Preventing large sojourn times using SMART scheduling

Preventing large sojourn times using SMART scheduling

Misja Nuyens[*], Adam Wierman[†] and Bert Zwart [‡ §]

November 14, 2005

Abstract

Recently, the class of SMART scheduling policies (disciplines) has been introduced in order to formalize the common heuristic of "biasing toward small jobs." We study the tail of the sojourn-time (response-time) distribution under both SMART policies and the Foreground-Background policy (FB) in the GI/GI/1 queue. We prove that these policies behave very well under heavy-tailed service times. Specifically, we show that the sojourn-time tail under FB and all SMART policies is similar to that of the service time tail, up to a constant, which makes the SMART class superior to FCFS. In contrast, for light-tailed service times, we prove that the sojourn-time tail under FB and SMART is larger than that under FCFS. However, we show that the sojourn-time tail for a job of size y under FB and all SMART policies still outperforms FCFS as long as y is not too large.

Subject classifications: Queues: Priority, Limit Theorems. Probability: Stochastic model applications

Area of review: Stochastic models

[*]Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands, mnuyens@few.vu.nl

[†]Computer Science Department, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA, USA, acw@cs.cmu.edu

[‡]CWI, P.O. Box 94079, 1090 GB Amsterdam

[§]Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, zwart@win.tue.nl

1 Introduction

Scheduling policies (disciplines) that bias toward small job sizes (service requirements) have recently received attention across a number of computer application areas. For instance, variants of Shortest-Remaining-Processing-Time (SRPT), Foreground-Background (FB), and Preemptive-Shortest-Job-First (PSJF) have been suggested for use in web servers (Cherkasova (1998), Harchol-Balter et al. (2003), Rawat and Kshemkalyani (2003)), routers (Rai et al. (2004), Yang and Veciana (2002)), and databases (McWherter et al. (2004)). As a result of the attention given to size-based policies by computer systems researchers, there has been a resurgence in analytical work studying these policies, see Aalto, Ayesta and Nyberg-Oksanen (2004), Borst et al. (2003), Mandjes and Nuyens (2005), Núñez Queija (2002), and Wierman and Harchol-Balter (2003), with SRPT and FB dominating the literature.

However, SRPT and FB are idealized versions of the policies implemented by practitioners. The intricacies of computer systems force the use of more complex hybrid policies in practice, see McWherter et al. (2004), Rai et al. (2004), and Rawat and Kshemkalyani (2003). These more complex policies are built around the heuristic of "biasing toward jobs with small sizes or remaining sizes." The SMART classification (see Definition 1 below), introduced in Wierman, Harchol-Balter and Osogami (2005), formalizes this heuristic. The SMART classification includes policies such as SRPT and PSJF, but does not include FB, Processor-Sharing (PS), or First-Come-First-Served (FCFS). To this point, it has been proven that all SMART policies have mean sojourn time (response time) within a factor of two of optimal (Wierman, Harchol-Balter and Osogami (2005)). However, beyond the mean sojourn time, little is known about the behavior of SMART policies.

Though the SMART classification does not include FB, there are many similarities between FB and SMART policies. At every point in time, FB shares the server evenly among all the jobs in the system with the smallest age (least attained service), so that small jobs will have the server mostly to themselves. Furthermore, under distributions with decreasing failure rate, the age of a job is a good indicator of its remaining size. Since FB attempts to bias toward small (remaining) job sizes without knowledge of job sizes, it can be viewed a "poor man's SMART policy". However, without knowledge of job sizes, FB cannot bias as strongly towards small jobs as SMART policies. Therefore it is interesting to contrast the behavior of SMART policies with the behavior of FB .

In this work, we focus on the behavior of the sojourn-time tail of FB and SMART policies in a GI/GI/1 queue under both heavy-tailed and light-tailed service distributions. We characterize the likelihood of large sojourn times under FB and SMART policies. For these policies such an analysis is especially important because it can quell fears that large jobs suffer "starvation" as a result of the bias toward small jobs.

We prove two main results. First, we show that for a large class of heavy-tailed service distributions, both FB and SMART policies have a sojourn-time tail that is similar to that of the service

distribution, up to a multiplicative constant (Theorem 2). This result is encouraging, since in many computer applications service distributions tend to be heavy-tailed, see Barford and Crovella (1998), Downey (2001), Leland et al. (1993), and Peterson (1996). Second, for a large class of light-tailed service distributions having no mass at the endpoint of the distribution, both FB and SMART policies have a sojourn-time tail that is equal, on a logarithmic scale, to that of the busy period (Theorem 3). Interestingly, when the service distribution is allowed to have mass at the endpoint, FB still has a sojourn-time tail that is equal to that of the busy period, while some SMART policies can have a lighter tail.

Theorems 2 and 3 illustrate a trade-off that seems to be a general tendency: policies that have (near) optimal sojourn-time tail behavior under heavy-tailed service distributions, can behave poorly under light-tailed service distributions. In particular, it seems unlikely that any policy can obtain the "best of both worlds." The service distribution must thus play a key role in the choice of the scheduling policy for a particular application.

A more detailed look at the picture sketched above reveals the following: the poor behavior of the sojourn time under SMART disciplines under light-tailed service distributions is merely caused by the behavior of the largest jobs. In fact, the tail behavior of the sojourn time of a job of size y under any service distribution (light or heavy-tailed) is still better than that under FCFS, provided that y is not too large. *Using a policy from the* SMART *class is therefore especially attractive when the tail of the service-time distribution is not known in advance.*

The results we have described so far illustrate the similarities of FB and SMART with respect to the sojourn-time tail under both heavy and light-tailed service distributions. However, one expects SMART policies to provide smaller sojourn times than FB, since FB does not use knowledge of the job sizes to make scheduling decisions. This expectation is confirmed in Theorem 5, where we show that the conditional sojourn time for a job of size x is stochastically larger under FB than under any SMART policy in an M/GI/1 queue.

This paper is organized as follows. We introduce the SMART classification in Section 2. Next, in Section 3, we introduce our notation and define the classes of service-time distributions we study. In Section 4 we present and discuss the main results of the paper. The analysis begins in Section 5 where we study the case of light-tailed service distributions, and continues in Section 6 where we study the case heavy-tailed service distributions. Then, in Section 7, we restrict the analysis to the M/GI/1 setting and derive stochastic bounds relating the sojourn time under FB and SMART disciplines. Finally, we conclude in Section 8.

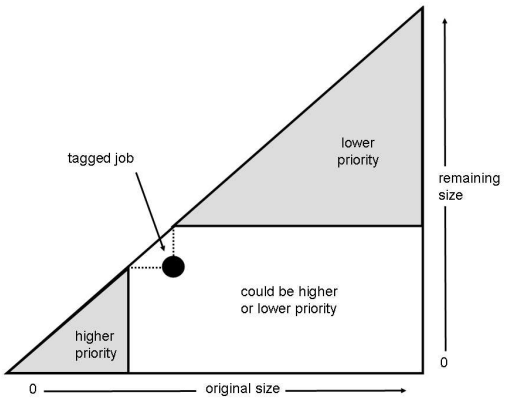

Figure 1: *This diagram illustrates the priority structure induced by the Bias Property in Definition 1. Furthermore, the Consistency and Transitivity properties guarantee that, upon arrival, a job will find at most one job with higher priority in the white region (see Wierman, Harchol-Balter and Osogami (2005)).*

2 Defining SMART policies

In Wierman, Harchol-Balter and Osogami (2005), the class of `SMART` policies is defined as follows. In the definition, we denote jobs by a, b, or c, where job a has remaining size r_a and original size s_a. We also define job a to have priority over job b if job b can never run while job a is in the system.

