm{I}_Q = \Omega((1-\rho)^{-(1-\epsilon)})$. This follows from (7.12) and Lemma 7.6, similarly to the proof of Theorem 7.9. □

7.4. Extending heavy-traffic analysis from M-SERPT to Gittins and M-Gittins

Having characterized heavy-traffic scaling under M-SERPT, we now do the same for Gittins and M-Gittins. Our first result shows that the mean waiting and residence times of Gittins and M-Gittins have the same heavy-traffic scaling as that of M-SERPT. Note that the precondition holds for all of the job size distributions we consider in Section 7.3.[20]

Theorem 7.11. *In the $\rho \to 1$ limit,*

$$\mathbf{E}[R^{\text{Gittins-1}}], \mathbf{E}[R^{\text{M-Gittins-1}}] = O(\mathbf{E}[R^{\text{M-SERPT-1}}]),$$

and if $\mathbf{E}[R^{\text{M-SERPT-1}}] = O(\mathbf{E}[Q^{\text{M-SERPT-1}}])$, then

$$\mathbf{E}[Q^{\text{Gittins-1}}], \mathbf{E}[Q^{\text{M-Gittins-1}}] = \Theta(\mathbf{E}[Q^{\text{M-SERPT-1}}]).$$

[20] With some extra effort, one can show it also holds for $X \in \mathsf{OR}(-2, -1)$.

Proof. The residence time result follows immediately from results of Scully et al. [9, Eq. (3.8) and Proposition 4.8], which imply

$$\mathbf{E}[R^{\text{Gittins-1}}] \leq \mathbf{E}[R^{\text{M-Gittins-1}}] \leq \mathbf{E}[R^{\text{M-SERPT-1}}].$$

For waiting time, we first invoke further results of Scully et al. [9, Proposition 4.7 and Lemma 5.6], which imply

$$\mathbf{E}[Q^{\text{Gittins-1}}] \geq \mathbf{E}[Q^{\text{M-Gittins-1}}] \geq \frac{\mathbf{E}[Q^{\text{M-SERPT-1}}]}{2}.$$

It thus suffices to show $\mathbf{E}[Q^{\text{Gittins-1}}] = O(\mathbf{E}[Q^{\text{M-SERPT-1}}])$. Because Gittins minimizes mean response time [2–4], we have

$$\mathbf{E}[Q^{\text{Gittins-1}}] \leq \mathbf{E}[T^{\text{Gittins-1}}] \leq \mathbf{E}[T^{\text{M-SERPT-1}}] = \mathbf{E}[Q^{\text{M-SERPT-1}}] + \mathbf{E}[R^{\text{M-SERPT-1}}],$$

so the result follows from the $\mathbf{E}[R^{\text{M-SERPT-1}}] = O(\mathbf{E}[Q^{\text{M-SERPT-1}}])$ precondition. □

Our final heavy-traffic result shows that for certain job size distributions, under M-Gittins, mean waiting time dominates mean inflated residence time. The conditions are the same as those shown for M-SERPT over the course of Section 7.3, except QDHR ∪ QIMRL is replaced by QDHR.

Theorem 7.12. *If*

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

then in the $\rho \to 1$ *limit,*

$$\mathbf{E}[S^{\text{M-Gittins-1}}] = o(\mathbf{E}[Q^{\text{M-Gittins-1}}]).$$

More specifically, $\mathbf{E}[S^{\text{M-Gittins-1}}]$ *obeys the same scaling bounds as shown for* $\mathbf{E}[S^{\text{M-SERPT-1}}]$ *in Theorems 7.5, 7.9 and 7.10.*

Proof. The proof is very similar to the proofs of analogous results for M-SERPT (Theorems 7.5, 7.9 and 7.10), so we just describe the differences.

- If $X \in \mathsf{OR}(-\infty, -2)$, we follow the same proof as Theorem 7.9 and the lemmas it requires, except we use Theorem 6.4 to bound $y_x^{\text{M-Gittins}}$ and $z_x^{\text{M-Gittins}}$.
- If $X \in \mathsf{MDA}(\Lambda) \cap \mathsf{QDHR}$, we follow the same proof as Theorem 7.10 and the lemmas it requires, except we use Theorem 6.5 to bound $y_x^{\text{M-Gittins}}$ and $z_x^{\text{M-Gittins}}$.
- If $X \in \mathsf{Bounded}$, we follow the same proof as Theorem 7.5, except we use a result of Aalto et al. [3, Proposition 9] to justify the existence of the critical age a_*. □

8. Conclusion

We study optimal scheduling in the M/G/k to minimize mean response time. This problem is solved by the Gittins policy for the single-server $k = 1$ case but was previously open for the much more difficult multiserver case. We introduce a new variant of Gittins called *M-Gittins* (Definition 2.4) and show that it minimizes mean response time in the heavy-traffic M/G/k for a large class of finite-variance job size distributions (Theorem 3.1). We also show that the simple and practical M-SERPT policy is a 2-approximation for mean response time in the heavy-traffic M/G/k under similar conditions (Theorem 3.2). As a byproduct of our M/G/k study, we obtain results characterizing the heavy-traffic scaling of M/G/1 mean response time under Gittins, M-Gittins, and M-SERPT (Theorem 3.3).

