

# Is tail-optimal scheduling possible?

Adam Wierman

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

Bert Zwart

CWI Amsterdam, VU University Amsterdam, Eurandom & Georgia Tech. bertz@cwi.nl

This paper focuses on the competitive analysis of scheduling disciplines in a large deviations setting. Though 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-tailed and heavy-tailed job size distributions. We prove that no such work-conserving, non-anticipatory, non-learning policy exists and thus that a policy must learn (or know) the job size distribution in order to optimize the sojourn time tail.

*Key words:* scheduling; queuing; large deviations; competitive analysis

---

## 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 which may be attacked with a variety of techniques including optimization and stochastics/queuing, see for example Harchol-Balter (2007), Pinedo (2008).

Many results in scheduling either focus on (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-tailed 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) and Pakes (1975)), preemptive Last Come First Served (LCFS) (Meyer and Teugels

(1980) and Zwart (2001)), Processor Sharing (PS) (Borst et al. (2006)), Shortest Remaining Processing Time (SRPT) (Nuyens and Zwart (2006) and 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, 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 in, for example, 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. Section 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. Richter 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 and vice versa. 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 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, and vice versa. 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 (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  $\rho$  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 is known to be non-optimal.

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 Section 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 Section 3, we present and prove the main result of the paper. Finally, in Section 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. We consider all policies  $\pi$  that satisfy the following three conditions:

- (i)  $\pi$  is *work-conserving*: the scheduling policy always has the server working at speed 1 whenever work is present in the system.
- (ii)  $\pi$  is *non-anticipative*: 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)  $\pi$  is *non-learning*: the scheduler cannot learn or know anything about the distribution of the inter-arrival times and job sizes in the sense that the scheduling discipline is not allowed to depend on information about jobs arriving in previous busy periods.

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. The first two assumptions are standard and allow 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 non-standard, 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.

The third condition is also technically convenient since it provides a tractable representation for the steady-state sojourn time  $V_\pi$  under policy  $\pi$ . Assume that the system is empty when customer 1 arrives. Let  $V_{\pi,i}$  be the sojourn time of the  $i$ th customer under scheduling discipline  $\pi$ . Our assumptions imply that  $V_{\pi,i}, i \geq 1$  is a regenerative process, and the steady-state sojourn time  $V_\pi$  satisfies

$$P(V_\pi > t) = \frac{1}{E[N]} E \left[ \sum_{i=1}^N I(V_{\pi,i} > t) \right], \quad (2.1)$$

where  $N$  is the number of customers in a busy period, and  $I(G)$  is the indicator function of the event  $G$ . We introduce some additional notation: denote a generic job size by  $B$  and its mean by  $\beta$ , a generic interarrival time by  $A$ , the arrival rate by  $\lambda$ , and the load by  $\rho = \lambda\beta < 1$ . Under these conditions,  $V_\pi$  is a.s. finite. Note that there is a one-to-one correspondence between a specific choice of job size and interarrival time distributions and the probability measure  $P(\cdot)$ .

### 2.2. Tail optimality

The major focus of the paper is how to choose  $\pi$  such that the sojourn time tail  $P(V_\pi > t)$  converges to 0 as fast as possible as  $t \rightarrow \infty$ . 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  $\pi_0$  is weakly (tail-)competitive for a class  $\mathcal{P}$  of interarrival time distributions and job size distributions, if

$$\limsup_{t \rightarrow \infty} \frac{P(V_{\pi_0} > t)^{1+\epsilon}}{P(V_{\pi} > t)} < \infty \quad (2.2)$$

holds for every  $\epsilon > 0$ , every  $P \in \mathcal{P}$  and every work-conserving, non-anticipative, non-learning scheduling policy  $\pi$ .  $\pi_0$  is called (tail-)competitive if the same property holds for  $\epsilon = 0$ , and strongly (tail-)competitive if additionally the limsup is bounded by 1 for  $\epsilon = 0$ .

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 for example 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:

$$P(V_{\pi} > t) \geq P(B > t), \quad (2.3)$$

$$P(V_{\pi} > t) \geq \frac{1}{E[N]} P(C_{max} > t). \quad (2.4)$$

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 (2.1), see Boxma and Zwart (2007) for details. An approach for proving optimality of  $\pi_0$  is to analyze the tail behavior of  $V_{\pi_0}$ , and then to compare 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  $\pi_0$ .