Definition 1 *A work conserving policy* `P` *belongs to the class* `SMART`*, notation* `P` $\in$ `SMART`*, if it obeys the following properties.*

Bias Property: *If $r_b > s_a$, then job a has priority over job b.*

Consistency Property: *If job a ever receives service while job b is in the system, then at all times thereafter job a has priority over job b.*

Transitivity Property: *If an arriving job b preempts job c, then thereafter, until job c receives service, every arrival a with size $s_a < s_b$ is given priority over job c.*

Figure 1 illustrates the priority structure in the definition of `SMART`. This definition formalizes the heuristic of biasing toward jobs that are (originally) short or have small remaining service requirements. Furthermore, this heuristic is validated by Theorem 5.1 of Wierman, Harchol-Balter and Osogami (2005), which states that all `SMART` policies have sojourn time within a factor of 2 of optimal. Thus, `SMART` policies have provably "`SMA`ll `R`esponse `T`imes".

The first thing to notice about the class of SMART policies is that it includes many common policies. The SMART class includes, for example, SRPT and PSJF, which always serves the job with the smallest original size. Furthermore, it is easy to prove that the SMART class includes all policies that assign to a job the product of its remaining size raised to the ith power and its original size raised to the jth power, for $i, j > 0$, and give priority to the job with lowest product. The SMART class also includes a range of policies with more complicated priority schemes. Furthermore, it is important to point out that SMART includes many policies that do not maintain a static priority scheme, e.g., a SMART policy may switch from using the SRPT rule to using the PSJF rule over time.

Despite its breadth, many policies are excluded from SMART. Since the class of SMART policies only includes preemptive policies, it does not include Shortest-Job-First (SJF), which non-preemptively serves the smallest job in the system. Nor does it include any age based policies, like FB. For more details on which policies are included in SMART , and a discussion of the motivation for each of the three properties in the definition of SMART, see Wierman, Harchol-Balter and Osogami (2005).

3 Assumptions

Throughout the paper, unless otherwise stated, we consider a stationary preemptive-resume GI/GI/1 queue with generic service time B, having $E[B] < \infty$, and generic interarrival time A. The system load ρ satisfies $\rho = E[B]/E[A] < 1$. Let F be the service distribution and $\overline{F} = 1 - F$ its tail. Define its (right) endpoint $x_F \stackrel{\text{def}}{=} \sup\{x : F(x) < 1\}$. Define $f(x) \sim g(x)$ to mean that $\lim_{x\to\infty} f(x)/g(x) = 1$. Let Φ_X denote the moment generating function of a random variable X, i.e., $\Phi_X(s) = E[e^{sX}]$.

Let $V^{\mathtt{P}}$ denote the sojourn time (response time) under the policy P; this is the time between the arrival and the departure of a job. Let $V(x)^{\mathtt{P}}$ be the conditional sojourn time for a job of size x under policy P. Denote the *waiting time*, the time a job waits before its service starts, by $W^{\mathtt{P}}$, and let $W(x)^{\mathtt{P}}$ be the waiting time of a job of size x. Finally, let $R^{\mathtt{P}}$ be the *residence time* of a job, i.e., the time the job spends in the system after his service has started, and let $R(x)^{\mathtt{P}}$ denote the residence time of a job of size x. Hence, we may write

$$V(x)^{\mathtt{P}} = R(x)^{\mathtt{P}} + W(x)^{\mathtt{P}}.$$

We consider two classes of service distributions in this work. The class of heavy-tailed distributions that we study are those of intermediate regular variation at infinity:

Definition 2 *We say that the tail $\overline{F}(x)$ of a distribution is of* ***intermediate regular variation at infinity****, $\overline{F} \in \mathcal{IR}$, when*

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

This class includes all regularly varying tails and thus includes, for example, Pareto distributions.

The class of light tailed distributions we study obeys the following assumptions:

Assumption A: $\Phi_B(s) < \infty$ *for some* $s > 0$.

Assumption B: $P(B = x_F) = 0$.

Note that the distributions that satisfy both of these assumptions include light-tailed distributions with infinite endpoints (e.g., exponential, gamma, and certain Weibull distributions), as well as all continuous distributions with finite support (e.g., uniform and beta distributions).

When studying the case of light-tailed service, we will describe the logarithmic behavior of the tail of the sojourn time distribution using the *decay rate*.

Definition 3 *The* ***(asymptotic) decay rate*** $\gamma(X)$ *of a random variable* X *is defined by*

$$\gamma(X) = \lim_{x\to\infty} \frac{-\log P(X > x)}{x},$$

given that the limit exists.

Informally, for large x, one may write $P(X > x) \approx e^{-\gamma(X)x}$. It should be noted that a smaller decay rate corresponds to a larger tail of the distribution.

In both the light and heavy tailed case, our analysis will depend heavily on the use of busy periods. Let L be the length of a generic busy period, and let L_x be the length of a busy period in the queue with generic service time $BI(B < x)$ and generic interarrival time A. So L_x is the busy period made up of arrivals with service times less than x. Furthermore, let $L_x(y)$ be an L_x busy period that is started by a job of size y. Define $x \wedge y \stackrel{\text{def}}{=} \min(x, y)$. Let $\widetilde{L_x}$ be the length of a busy period with job sizes $B \wedge x$ and generic interarrival time A, and define $\widetilde{L_x}(y)$ similarly. The decay rate of the busy period can be expressed in terms of the moment generating functions of A and B. The following result is taken from Nuyens & Zwart (submitted).

Lemma 1 *For* $0 < x \leq \infty$,

$$\gamma(L_x) = \sup_{s\geq 0}\left[s + \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_{BI(B<x)}(s)}\right)\right],$$
$$\gamma(\widetilde{L_x}) = \sup_{s\geq 0}\left[s + \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_{B\wedge x}(s)}\right)\right].$$

In particular,

$$\gamma(L) = \sup_{s\geq 0}\left[s + \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(s)}\right)\right].$$

The expressions for $\gamma(L), \gamma(L_x)$ and $\gamma(\widetilde{L_x})$ can in general be solved numerically. If the arrival process is Poisson with rate λ say, we get $\gamma(L) = \sup_{s \geq 0} [s - \lambda(\Phi_B(s) - 1)]$. Specializing this to the $M/M/1$ queue, where the service times have an exponential distribution with rate μ, we get the explicit expression $\gamma(L) = \mu(1 - \sqrt{\rho})^2$.

4 Results and discussion

We now state and discuss the main contributions of this work.

Theorem 2 *In the GI/GI/1 queue with* $\mathtt{P} \in \mathtt{SMART}$ *or* $\mathtt{P} = \mathtt{FB}$*, if* $\overline{F} \in \mathcal{IR}$*, then*

$$P(V^{\mathtt{P}} > x) \sim P(B > (1-\rho)x), \text{ as } x \to \infty. \tag{1}$$

Since $\mathtt{SRPT} \in \mathtt{SMART}$, Theorem 2 can be viewed as a generalization of the recent results that show that relation (1) holds in the M/GI/1 setting for SRPT (Núñez Queija (2002)) and for FB (Núñez Queija (2002) and Nuyens (2004)). Furthermore, relation (1) has been shown to hold for a number of queues with the processor sharing discipline, see (Borst, Nunez and Zwart (submitted)) for an overview.

Although it has not been proven, the sojourn time tail of SMART policies seems to be optimal: there seem to be no policies with a smaller tail of the sojourn-time distribution than described in (1). In any case, this behavior is "near optimal" in the sense that no policy can have a sojourn time tail more than a multiplicative constant smaller. Furthermore, for heavy-tailed distributions, the tail of the sojourn time under a SMART policy is much smaller than the tail under FCFS: for FCFS and all other non-preemptive policies, the sojourn-time tail is 'one degree heavier' than the tail of the service distribution, i.e., it is of the order $xP(B > x)$, see Borst et al. (2003) for a survey.

For the second main result, recall that L is the length of a generic busy period.