A natural question to ask is whether the conditions under which we prove M-Gittins's optimality can be relaxed, particularly the QDHR and Bounded assumptions. The difficulty lies in the fact that for some job size distributions, the bound in Theorem 5.1 is not strong enough because inflated residence time is infinite. It is possible that the techniques used by Köllerström [10,11] to analyze the heavy-traffic M/G/k under FCFS could be helpful, seeing as FCFS has infinite inflated residence time.

Another major open question is analyzing the performance of M-Gittins outside of the heavy-traffic limit. In the single-server case, one can generalize the techniques of Scully et al. [9] to show that M-Gittins is a 3-approximation for M/G/1 mean response time at all loads. However, the multiserver case remains open.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

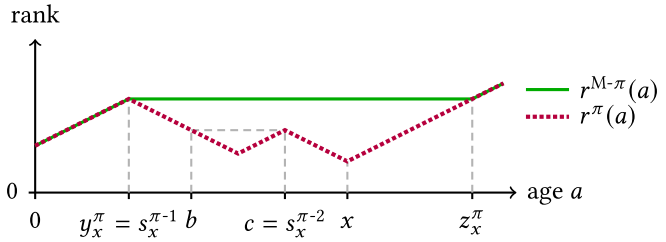


Fig. A.1. Age cutoffs for nonmonotonic rank functions.

Appendix A. Difficulty of M/G/k analysis for nonmonotonic rank functions

In this appendix we explain why Theorem 5.1 does not readily generalize to SOAP policies with nonmonotonic rank functions.

Recall that the proof of Theorem 5.1 considers a tagged job J of size x and considers several categories of work completed while J is in the system. Our focus here is on relevant work, which is work on jobs that are prioritized ahead of J. Let $s_x^{\pi\text{-}k}$ be the maximum age at which a new job, namely one that arrives after J, can contribute relevant work under π-k. When π is monotonic, $s_x^{\pi\text{-}k}$ does not depend on the number of servers k. Specifically, we have $s_x^{\pi\text{-}k} = y_x^\pi$. The problem for nonmonotonic SOAP policies π is that, as we show below, we can have $s_x^{\pi\text{-}k} > s_x^{\pi\text{-}1}$ when $k \geq 2$.

The following discussion uses definitions of y_x^π and z_x^π generalized to all SOAP policies π.

- If π is monotonic, then y_x^π and z_x^π are given by Definition 4.1.
- If π is nonmonotonic, we can define y_x^π and z_x^π in terms of a monotonic SOAP policy related to π [9]. Specifically, letting M-π be the monotonic SOAP policy with rank function

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

we define $y_x^\pi = y_x^{\text{M-}\pi}$ and $z_x^\pi = z_x^{\text{M-}\pi}$.

Consider the example SOAP policy π and tagged job size x shown in Fig. A.1. In the single-server $k = 1$ case, we have $s_x^{\pi\text{-}1} = y_x^\pi$. To see why, consider the moment a new job J' reaches age y_x^π while the tagged job J is still in the system. For this to occur, it must be that J is also at age y_x^π, because otherwise J would have priority over J'. With both J and J' at the same rank, the FCFS tiebreaker prioritizes J. Thereafter, J never has rank worse than $r^\pi(y_x^\pi)$, so J' remains stuck at age y_x^π and is never prioritized over J.

We now reconsider the same example from Fig. A.1 but with $k \geq 2$ servers. The key difference is that because there are multiple servers, J' can receive service even while J has better rank because J and J' can occupy different servers simultaneously. This means J' no longer gets stuck at age y_x^π. In particular, if J reaches age c and J' passes age b, then J' contributes relevant work between ages b and c. Therefore, $s_x^{\pi\text{-}k} = c > s_x^{\pi\text{-}1}$ for $k \geq 2$.

The bound in Theorem 5.1 follows from assuming that every new job J' will contribute relevant work until it completes or reaches age $s_x^{\pi\text{-}k}$. This is a worst-case estimate, because the tagged job J might complete before J' completes or reaches age $s_x^{\pi\text{-}k}$. When π is monotonic, we have $s_x^{\pi\text{-}k} = s_x^{\pi\text{-}1}$, so this overestimate is tight enough to compare the mean response times under π-k and π-1. However, when π is nonmonotonic, it may be that $s_x^{\pi\text{-}k} > s_x^{\pi\text{-}1}$, as explained above, so we do not obtain a tight comparison between the π-k and π-1 systems. This suggests generalizing Theorem 5.1 to nonmonotonic SOAP policies requires not relying as heavily on worst-case quantities like $s_x^{\pi\text{-}k}$.