### 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.  $L(ax)/L(x) \rightarrow 1$  as  $x \rightarrow \infty$  for every  $a > 0$ ) and  $\alpha > 1$  is a constant. Regularly varying distributions are a generalizations 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  $\Phi_A$  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  $\lambda$ ).  $\Psi(\theta)$  is strictly convex if either  $A$  or  $B$  is non-deterministic. Now, we can state the large deviations results for FCFS and LCFS:

$$\lim_{t \rightarrow \infty} \frac{-\log P(V_F > t)}{t} = \gamma_F := \sup\{\theta : \Psi(\theta) - \theta \leq 0\}, \quad (2.5)$$

$$\lim_{t \rightarrow \infty} \frac{-\log P(V_L > t)}{t} = \gamma_L := \sup_{\theta \geq 0}\{\theta - \Psi(\theta)\}. \quad (2.6)$$

These theorems hold without any regularity conditions on  $\Psi$ , as is shown in, for example, Nuyens and Zwart (2006); see also Asmussen (2003), Glynn and Whitt (1994), Palmowski and Rolski (2006). From the strict convexity of  $\Psi(\theta) - \theta$ , and the fact that  $\Psi'(0) = \rho$ , it follows that

$$\gamma_L < (1 - \rho)\gamma_F. \quad (2.7)$$

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 under the condition that Cramér's 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 (2.4) it follows that, if Cramér's condition is satisfied, then

$$\limsup_{t \rightarrow \infty} \frac{P(V_F > t)}{P(V_\pi > t)} < \infty \quad (2.8)$$

for any scheduling discipline  $\pi$ , cf. Boxma and Zwart (2007).

In contrast to the optimality of  $\gamma_F$ , the decay rate  $\gamma_L$  is the smallest possible decay rate. To see this, note that  $V_\pi$  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  $\gamma_L$  in Nuyens et al. (2008).

Interestingly, many other common policies have been shown to have decay rate equal to  $\gamma_L$ . 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  $\gamma_L$  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:

$$P(V_F > x) \sim \frac{\rho}{1 - \rho} \frac{1}{\alpha - 1} t P(B > t), \quad (2.9)$$

$$P(V_L > t) \sim E[N] P(B > t(1 - \rho)), \quad (2.10)$$

$$P(V_{PS} > t) \sim P(V_{SRPT} > t) \sim P(B > t(1 - \rho)), \quad (2.11)$$

where  $f(x) \sim g(x)$  denotes  $\lim_{x \rightarrow \infty} f(x)/g(x) = 1$ . 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, since  $P(B > t(1 - \rho)) \sim (1 - \rho)^{-\alpha} P(B > t)$ , 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 non-preemptive policies as is shown by Anantharam (1999). The reason is that, under any non-preemptive 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 (2.1).

### 3. Main result

The previous section reveals a clear dichotomy between the scheduling policies that perform well under light-tailed and heavy-tailed job size distributions. FCFS is weakly 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 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, non-anticipative, and non-learning scheduling policy  $\pi$  that is weakly competitive for both light-tailed and heavy-tailed job size 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 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 competitive, see (3.1) 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, since (3.1) is only proven to be necessary for probability measures under which job sizes are heavy-tailed. Using this construction, we show that (3.1) implies non-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

For a given scheduling discipline and  $t \geq 0$ , we define  $R(t)$  to be the service allocated in  $[0, t]$  to all jobs arriving in the system after time 0, if a job (called  $B_1$  for future convenience) of size at least  $y(1 - \rho)t$  arrived at an empty system at time 0. Here  $y > 1$  is a parameter. Observe that  $\limsup_{t \rightarrow \infty} R(t)/t \leq \rho$  a.s. The first major step is to show that

$$\forall \delta > 0, \quad \lim_{t \rightarrow \infty} P(R(t) \geq (\rho - \delta)t \mid B_1 > (1 - \rho)yt) = 1 \quad (3.1)$$

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), and 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 competitive. Intuitively, the reason that 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 (3.1) is a necessary condition, we construct a counterexample. Fix a scheduling policy and choose  $P(\cdot)$  such that  $P(B > t)$  is regularly varying with index  $\alpha > 2$  and  $\rho < 1$ . Additionally, suppose that (3.1) does not hold, i.e. there exist constants  $y > 1, \delta, \gamma > 0$  and a sequence  $(t_n), t_n \rightarrow \infty$  such that

$$P(R(t_n) \leq (\rho - \delta)t_n \mid B_1 > y(1 - \rho)t_n) > \gamma, n \geq 1. \quad (3.2)$$

We are now ready to state our first proposition.