Theorem 3 *In the GI/GI/1 queue with* $\mathtt{P} \in \mathtt{SMART}$*, if Assumption A holds, then*

$$\gamma(L) = \gamma(V^{\mathtt{FB}}) \leq \gamma(V^{\mathtt{P}}) \leq \gamma(V^{\mathtt{SRPT}}).$$

Furthermore, if both Assumptions A and B hold, then $\gamma(V^{\mathtt{P}}) = \gamma(V^{\mathtt{FB}}) = \gamma(L)$*. That is,*

$$\log P(V^{\mathtt{P}} > x) \sim \log P(V^{\mathtt{FB}} > x) \sim \log P(L > x), \text{ as } x \to \infty. \tag{2}$$

For light-tailed service times, $\gamma(L)$ is the smallest possible decay rate of the sojourn time, see Lemma 6. Hence, we can conclude that, on a logarithmic scale, the tail of V under SMART disciplines can be as large as possible. Theorem 3 generalizes the result obtained in Mandjes and Nuyens (2005) for the M/GI/1/FB queue. Since $\mathtt{SRPT} \in \mathtt{SMART}$, Theorem 3 can also be viewed as a generalization of a result in Nuyens and Zwart (submitted), where it is shown that (2) holds under Assumptions A

and B for GI/GI/1/SRPT. In addition, Nuyens and Zwart (submitted) show that, for distributions that satisfy Assumption A but not Assumption B, the decay rate of $V^{\texttt{SRPT}}$ in the GI/GI/1 queue lies strictly between that of L and that of the stationary workload in the queue, which is equal to the decay rate of the sojourn time under FCFS. Relation (2) has also recently been obtained for PS in Mandjes and Zwart (2004): in the GI/GI/1/PS queue, (2) holds under Assumptions A, B, and an additional condition that rules out service distributions with very light tails.

The above theorems emphasize the similarities in the sojourn-time distribution of SMART and FB. However, note that if Assumption A holds and Assumption B does not hold, the decay rate depends in a delicate way on the specific policy considered. The treatment of jobs with the same priority becomes crucial. That is, when jobs of the same priority are ordered in a FCFS manner (as in SRPT) the tail behavior can differ from when the server is shared among jobs with the same priority (as in FB).

It is important to note the contrast in the behavior of the sojourn-time tail of FB and SMART policies under heavy-tailed and light-tailed service distributions. This seems to be a general tendency: policies that behave (near) optimally for heavy-tailed service times, can behave very poorly for light-tailed distributions. However, a disclaimer should be added: this poor behavior is merely caused by the sojourn times of the largest jobs. For smaller jobs, using FB or a SMART policy may not be so bad, as illustrated by the following theorem.

Theorem 4 *Suppose* $\texttt{P} \in \texttt{SMART}$. *Let* y *be such that* $P(B = y) = 0$. *Then*

$$\gamma(V^{\texttt{P}}(y)) = \gamma(L_y). \tag{3}$$

Furthermore, for all y,

$$\gamma(V^{\texttt{FB}}(y)) = \gamma(\widetilde{L_y}). \tag{4}$$

This result unifies earlier results for the GI/GI/1 SRPT and the M/GI/1 FB in Mandjes and Nuyens (2005) and Nuyens and Zwart (submitted), and is important for several reasons. First of all, if y is not too large, $\gamma(L_y)$ is larger than the decay rate $\gamma_{\texttt{FCFS}}$ of the sojourn time under FCFS as is illustrated in Nuyens and Zwart (submitted). In particular, the threshold value y^* for which $\gamma(L_{y^*}) = \gamma_{\texttt{FCFS}}$ converges to infinity if the traffic is either light ($\rho \to 0$) or heavy ($\rho \to 1$). Thus, under very high or low loads, one can still say that the SMART class outperforms FCFS under light-tailed service times. In addition, (3) is still valid when service times are heavy tailed. This illustrates the robustness of policies in the SMART class, which is important when the specific shape of the service-time distribution is not known in advance.

Theorem 4 illustrates that the tail of the conditional response time is heavier under FB than under SMART policies. This is not surprising since FB does not use information about the job sizes or remaining sizes when scheduling. In fact, FB was intentionally excluded from SMART because

$E[V^{\mathtt{FB}}]$ can be made arbitrarily larger than $E[V^{\mathtt{P}}]$ for $\mathtt{P} \in \mathtt{SMART}$, see Wierman, Harchol-Balter and Osogami (2005). In Section 7, we prove the following stochastic bound illustrating that SMART policies outperform FB in the M/GI/1 setting:

Theorem 5 *In an M/GI/1 queue, for all* $\mathtt{P} \in \mathtt{SMART}$*:*

$$V(x)^{\mathtt{P}} \leq_{st} R(x)^{\mathtt{PSJF}} + W(x)^{\mathtt{SRPT}} \leq_{st} V(x)^{\mathtt{FB}}.$$

5 Light-tailed service demands

In this section we prove Theorems 3 and 4. Theorem 3 follows from Lemmas 8, 9, 10 and 12, while Theorem 4 follows from Lemmas 11 and 12.

We start by upper bounding the tail of $V^{\mathtt{P}}$ under all work-conserving disciplines P using the observation that $V^{\mathtt{P}} \leq L^*$, where $L^* \stackrel{d}{=} L(Q+B)$ is the length of the busy period starting with the amount of work $Q+B$, Q is the steady-state amount of work in the system (upon customer arrivals), and B is a generic service time. Furthermore, $L(\cdot)$, Q and B are independent, i.e., L^* is a residual busy period.

It can be shown that the decay rates of L^* and L coincide. Moreover, we have the following upper bound:

Lemma 6 *For all work-conserving disciplines* P*,*

$$\limsup_{x\to\infty} \frac{1}{x} \log P(V^{\mathtt{P}} > x) \leq -\gamma(L^*) = -\gamma(L). \tag{5}$$

Proof The first inequality follows from the above observation that $V^{\mathtt{P}} \leq L^*$. The equality $\gamma(L^*) = \gamma(L)$ is trivial in the $M/G/1$ queue, but for the $GI/GI/1$ queue we need additional arguments. Let V be the steady-state virtual waiting time in the $GI/GI/1$ queue. Then $[L(V) \mid L(V) > 0] \stackrel{d}{=} [L(V) \mid V > 0]$ has density $P(L > x)/E[L]$. By Lemma 3.2 in Abate and Whitt (1997), L and $L(V)$ have the same decay rate. However, we are interested in the decay rate of $L(Q+B)$. In the $M/G/1$ case, we could apply PASTA. In the general case, we note that $V \stackrel{d}{=} (Q+B-A^*)^+$, with A^* a residual interarrival time. Therefore, V is stochastically smaller than $Q+B$. Consequently, $\gamma(L^*) = \gamma(L(Q+B)) \leq \gamma(L(V)) = \gamma(L)$. To prove the upper bound, let $A(t)$ be the total amount of work fed into the system between time 0 and t. Using the Chernov bound, we find

$$P(L(Q+B) > x) \leq P(A(x) - x + Q + B > 0) \leq E[e^{sQ}]E[e^{sB}]E[e^{s(A(x)-x)}].$$

The proof is now completed by minimizing the last factor over s, and showing that for the optimizing argument s^*, we have $E[e^{s^*Q}] < \infty$ and $E[e^{s^*B}] < \infty$. Since this is exactly what is done in Proposition 3.1 of Mandjes & Zwart (2004), we refer to that work for the remaining supporting

arguments. □

The following lemma, which is Proposition 2.2 in Nuyens and Zwart (submitted), will play a key role in our arguments.

Lemma 7 *For a GI/GI/1 queue under Assumption A,* $\gamma(L_x) \downarrow \gamma(L)$ `and` $\gamma(\widetilde{L_x}) \downarrow \gamma(L)$ *as* $x \uparrow x_F$.

After these two preliminary lemmas, we are now ready to prove Theorems 3 and 4. we start by analyzing the behavior of `SMART` . The techniques we apply are similar to those applied for `SRPT` in Nuyens and Zwart (submitted). We start by doing the analysis in the simplest case: both Assumptions A and B hold, and the service distribution is unbounded.