Appendix B. New formulas for mean waiting and residence times

In this appendix we prove the following new formulas for mean waiting, residence, and inflated residence times.

Theorem B.1. *Under any monotonic SOAP policy π,*

$$\mathbf{E}[Q^{\pi\text{-}1}] = \int_0^\infty \left(\left(\frac{\overline{F}(y_x^\pi)}{\overline{\rho}(y_x^\pi)} + \frac{\overline{F}(z_x^\pi)}{\overline{\rho}(z_x^\pi)} \right) \frac{\lambda\tau(z_x^\pi)\overline{F}(x)}{\overline{\rho}(x)^2} + \frac{\lambda x \overline{F}(y_x^\pi)\overline{F}(x)}{\overline{\rho}(y_x^\pi)^2} \right) \mathrm{d}x.$$

Theorem B.2. *Under any monotonic SOAP policy π,*

$$\mathbf{E}[R^{\pi\text{-}1}] = \int_0^\infty \left(\frac{\lambda z_x^\pi \overline{F}(x)\overline{F}(z_x^\pi)}{\overline{\rho}(y_x^\pi)\overline{\rho}(z_x^\pi)} + \frac{\overline{F}(x)}{\overline{\rho}(y_x^\pi)} \right) \mathrm{d}x.$$

Theorem B.3. *Under any monotonic SOAP policy π,*

$$\mathbf{E}[S^{\pi\text{-}1}] = \int_0^\infty \left(\frac{\lambda z_x^\pi \overline{F}(x)\overline{F}(z_x^\pi)}{\overline{\rho}(y_x^\pi)\overline{\rho}(z_x^\pi)} + \frac{\overline{F}(y_x^\pi)}{\overline{\rho}(y_x^\pi)} \right) \mathrm{d}x.$$

Proving these results requires new technical machinery for, roughly speaking, performing integration by parts on expressions involving y_x^π and z_x^π, such as those in (4.4). Appendix B.1 introduces the general technical machinery, which Appendix B.2 then applies to prove the above results.

Throughout this appendix, ∂ denotes the derivative operator, and $[t_1, \ldots, t_n \mapsto \text{RHS}]$ denotes the function that maps variables $t_1, \ldots, t_n$ to expression RHS.

B.1. Integration by parts with hills and valleys

Definition B.4. A *hill-valley partition* of $\mathbb{R}_+$ is a sequence

$$0 = u_0 \le v_0 < u_1 < v_1 < u_2 < v_2 < \ldots.$$

Intervals of the form $(u_i, v_i]$ are called *valleys*, and intervals of the form $(v_i, u_{i+1}]$ are called *hills*.[21]

Definition B.5. Functions $y, z : \mathbb{R}_+ \to \mathbb{R}_+$ are a *hill-valley pair* for a given hill-valley partition if for each valley $(u_i, v_i]$,

$$y(x) = u_i, z(x) = v_i, \quad \text{for all } x \in (u_i, v_i],$$

and for each hill $(v_i, u_{i+1}]$,

$$y(x) = x, z(x) = x, \quad \text{for all } x \in (v_i, u_{i+1}].$$

For compactness, we write $y_x = y(x)$ and $z_x = z(x)$.

It is simple to check that for any monotonic SOAP policy π, the pair y^π, z^π (Definition 4.1) is a hill-valley pair.

Definition B.6. For functions $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, we define the *difference ratio operator* Δ as follows:

$$\Delta\Phi(\langle u, v\rangle) = \begin{cases} \dfrac{\Phi(v) - \Phi(u)}{v - u} & \text{if } u \ne v \\ \partial\Phi(u) & \text{if } u = v, \end{cases}$$

where ∂ is the derivative operator. Similarly, for functions with multiple arguments, Δ_i is a version of Δ that works on the ith argument:

$$\Delta_i\Phi(\ldots, \langle u, v\rangle, \ldots) = \Delta[t \mapsto \Phi(\ldots, t, \ldots)](\langle u, v\rangle).$$

Like ∂, it is easily seen that Δ is a linear operator. When applied to polynomials, Δ elegantly generalizes ∂. For example,

$$\Delta\left[t \mapsto \frac{1}{t}\right](\langle u, v\rangle) = \frac{1}{uv}. \tag{B.1}$$

The Δ operator also obeys various chain-rule-like identities. We highlight the two we use below.