PROPOSITION 1. Consider an  $M/G/1$  queue operating under  $P(\cdot)$ . Let  $\pi$  be a discipline satisfying (3.2). Then

$$\liminf_{n \rightarrow \infty} \frac{P(V_\pi > t_n \delta/4)}{\sqrt{t_n} P(B > t_n)} > 0. \quad (3.3)$$

Thus,  $\pi$  is not weakly competitive with PS under  $P(\cdot)$ . Consequently, if  $\pi$  is weakly competitive for heavy-tailed job sizes, then (3.1) must hold for all  $y > 1$  and all  $\delta > 0$ .

**Proof:** Denote the event in (3.2) by  $F_n$ . Fix  $\eta > 0$  and let  $E_n$  be the event that there are at least  $(\lambda - \eta)t_n$  and at most  $(\lambda + \eta)t_n$  class-1 arrivals in  $[0, t_n]$ . Suppose in addition that under  $E_n$  all these arrivals have service requirements that are bounded from above by  $\sqrt{t_n} \delta/4$ . We see that

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

The first probability converges to 1 in view of the law of large numbers for Poisson processes. The second probability converges to 1 as well, since the assumption  $\alpha > 2$  implies  $P(B \leq \sqrt{t_n} \delta/4) = 1 - o(1/t_n)$  as  $n \rightarrow \infty$ . Thus,  $P(E_n) \rightarrow 1$ .

Let  $Y(t_n)$  be the amount of work offered to the system in  $(0, t_n]$ . Since  $P(E_n) \rightarrow 1$ ,  $Y(t_n)/t_n \rightarrow \rho$  in  $P(\cdot | E_n)$  probability as  $n \rightarrow \infty$ . Thus, it follows that  $G_n = E_n \cap F_n \cap \{Y(t_n) > (\rho - \delta/2)t_n\}$  satisfies  $P(G_n) \geq \gamma/2$  for  $n$  sufficiently large. Note that the length of the busy period under  $G_n$  is at least  $t_n$ , and that the amount of work of the customers arriving after time 0 at time  $t_n$  (given by  $W(t_n)$ ) is at least  $(\delta/2)t_n$ . Since the remaining service requirement of each customer is at most  $\sqrt{t_n} \delta/4$  at time  $t_n$ , the amount of customers  $Q(t_n)$  that arrived after time 0 and is still in the system at time  $t_n$  is at least  $(\delta/2)t_n / \sqrt{\delta t_n/4} = \sqrt{\delta t_n}$ .

Using (2.1) we obtain

$$P(V_\pi > (\delta/4)t_n) \geq \frac{1}{E[N]} E \left[ I(G_n) I(B_1 > y(1 - \rho)t_n) \sum_{i=1}^N I(V_{\pi,i} > (\delta/4)t_n) \right]. \quad (3.4)$$

The above considerations imply that the last expression is equal to

$$\frac{1}{E[N]} E \left[ I(G_n) I(W(t_n) > (\delta/2)t_n) I(Q(t_n) \geq \sqrt{\delta t_n}) I(B_1 > y(1 - \rho)t_n) \sum_{i=1}^N I(V_{\pi,i} > (\delta/4)t_n) \right]. \quad (3.5)$$

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  $\sum_{i=1}^N I(V_{\pi,i} > (\delta/4)t_n)$  can be lower bounded by  $\sqrt{\delta t_n/4}$  in the last expression, implying that, for large enough  $n$ ,

$$P(V_\pi > (\delta/4)t_n) \geq \frac{1}{E[N]} \sqrt{\delta t_n/4} (\gamma/2) P(B_1 > y(1 - \rho)t_n). \quad (3.6)$$

□

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

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

by using the same argument. Taking  $z$  arbitrary close to 1 provides the interesting interpretation that if the necessary condition (3.1) 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 operating under  $P(\cdot)$ . Let  $\pi$  be a discipline satisfying (3.2). Then for all  $\epsilon > 0$  there exists a  $P$  such that

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

### 3.2. Light-tails: Non-competitiveness

Given the necessary condition for a scheduling policy to be 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 (3.1) implies non-competitiveness in the light-tailed example we construct. This change of measure argument is necessary since (3.1) is only shown to be necessary for heavy-tailed job sizes.

