OPERATIONS RESEARCH
Vol. 60, No. 5, September–October 2012, pp. 1249–1257
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1086

Is Tail-Optimal Scheduling Possible?

Adam Wierman
Department of Computer Science, California Institute of Technology, Pasadena, California 91125, adamw@caltech.edu

Bert Zwart
VU University Amsterdam, Eurandom, Georgia Institute of Technology, and CWI Amsterdam, 1098 XG Amsterdam, The Netherlands, bertz@cwi.nl

This paper focuses on the competitive analysis of scheduling disciplines in a large deviations setting. Although there are policies that are known to optimize the sojourn time tail under a large class of heavy-tailed job sizes (e.g., processor sharing and shortest remaining processing time) and there are policies known to optimize the sojourn time tail in the case of light-tailed job sizes (e.g., first come first served), no policies are known that can optimize the sojourn time tail across both light- and heavy-tailed job size distributions. We prove that no such work-conserving, nonanticipatory, nonlearning policy exists, and thus that a policy must learn (or know) the job size distribution in order to optimize the sojourn time tail.

Subject classifications: scheduling; queueing; large deviations; competitive analysis.
Area of review: Stochastic Models.
History: Received August 2009; revisions received April 2011, February 2012; accepted April 2012. Published online in *Articles in Advance* October 9, 2012.

1. Introduction

The analysis and design of scheduling disciplines (a.k.a. policies) is a core area of operations research with a long history of both theoretical and applied research. From an applied perspective, scheduling policies are fundamental pieces of network designs, computer systems, manufacturing systems, etc., and understanding their performance analytically has been an important problem for decades. From the theoretical side, scheduling provides an important set of problems that can be attacked with a variety of techniques, including optimization and stochastics/queueing; see, for example, Harchol-Balter (2007), Pinedo (2008).

Many results in scheduling focus on either (i) worst-case competitive analysis of policies under arbitrary workloads, or (ii) average case behavior in a random environment. Both styles of analysis have been extremely fruitful. However, system designers are sometimes also interested in more than just good performance in expectation, and worst-case performance can be too pessimistic for design purposes. Often, an understanding of the distribution of performance measures such as sojourn time (a.k.a. response time, flow time) is crucial. But, unfortunately, exact distributional analysis of sojourn time is usually difficult.

Nevertheless, it is often possible to obtain information about the tail of the distribution of the sojourn time of a job in a random environment using asymptotic techniques, such as large deviations. Such analysis provides insights into both the frequency and nature of excessively large sojourn times, which is often the type of information system designers are looking for. Indeed, the large deviations analysis of scheduling policies has provided insight in many areas of computer system and network design where information about the tail is essential, such as buffer sizing (Wischik and McKeown 2005, Jelenkovic and Momcilovic 2003), effective bandwidths (Kelly 1996, Whitt 1993), and ruin probabilities (Asmussen 2000).

The large deviations analysis of scheduling policies has grown from the analysis of a few scheduling policies in simple models to the point where now the state-of-art provides analysis of almost all common scheduling policies under general arrival processes and large classes of both light- and heavy-tailed job sizes. For example, results exist for the GI/GI/1 queue under both heavy- and light-tailed job sizes for first come first served (FCFS) (Asmussen 2003, Borovkov 1976, Cohen 1973, Pakes 1975), preemptive last come first served (LCFS) (Meyer and Teugels 1980, Zwart 2001), processor sharing (PS) (Borst et al. 2006), shortest remaining processing time (SRPT) (Nuyens and Zwart 2006, Nuyens et al. 2008), and other disciplines. Complete surveys can be found in Borst et al. (2003) and Boxma and Zwart (2007).

The central focus of the current paper is on designing scheduling disciplines that are optimal in the context of the sojourn time tail, i.e., scheduling disciplines that prevent long sojourn times in an asymptotically optimal way. To this end, some optimality results exist in the literature. Ramanan and Stolyar (2001) have shown that the tail of sojourn time under FCFS is asymptotically optimal when job sizes are light-tailed, and this has been extended to end-to-end delays in networks by Stolyar (2003). On the other hand, the performance of FCFS is very poor if

job sizes are heavy-tailed, as observed, for example, in Anantharam (1999). In contrast, the tail of sojourn time under SRPT is asymptotically optimal when job sizes are heavy-tailed—specifically, regularly varying—but is very poor when job sizes are light-tailed; see, for example, Nuyens et al. (2008).

These as well as other results in the literature (cf. §2) reveal an interesting dichotomy: scheduling policies that perform well (in a large deviations sense) under heavy-tailed workloads perform poorly under light-tailed workloads, and vice versa. (Note that a similar dichotomy exists in a stochastic ordering sense; cf. Righter et al. 1990.) From this dichotomy has emerged an interesting fundamental question: *Does there exist a scheduling policy that is optimal for the sojourn time tail under all job size distributions*?

This question is the focus of the current paper. We state the question in more formal terms and then rigorously prove that the answer is "no" (Theorem 3). Specifically, we prove that any policy that cannot learn the job size distribution and is optimal for regularly varying job sizes is far from optimal under light-tailed job sizes. Thus, it is impossible (without learning or knowing the job size distribution) to schedule optimally under both heavy-tailed and light-tailed job sizes. In fact, our proof also illustrates that if a policy has an optimal sojourn time tail under light-tailed job sizes, then it has the heaviest possible sojourn time tail under regularly varying job sizes. This result highlights the fact that scheduling to optimize the sojourn time tail is fundamentally harder than scheduling to optimize the mean sojourn time, which can be done optimally using SRPT, regardless of the job size and interarrival time distributions (Schrage 1968).

The major insights offered by our analysis are necessary conditions for a scheduling discipline to be optimal for heavy tails and for light tails. For heavy tails, a necessary condition for a scheduling discipline to be optimal is to limit the impact that a single large job can have (cf. the "principle of a big jump" for the GI/GI/1 FCFS queue; see, for example, Zachary 2004). Specifically, it is necessary that the system remains rate-stable if a job of infinite size is added to the system. This implies that huge jobs cannot receive a long-term service rate that is larger than the "spare capacity" $1-\rho$, where ρ is the system load. This property is shown to be incompatible with the optimality requirements for light tails. Essentially, it implies that small jobs have priority over huge jobs, implying that a huge job needs to wait for a busy period of small jobs. Thus, any service discipline that is optimal for heavy tails essentially behaves like SRPT and LCFS for light tails, which are known to be nonoptimal.