Lemma B.7. *Let $\Phi, \Psi : \mathbb{R} \to \mathbb{R}$ be differentiable. For all $u, v \in \mathbb{R}$,*

$$\Delta[t \mapsto \Phi(\Psi(t))](\langle u, v\rangle) = \Delta\Phi(\langle \Psi(u), \Psi(v)\rangle)\, \Delta\Psi(\langle u, v\rangle).$$

Proof. If $u = v$, this is the chain rule. If $u \ne v$ but $\Psi(u) = \Psi(v)$, then both sides are 0. If $\Psi(u) \ne \Psi(v)$, then the result follows by a simple computation. □

Lemma B.8. *Let $\Phi : \mathbb{R}^2 \to \mathbb{R}$ be differentiable. For all $u, v \in \mathbb{R}$,*

$$\Delta[t \mapsto \Phi(t, t)](\langle u, v\rangle) = \Delta_2\Phi(u, \langle u, v\rangle) + \Delta_1\Phi(\langle u, v\rangle, v).$$

Proof. If $u = v$, this is the multivariable chain rule. If $u \ne v$,

$$\begin{aligned}(v - u)\,\Delta[t \mapsto \Phi(t, t)](\langle u, v\rangle) &= \Phi(v, v) - \Phi(u, u) \\ &= \Phi(v, v) - \Phi(u, v) + \Phi(u, v) - \Phi(u, u) \\ &= (v - u)(\Delta_1\Phi(\langle u, v\rangle, v) + \Delta_2\Phi(u, \langle u, v\rangle)).\end{aligned}$$ □

[21] We borrow the terms "hill" and "valley" from Scully et al. [9], who use a similar concept to analyze SOAP policies, but this definition is abstracted away from the details of SOAP. As a corner case, we consider the first hill or valley to also include 0.

The most important result of this appendix is the following lemma, which formulates a version of integration by parts that works for hill-valley pairs despite their discontinuity.

Lemma B.9. *Let y, z be a hill-valley pair, $\Phi : \mathbb{R}_+^3 \to \mathbb{R}$ be differentiable, $P : \mathbb{R}_+ \to \mathbb{R}$ be differentiable, and $\overline{P}(x) = c - P(x)$ for some $c \in \mathbb{R}$. If*

$$\overline{P}(0)\Phi(0, 0, z_0) = 0,$$
$$\lim_{x\to\infty} \overline{P}(x)\Phi(y_x, x, z_x) = 0,$$

then

$$\int_0^\infty \Phi(y_x, x, z_x)\, \partial P(x)\, \mathrm{d}x = \int_0^\infty \Big(\overline{P}(y_x)\, \Delta_3\Phi(y_x, y_x, \langle y_x, z_x\rangle) + \overline{P}(x)\, \partial_2\Phi(y_x, x, z_x) + \overline{P}(v)\, \Delta_1\Phi(\langle y_x, z_x\rangle, z_x, z_x)\Big)\, \mathrm{d}x.$$

Proof. For each valley $(u, v]$,

$$\begin{aligned}
&\int_u^v \Phi(y_x, x, z_x)\, \partial P(x)\, \mathrm{d}x \\
&\quad= \int_u^v \overline{P}(x)\, \partial_2\Phi(u, x, v)\, \mathrm{d}x + \overline{P}(u)\Phi(u, u, v) - \overline{P}(v)\Phi(u, v, v) \\
&\quad= \int_u^v \overline{P}(x)\, \partial_2\Phi(u, x, v)\, \mathrm{d}x + \overline{P}(u)\Phi(u, u, u) - \overline{P}(v)\Phi(v, v, v) \\
&\qquad+ (v - u)\overline{P}(u)\, \Delta_3\Phi(u, u, \langle u, v\rangle) + (v - u)\overline{P}(v)\, \Delta_1\Phi(\langle u, v\rangle, v, v) \\
&\quad= \int_u^v \Big(\overline{P}(u)\, \Delta_3\Phi(u, u, \langle u, v\rangle) + \overline{P}(x)\, \partial_2\Phi(u, x, v) + \overline{P}(v)\, \Delta_1\Phi(\langle u, v\rangle, v, v)\Big)\, \mathrm{d}x \\
&\qquad+ \overline{P}(u)\Phi(u, u, u) - \overline{P}(v)\Phi(v, v, v) \\
&\quad= \int_u^v \Big(\overline{P}(y_x)\, \Delta_3\Phi(y_x, y_x, \langle y_x, z_x\rangle) + \overline{P}(x)\, \partial_2\Phi(y_x, x, z_x) + \overline{P}(v)\, \Delta_1\Phi(\langle y_x, z_x\rangle, z_x, z_x)\Big)\, \mathrm{d}x \\
&\qquad+ \overline{P}(u)\Phi(u, u, u) - \overline{P}(v)\Phi(v, v, v).
\end{aligned}$$