Thus, we construct a specific probability measure under which service times are light-tailed. To help distinguish the light-tailed 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  $\tilde{B}$  be a service time distribution which is regularly varying with index  $-\alpha, \alpha > 2$ . Let  $\tilde{\beta}$  be its mean, and let  $\tilde{\lambda}$  be an arrival rate such that  $\tilde{\lambda}\tilde{\beta} = 1 - \epsilon$  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  $\tilde{\Phi}(\theta) = \tilde{E}[e^{\theta\tilde{B}}]$  be the MGF of  $\tilde{B}$ . Note that this MGF is finite iff  $\theta \leq 0$ . Next, define a parameter  $s \in (0, \tilde{\lambda})$ . Set  $\lambda_s = \tilde{\lambda} - s$ , and let  $B_s$  be a random variable with MGF  $\tilde{\Phi}(\theta - s)/\tilde{\Phi}(-s)$ . Its mean  $\beta_s$  is given by  $\tilde{\Phi}'(-s)/\tilde{\Phi}(-s)$ . The corresponding load  $\rho_s = \lambda_s\beta_s$ . Note that  $\rho_s$  is continuous and strictly decreasing in  $s$ , and that  $\rho_s \rightarrow 0$  as  $s \uparrow \tilde{\lambda}$ . 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  $\Phi$  be the MGF of  $B$ , and note that  $\Phi(\theta) = \tilde{\Phi}(\theta - s^*)/\tilde{\Phi}(-s^*)$ .

From the construction of  $\lambda$  and  $B$  we have the following properties of  $\gamma_F$  and  $\gamma_L$ . Recall these are the fastest and slowest decay rates achievable.

LEMMA 1. Given the construction of  $\lambda$  and  $B$ , we have  $\gamma_F = s^*$ .

**Proof:** For the  $M/G/1$  queue, (2.5) specializes to

$$\gamma_F = \sup\{\theta : \lambda(\Phi(\theta) - 1) \leq \theta\}. \quad (3.7)$$

Since  $\Phi(\theta) = \infty$  if  $\theta > s^*$ ,  $\gamma_F \leq s^*$ . Next, observe that by convexity  $\frac{1 - \tilde{\Phi}(-s^*)}{s^*} \leq \tilde{\beta}$ , which implies

$$\tilde{\lambda} \frac{1 - \tilde{\Phi}(-s^*)}{s^*} \leq \tilde{\rho} \leq 1, \quad (3.8)$$

and

$$\frac{\lambda}{\lambda + s^*} \Phi(s^*) = \frac{\tilde{\lambda} - s^*}{\tilde{\lambda}\tilde{\Phi}(-s^*)} \leq 1, \quad (3.9)$$

where the last inequality is equivalent to (3.8). Returning to (3.7), we complete the proof as follows:

$$\lambda(\Phi(s^*) - 1) = (\tilde{\lambda} - s^*) \left( \frac{1}{\tilde{\Phi}(s^*)} - 1 \right) \leq \tilde{\lambda}(1 - \tilde{\Phi}(-s^*)) \leq s^*,$$

where the second line follows from (3.9) and the third line follow from (3.8). Thus,  $\gamma_F \geq s^*$

□

LEMMA 2. Given the construction of  $\lambda$  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  $\gamma_L$ . The key observation behind this is that (i) the left derivative  $\Psi'(s) \leq \tilde{\rho} < 1$ , for  $s \leq s^*$ , and (ii)  $\Psi(s) = \infty$  for  $s > s^*$ . The second observation is trivial, while the first follows from (3.9) and

$$\Psi'(s) = \lambda\Phi'(s) \leq \lambda\Phi'(s^*) = \frac{\tilde{\lambda} - s^*}{\tilde{\lambda}\tilde{\Phi}(-s^*)}\tilde{\rho} \leq 1.$$

□

Note that combining the above lemmas with the fact that  $\gamma_L < (1 - \rho)\gamma_F$  implies

$$s^* - \lambda(\Phi(s^*) - 1) < (1 - \rho)s^*, \quad (3.10)$$

which we need to use later in the proof. We are now ready to state our second proposition.