The proof is actually based on these insights and first focuses on the case of heavy-tailed job sizes. We formalize the above intuition, leading to a necessary condition for a scheduling policy to have an optimal sojourn time tail. After that, an exponential change of measure argument is used to construct a light-tailed input process for which any scheduling policy that satisfies the necessary condition for heavy tails is suboptimal. The change of measure construction is a technically crucial part of the argument because it allows the avoidance of structural consistency assumptions about the scheduling policy across differing stochastic input processes.

The remainder of the paper is organized as follows. In §2, we formally introduce the model and notation of the paper. Additionally, we discuss the relevant prior literature on large deviations and scheduling and provide a framework for the competitive analysis of scheduling disciplines in a large deviations setting. Then in §3 we present and prove the main result of the paper. Finally, in §4 we conclude with a discussion of some interesting new directions motivated by the impossibility result in this paper.

2. Preliminaries

In this section we will (i) introduce the model and class of scheduling policies we consider, (ii) define a competitive analysis framework for studying optimality in a large deviations setting, and (iii) survey background large deviations results about common scheduling policies.

2.1. Scheduling Policies

In this paper, we analyze scheduling policies for the GI/GI/1 queue, i.e., the single server queue with renewal arrivals and i.i.d. service times, and we use V_π to denote the stationary sojourn time under policy π. We focus on policies π that satisfy the following three conditions:

(i) π is *work-conserving*: the scheduling policy always has the server working at speed 1 whenever work is present in the system.

(ii) π is *nonanticipative*: a scheduling decision at time t does not depend on information about customers that arrive beyond time t. (We do allow the scheduler to use the sizes of jobs on and after arrival.)

(iii) π is *nonlearning*: the scheduling decisions cannot depend on information about previous busy periods. That is, a scheduling decision on a sample path cannot change when the history before the current busy period is changed.

The first two assumptions are standard and allow a policy to exploit detailed information, such as past and/or remaining service requirements of individual jobs. The third condition is formulated in such a way that a scheduling discipline cannot be driven by data from the (distant) past. It is nonstandard but is satisfied by all common policies and even many adaptive policies, such as the one in Jelenkovic et al. (2007). The third condition is important because it creates a setting in which the scheduler is not aware of the job size distribution.

Our assumptions are identical to those in Ramanan and Stolyar (2001), who studied the optimality of FCFS and are satisfied by all common policies, including FCFS, LCFS, SRPT, PS, and many others. Section 6 of that reference

also contains a mathematically rigorous definition of the class of scheduling disciplines. We sketch the framework here for convenience. Condition (i) is easy and makes the workload process $W(\cdot)$ independent of the scheduling discipline. The idea is now to define the state process $\mathscr{W}(\cdot)$ as follows. Let $b(t)$ be the time elapsed since the beginning of the busy period active at t, then $\mathscr{W}(t) = \{W((b(t)+s)\wedge t), s \geqslant 0\}$. In words: the state at time t is the entire history of the workload back from time t to the beginning of the busy period. Then π is defined sample pathwise: the scheduling discipline at time t is a function of $\mathscr{W}(t)$ and therefore does not depend on anything else, e.g., not on t (which rules out the construction of scheduling disciplines that depend on functions of t). Ramanan and Stolyar do not assume the input is of $GI/GI/\cdot$ type and take care in constructing a stationary version of $\mathscr{W}(\cdot)$. In our case, $W(\cdot)$ is a regenerative process, with the beginning of busy periods being regeneration points. Thus, also $\mathscr{W}(\cdot)$ is regenerative. Let $V_{\pi,i}$ be the sojourn time of the ith customer under scheduling discipline π. Assume that the system is empty when customer 1 arrives. The construction sketched above implies that $V_{\pi,i}$, $i \geqslant 1$ is a regenerative process, and the steady-state sojourn time V_π satisfies

$$P(V_\pi > t) = \frac{1}{E[N]} E\left[\sum_{i=1}^{N} I(V_{\pi,i} > t)\right], \tag{1}$$

where N is the number of customers in a busy period and $I(G)$ is the indicator function of the event G; see, for example, of Asmussen (2003, p. 171, (1.5)).

We additionally introduce some notation: denote a generic job size by B and its mean by β, a generic interarrival time by A, the arrival rate by λ, and the load by $\rho = \lambda\beta < 1$. Importantly, under these conditions V_π is a.s. finite.

2.2. Tail Optimality

The major focus of the paper is how to choose π such that the sojourn time tail $P(V_\pi > t)$ converges to 0 as fast as possible as $t \to \infty$. That is, we are interested in scheduling disciplines that avoid long sojourn times in an optimal way. Motivated by this, we define a notion of optimality of scheduling policies with respect to the sojourn time tail.

Definition 1. A scheduling discipline π_0 is *weakly tail-competitive* for a class $\mathscr{P}$ of interarrival time distributions and job size distributions, if

$$\limsup_{t\to\infty} \frac{P(V_{\pi_0} > t)^{1+\tau}}{P(V_\pi > t)} < \infty \tag{2}$$

holds for every $\tau > 0$, every $P \in \mathscr{P}$ and every work-conserving, nonanticipative, nonlearning scheduling policy π. π_0 is called *tail-competitive* if the same property holds for $\tau = 0$, and *strongly tail-competitive* if additionally the lim sup is bounded by 1 for $\tau = 0$.

A related definition is proposed in Boxma and Zwart (2007). We would like to point out that the notion of optimality we propose strikes a balance between the average-case behavior and worst-case behavior of scheduling algorithms; these two notions are more prevalent in the scheduling literature. For another optimality notion of scheduling policies, see Koutsoupias and Papadimitriou (2000).