Lemma 8 *In the GI/GI/1 queue with* $\mathtt{P} \in \mathtt{SMART}$*, if Assumptions A and B hold, and* $x_F = \infty$*, then* $\gamma(V^{\mathtt{P}}) = \gamma(L)$*. That is,*

$$\log P(V^{\mathtt{P}} > x) \sim \log P(L > x), \text{ as } x \to \infty.$$

Proof Let A_1 be the first arrival after that of a tagged customer with size B_0. Let a be such that $P(A < a) > 0$ and $y < x_F - a$. Then for all $P \in \mathtt{SMART}$,

$$\begin{aligned} P(V^{\mathtt{P}} \geq x) &\geq P(V(B_0)^{\mathtt{P}} > x, A_1 < a, B_0 > y + a) \\ &= P(A_1 < a, B_0 > y + a) P(V(B_0)^{\mathtt{P}} > x | A_1 < a, B_0 > y + a). \end{aligned}$$

Conditional on $B_0 > y + a$ and $A_1 < a$, the tagged job has remaining service time larger than y when the new job arrives. The Bias Property implies that this new job has higher priority than the tagged job if its service times is smaller than y. Furthermore, all jobs with service time smaller than y that arrive while the new job is in the system will also have higher priority than the tagged job. Thus, conditional on $B_0 > y + a$ and $A_1 < a$, we have $V(B_0)^{\mathtt{P}} \geq_{st} L_y$. Hence,

$$P(V^{\mathtt{P}} \geq x) \geq P(A_1 < a, B_0 > y + a) P(L_y > x).$$

Since $P(A_1 < a, B_0 > y + a) > 0$, the existence of $\gamma(L_y)$ implies that

$$\liminf_{x \to \infty} \frac{1}{x} \log P(V^{\mathtt{P}} \geq x) \geq \liminf_{x \to \infty} \frac{1}{x} P(L_y > x) = -\gamma(L_y). \tag{6}$$

To prove the lemma, it suffices to show that the liminf result corresponding to (5) holds. Letting y go to ∞ in (6), and applying Lemma 7, yields

$$\liminf_{x \to \infty} \frac{1}{x} \log P(V^{\mathtt{P}} \geq x) \geq -\gamma(L).$$

This completes the proof. □

We now relax the assumption that the service distribution is unbounded. This relaxation results in the need for a more involved argument.

Lemma 9 *In the GI/GI/1 queue with* $\mathtt{P} \in \mathtt{SMART}$*, if Assumptions A and B hold, and* $x_F < \infty$*, then* $\gamma(V^{\mathtt{P}}) = \gamma(L)$*.*

Proof If $P(A < a) > 0$ for all $a > 0$, then the result follows from (6) and Lemma 7, as in the proof of Lemma 8. However, this may not be the case, so we need a different construction.

By definition of x_F, there exists a decreasing sequence $\{\varepsilon_n\}$ such that $\varepsilon_n \to 0$ as $n \to \infty$, and $P(x_F - \varepsilon_n < B < x_F - \varepsilon_n/2) > 0$ for all n. Since $P(B > A) > 0$, we can assume that ε_1 is such that $P(A < x_F - 2\varepsilon_1) > 0$. Let Z_n be the event that the last $\lfloor x_F/\varepsilon_n \rfloor$ customers that arrived before the tagged customer had a service time in the interval $[x_F - \varepsilon_n, x_F - \varepsilon_n/2]$, and that the last $\lfloor x_F/\varepsilon_n \rfloor$ inter-arrival times were smaller than $x_F - 2\varepsilon_n$. By definition of ε_n, we have $P(Z_n) > 0$ for all n.

Furthermore, the Bias Property guarantees that, on the event Z_n, there is a customer with remaining service time larger than $k\varepsilon_n$ after the kth of the inter-arrival times. Hence, at the arrival of the tagged customer (after $k = \lfloor x_F/\varepsilon_n \rfloor$ arrivals), there is a customer in the system with remaining service time in the interval $[x_F - \varepsilon_n, x_F - \varepsilon_n/2]$. If the tagged customer has service time $B_0 > x_F - \varepsilon_n/2$, his sojourn time satisfies $V^{\mathtt{P}} \geq L_{x_F - \varepsilon_n}$. Consequently, for all $n \in \mathbb{N}$,

$$P(V^{\mathtt{P}} > x) \geq P(Z_n)P(B_0 > x_F - \varepsilon_n/2)P(L_{x_F-\varepsilon_n} > x).$$

Thus, for $\mathtt{P} \in \mathtt{SMART}$, we have

$$\liminf_{x\to\infty} \frac{1}{x} \log P(V^{\mathtt{P}} > x) \geq -\gamma(L_{x_F-\varepsilon_n}).$$

As $n \to \infty$, and hence $\varepsilon_n \downarrow 0$, Lemma 7 implies that $\gamma(L_{x_F-\varepsilon_n}) \to \gamma(L)$. Using Lemma 6 completes the proof. □

Finally, we relax Assumption B. In contrast to FB, the sojourn-time tail of SMART policies can improve when there is mass in the endpoint of the service distribution. This is not surprising since many SMART policies, e.g., SRPT, are equivalent to FCFS in the GI/D/1 queue. However, SMART also includes policies where, like FB, jobs of the same size are not served in FCFS order. Thus, the SMART policies have a range of possible sojourn-time tails in this setting.

Lemma 10 *In the GI/GI/1 queue, under Assumption A, for all* $\mathtt{P} \in \mathtt{SMART}$*,*

$$\gamma(V^{\mathtt{FB}}) \leq \gamma(V^{\mathtt{P}}) \leq \gamma(V^{\mathtt{SRPT}}). \tag{7}$$

Proof By Theorem 3, we only need to deal with the case that Assumption B does not hold, i.e., $P(B = x_F) > 0$. The first inequality follows from Lemma 6 and Lemma 12. For the second inequality, note that $P(V^{\mathtt{P}} > x) \geq P(V(x_F)^{\mathtt{P}} > x)P(B = x_F)$. Thus, since $P(B = x_F) > 0$, $\gamma(V^{\mathtt{P}}) \leq \gamma(V(x_F)^{\mathtt{P}})$. Furthermore, for all $\mathtt{P} \in \mathtt{SMART}$, $V(x_F)^{\mathtt{P}} \geq_{st} W(x_F)^{\mathtt{P}} \geq_{st} W_2(x_F)$, where

$W_2(x_F)$ is the waiting time of a low priority job in a 2-class priority queue where the high-priority class includes all jobs smaller than x_F. To complete the proof, we apply Theorems 3.1 and 4.2 of Nuyens and Zwart (submitted), which state that $\gamma(W_2(x_F)) = \gamma(V^{\texttt{SRPT}})$. $\square$

We end this section with the analysis of the conditional sojourn time under SMART policies. Again, this analysis is more complex than the case of FB; however the approach is similar to that applied for SRPT in Nuyens and Zwart (submitted).

Lemma 11 *In the GI/GI/1 queue with* $\texttt{P} \in \texttt{SMART}$*, if* $P(B = y) = 0$*, then*

$$\gamma(V(y)^{\texttt{P}}) = \gamma(L_y). \tag{8}$$

Proof For the lower bound, we remark that $V^{\texttt{P}}(y) \geq_{st} L_y^*$ for all $\texttt{P} \in \texttt{SMART}$. By Lemma 6, this residual busy period has decay rate $\gamma(L_y)$.

For the upper bound, we use Lemma 4.1 of Wierman, Harchol-Balter and Osogami (2005), which states that at any point in time, at most one customer with original service time larger than y has remaining service time smaller than y. Denoting by Q_y the stationary workload, upon arrival instants, made up of customers with service time smaller than y, we can bound

$$V^{\texttt{P}}(y) \leq_{st} L_y(Q_y + y + y).$$

Denoting the amount of work brought by customers (with size smaller than y) entering the queue in the interval $[0, x]$ by $A_y(x)$, the Chernov bound yields that for all $s \geq 0$,

$$\begin{aligned} P(V^{\texttt{P}}(y) > x) &\leq P(L_y(Q_y + 2y) > x) \leq P(Q_y + 2y + A_y(x) > x) \\ &= P(\exp(s(Q_y + 2y + A_y(x)) > e^{sx}) \leq e^{-sx} e^{2sy} E e^{sQ_y} E e^{sA_y(x)}. \end{aligned}$$

Hence, for all $s < \gamma(Q_y)$, we have

$$\limsup_{x\to\infty} \frac{1}{x} \log P(V^{\texttt{P}}(y) > x) \leq -s + \limsup_{x\to\infty} \frac{1}{x} \log E e^{sA_y(x)} = -s - \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_{BI(B<y)}(s)}\right),$$

where the equality follows from Lemma 2.1 in Mandjes and Zwart (2004). Taking the infimum over all $s \in [0, \gamma(Q_y))$ yields

$$\limsup_{x\to\infty} \frac{1}{x} \log P(V^{\texttt{P}}(y) > x) \leq -\sup_{0\leq s<\gamma(Q_y)} \left[s + \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_{BI(B<y)}(s)}\right) \right] = -\gamma(L_y^*),$$

where the equality follows from equation (5.1) in Nuyens and Zwart (submitted). Noting that L_y and L_y^* have the same decay rate yields the desired upper bound, and completes the proof. $\square$

The proof of Lemma 11 can be adapted to give the following result for FB.