**PROPOSITION 2.** *Consider a scheduling discipline  $\pi$  that satisfies (3.1) under  $\tilde{P}$  for  $y = 1 + \epsilon$  and  $\delta = \epsilon^2$ , then  $P(V_\pi > t)/P(V_{\text{FCFS}} > t) \rightarrow \infty$  at an exponential rate for  $\epsilon \in (0, 1/4)$ .*

**Proof:** Consider the  $M/G/1$  queue with the above particular choice of  $\lambda$  and  $B$ . Let  $X(t)$  be the net amount of work entering the queue in the interval  $(0, t]$  (i.e. total amount of work minus  $t$ ). We can write  $E[e^{\theta X(t)}] = e^{t(\Psi(\theta) - \theta)}$ . Using a standard change of measure argument, cf. Chapter 13.4 of Asmussen (2003), and the lemmas above, we can write

$$P(V_\pi > t) = e^{-\gamma_L t} \tilde{E}[e^{-\gamma_F X(t)} I(V_\pi > t)]. \quad (3.11)$$

We can lower bound this by requiring that we enter an empty system under  $\tilde{P}$ , and adding some other events, which gives

$$P(V_\pi > t) \geq e^{-\gamma_L t} \epsilon \tilde{E}[e^{-\gamma_F X(t)} I(X(t) < 0) I(R(t)/t \geq 1 - \epsilon - \epsilon^2) I(B_1 > \epsilon(1 + \epsilon)t)]. \quad (3.12)$$

Observe that for large enough  $t$ ,  $X(t)/t \leq 0$  a.s. under  $\tilde{P}$ , and that we can lower bound  $e^{-\theta X(t)}$  by 1 if  $X(t) < 0$  (since  $\theta > 0$ ). From Condition (3.1) under  $\tilde{P}$  with  $y = 1 + \epsilon$  we then infer that for large enough  $t$ ,

$$P(V_\pi > t) \geq \epsilon e^{-(\gamma_L + (\epsilon + \epsilon^2)\gamma_F)t(1+o(1))}. \quad (3.13)$$

Recall that  $\rho > \epsilon + \epsilon^2$ . Thus,

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

which completes the proof.

□

An interesting observation about the above proof is that the logarithmic decay rate of any policy that satisfies (3.1) can be made arbitrarily close to the slowest possible possible decay rate (that of LCFS) since  $\gamma_F$  and  $\gamma_L$  both converge to strictly positive constants as  $\epsilon \rightarrow 0$ . Thus, if a policy is competitive in the case of regularly varying job sizes it has (nearly) the heaviest possible tail in the case of light-tailed job sizes. Further, if the policy is competitive in the case of light-tailed job sizes, then the necessary Condition (3.1) does not hold, and the remark after Proposition 1 implies that the policy has the heaviest possible tail in the case of regularly varying job sizes.

**COROLLARY 2.** *Consider a scheduling discipline  $\pi$  that satisfies (3.1) under  $\tilde{P}$ . Then for all  $\epsilon > 0$  there exists a  $P$  such that*

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

## 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-tailed 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 non-optimality 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.

Though this paper provides a negative result, the impossibility of tail-optimal scheduling, the result provides insights into the limitation 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: Though no policy can be competitive across heavy-tailed and light-tailed workloads, maybe it is possible for a policy to be  $\gamma$ -tail-competitive, in the sense that the optimality definition holds for all  $\epsilon > \gamma$ . Currently, no policy has been proven to have a non-trivial  $\gamma$ .
- (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-tailed and heavy-tailed job size distributions even in the face of time-varying workloads. Also, it may be the case that only information about some moments of the service time distribution is required to design a tail optimal scheduling policy.
- (iii) Finally, it seems possible to obtain some positive results. We conjecture that PS and SRPT are strongly competitive for regularly varying job sizes, and that FCFS is strongly 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?

**Acknowledgment.** Adam Wierman's research is partly supported by NSF CCF 0830511, Microsoft Research, and the Okawa Foundation. Bert Zwart's research is partly supported by NSF grants 0727400 and 0805979, an IBM faculty award, and a VIDI grant from NWO.

## References

- Anantharam, V. 1999. Scheduling strategies and long-range dependence. *Queueing Systems Theory Appl.* **33**(1-3) 73–89. Queues with heavy-tailed distributions.
- Asmussen, S. 2000. *Ruin probabilities, Advanced Series on Statistical Science & Applied Probability*, vol. 2. World Scientific Publishing Co. Inc., River Edge, NJ.
- Asmussen, S. 2003. *Applied Probability and Queues*. Springer.
- Bingham, N.H., C.M Goldie, J.L. Teugels. 1987. *Regular Variation*. Cambridge University Press.
- Borovkov, A. A. 1976. *Stochastic processes in queueing theory*. Springer-Verlag, New York. Translated from the Russian by Kenneth Wickwire, Applications of Mathematics, No. 4.
- Borst, S., O. Boxma, R. Nunez-Queija, B. Zwart. 2003. The impact of the service discipline on delay asymptotics. *Performance Evaluation* **54** 175–206.
- Borst, S., R. Nunez-Queija, B. Zwart. 2006. Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* **53**(1-2) 31–51.
- Boxma, O., B. Zwart. 2007. Tails in scheduling. *Performance Evaluation Review* **34**(4) 13–20.