Insight into the optimality of scheduling disciplines can be obtained from the following two simple lower bounds, which are independent of the scheduling discipline:

$$P(V_\pi > t) \geqslant P(B > t), \tag{3}$$

$$P(V_\pi > t) \geqslant \frac{1}{E[N]} P(C_{\max} > t). \tag{4}$$

Here $C_{\max}$ is the maximum amount of work in the system during a busy cycle. The first bound is trivial, and the second bound simply follows from (1); see Boxma and Zwart (2007) for details. An approach for proving optimality of π_0 is to analyze the tail behavior of V_{π_0} and then to compare it with the tail behavior of $C_{\max}$ or B. We now review existing results on the tail behavior of $P(V_{\pi_0} > t)$ for several choices of π_0.

2.3. Review of Results for Specific Scheduling Disciplines

There is a wide array of scheduling policies that have been studied in the literature. A comprehensive survey of the sojourn time tail behavior of various scheduling disciplines can be found in Borst et al. (2003) and Boxma and Zwart (2007). To keep the paper self-contained, we now present some of the results that are crucial to the goal of the paper.

We focus on two specific classes of job size distributions: light-tailed and heavy-tailed. We say that a job size B is *light-tailed* if $\Phi(\theta) = E[\exp\{\theta B\}] < \infty$ for some $\theta > 0$. For *heavy-tailed* job sizes, we consider the class of *regularly varying distributions*, which have $P(B > t) = L(t)t^{-\alpha}$ where L is a slowly varying function (i.e., $L(ax)/L(x) \to 1$ as $x \to \infty$ for every $a > 0$) and $\alpha > 1$ is a constant. Regularly varying distributions are a generalization of Pareto job sizes; see Bingham et al. (1987) for background.

Light tails. We focus on FCFS and (preemptive) LCFS. For FCFS, we write $V_\pi = V_F$ and for LCFS we set $V_\pi = V_L$. Let Φ_A be the MGF of A and set $\Psi(\theta) = -\Phi_A^{-1}(1/\Phi(\theta))$. (Note that $\Psi(\theta) = \lambda(\Phi(\theta) - 1)$ if the interarrival time distribution is exponential with rate λ.) $\Psi(\theta)$ is strictly convex if either A or B is nondeterministic. Now, we can state the large deviations results for FCFS and LCFS:

$$\lim_{t\to\infty} \frac{-\log P(V_F > t)}{t} = \gamma_F := \sup\{\theta\colon \Psi(\theta) - \theta \leqslant 0\}, \tag{5}$$

$$\lim_{t\to\infty} \frac{-\log P(V_L > t)}{t} = \gamma_L := \sup_{\theta\geqslant 0}\{\theta - \Psi(\theta)\}. \tag{6}$$

These theorems hold without any regularity conditions on Ψ, as is shown, for example, in Nuyens and Zwart (2006);

see also Asmussen (2003), Glynn and Whitt (1994), Palmowksi and Rolski (2006). From the strict convexity of $\Psi(\theta) - \theta$, and the fact that $\Psi'(0) = \rho$, it follows that

$$\gamma_L < (1-\rho)\gamma_F. \tag{7}$$

This inequality shows that for light tails, FCFS is better at preventing large sojourn times than LCFS. Indeed, Ramanan and Stolyar (2001) have shown that FCFS maximizes the decay rate, assuming that the input process satisfies a sample path large deviations principle. In our setting, this implies weak optimality. Optimality of FCFS can be guaranteed if Cramérs condition holds, i.e., if $\Phi_A(-\gamma_F)\Phi(\gamma_F) = 1$ and $\Phi'(\gamma_F) < \infty$. In this case, it is known that $P(C_{\max} > t) \sim KP(V_F > t)$ for a constant K (cf. Iglehart 1972). Combining this with (4) it follows that if Cramérs condition is satisfied, then

$$\limsup_{t\to\infty} \frac{P(V_F > t)}{P(V_\pi > t)} < \infty \tag{8}$$

for any scheduling discipline π; cf. Boxma and Zwart (2007).

In contrast to the optimality of γ_F, the decay rate γ_L is the smallest possible decay rate. To see this, note that V_π is by definition stochastically smaller than the total time to emptiness when starting from steady state, just after an arrival (i.e., a residual busy period). The decay rate of this random variable was shown to be γ_L in Nuyens et al. (2008).

Interestingly, many other common policies have been shown to have decay rate equal to γ_L. In particular, SRPT (Nuyens and Zwart 2006), PS (Mandjes and Zwart 2006), FB (Mandjes and Nuyens 2005, Nuyens and Wierman 2008), and more generally, all SMART policies (Wierman and Nuyens 2008) have a decay rate that coincides with γ_L under some mild regularity conditions. The intuition behind all these policies is that a large sojourn time is caused by a large service requirement. In addition, the corresponding customer will leave the system after a long busy period of small customers; see, for example, the proof in Nuyens et al. (2008).

Heavy tails. Under regularly varying job sizes and general interarrival times, the following results hold:

$$P(V_F > x) \sim \frac{\rho}{1-\rho}\frac{1}{\alpha-1}tP(B > t), \tag{9}$$

$$P(V_L > t) \sim E[N]P(B > t(1-\rho)), \tag{10}$$

$$P(V_{PS} > t) \sim P(V_{\text{SRPT}} > t) \sim P(B > t(1-\rho)), \tag{11}$$

where $f(x) \sim g(x)$ denotes $\lim_{x\to\infty} f(x)/g(x) = 1$. For FCFS, we refer to Borovkov (1976), Cohen (1973), and Pakes (1975). The tail behavior for LCFS was shown for Poisson arrivals by Meyer and Teugels (1980) and for renewal arrivals by Zwart (2001). The tail behavior of PS has been reviewed in Borst et al. (2006). For SRPT see Nuyens et al. (2008).