For each hill $(v, u]$,

$$\begin{aligned}
&\int_v^u \Phi(y_x, x, z_x)\, \partial P(x)\, \mathrm{d}x \\
&\quad= \int_v^u \overline{P}(x)\, \partial[t \to \Phi(t, t, t)](x)\, \mathrm{d}x + \overline{P}(v)\Phi(v, v, v) - \overline{P}(u)\Phi(u, u, u) \\
&\quad= \int_v^u \Big(\overline{P}(x)\, \partial_3\Phi(x, x, x) + \overline{P}(x)\, \partial_2\Phi(x, x, x) + \overline{P}(x)\, \partial_1\Phi(x, x, x)\Big)\, \mathrm{d}x \\
&\qquad+ \overline{P}(v)\Phi(v, v, v) - \overline{P}(u)\Phi(u, u, u) \\
&\quad= \int_v^u \Big(\overline{P}(y_x)\, \Delta_3\Phi(y_x, y_x, \langle y_x, z_x\rangle) + \overline{P}(x)\, \partial_2\Phi(y_x, x, z_x) + \overline{P}(v)\, \Delta_1\Phi(\langle y_x, z_x\rangle, z_x, z_x)\Big)\, \mathrm{d}x \\
&\qquad+ \overline{P}(v)\Phi(v, v, v) - \overline{P}(u)\Phi(u, u, u).
\end{aligned}$$

Summing the hill and valley expressions over all hills and valleys, most of the non-integral terms cancel out, and the two that remain are 0 by assumption:

$$\begin{aligned}
&\int_0^\infty \Phi(y_x, x, z_x)\, \partial P(x)\, \mathrm{d}x \\
&\quad= \int_0^\infty \Big(\overline{P}(y_x)\, \Delta_3\Phi(y_x, y_x, \langle y_x, z_x\rangle) + \overline{P}(x)\, \partial_2\Phi(y_x, x, z_x) + \overline{P}(v)\, \Delta_1\Phi(\langle y_x, z_x\rangle, z_x, z_x)\Big)\, \mathrm{d}x \\
&\qquad+ \overline{P}(0)\Phi(0, 0, z_0) - \lim_{x\to\infty} \overline{P}(x)\Phi(y_x, x, z_x). \quad \square
\end{aligned}$$

Our final two lemmas show that integrals using Δ can sometimes be turned into integrals using ∂.

Lemma B.10. *Let y, z be a hill-valley pair and $\Phi : \mathbb{R}_+^3 \to \mathbb{R}_+$ be differentiable with respect to its second argument. Then*

$$\int_0^\infty \Delta_2\Phi(y_x, \langle y_x, z_x\rangle, z_x)\, \mathrm{d}x = \int_0^\infty \partial_2\Phi(y_x, x, z_x)\, \mathrm{d}x.$$

Proof. For each valley $(u, v]$,

$$\begin{aligned}\int_u^v \Delta_2\Phi(y_x, \langle y_x, z_x\rangle, z_x)\,\mathrm{d}x &= \int_u^v \Delta_2\Phi(u, \langle u, v\rangle, v)\,\mathrm{d}x\\ &= (v-u)\,\Delta_2\Phi(u, \langle u, v\rangle, v)\\ &= \Phi(u, v, v) - \Phi(u, u, v)\\ &= \int_u^v \partial_2\Phi(u, x, v)\,\mathrm{d}x\\ &= \int_u^v \partial_2\Phi(y_x, x, z_x)\,\mathrm{d}x.\end{aligned}$$

For each hill $(v, u]$,

$$\begin{aligned}\int_v^u \Delta_2\Phi(y_x, \langle y_x, z_x\rangle, z_x)\,\mathrm{d}x &= \int_v^u \Delta_2\Phi(x, \langle x, x\rangle, x)\,\mathrm{d}x\\ &= \int_v^u \partial_2\Phi(x, x, x)\,\mathrm{d}x\\ &= \int_v^u \partial_2\Phi(y_x, x, z_x)\,\mathrm{d}x.\end{aligned}$$

Summing the hill and valley expressions over all hills and valleys yields the desired result. □

Lemma B.11. *Let y, z be a hill-valley pair and both $\Phi : \mathbb{R}_+^3 \to \mathbb{R}$ and $\Psi : \mathbb{R}_+ \to \mathbb{R}$ be differentiable. Then*

$$\int_0^\infty \Delta[t \mapsto \Phi(y_x, \Psi(t), z_x)](\langle y_x, z_x\rangle)\,\mathrm{d}x = \int_0^\infty \Delta_2\Phi(y_x, \langle \Psi(y_x), \Psi(z_x)\rangle, z_x)\,\partial\Psi(x)\,\mathrm{d}x.$$

Proof. We compute

$$\begin{aligned}&\int_0^\infty \Delta[t \mapsto \Phi(y_x, \Psi(t), z_x)](\langle y_x, z_x\rangle)\,\mathrm{d}x\\ &\quad= \int_0^\infty \Delta_2\Phi(y_x, \langle \Psi(y_x), \Psi(z_x)\rangle, z_x)\,\Delta\Psi(\langle y_x, z_x\rangle)\,\mathrm{d}x && \text{[by Lemma B.7]}\\ &\quad= \int_0^\infty \Delta_2\Big[u, t, v \mapsto \Delta_2\Phi(u, \langle \Psi(u), \Psi(v)\rangle, v)\cdot\Psi(t)\Big](y_x, \langle y_x, z_x\rangle, z_x)\,\mathrm{d}x\\ &\quad= \int_0^\infty \Delta_2\Phi(y_x, \langle \Psi(y_x), \Psi(z_x)\rangle, z_x)\,\partial\Psi(x)\,\mathrm{d}x. && \text{[by Lemma B.10]}\ \square\end{aligned}$$