Lemma 12 *In the GI/GI/1 queue, under Assumption A, $\gamma(V^{\mathtt{FB}}(y)) = \gamma(\widetilde{L_y})$ for all y. Furthermore, $\gamma(V^{\mathtt{FB}}) = \gamma(\widetilde{L})$.*

Proof In the queue with generic service time $B \wedge y$, we have $V^{\mathtt{FB}} \stackrel{d}{=} \widetilde{L_y}^*$. Hence, $V^{\mathtt{FB}}(y) \geq_{st} \widetilde{L_y}^*$. Furthermore, $V^{\mathtt{FB}}(y) \leq_{st} L_y(Q_y + y)$, where Q_y is the stationary workload in the queue upon arrival instants. Using these two bounds for $V^{\mathtt{FB}}(y)$, we can apply the proof of Lemma 12, replacing $BI(B < y)$ with $B \wedge y$, to get

$$\gamma(V^{\mathtt{FB}}(y)) = \gamma(\widetilde{L}_y^*). \tag{9}$$

To prove the second part of the lemma, we consider two cases. If $P(B = x_F) > 0$, then using (9) with $y = x_F$ yields the desired result. If $P(B = x_F) = 0$, then the result follows from applying Lemma 7. □

6 Heavy-tailed service demands

In this section we prove Theorem 2. To do so, we will use the following sufficient conditions and theorem from Guillemin, Robert and Zwart (2004), see also Borst, Nunez & Zwart (submitted):

Condition 1 *For some $g > 0$, $V(x)/x \to g$ a.s. as $x \to \infty$.*

Condition 2 *There exists a constant k such that*

$$P(V(x) > kx) = o(\overline{F}(x)).$$

Theorem 13 *If Conditions 1 and 2 hold, $\overline{F} \in \mathcal{IR}$, and $E[B^p] < \infty$ for some $p > 1$, then*

$$P(V > gx) \sim P(B > x) \text{ as } x \to \infty.$$

We will prove Theorem 2 in two steps: first we show that Condition 1 holds for FB and all $\mathtt{P} \in \mathtt{SMART}$, and then that Condition 2 holds.

Lemma 14 *For all $\mathtt{P} \in \mathtt{SMART}$ and $\mathtt{P} = \mathtt{FB}$, we have that $V(x)^{\mathtt{P}}/x \to 1/(1-\rho)$ a.s. as $x \to \infty$.*

It is not surprising that Condition 1 holds: as noted in Núñez Queija (2002), it seems that such a result holds for all policies under which the system remains stable when a permanent customer is added, including FB and all SMART policies. The intuition is that the fraction of service capacity given to a non-permanent customer must converge to ρ for the system to remain stable; thus the permanent customer must receive a fraction $1-\rho$ of the service capacity during its sojourn time.

Proof We will prove the result by showing upper and lower bounds on the limit.

To prove the lower bound, we first derive a stochastic lower bound for the sojourn time of $\mathtt{P} = \mathtt{FB}$ and $\mathtt{P} \in \mathtt{SMART}$ in terms of a single busy period. For FB we have

$$V(x)^{\mathtt{FB}} \geq_{st} \widetilde{L_x}(x) \geq_{st} L_x(x) \geq_{st} L_{\epsilon x}((1-\epsilon)x), \qquad 0 < \epsilon < 1. \tag{10}$$

Furthermore, for $\mathtt{P} \in \mathtt{SMART}$, the Bias property guarantees that until the tagged job has received $(1-\epsilon)x$ units of service, all arriving jobs smaller than ϵx receive priority. Hence,

$$V(x)^{\mathtt{P}} \geq_{st} L_{\epsilon x}((1-\epsilon)x), \qquad 0 < \epsilon < 1. \tag{11}$$

To understand the length of the busy period, we will analyze a PLCFS system. Define $S_y(t)$ to be the service given in time t to a permanent customer arriving in an empty queue at time 0 when the generic service time is $BI_{[B<y]}$. Denoting the inverse of L_y by L_y^{-1}, we have for all x and t,

$$P(S_y(t) > x) = P(L_y(x) < t) = P(L_y^{-1}(t) > x).$$

Hence, S_y is stochastically equal to L_y^{-1}, so that

$$\lim_{x\to\infty} \frac{L_y(x)}{x} = \lim_{x\to\infty} \frac{x}{L_y^{-1}(x)} = \lim_{x\to\infty} \frac{x}{S_y(x)} \qquad a.s.$$

Note that this also holds if y is a function of x. Furthermore, define $S(t) = \lim_{y\to\infty} S_y(t)$. Then

$$\lim_{t\to\infty} \frac{S(t)}{t} = 1-\rho \text{ a.s.} \tag{12}$$

Set $z = (1-\epsilon)x$. From (10), (11) and (12), it follows that for all $0 < \varepsilon < 1$,

$$\begin{aligned}
\lim_{x\to\infty} \frac{V(x)^{\mathtt{P}}}{x} &\geq \lim_{x\to\infty} \frac{L_{\varepsilon x}((1-\varepsilon)x)}{x} \\
&= (1-\varepsilon) \lim_{z\to\infty} \frac{L_{\varepsilon z/(1-\varepsilon)}(z)}{z} \\
&= (1-\varepsilon) \lim_{z\to\infty} \frac{z}{S_{\varepsilon z/(1-\varepsilon)}(z)} \\
&= \frac{1-\varepsilon}{1-\rho} \qquad a.s.
\end{aligned}$$

The final equality follows because for any constant c, there exists a $z(c)$ such that for all $z > z(c)$, $S_c(z) \leq_{st} S_{\epsilon z/(1-\epsilon)}(z) \leq_{st} S(z)$. This completes the proof of the lower bound.

We now move to the upper bound. For all $\mathtt{P} \in \mathtt{SMART}$ and $\mathtt{P} = \mathtt{FB}$,

$$V(x)^{\mathtt{P}} \leq_{st} L(x+Q)$$

where Q is the steady state work in the system upon customer arrivals. Again using a PLCFS system, the events $\{S(t) - Q = x\}$ and $\{L(x+Q) = t\}$ have the same probability. Arguing as above, and using that

$$\lim_{t\to\infty} \frac{S(t)-Q}{t} = 1-\rho \text{ a.s.},$$

completes the proof of the upper bound. □

Before proving that Condition 2 holds for SMART and FB, we prove an auxiliary result. A similar result has been shown before for the workload in the $M/G/1$ queue in Jelenkovic and Momcilovic (2003). A key ingredient to the proof of this auxiliary result is the following lemma, which is due to Resnick and Samorodnitsky (1999).