There are two important observations about these results that we would like to highlight. First, because $P(B > t(1-\rho)) \sim (1-\rho)^{-\alpha}P(B > t)$, PS, SRPT, and PLCFS are within a constant of optimal. Second, notice that FCFS has a sojourn time tail that is one degree heavier than optimal. In fact, the sojourn time tail of FCFS is as heavy as possible, up to a constant factor. The same holds for all other nonpreemptive policies, as is shown by Anantharam (1999). The reason is that under any nonpreemptive policy, a job of size x will cause of the order x other customers to wait for a long time. This quickly leads to a lower bound of the order $xP(B > x)$, using (1).

3. Main Result

The previous section reveals a clear dichotomy between the scheduling policies that perform well under light- and heavy-tailed job size distributions. FCFS is weakly tail-competitive under light-tailed job sizes but is far from optimal under heavy-tailed job sizes, whereas the opposite is true for LCFS, SRPT, and PS. This motivates the question: does there exist a scheduling policy that is weakly tail-competitive across all job size distributions? The main contribution of this paper is to prove that the answer is "no."

THEOREM 1. *There does not exist a work conserving, nonanticipative, and nonlearning scheduling policy π that is weakly tail-competitive for any $\mathcal{P}$ that contains all P having a job size distribution that is either light-tailed or regularly varying with $\alpha > 2$, and all exponential interarrival time distributions.*

The remainder of this section proves this result, which follows from Propositions 1 and 2 below. In particular, we construct two counterexamples, and for this it suffices to assume that interarrival times are exponentially distributed. Thus, throughout the analysis, we consider an $M/G/1$ queue. The final condition in the theorem could be weakened by generalizing all proofs below to general interarrival times. Although this is feasible, we decided not to pursue this to keep the proofs relatively clean.

The structure of the proof, and the remainder of the section, is as follows.

We first focus on the case of heavy-tailed job sizes. We derive a necessary condition for a scheduling policy to be tail-competitive; see (12) below. This condition is a formalization of the property that a scheduling discipline needs to be stable in the presence of an infinite-sized job.

After that, we construct a probability measure under which the job sizes are light-tailed using an exponential change of measure starting from the probability measure corresponding to the system with heavy-tailed job sizes. This construction is crucial, because (12) is proven to be necessary only for probability measures under which job sizes are heavy-tailed. Using this construction, we show that (12) implies non-tail-competitiveness for light-tailed service times. Our proof reveals the insight that optimality for light tails requires large jobs to have a sufficiently large service rate during their sojourn.

3.1. Heavy Tails: A Necessary Condition

Throughout, we assume that a job, number job 1, enters an empty system at time 0. The size of this job will be denoted by B_1. For a given scheduling discipline and $t > 0$, we define $R(t)$ to be the service allocated in $[0, t]$ to all jobs arriving in the system after time 0. Observe that $\limsup_{t\to\infty} R(t)/t \leqslant \rho$ a.s. The first major step is to show that

$$\forall\, y > 1,\ \forall\, \delta > 0,$$
$$\lim_{t\to\infty} P(R(t) \geqslant (\rho-\delta)t \mid B_1 > (1-\rho)yt) = 1 \tag{12}$$

is a necessary condition for optimality if job sizes are heavy-tailed.

This condition serves as a formalization of the statement "the scheduling policy must guarantee that the system remains stable in the presence of an infinite-sized job." This informal statement was provided as an intuition for the sojourn time tail of, for example, SRPT and PS (cf. Borst et al. 2006, Boxma and Zwart 2007, Nuyens et al. 2008), and we show here that it can be formalized and proven to be a necessary condition for a scheduler to be weakly tail-competitive. Intuitively, the reason this is a necessary condition stems from the so-called "principle of a single big jump" (see, for example, Zachary 2004) for heavy-tailed distributions, which states that the most likely rare event for heavy-tailed distributions is the arrival of a very large job. Thus, for a scheduling policy to do well, it must isolate the impact of the arrival of a single large job.

To prove that (12) is a necessary condition, we construct a counterexample. To do this, we fix a scheduling policy and choose P such that B is regularly varying with index $\alpha > 2$ and $\rho < 1$. Additionally, suppose that (12) does not hold, i.e., there exist constants $y > 1$, $\delta > 0$, $\gamma > 0$, and a sequence (t_n), $t_n \to \infty$ such that

$$P(R(t_n) \leqslant (\rho-\delta)t_n \mid B_1 > y(1-\rho)t_n) > \gamma, \quad n \geqslant 1. \tag{13}$$

We are now ready to state our first proposition.

Proposition 1. *Consider an $M/G/1$ queue operating under P such that B is regularly varying with index $\alpha > 2$ and $\rho < 1$. Let π be a discipline satisfying* (13) *for some $y > 1$, $\delta > 0$, $\gamma > 0$, and a sequence (t_n), $t_n \to \infty$. Then*

$$\liminf_{n\to\infty} \frac{P(V_\pi > t_n\delta/4)}{\sqrt{t_n}P(B > t_n)} > 0. \tag{14}$$

Thus, if π does not satisfy (12) *under P, then π is not weakly tail-competitive under P.*

An immediate consequence of Proposition 1 is that for π to be weakly tail-competitive for heavy-tailed job sizes, (12) must hold for all P such that B is regularly varying with index $\alpha > 2$ and $\rho < 1$. Note that the proof below uses only the fact that some $\alpha > 2$ moments of the job size exist.

Proof. Let $N(t)$ be the number of arrivals in $(0, t]$. Fix $\eta > 0$ and let E_n be the event that $N(t_n) \in ((\lambda-\eta)t_n, (\lambda+\eta)t_n)$ and that all these arrivals have service requirements smaller than $\sqrt{t_n\delta/4}$. We can lower bound $P(E_n)$ as follows:

$$P(E_n) \geqslant P\big(N(t_n) \in \big((\lambda-\eta)t_n, (\lambda+\eta)t_n\big)\big) \cdot P\big(B \leqslant \sqrt{t_n\delta/4}\big)^{\lceil(\lambda+\eta)t_n\rceil}.$$

Notice that the first probability on the right-hand side converges to 1 as $n \to \infty$ in view of the law of large numbers for Poisson processes. The second probability on the righthand side also converges to 1 as $n \to \infty$, because the assumption that $\alpha > 2$ implies $P(B \leqslant \sqrt{t_n\delta/4}) = 1 - o(1/t_n)$ as $n \to \infty$. Thus, $P(E_n) \to 1$.