B.2. Proofs of new formulas

We now apply the theory developed in Appendix B.1 to prove Theorems B.1–B.3. Throughout the proofs, y_x and z_x refer to y_x^π and z_x^π, respectively. Recall that y, z form a hill-valley pair (Definition B.5) under any monotonic SOAP policy π.

Proof of Theorem B.1. We compute

$$\begin{aligned}&\mathbf{E}[Q^{\pi\text{-}1}] = \int_0^\infty \frac{\tau(z_x)}{\overline{\rho}(y_x)\overline{\rho}(z_x)}\,\mathrm{d}F(x) && \text{[by (4.4)]}\\ &= \int_0^\infty \left(\frac{\overline{F}(y_x)}{\overline{\rho}(y_x)}\,\Delta\left[t \mapsto \frac{\tau(t)}{\overline{\rho}(t)}\right](\langle y_x, z_x\rangle) + \frac{\overline{F}(z_x)\tau(z_x)}{\overline{\rho}(z_x)}\,\Delta\left[t \mapsto \frac{1}{\overline{\rho}(t)}\right](\langle y_x, z_x\rangle)\right)\mathrm{d}x && \text{[by Lemma B.9]}\\ &= \int_0^\infty \left(\frac{\overline{F}(y_x)}{\overline{\rho}(y_x)^2} + \frac{\overline{F}(y_x)\tau(z_x)}{\overline{\rho}(y_x)}\,\Delta\left[t \mapsto \frac{1}{\overline{\rho}(t)}\right](\langle y_x, z_x\rangle)\right.\\ &\qquad\left. + \frac{\overline{F}(z_x)\tau(z_x)}{\overline{\rho}(z_x)}\,\Delta\left[t \mapsto \frac{1}{\overline{\rho}(t)}\right](\langle y_x, z_x\rangle)\right)\mathrm{d}x && \text{[by Lemma B.8]}\\ &= \int_0^\infty \left(\frac{\overline{F}(y_x)}{\overline{\rho}(y_x)\overline{\rho}(y_x)}\,\partial\tau(x) + \tau(z_x)\left(\frac{\overline{F}(y_x)}{\overline{\rho}(y_x)} + \frac{\overline{F}(z_x)}{\overline{\rho}(z_x)}\right)\partial\left[t \mapsto \frac{1}{\overline{\rho}(t)}\right](x)\right)\mathrm{d}x, && \text{[by Lemma B.10]}\end{aligned}$$

which equals the desired result by (4.5). □

Proof of Theorem B.2. We compute

$$\begin{aligned}
\mathbf{E}[R^{\pi^{-1}}] &= \int_0^\infty \frac{x}{\overline{\rho}(y_x)}\,\mathrm{d}F(x) && \text{[by (4.4)]}\\
&= \int_0^\infty \left(z_x \overline{F}(z_x)\, \Delta\left[t \mapsto \frac{1}{\overline{\rho}(t)} \right](\langle y_x, z_x \rangle) + \frac{\overline{F}(x)}{\overline{\rho}(y_x)} \right) \mathrm{d}x && \text{[by Lemma B.9]}\\
&= \int_0^\infty \left(\frac{-z_x \overline{F}(z_x)}{\overline{\rho}(y_x)\overline{\rho}(z_x)}\, \partial\overline{\rho}(x) + \frac{\overline{F}(x)}{\overline{\rho}(y_x)} \right) \mathrm{d}x, && \text{[by (B.1), Lemma B.11]}
\end{aligned}$$

which equals the desired result by (4.5). □

Proof of Theorem B.3. Very similarly to the proof of Theorem B.2, we compute

$$\begin{aligned}
\mathbf{E}[S^{\pi^{-1}}] &= \int_0^\infty \frac{z_x}{\overline{\rho}(y_x)}\,\mathrm{d}F(x) && \text{[by (4.6)]}\\
&= \int_0^\infty \left(z_x \overline{F}(z_x)\, \Delta\left[t \mapsto \frac{1}{\overline{\rho}(t)} \right](\langle y_x, z_x \rangle) + \frac{\overline{F}(y_x)}{\overline{\rho}(y_x)} \right) \mathrm{d}x && \text{[by Lemma B.9]}\\
&= \int_0^\infty \left(\frac{-z_x \overline{F}(z_x)}{\overline{\rho}(y_x)\overline{\rho}(z_x)}\, \partial\overline{\rho}(x) + \frac{\overline{F}(y_x)}{\overline{\rho}(y_x)} \right) \mathrm{d}x, && \text{[by (B.1), Lemma B.11]}
\end{aligned}$$