Lemma 15 *Let $S_n = X_1 + \cdots + X_n$ be a random walk with i.i.d. step sizes such that $E[X_1] < 0$ and $E[(X_1^+)^p] < \infty$ for some $p > 1$. Then, for any $\alpha < \infty$, there exist $c, k^* > 0$ such that for any n, x and $k > k^*$,*

$$P(S_n > kx \mid X_i < x, i \leq n) \leq cx^{-\alpha}.$$

Lemma 16 *Let X_i be i.i.d. random variables with $E[(X_i^+)^p] < \infty$ for some $p > 1$. Let $S_n(y) = \sum_{i=1}^n (X_i \wedge y)$. Define $M(y) = \sup_n S_n(y)$. For every $\beta > 0$, there exists a $k > 0$ such that $P(M(x) > kx) = \mathrm{o}(x^{-\beta})$.*

Proof Let $\beta > 0$. For fixed $y \geq 1$, we write the standard geometric random sum decomposition

$$M(y) \stackrel{d}{=} \sum_{i=1}^{N(y)} H_i(y),$$

with $N(y)$ the number of ladder heights, and $H_i(y)$ the ith overshoot; for details see e.g. Chapter VIII of Asmussen (2003). By a sample-path comparison, it follows that

$$M(y) \stackrel{st}{\leq} \sum_{i=1}^{N(\infty)} [H_i(\infty) \wedge y].$$

Writing $H_i = H_i(\infty)$, we have for any $k, \gamma > 0$,

$$P(M(x) > kx) \leq P(N(\infty) > \lfloor x^\gamma \rfloor) + P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor} (H_i \wedge x) > kx \Big), \tag{13}$$

where $\lfloor z \rfloor$ is the largest integer smaller than or equal to z. Since the number of overshoots is geometrically distributed, the first term in (13) behaves like $\exp(-c\lfloor x^\gamma \rfloor)$ for some $c > 0$. Since this decays faster than any power tail for any $\gamma > 0$, it suffices to consider the second term.

Let $0 < q < \min\{1, p-1\}$. Since the tail of H_i is one degree heavier than that of the X_k (see Theorem 2.1 in Chapter VIII of Asmussen (2003)), we have $EH_i^q < EH_i^{p-1} < \infty$. Hence, $H_i^q - 2E[H_i^q]$ satisfies the assumption of Lemma 15. Take $\gamma \in (0, q)$. Since y^q is a concave function

in y, we have

$$\begin{aligned} P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor}(H_i \wedge x) > kx\Big) &= P\Big(\big[\sum_{i=1}^{\lfloor x^\gamma \rfloor}(H_i \wedge x)\big]^q > (kx)^q\Big) \\ &\le P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor}(H_i \wedge x)^q > (kx)^q\Big) \\ &= P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor}((H_i \wedge x)^q - 2E[H_i^q]) > (kx)^q - 2\lfloor x^\gamma \rfloor E[H_i^q]\Big). \end{aligned}$$

To apply Lemma 15, we need conditioned, and not truncated random variables. Choose an integer $l > \beta/(q-\gamma)$. Considering the event that at least l of the H_i are larger than x, and its complement, we find

$$\begin{aligned} &P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor}((H_i \wedge x)^q - 2E[H_i^q]) > (kx)^q - 2\lfloor x^\gamma \rfloor E[H_i^q]\Big) \\ &\le \binom{\lfloor x^\gamma \rfloor}{l} P(H_i > x)^l + P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor}(H_i^q - 2E[H_i^q]) > (kx)^q - 2\lfloor x^\gamma \rfloor E[H_i^q] \,\Big|\, \#\{i : H_i > x\} < l\Big) \\ &\le x^{\gamma l} P(H_i > x)^l + P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor - l}(H_i^q - 2E[H_i^q]) > (kx)^q - lx^q - 2\lfloor x^\gamma \rfloor E[H_i^q] \,\Big|\, H_i \le x\Big). \end{aligned} \tag{14}$$

We complete the proof by showing that both terms in (14) are $o(x^{-\beta})$.

Since $E[H_i^q] < \infty$, we know that $P(H_i > x) = o(x^{-q})$. Hence, since $0 < \gamma < q$, and $l > \beta/(q-\gamma)$, we have $x^{\gamma l} P(H_i > x)^l = o(x^{\gamma l} x^{-ql}) = o(x^{-\beta})$. Let $\overline{k} > 0$. Since $q > \gamma$, for k large enough, the second term in (14) is smaller than

$$P\Big(\sum_{i=1}^{\lfloor x^\gamma \rfloor - l}(H_i^q - 2E[H_i^q]) > \overline{k}\big(x^q - 2E[H_i^q]\big) \,\Big|\, H_i^q - 2E[H_i^q] \le x^q - 2E[H_i^q]\Big). \tag{15}$$

Applying Lemma 15 with a suitable choice of $\overline{k}$, there exist $c > 0$ and $\eta > \beta/q$ such that (15) is smaller than

$$c\big(x^q - 2E[H_i^q]\big)^{-\eta} \sim c\big(x^q\big)^{-\eta} = o(x^{-\beta}), \qquad x \to \infty.$$

This completes the proof. □

Consider now a $GI/GI/1$ queue with the same interarrival-time distribution as before, but with generic service time $B \wedge x$. Set $A_x(t) = \sum_{i=1}^{K(t)}(B_i \wedge x)$, with $K(t)$ the number of arrivals in $(0,t]$, so $A_x(t)$ is the work entering the queue in the time interval $(0,t]$. Furthermore, at the beginning of each busy period an initial setup time x is added. Let $\widetilde{L}_x^{*s}$ be the residual busy period after the arrival of a customer of size x. Then $\widetilde{L}_x^{*s}$ can be represented as follows:

$$\widetilde{L}_x^{*s} = \inf\{t : x + \widetilde{Q}_x^s + A_x(t) - t = 0\},$$

with $\widetilde{Q}_x^s$ the steady-state workload upon customer arrivals in this queue, including the effect of the initial set-up.

Furthermore, let $D_x^{\mathtt{P}}(t)$ be the stochastic processes of work under policy $\mathtt{P}$ that would have priority over an arriving job of size x at time t.

Lemma 17 *For* $\mathtt{P} = \mathtt{FB}$ *and all* $\mathtt{P} \in \mathtt{SMART}$*, we have*

$$V(x)^{\mathtt{P}} \leq_{st} \widetilde{L}_x^{*s}.$$

Proof The bound holds for FB, since the residual busy period bounds $V(x)^{\mathtt{FB}}$ if the setup time were not included.

To see that the residual busy period also bounds $V(x)^P$ for $\mathtt{P} \in \mathtt{SMART}$, note that the process $D_x^{\mathtt{P}}(t)$ consists of two types of busy periods: (i) busy periods started by a job of original size $> x$ that now has remaining size $\leq x$ and (ii) busy periods started by a job of original size $\leq x$. In both cases, the Bias Property prevents any job with remaining size $> x$ from receiving service during the busy period; thus only new arrivals of size $\leq x$ can contribute once the busy period is started. Since the initial job under $\mathtt{P} \in \mathtt{SMART}$ is necessarily smaller than the setup x of $\widetilde{L}_x^{*s}$, and the arrivals during the busy period are stochastically larger in $\widetilde{L}_x^{*s}$, the residual length of both of these busy periods is stochastically smaller than $\widetilde{L}_x^{*s}$. □

The following lemma implies that Condition 2 holds for FB and SMART .