Next, define $Y(t)$ as the amount of work offered to the system in $(0, t]$. An immediate consequence of $P(E_n) \to 1$ is that $Y(t_n)/t_n \to \rho$ in $P(\cdot \mid E_n)$ probability as $n \to \infty$.

Now, let $F_n = \{R(t_n) \leqslant (\rho-\delta)t_n\}$ and define $G_n = E_n \cap F_n \cap \{Y(t_n) > (\rho-\delta/2)t_n\}$. From (13) and our analysis of $P(E_n)$ and $Y(t_n)$ above, we have that

$$P(G_n \mid B_1 > y(1-\rho)t_n) \geqslant \gamma/2$$

for n sufficiently large.

For the remainder of the proof we focus on what happens at time t_n and after. Let $W(t_n)$ be the amount of work in the queue at time t_n made up by the customers that arrived after time 0 and before time t_n. By the construction of G_n, we have that

$$W(t_n) \geqslant (\delta/2)t_n.$$

Because the remaining service requirement of each customer is at most $\sqrt{t_n\delta/4}$ under even E_n, this also gives a bound on $Q(t_n)$, which denotes the number of jobs that arrived after time 0 and are still in the queue at time t_n. In particular, we have that

$$Q(t_n) \geqslant (\delta/2)t_n/\sqrt{\delta t_n/4} = \sqrt{\delta t_n}.$$

To complete the proof, we combine the above bounds on $W(t_n)$ and $Q(t_n)$ with Equation (1). Specifically, using (1) we obtain

$$P\big(V_\pi > (\delta/4)t_n\big) \geqslant \frac{1}{E[N]} E\left[I(G_n) I\big(B_1 > y(1-\rho)t_n\big) \sum_{i=1}^{N} I\big(V_{\pi,i} > (\delta/4)t_n\big)\right].$$

The derived bounds on $W(t_n)$ and $Q(t_n)$ imply that the last expression is equal to

$$\frac{1}{E[N]} E\left[I(G_n) I\big(W(t_n) > (\delta/2)t_n\big) I\Big(Q(t_n) \geqslant \sqrt{\delta t_n}\Big) \cdot I\big(B_1 > y(1-\rho)t_n\big) \sum_{i=1}^{N} I\big(V_{\pi,i} > (\delta/4)t_n\big)\right]. \tag{15}$$

Consider now the evolution of the workload process between time t_n and $t_n(1+\delta/4)$. The work that is present in the system at both of these times amounts to a total mass of at least $\delta t_n/2 - \delta t_n/4 = \delta t_n/4$. The number of different customers corresponding to this work is at least $\sqrt{\delta t_n/4}$, and each of these customers stayed in the system at least $\delta t_n/4$ time units. We therefore conclude that $\sum_{i=1}^{N} I(V_{\pi,i} > (\delta/4)t_n) \geqslant \sqrt{\delta t_n/4}$ given the conditional events in (15). Thus, we have, for large enough n,

$$P\big(V_\pi > (\delta/4)t_n\big) \geqslant \frac{1}{E[N]}\sqrt{\delta t_n/4}(\gamma/2)P\big(B_1 > y(1-\rho)t_n\big), \tag{16}$$

which completes the proof. □

An interesting observation about the above proof is that the $\sqrt{t_n}$ is not special. Imposing $B_i \leqslant t_n^{1-z}$ with $z \in (0,1)$ and $\alpha > 1/(1-z)$ generalizes the proposition to state that

$$\liminf_{n\to\infty} \frac{P(V_\pi > t_n\delta/4)}{t_n^z P(B > t_n)} > 0$$

by using the same argument. Taking z arbitrarily close to 1 provides the interesting interpretation that if the necessary condition (12) does not hold, then the tail is arbitrarily close to the tail of FCFS, which is the heaviest tail possible, up to a constant, for any work-conserving policy.

Corollary 1. *Consider an $M/G/1$ queue. Let π be a discipline satisfying* (13) *for all P such that B is regularly varying with index $\alpha > 2$ and $\rho < 1$. Then, for all $\tau > 0$ there exists a P such that*

$$\limsup_{t\to\infty} \frac{P(V_F > t)^{1+\tau}}{P(V_\pi > t)} < \infty.$$

3.2. Light-Tails: Non-Tail-Competitiveness

Given the necessary condition for a scheduling policy to be tail-competitive under regularly varying job sizes, we now construct a probability measure using an exponential change of measure under which the job size distribution is light-tailed starting from the measure corresponding to the regularly varying job size distribution. We then show that (12) implies non-tail-competitiveness in the light-tailed example we construct. This change of measure argument is necessary because (12) is shown to be necessary only for heavy-tailed job sizes.

Thus, we construct a specific probability measure under which service times are light-tailed. *To help distinguish the light- and heavy-tailed examples, from this point forward we add tildes when referring to the setting in which service times are regularly varying.* Specifically, let B^* be a service time distribution that is regularly varying with index $\alpha > 2$. Let β^* be its mean, and let λ^* be an arrival rate such that $\lambda^*\beta^* = 1-\epsilon$ for some $\epsilon \in (0, 1/4)$. Note that in the heavy-tailed example, any value of the load was allowed, so this is not a restriction.

To construct the arrival rate and service time distribution in the light-tailed case, we proceed as follows. Let $\Phi^*(\theta)$ be the MGF of B^*. Note that this MGF is finite if and only if $\theta \leqslant 0$. Next, consider an arrival rate λ_s and a random variable B_s parameterized by $s \in (0, \tilde{\lambda})$ as follows. Define $\lambda_s = \lambda^* - s$ and B_s by the MGF $\Phi^*(\theta - s)/\Phi^*(-s)$. Its mean, β_s, is given by $\Phi^{*\prime}(-s)/\Phi^*(-s)$. The corresponding load is $\rho_s = \lambda_s\beta_s$. Because $\log\Phi(s)$ is strictly convex and continuously differentiable on $(-\infty, 0)$ (see Ganesh et al. 2004, p. 28), ρ_s is continuous and strictly decreasing in s. In addition, $\rho_s \to 0$ as $s \uparrow \lambda^*$. Now, pick s^* such that $\rho_{s^*} \in (\epsilon+\epsilon^2, 1-\epsilon-\epsilon^2)$, and define $\lambda = \lambda_{s^*}$, and $B = B_{s^*}$. Let Φ be the MGF of B, and note that $\Phi(\theta) = \Phi^*(\theta - s^*)/\Phi^*(-s^*)$.