which equals the desired result by (4.5). □

References

[1] Linus Schrage, A proof of the optimality of the shortest remaining processing time discipline, Oper. Res. 16 (3) (1968) 687–690.
[2] Samuli Aalto, Urtzi Ayesta, Rhonda Righter, On the Gittins index in the M/G/1 queue, Queueing Syst. 63 (1) (2009) 437–458.
[3] Samuli Aalto, Urtzi Ayesta, Rhonda Righter, Properties of the Gittins index with application to optimal scheduling, Probab. Engrg. Inform. Sci. 25 (03) (2011) 269–288.
[4] John C. Gittins, Kevin D. Glazebrook, Richard Weber, Multi-armed Bandit Allocation Indices, John Wiley & Sons, 2011.
[5] Hanhua Feng, Vishal Misra, Mixed scheduling disciplines for network flows, in: ACM SIGMETRICS Performance Evaluation Review, vol. 31(2), ACM, 2003, pp. 36–39.
[6] Isaac Grosof, Ziv Scully, Mor Harchol-Balter, SRPT for multiserver systems, Perform. Eval. 127–128 (2018) 154–175, http://dx.doi.org/10.1016/j.peva.2018.10.001, http://www.sciencedirect.com/science/article/pii/S0166531618302773.
[7] Minghong Lin, Adam Wierman, Bert Zwart, The average response time in a heavy-traffic SRPT queue, in: ACM SIGMETRICS Performance Evaluation Review, vol. 38(2), ACM, 2010, pp. 12–14.
[8] Bart Kamphorst, Bert Zwart, Heavy-traffic analysis of Sojourn time under the foreground–background scheduling policy, Stoch. Syst. 10 (1) (2020) 1–28, http://dx.doi.org/10.1287/stsy.2019.0036.
[9] Ziv Scully, Mor Harchol-Balter, Alan Scheller-Wolf, Simple near-optimal scheduling for the M/G/1, Proc. ACM Meas. Anal. Comput. Syst. 4 (1) (2020) 11:1–11:29, http://dx.doi.org/10.1145/3379477.
[10] Julian Köllerström, Heavy traffic theory for queues with several servers. I, J. Appl. Probab. 11 (3) (1974) 544–552, http://www.jstor.org/stable/3212698.
[11] Julian Köllerström, Heavy traffic theory for queues with several servers. II, J. Appl. Probab. 16 (2) (1979) 393–401, http://www.jstor.org/stable/3212906.
[12] Nikhil Bansal, Bart Kamphorst, Bert Zwart, Achievable performance of blind policies in heavy traffic, Math. Oper. Res. 43 (3) (2018) 949–964.
[13] Yan Chen, Jing Dong, Scheduling with service-time information: The power of two priority classes, 2020, Preprint.
[14] Ziv Scully, Mor Harchol-Balter, Alan Scheller-Wolf, SOAP: One clean analysis of all age-based scheduling policies, Proc. ACM Meas. Anal. Comput. Syst. 2 (1) (2018) 16:1–16:30, http://dx.doi.org/10.1145/3179419.
[15] Moshe Shaked, J. George Shanthikumar, Stochastic Orders, Springer Science & Business Media, 2007.
[16] S. Aalto, U. Ayesta, On the nonoptimality of the foreground-background discipline for IMRL service times, J. Appl. Probab. 43 (2) (2006) 523–534.
[17] Rhonda Righter, J. George Shanthikumar, Genji Yamazaki, On extremal service disciplines in single-stage queueing systems, J. Appl. Probab. 27 (2) (1990) 409–416.
[18] Rhonda Righter, J. George Shanthikumar, Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures, Probab. Engrg. Inform. Sci. 3 (3) (1989) 323–333.
[19] Samuli Aalto, Urtzi Ayesta, Mean delay analysis of multi level processor sharing disciplines, in: INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings, IEEE, 2006, pp. 1–11.
[20] N. Bingham, C. Goldie, J. Teugels, Regular Variation, Cambridge University Press, Cambridge, 1987.
[21] Sidney I. Resnick, Extreme Values, Regular Variation and Point Processes, Springer, 2013.
[22] Kevin D. Glazebrook, José Niño-Mora, Parallel scheduling of multiclass M/M/m queues: Approximate and heavy-traffic optimization of achievable performance, Oper. Res. 49 (4) (2001) 609–623.
[23] Kevin D. Glazebrook, An analysis of Klimov's problem with parallel servers, Math. Methods Oper. Res. 58 (1) (2003) 1–28.
[24] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Trans. Netw. 5 (6) (1997) 835–846.
[25] Martin F. Arlitt, Carey L. Williamson, Web server workload characterization: The search for invariants, ACM SIGMETRICS Perform. Eval. Rev. 24 (1) (1996) 126–137.
[26] Kihong Park, Walter Willinger, Self-similar network traffic: An overview, Self-Similar Netw. Traffic Perform. Eval. (2000) 1–38.
[27] Mor Harchol-Balter, Allen B. Downey, Exploiting process lifetime distributions for dynamic load balancing, ACM Trans. Comput. Syst. 15 (3) (1997) 253–285, http://dx.doi.org/10.1145/263326.263344.
[28] Bala Kalyanasundaram, Kirk R. Pruhs, Minimizing flow time nonclairvoyantly, in: Proceedings 38th Annual Symposium on Foundations of Computer Science, IEEE, 1997, pp. 345–352.