- Cohen, J. W. 1973. Some results on regular variation for distributions in queueing and fluctuation theory. *J. Applied Probability* **10** 343–353.
- Glynn, P. W., W. Whitt. 1994. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. J. Galambos, J. Gani, eds., *Studies in Applied Probability*. Applied Probability Trust, 131–156.
- Harchol-Balter, M. 2007. New perspectives on scheduling. *Performance Evaluation Review* **34**(4).
- Iglehart, D. L. 1972. Extreme values in the  $GI/G/1$  queue. *Ann. Math. Statist.* **43** 627–635.
- Jelenkovic, P. R., X. Kang, J. Tan. 2007. Adaptive and scalable comparison of scheduling. *Proc. of ACM Sigmetrics*.
- Jelenkovic, P. R., P. Momcilovic. 2003. Asymptotic loss probability in a finite buffer fluid queue with heterogeneous heavy-tailed on-off processes. *Annals of Applied Probability* **13**(2) 576–603.
- Kelly, F. P. 1996. Notes on effective bandwidths. F. P. Kelly, S. Zachary, I. B. Ziedins, eds., *Stochastic networks: Theory and applications*. Oxford University Press, 141–168.
- Koutsoupias, E., C. H. Papadimitriou. 2000. Beyond competitive analysis. *SIAM Journal on Computing* **30**(1) 300–317.
- Mandjes, M., M. Nuyens. 2005. Sojourn times in the  $M/G/1$  FB queue with light-tailed service times. *Probability in the Engineering and Information Sciences* **19** 351–361.
- Mandjes, M., B. Zwart. 2006. Large deviations of sojourn times in processor sharing queues. *Queueing Systems* **52**(4) 237–250.
- Meyer, A. De, J.L. Teugels. 1980. On the asymptotic behaviour of the distributions of the busy period and the service time in  $M/G/1$ . *J. Applied Probability* **17** 802–813.
- Nuyens, M., A. Wierman. 2008. The foreground-background queue: A survey. *Performance Evaluation* **65**(3-4) 286–307.
- Nuyens, M., A. Wierman, B. Zwart. 2008. Preventing large sojourn times using SMART scheduling. *Operations Research* **56**(1) 88–101.
- Nuyens, M., B. Zwart. 2006. A large-deviations analysis of the  $GI/GI/1$  SRPT queue. *Queueing Systems* **54**(2) 85–97.
- Pakes, A. G. 1975. On the tails of waiting-time distributions. *J. Applied Probability* **12** 555–564.
- Palmowski, Z., T. Rolski. 2006. On busy period asymptotics in the  $GI/G/1$  queue. *Advances in Applied Probability* **83**(1-2) 92–103.
- Pinedo, M. L. 2008. *Scheduling*. 3rd ed. Springer, New York. Theory, algorithms, and systems, With 1 CD-ROM (Windows, Macintosh and UNIX).
- Ramanan, K., A. L. Stolyar. 2001. Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability* **11** 1–48.
- Righter, R., J. G. Shanthikumar, G. Yamazaki. 1990. On external service disciplines in single stage queueing systems. *J. of Applied Probability* **27** 409–416.
- Schrage, L. E. 1968. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research* **16** 678–690.
- Stolyar, A. L. 2003. Control of end-to-end delay tails in a multiclass network: LWDF discipline optimality. *Annals of Applied Probability* **13**(3) 1151–1206.
- Whitt, W. 1993. Tail probabilities with statistical multiplexing and effective bandwidths in multiclass queues. *Telecommunication Systems* **2** 71–107.
- Wierman, A., M. Nuyens. 2008. Scheduling despite inexact job-size information. *Proc. of ACM Sigmetrics*.
- Wischik, D., N. McKeown. 2005. Part i: Buffer sizes for core routers. *ACM SIGCOMM Computer Communication Review* **35**(3) 75–78.
- Zachary, S. 2004. A note on Veraverbeke’s theorem. *Queueing Syst.* **46**(1-2) 9–14.
- Zwart, A. P. 2001. Tail asymptotics for the busy period in the  $GI/G/1$  queue. *Mathematics of Operations Research* **26**(3) 485–493.