From the construction of λ and B we have the following properties of γ_F and γ_L. Recall these are the fastest and slowest decay rates achievable.

Lemma 1. *Given the construction of λ and B above, we have $\gamma_F = s^*$.*

Proof. For the M/G/1 queue, (5) specializes to

$$\gamma_F = \sup\{\theta\colon \lambda(\Phi(\theta) - 1) \leqslant \theta\}. \tag{17}$$

Because $\Phi(\theta) = \infty$, if $\theta > s^*$, then $\gamma_F \leqslant s^*$. Next, observe that by convexity $(1 - \Phi^*(-s^*))/s^* \leqslant \beta^*$, which implies

$$\lambda^* \frac{1 - \Phi^*(-s^*)}{s^*} \leqslant \rho^* \leqslant 1, \tag{18}$$

and

$$\frac{\lambda}{\lambda + s^*}\Phi(s^*) = \frac{\lambda^* - s^*}{\lambda^*\Phi^*(-s^*)} \leqslant 1, \tag{19}$$

where the last inequality is equivalent to (18). Returning to (17), we complete the proof as follows:

$$\begin{aligned}\lambda(\Phi(s^*) - 1) &= (\lambda^* - s^*)\left(\frac{1}{\Phi^*(-s^*)} - 1\right)\\ &\leqslant \lambda^*(1 - \Phi^*(-s^*))\\ &\leqslant s^*,\end{aligned}$$

where the second line follows from (19) and the third line from (18). Thus, $\gamma_F \geqslant s^*$. □

Lemma 2. *Given the construction of λ and B, we have $\gamma_L = \gamma_F - \Psi(\gamma_F)$.*

Proof. To prove this lemma, we show that s^* is the optimizing value of the program that determines γ_L. The key observation behind this is that (i) the left derivative $\Psi'(s)$ satisfies $\Psi'(s) \leqslant \rho^* < 1$, for $s \leqslant s^*$, and (ii) $\Psi(s) = \infty$ for $s > s^*$. The second observation is trivial, while the first follows from (19) and

$$\Psi'(s) = \lambda\Phi'(s) \leqslant \lambda\Phi'(s^*) = \frac{\lambda^* - s^*}{\lambda^*\Phi^*(-s^*)}\rho^* \leqslant 1. \quad \square$$

To state and prove the main result of this section we need a change of measure argument which is standard (see, for example, Mandjes and Zwart 2006, where it is spelled out for the GI/GI/1 PS queue), but detailed here to save the reader some work. (We will, however, be brief and refer to Chapter XIII of Asmussen 2003 for more background.) Let $N(t)$ be the number of arrivals in $(0, t]$, and let $B_i, i \geqslant 1$ be the sizes of the arriving jobs from $(0, t]$. Set $X(t) = \sum_{i=1}^{N(t)} B_i - t$. Let for $\mathcal{F}_t$ be the σ-field generated by $X(t)$ as well as B_0 (the job size of the job entering the system at time 0) and any other information available at time 0 (i.e., all information available up to time t), and define $M(t) = \exp\{\gamma_F X(t) + \gamma_L t\}$. $M(t), t \geqslant 0$ is a mean one martingale w.r.t. $\mathcal{F}_t, t \geqslant 0$ (actually, it is an example of the so-called Wald martingale); the mean one property follows from the identity $E[e^{\theta X(t)}] = e^{t(\Psi(\theta)-\theta)}$ and the above lemmas. This yields the martingale property, together with the fact that $X(t), t \geqslant 0$ has stationary independent increments.

This allows us to define a probability measure $\tilde{P}$ as follows. Given an event $A \in \mathcal{F}_t$, $\tilde{P}(A) = E[M(t)I(A)]$. It can be verified that the job sizes $B_i, i \geqslant 1$ have the same distribution as B^* under $\tilde{P}$, and $N(t), t \geqslant 0$ is a Poisson process with rate λ^* under $\tilde{P}$; see, for example, Asmussen (2003, Exercise XIII 3.5). Our construction does not change the distribution of B_0 when changing from P to $\tilde{P}$.

We are now ready to state our second proposition.

Proposition 2. *Consider a scheduling discipline π that satisfies* (12) *under $\tilde{P}$ such that $\tilde{B}$ is regularly varying with $\alpha > 2$. Then there exists a $\tau > 0$ and a P such that B is light-tailed, for which*

$$\liminf_{t\to\infty} \frac{P(V_\pi > t)^{1+\tau}}{P(V_F > t)} = \infty.$$

Thus, if π satisfies (12) *under $\tilde{P}$, then π is not weakly tail-competitive under P.*

Proof. To begin, note that by supposition, (12) holds for all $y > 1$ and $\delta > 0$, so we can fix $y = 1 + \epsilon$ and $\delta = \epsilon^2$ for $\epsilon \in (0, 1/4)$. Additionally, we consider P corresponding to a light-tailed M/GI/1 queue with B and λ chosen as defined earlier in this section.