[29] Luca Becchetti, Stefano Leonardi, Nonclairvoyant scheduling to minimize the total flow time on single and parallel machines, J. ACM 51 (4) (2004) 517–539.
[30] Linus E. Schrage, Louis W. Miller, The queue M/G/1 with the shortest remaining processing time discipline, Oper. Res. 14 (4) (1966) 670–684.
[31] Linus E. Schrage, The queue M/G/1 with feedback to lower priority queues, Manage. Sci. 13 (7) (1967) 466–474.
[32] Natalia Osipova, Urtzi Ayesta, Konstantin Avrachenkov, Optimal policy for multi-class scheduling in a single server queue, in: Teletraffic Congress, 2009. ITC 21 2009. 21st International, IEEE, 2009, pp. 1–8.
[33] Leonard Kleinrock, Queueing Systems, Volume 2: Computer Applications , vol. 66, Wiley New York, 1976.
[34] Adam Wierman, Mor Harchol-Balter, Takayuki Osogami, Nearly insensitive bounds on SMART scheduling, in: ACM SIGMETRICS Performance Evaluation Review, vol. 33(1), ACM, 2005, pp. 205–216.
[35] Mor Harchol-Balter, Performance Modeling and Design of Computer Systems: Queueing Theory in Action, first ed., Cambridge University Press, New York, NY, USA, 2013.

Ziv Scully is a graduate student in Computer Science at Carnegie Mellon advised by Mor Harchol-Balter and Guy Blelloch. He graduated from MIT in 2016 with a BS in Mathematics with Computer Science. He is the recipient of an NSF Graduate Fellowship and an ARCS Foundation scholarship. Ziv's research focus is optimizing and analyzing computer systems and algorithms from a stochastic perspective, including job scheduling, load balancing, combinatorial optimization under uncertainty, and parallel algorithms. Recent publications of his have been recognized with awards from the INFORMS Applied Probability Society (Best Student Paper Prize finalist, 2018), IFIP PERFORMANCE (Best Student Paper Award winner, 2018), and ACM SIGMETRICS (Outstanding Student Paper Award winner, 2019; Best Video Award winner, 2020).

Isaac Grosof is a graduate student at Carnegie Mellon studying Computer Science, advised by Mor Harchol-Balter. Isaac graduated from MIT in 2017 with a BS and an MEng in Computer Science. Isaac's research focus is on the design, analysis and optimization of computer systems from a stochastic perspective, especially combining worst-case and stochastic techniques, and pushing the boundaries of tractable analysis. Isaac's publications have received the Best Student Paper award from IFIP Performance in 2018 and the Outstanding Student Paper award from ACM SIGMETRICS in 2019.

Mor Harchol-Balter is the Bruce J. Nelson Professor of Computer Science at Carnegie Mellon. She received her Ph.D. from U.C. Berkeley in 1996 under the direction of Manuel Blum. She joined CMU in 1999, and served as the Head of the PhD program from 2008-2011. Mor is a Fellow of both ACM and IEEE. She is a recipient of the McCandless Junior Chair, the NSF CAREER award, and several teaching awards, including the Herbert A. Simon Award and Spira Teaching Award. She is a recipient of dozens of Industrial Faculty Awards including multiple awards from Google, Microsoft, IBM, EMC, Facebook, Intel, Yahoo!, and Seagate. Mor's work focuses on designing new resource allocation policies, including load balancing policies, power management policies, and scheduling policies, for distributed systems. Mor is heavily involved in the SIGMETRICS/PERFORMANCE research community, where she has received many paper awards (SIGMETRICS '19, PERFORMANCE '18, INFORMS APS '18, EUROSYS '16, MASCOTS '16, SIGMETRICS '03, SIGMETRICS '96). She is also the author of a popular textbook, Performance Analysis and Design of Computer Systems , published by Cambridge University Press, which bridges Operations Research and Computer Science. Mor is best known for her enthusiastic keynote talks (recent examples: QTNA '19, YEQT '18, MIT LIDS '17, CANQUEUE '16, SIGMETRICS '16, ICDCS '15, GREENMETRICS '14) and her many PhD students, most of whom are professors at top academic institutions.