Lemma 18 *For every* $\beta > 0$*, there exists a constant* k *such that*

$$P(V(x)^{\mathtt{P}} > kx) = o(x^{-\beta}), \qquad x \to \infty \tag{16}$$

for all $\mathtt{P} \in \mathtt{SMART}$ *and for* $\mathtt{P} = \mathtt{FB}$*. As a consequence, Condition 2 holds for* $\mathtt{P} = \mathtt{FB}$ *and* $\mathtt{P} \in \mathtt{SMART}$*.*

Proof Let $\mathtt{P} = \mathtt{FB}$ or $\mathtt{P} \in \mathtt{SMART}$. We will bound $V(x)^{\mathtt{P}}$ using the residual busy period $\widetilde{L}_x^{*s}$ as per Lemma 17. Furthermore, define

$$U_x^c \stackrel{\text{def}}{=} \sup_{t>0}[A_x(t) - ct + x] = x + \sup_{t>0}[A_x(t) - ct]. \tag{17}$$

Then $U_x^1 = \widetilde{Q}_x^s$. For $(1-\rho)/2 < \delta < 1/2$ and $k > 1/\delta$, we have by Lemma 17,

$$\begin{aligned}
P(V(x)^{\mathtt{P}} > kx) &\leq P(\widetilde{L}_x^{*s} > kx) \\
&\leq P(x + U_x^1 + A_x(kx) - kx > 0) \\
&\leq P(U_x^1 + A_x(kx) - (1-2\delta)kx + x > \delta kx) \\
&\leq P(U_x^1 > \delta kx/2) + P(A_x(kx) - (1-2\delta)kx + x > \delta kx/2) \\
&\leq P(U_x^{1-2\delta} > \delta kx/2) + P(\sup_{t>0}[A_x(t) - (1-2\delta)t + x] > \delta kx/2) \\
&= 2P(U_x^{1-2\delta} > \delta kx/2).
\end{aligned}$$

By taking $\overline{k} = k\delta/2$, and $c = 1 - 2\delta$, it suffices to show that there exists a $\overline{k} > 1$ such that

$$P(U_x^c > \overline{k}x) = P(U_x^c - x > (\overline{k} - 1)x) = o(x^{-\beta}). \tag{18}$$

We complete the proof by viewing $U_x^c - x$ in terms of a random walk. Since the supremum in (17) is attained at arrival instants, we may write

$$U_x^c - x = \sup_n \sum_{i=1}^{n} (B_i \wedge x) - cA_i \leq \sup_n \sum_{i=1}^{n} [(B_i - cA_i) \wedge x],$$

where A_i is the time between the $(i-1)$st arrival and the ith arrival. Since $E[B_i - (1-2\delta)A_i] < E[B_i - \rho A_i] = 0$ and $E[((B_i - (1-2\delta)A_i)^+)^p] \leq E[B_i^p] < \infty$, we may apply Lemma 16, and (18) follows.

To show that FB and P $\in$ SMART obey Condition 2, note that since $\overline{F} \in \mathcal{IR}$, there exists a $\beta > 0$ such that $x^{-\beta} = o(P(B > x))$. Take this β and choose k as in (16). Condition 2 now follows. $\square$

Combining Lemmas 14 and 18 guarantees that FB and all P $\in$ SMART obey Conditions 1 and 2. Theorem 2 follows from applying Theorem 13.

7 Stochastic bounds for M/GI/1

In this section we derive the stochastic bounds on the sojourn time of SMART policies given in Theorem 5. We limit ourselves to the M/GI/1 setting due to the following properties that depend on memoryless arrivals and are necessary in the proofs: PASTA (Poisson Arrivals See Time Averages) and the linearity of busy periods, i.e., $L(x+y) = L(x) + L(y)$. Three policies that will be particularly important to the analysis are SRPT, PSJF, and FB; therefore we begin by providing a useful characterization for the conditional sojourn time of these policies.

Let $D_x^{\mathtt{P}}(t)$ again be the stochastic processes of work under policy P that would have priority over an arriving job of size x at time t, and let $D_x^{\mathtt{P}}$ denote its stationary version. Note that under SMART policies, $D_x^{\mathtt{P}}$ may in general depend on the behavior of the system after x arrives. However, it was shown in Wierman, Harchol-Balter and Osogami (2005) that $D_x^{\mathtt{P}} \leq_{st} D_x^{\mathtt{SRPT}}$, and $D_x^{\mathtt{SRPT}}$ depends only on the state of the system at the arrival of the tagged job.

We start by describing the waiting time under SRPT. Under SRPT, the waiting time of a job of size x is distributed like a busy period with an initial customer of size $D_x^{\mathtt{SRPT}}$ and generic service time $BI_{[B<x]}$, i.e.,

$$W(x)^{\mathtt{SRPT}} \stackrel{d}{=} L_x(D_x^{\mathtt{SRPT}}). \tag{19}$$

Under PSJF , the service of a job is interrupted only by all jobs with original size smaller than x, the residence time is distributed like a busy period with an initial job of size x:

$$R(x)^{\mathtt{PSJF}} \stackrel{d}{=} L_x(x). \tag{20}$$

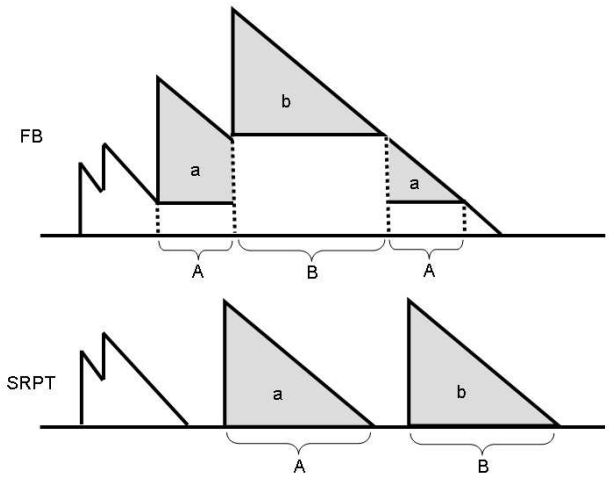

Figure 2: *This diagram illustrates the main idea in the proof of Proposition 19. The two figures show* $D_x^{\mathtt{P}}(t)$ *under* FB *and* SRPT*. The shaded jobs are large jobs. We have argued that the arrival processes are stochastically identical while large jobs are in the system, so none of these arrivals are drawn. In the proof, we pair up the times* T_a *when large job* a *is the most recent large arrival and times* T_b *when job* b *is the most recent large arrival. This figure illustrates that, since all large jobs arrive when* $D_x^P = 0$ *under* SRPT*,* $D_x^{\mathtt{SRPT}} \leq_{st} D_x^{\mathtt{FB}}$ *given that a large job is in the system.*

Under FB, the sojourn time of a job of size x is stochastically the same as the length of a busy period with generic service time $B \wedge x$ and an initial customer of size $x + D_x^{\mathtt{FB}}$:

$$V(x)^{\mathtt{FB}} \quad \stackrel{d}{=} \quad \widetilde{L}_x(x + D_x^{\mathtt{FB}}). \tag{21}$$

The following proposition is needed to prove the second inequality in Theorem 5.

Proposition 19 *In the M/G/1 queue, we have for all* x *and all* $\mathtt{P} \in$ SMART*,* $D_x^{\mathtt{P}} \leq_{st} D_x^{\mathtt{SRPT}} \leq_{st} D_x^{\mathtt{FB}}$*.*

Proof The first inequality was proven in Theorem 4.1 in Wierman, Harchol-Balter and Osogami (2005). To prove the second inequality, let us first describe $D_x^{\mathtt{FB}}(t)$ and $D_x^{\mathtt{SRPT}}(t)$. The process $D_x^{\mathtt{FB}}(t)$ is simple: it is the work(load) process of an M/GI/1 queue with service distribution $B \wedge x$ and arrival rate λ. The process $D_x^{\mathtt{SRPT}}(t)$ is more complex. Under SRPT, only arrivals of size smaller than or equal to x (call these small jobs) immediately add their size to $D_x^{\mathtt{SRPT}}(t)$. An arrival of size larger than x (call this a large job) will contribute x to $D_x^{\mathtt{SRPT}}(t)$ the moment its remaining size drops to x. Thus, small arrivals form a Poisson process with rate λ and service distribution $BI_{[B \leq x]}$, but large arrivals do not form a Poisson process. In fact, a large arrival can only add to $D_x^{\mathtt{SRPT}}(t)$ when $D_x^{\mathtt{SRPT}}(t) = 0$. While $D_x^{\mathtt{SRPT}}(t) > 0$, the only arrivals are small arrivals, so the arrival process during those periods is Poisson.

The processes $D_x^{\texttt{SRPT}}(t)$ and $D_x^{\texttt{FB}}(t)$ are both work-conserving, so we can view them as the work processes of Preemptive-Last-Come-First-Served (PLCFS) systems. Call these PLCFS systems the *transformed* FB *and* SRPT *systems*. Note that the two transformed systems have the same load: every large arrival to the original system will eventually contribute x work to both transformed systems, though (possibly) at different points in time.