Recall $X(t)$ to be the total amount of work that arrives in $(0, t]$ minus t. The event $\{V_\pi > t\} \in \mathcal{F}_t$. By Asmussen's (2003, Theorem XIII.3.2) we obtain

$$\begin{aligned} P(V_\pi > t) &= \tilde{E}[M(t)^{-1} I(V_\pi > t)] \\ &= e^{-\gamma_L t} \tilde{E}[e^{-\gamma_F X(t)} I(V_\pi > t)]. \end{aligned} \tag{20}$$

To proceed, we lower bound the right-hand side of (20). To obtain a lower bound we require that the tagged job enters an empty system under $\tilde{P}$ (which probability equals ϵ under $\tilde{P}$), and we add indicator functions of some other events. In particular, we require the large sojourn time happening as the result of an arrival at time 0 of size $> \epsilon(1+\epsilon)t$ followed by $1 - \epsilon - \epsilon^2$ service being given to other arrivals in $(0, t]$. Note that the latter two events imply $\{V_\pi > t\}$, so that we obtain

$$\begin{aligned} P(V_\pi > t) \geqslant e^{-\gamma_L t} \epsilon \tilde{E}\big[& e^{-\gamma_F X(t)} I(X(t) < 0) I(R(t)/t \\ & \geqslant 1 - \epsilon - \epsilon^2) I(B_0 > \epsilon(1+\epsilon)t)\big]. \end{aligned} \tag{21}$$

We proceed by further lower bounding the right-hand side of (21). First observe that for large enough t, $X(t)/t < 0$ a.s. under $\tilde{P}$, so $I(X(t) < 0) \to 1$ a.s. We can lower bound $e^{-\gamma_F X(t)}$ by 1 when $X(t) < 0$ because $\gamma_F > 0$. Next, we bound the remaining terms on the right-hand side of the previous equation using Condition (12), which is assumed to hold for π under $\tilde{P}$. Specifically, recall that $y = 1 + \epsilon$, $\delta = \epsilon^2$, $\tilde{\rho} = 1 - \epsilon$. This combined with the fact that

$$P(B_0 > (\epsilon + \epsilon^2)t) = e^{-s^*(\epsilon+\epsilon^2)t(1+o(1))} = e^{-\gamma_F(\epsilon+\epsilon^2)t(1+o(1))}$$

yields that, for large enough t,

$$P(V_\pi > t)^{1+\tau} \geqslant (1 + o(1))\epsilon e^{-(\gamma_L + (\epsilon+\epsilon^2)\gamma_F)t(1+o(1))(1+\tau)}. \tag{22}$$

Finally, we can apply the above bound to understand $\liminf_{t\to\infty} P(V_\pi > t)^{1+\tau}/P(V_F > t)$. Note that combining (7) with the fact that $\rho > \epsilon + \epsilon^2$ gives

$$\gamma_L + (\epsilon + \epsilon^2)\gamma_F < (1 - \rho)\gamma_F + (\epsilon + \epsilon^2)\gamma_F < \gamma_F.$$

Because the inequalities above are strict, it follows that there exists a $\tau > 0$ such that $(\gamma_L + (\epsilon+\epsilon^2)\gamma_F)(1+\tau) < \gamma_F$, which completes the proof. □

An interesting observation about the above proof is that the logarithmic decay rate of any policy that satisfies (12) can be made arbitrarily close to the slowest possible decay rate (that of LCFS) because γ_F and γ_L both converge to strictly positive constants as $\epsilon \to 0$. Thus, if a policy is weakly tail-competitive in the case of regularly varying job sizes, it follows that it has (nearly) the heaviest possible tail in the case of light-tailed job sizes.

Corollary 2. *Consider a scheduling discipline π that satisfies* (12) *under $\tilde{P}$ such that $\tilde{B}$ is regularly varying with $\alpha > 2$. Then for all $\tau > 0$ there exists a P such that B is light-tailed under which*

$$\limsup_{t\to\infty} \frac{P(V_L > t)^{1+\tau}}{P(V_\pi > t)} < \infty.$$

Finally, note that if the policy is tail-competitive in the case of light-tailed job sizes, then we can conclude that the necessary Condition (12) does not hold for any regularly varying job size distributions with $\alpha > 2$. Thus, Corollary 1 implies that the policy has (nearly) the heaviest possible tail in the case of regularly varying job sizes.

4. Concluding Remarks

The main result of this paper is that it is impossible for a scheduling policy to be weakly tail-competitive for both light- and heavy-tailed job sizes. Our analysis shows that to be optimal for heavy tails, one has to make sure that small jobs can pass long jobs. However, this causes large jobs to wait for a busy period of small jobs, which yields nonoptimality for light tails. Moreover, if the optimality criterion for heavy tails is not satisfied, it is possible to construct examples exhibiting (close to) worst-case behavior. In addition, scheduling policies that are optimal for heavy tails can show worst-case behavior under light-tailed input.

Although this paper provides a negative result, the impossibility of tail-optimal scheduling, the result provides insights into the limitations of scheduling policies when it comes to preventing large sojourn times and also serves to motivate a number of interesting follow-up research questions.

(i) One problem of particular interest is motivated by the notion of tail-competitiveness that we introduce here: although no policy can be tail-competitive across heavy- and light-tailed workloads, maybe it is possible for a policy to be γ-tail-competitive, in the sense that the optimality definition holds for all $\epsilon > \gamma$. Currently, no nonlearning policy has been proven to have a nontrivial γ, i.e., no nonlearning policy can even have better-than-worst-case performance under both light- and heavy-tailed workloads.

(ii) A second topic is concerned with the design and analysis of learning policies that optimize the sojourn time tail across all job size distributions. For example, how can a policy be designed so that it can quickly differentiate between light- and heavy-tailed job size distributions even in the face of time-varying workloads? Interestingly, Nair et al. (2010) have recently provided a policy, called limited processor sharing (LPS), that achieves this goal by learning only information about the *mean* job size. However, LPS only achieves optimality under a limited class of heavy-tailed and light-tailed workloads. It would be interesting to understand whether it is possible to be tail-competitive by using just the mean job size.

(iii) Finally, it seems possible to obtain some positive results. We conjecture that PS and SRPT are strongly tail-competitive for regularly varying job sizes, and that FCFS is strongly tail-competitive for light-tailed job sizes. A natural follow-up question is whether such optimality conditions hold for larger classes of distributions (for example lognormal and Weibull distributions). In particular, what is the largest set of distributions for which SRPT optimizes the sojourn time tail? What about PS or FCFS?