Let us start by making a few observations about the sojourn times of large jobs in these two systems. In the transformed SRPT system, the sojourn time of a large job is stochastically equal to $L_x(x)$, while in the transformed FB system, it is distributed like $\widetilde{L_x}(x)$. However, by Lemma 20 below, the time that a large job is the most recent large arrival in the transformed FB system is distributed like $L_x(x)$ as well. This allows us to pair up the times that large jobs are most recent large arrival in the system, since in both transformed systems a large job will be the most recent large job in the system for stochastically identical lengths of time. This pairing is illustrated by Figure 2.

Using this pairing, we will first compare $D_x^{\texttt{FB}}$ and $D_x^{\texttt{SRPT}}$ conditional on the presence of a large job. In the transformed SRPT system, $D_x^{\texttt{SRPT}}(t) = 0$ at every time t when a large job arrives, whereas large jobs arrive to the transformed FB system at Poisson points, at which the transformed system is not necessarily idle. We can conclude that conditional on the presence of a large arrival in the transformed system, $D_x^{\texttt{SRPT}} \leq_{st} D_x^{\texttt{FB}}$.

Second, when no large arrival is in the system, the transformed FB and SRPT systems have stochastically identical work processes: only small jobs arrive, according to a Poisson process. Since the periods that large jobs are in the system are stochastically identical in length, the same holds for the periods that no large jobs are in system. Since both systems will achieve steady state, it must hold that $D_x^{\texttt{SRPT}} \leq_{st} D_x^{\texttt{FB}}$. Since we are only concerned with the work seen by a Poisson arrival x to a regenerative system that is convergent, we can apply PASTA, cf. Wolff (1989). This completes the proof. □

Lemma 20 *Let M denote the time that a tagged large job is the most recent large arrival present in the M/G/1/*PLCFS *queue. Then $M \stackrel{d}{=} L_x(x)$.*

Proof During the sojourn time of the tagged job of size x, the arrival of another job of size x just creates a sub-busy period that does not contribute. Since the arrival process is Poisson and all small arrivals contribute to M, we have $M \stackrel{d}{=} L_x(x)$. □

We are now ready to give the proof of Theorem 5.

Proof of Theorem 5 Theorem 4.1 in Wierman, Harchol-Balter and Osogami (2005) states that

$$V(x)^{\texttt{P}} \leq_{st} L_x(x + D_x^{\texttt{SRPT}}).$$

Since $L_x(Y) \leq_{st} \widetilde{L_x}(Y)$ for all random variables Y, Proposition 19 implies

$$V(x)^{\mathtt{P}} \leq_{st} L_x(x + D_x^{\mathtt{SRPT}}) \leq_{st} \widetilde{L_x}(x + D_x^{\mathtt{FB}}). \tag{22}$$

The proof of the theorem is completed by rewriting (22) using the linearity of busy periods and (19), (20), and (21). □

8 Conclusion

The SMART classification represents an emerging style of research based on analyzing large groups of policies instead of individual disciplines. Recent papers such as Núñez Queija (2002), Wierman and Harchol-Balter (2003, 2005), and Wierman, Harchol-Balter and Osogami (2005), have attempted to uncover the effect of general scheduling heuristics and mechanisms on performance, thus adding structure to the space of scheduling policies that cannot be obtained through the analysis of individual policies. Beyond the theoretical motivation for studying classifications of scheduling policies, there are also practical reasons. Namely, in practice, system designers can never implement the idealized policies (such as SRPT, PS, and FB) that are the focus of theoretical research. By analyzing classifications of policies, the hope is that theoretical results can be obtained for the unique, hybrid policies that are actually implemented in practice.

In this paper, we have analyzed the GI/GI/1 tail behavior of the sojourn time under both FB and SMART policies. We have proven that both FB and SMART policies have (near) optimal sojourn-time tails under heavy-tailed service distributions, and still outperform FCFS under light-tailed service distributions provided the service time of a customer is not too large. These analyses can be viewed as a formal verification that the heuristic of "biasing toward small jobs" is appropriate for many computer system applications, where service distributions tend to be heavy-tailed. Furthermore, we have derived stochastic bounds that relate the conditional sojourn times of FB and SMART policies in the M/GI/1 setting.

References

S. Aalto, U. Ayesta, and E. Nyberg-Oksanen. 2004. Two-level processor-sharing scheduling disciplines: Mean delay analysis. In *Proceedings of ACM Sigmetrics-Performance.*

S. Asmussen. 2003. *Applied Probability and Queues,* second edition. Springer, New York.

J. Abate and W. Whitt. 1997. Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. *Queueing Systems* **25**, 173–233.

P. Barford and M. Crovella. 1998. Generating representative web workloads for network and server performance evaluation. In *Proceedings of ACM Sigmetrics.*

S. Borst, O. Boxma, R. Núñez Queija, and B. Zwart. 2003. The impact of the service discipline on delay asymptotics. *Performance Evaluation* **54**: 175–206.

S. Borst, R. Núñez Queija, and B. Zwart. *Submitted*. Sojourn time asymptotics in Processor Sharing queues.

L. Cherkasova. 1998. Scheduling strategies to improve response time for web applications. In *High-performance computing and networking: international conference and exhibition*, 305–314.

A. Downey. 2001. Evidence for long-tailed distributions in the internet. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*.

F. Guillemin, Ph. Robert, and B. Zwart. 2004. Tail asymptotics for processor sharing queues. *Advances in Applied Probability* **36**, 525–543.

M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. 2003. Implementation of SRPT scheduling in web servers. *ACM Transactions on Computer Systems*, **21**(2).

P. Jelenkovic and P. Momcilovic. 2003. Large deviation analysis of subexponential waiting times in a Processor Sharing queue. *Mathematics of Operations Research* **28**, 587–608.

W. Leland, M. Taqqu, W. Willinger, and D. Wilson. 1993. On the self-similar nature of ethernet traffic. In *Proceedings of SIGCOMM '93*, 183–193.

M. Mandjes and M. Nuyens. 2005. Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the engineering and informational sciences* **19**(3): 351–361.

M. Mandjes and B. Zwart. (2004). Large deviations for waiting times in processor sharing queues. *Queueing Systems*, to appear.
Preprint version available at `http://www.cwi.nl/ftp/CWIreports/PNA/PNA-E0410.pdf`

D. McWherter, B. Schroeder, N. Ailamaki, and M. Harchol-Balter. 2004 Priority mechanisms for OLTP and transactional web applications. In *International Conference on Data Engineering*.

R. Núñez Queija. 2002. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research* **113**: 101–117.

M. Nuyens and B. Zwart. *Submitted*. A large-deviations analysis of the $GI/GI/1$ SRPT queue.
Preprint version available at `http://www.few.vu.nl/~mnuyens/publications/srpt.html`

M. Nuyens. 2004. *The Foreground-Background Queue*. PhD thesis, University of Amsterdam.

D. Peterson. 1996. Data center I/O patterns and power laws. In *CMG Proceedings*.

I. Rai, G. Urvoy-Keller, M. Vernon, and E. Biersack. 2004. Performance modeling of LAS based scheduling in packet switched networks. In *Proceedings of ACM Sigmetrics-Performance*.

M. Rawat and A. Kshemkalyani. 2003. SWIFT: Scheduling in web servers for fast response time. In *Symp. on Network Computing and App.*.

S. Resnick and G. Samorodnitsky. 1999. Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems* **33**, 43–71.

A. Wierman and M. Harchol-Balter. 2003. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM Sigmetrics*.

A. Wierman and M. Harchol-Balter. 2005 Classifying scheduling policies with respect to higher moments of conditional response time. In *Proceedings of ACM Sigmetrics*.

A. Wierman, M. Harchol-Balter, and T. Osogami. 2005. Nearly insensitive bounds for SMART scheduling. In *Proceedings of ACM Sigmetrics*.

R. Wolff. 1989 *Stochastic Modeling and the Theory of Queues*. Prentice Hall.

S. Yang and G. de Veciana. 2004. Enhancing both network and user performance for networks supporting best effort traffic. *Transactions on Networking* **12**: 349–360.