Acknowledgments

Adam Wierman's research is partly supported by the National Science Foundation Computing and Communication Foundations [Grant 0830511], Microsoft Research, and the Okawa Foundation. Bert Zwart's research is partly supported by the National Science Foundation [Grants 0727400 and 0805979], an IBM faculty award, and a VIDI grant from the Netherlands Organisation for Scientific Research.

References

Anantharam V (1999) Scheduling strategies and long-range dependence. *Queueing Systems Theory Appl.* 33(1–3):73–89.

Asmussen S (2000) *Ruin Probabilities, Advanced Series on Statistical Science, and Applied Probability*, Vol. 2 (World Scientific Publishing Co. Inc., River Edge, NJ).

Asmussen S (2003) *Applied Probability and Queues* (Springer).

Bingham NH, Goldie CM, Teugels JL (1987) *Regular Variation* (Cambridge University Press, Cambridge, UK).

Borovkov AA (1976) *Stochastic Processes in Queueing Theory* (Springer-Verlag, New York). [Translated from the Russian by Kenneth Wickwire, *Applications of Mathematics*, No. 4.]

Borst S, Nunez-Queija R, Zwart B (2006) Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* 53(1–2):31–51.

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

Boxma O, Zwart B (2007) Tails in scheduling. *Performance Evaluation Rev.* 34(4):13–20.

Cohen JW (1973) Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.* 10:343–353.

Ganesh A, O'Connell N, Wischik D (2004) *Big Queues* (Springer, New York).

Glynn PW, Whitt W (1994) Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. Galambos J, Gani J, eds. *Studies in Applied Probability* (Applied Probability Trust, University of Michigan, Ann Arbor) 131–156.

Harchol-Balter M (2007) New perspectives on scheduling. *Performance Evaluation Rev.* 34(4).

Iglehart DL (1972) Extreme values in the $GI/G/1$ queue. *Ann. Math. Statist.* 43(2):627–635.

Jelenkovic PR, Momcilovic P (2003) Asymptotic loss probability in a finite buffer fluid queue with heterogeneous heavy-tailed on-off processes. *Ann. Appl. Probab.* 13(2):576–603.

Jelenkovic PR, Kang X, Tan J (2007) Adaptive and scalable comparison of scheduling. *Proc. ACM Sigmetrics* (ACM, New York).

Kelly FP (1996) Notes on effective bandwidths. Kelly FP, Zachary S, Ziedins IB, eds. *Stochastic Networks: Theory and Applications* (Oxford University Press, Oxford, United Kingdom) 141–168.

Koutsoupias E, Papadimitriou CH (2000) Beyond competitive analysis. *SIAM J. Comput.* 30(1):300–317.

Mandjes M, Nuyens M (2005) Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probab. Engrg. Inform. Sci.* 19:351–361.

Mandjes M, Zwart B (2006) Large deviations of sojourn times in processor sharing queues. *Queueing Systems* 52(4):237–250.

Meyer AD, Teugels JL (1980) On the asymptotic behaviour of the distributions of the busy period and the service time in M/G/1. *J. Appl. Probab.* 17:802–813.

Nair J, Wierman A, Zwart B (2010) Tail-robust scheduling via limited processor sharing. *Performance Evaluation* 14(11):978–996.

Nuyens M, Wierman A (2008) The foreground-background queue: A survey. *Performance Evaluation* 65(3–4):286–307.

Nuyens M, Zwart B (2006) A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems* 54(2):85–97.

Nuyens M, Wierman A, Zwart B (2008) Preventing large sojourn times using SMART scheduling. *Oper. Res.* 56(1):88–101.

Pakes AG (1975) On the tails of waiting-time distributions. *J. Appl. Probab.* 12:555–564.

Palmowksi Z, Rolski T (2006) On busy period asymptotics in the GI/G/1 queue. *Adv. Appl. Probab.* 83(1–2):92–103.

Pinedo ML (2008) *Scheduling: Theory, Algorithms, and Systems*, 3rd ed. (Springer, New York).

Ramanan K, Stolyar AL (2001) Largest weighted delay first scheduling: Large deviations and optimality. *Ann. Appl. Probab.* 11:1–48.

Righter R, Shanthikumar JG, Yamazaki G (1990) On external service disciplines in single stage queueing systems. *J. Appl. Probab.* 27:409–416.

Schrage LE (1968) A proof of the optimality of the shortest remaining processing time discipline. *Oper. Res.* 16(3):687–690.

Stolyar AL (2003) Control of end-to-end delay tails in a multiclass network: LWDF discipline optimality. *Ann. Appl. Probab.* 13(3): 1151–1206.

Whitt W (1993) Tail probabilities with statistical multiplexing and effective bandwidths in multiclass queues. *Telecomm. Systems* 2:71–107.

Wierman A, Nuyens M (2008) Scheduling despite inexact job-size information. *Proc. ACM Sigmetrics* (ACM, New York).

Wischik D, McKeown N (2005) Part i: Buffer sizes for core routers. *ACM SIGCOMM Comput. Comm. Rev.* 35(3):75–78.

Zachary S (2004) A note on Veraverbeke's theorem. *Queueing Systems* 46(1–2):9–14.

Zwart AP (2001) Tail asymptotics for the busy period in the GI/G/1 queue. *Math. Oper. Res.* 26(3):485–493.

Adam Wierman is an assistant professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology, where he is a member of the Rigorous Systems Research Group (RSRG). His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the ACM SIGMETRICS Rising Star Award in 2011 and has received best paper awards at ACM SIGMETRICS, IFIP Performance, IEEE INFOCOM, and ACM GREENMETRICS. He has also received multiple teaching awards, including the Associated Students of the California Institute of Technology (ASCIT) Teaching Award.

Bert Zwart held appointments at INRIA, Eindhoven, and Georgia Institute of Technology before moving to the Center of Mathematics and Computer Science (CWI) in Amsterdam, where he leads the Probability and Stochastic Networks group. He has a professor position at VU University Amsterdam and is also affiliated at EURANDOM and Georgia Tech. His honors include an IBM faculty award, the Erlang Prize, and VENI and VIDI awards from the Dutch Science Foundation (NWO).