Tails in scheduling

Onno Boxma
EURANDOM &
Eindhoven University of Technology
Department of Mathematics & Computer Science
PO Box 513, 5600 MB Eindhoven, The Netherlands
boxma@win.tue.nl

Bert Zwart
Georgia Institute of Technology
H. Milton Stewart School of Industrial and Systems Engineering 765 Ferst Drive
Atlanta, GA 30332-0205
bertzwart@gatech.edu

ABSTRACT

This paper gives an overview of recent research on the impact of scheduling on the tail behavior of the response time of a job. We cover preemptive and non-preemptive scheduling disciplines, consider light-tailed and heavy-tailed distributions, and discuss optimality properties. The focus is on results, intuition and insight rather than methods and techniques.

1. INTRODUCTION

The response time (a.k.a. sojourn time) of a job is the time from when the job arrives until it departs. While mean response time is a very common metric of study, the tail of response time is equally important, as it gives insight into the occurrence of unusually long delays. This insight can be helpful in designing systems such that the impact of long delays is limited. The recent surge of activity in this area is further stimulated by the growing awareness that job sizes in computer and communication networks can be extremely variable, leading to undesirable properties such as long-range dependence and heavy-tailed response times; see [44] for a survey volume on this topic. Additional motivation is the fact that unusually long delays have significant impact on the performance of the system as perceived by a user. Users that wait for a long time may abandon the system which negatively impacts the performance.

In this paper we focus on the impact of *scheduling* of the jobs on response time tail behavior. An important goal of scheduling is to prevent extreme delays, and we give an overview of results in the queueing literature to indicate how a specific scheduling discipline affects long delays. The paper is not methodologically oriented (we refer to [9] for this). Our emphasis is on results, intuition and insight. We try to explain how long delays occur most likely, and how these delays depend on the load and on the statistical assumptions underlying the model.

Throughout the paper we distinguish between light-tailed and heavy-tailed service requirements (also in some places called job sizes), as the methods, results and insights in these two cases are completely different. Queueing models with light-tailed service requirement distributions, like the exponential distribution and more generally phase-type distributions, typically also have light-tailed sojourn time distributions. Queueing models with heavy-tailed service distributions, like the Pareto distribution, typically also have heavy-tailed sojourn time distributions. We give several examples in which the service requirement tail follows a power law ($x^{-\alpha}$), yielding a sojourn time tail that also follows a power law, often also $x^{-\alpha}$ or the even heavier $x^{1-\alpha}$. These are in fact the best and worst possible cases for a large class of scheduling policies. This will follow from the observation that a sojourn time is always bounded from below by its own service time, and from above by the remainder of the busy period in which the tagged customer arrived.

The impact of scheduling on the tail behavior of job response times is a topic with many aspects – too many to cover in this brief survey. We shall pay some attention to preemptive and non-preemptive policies, and to optimality issues. But we shall not discuss the case of multiple customer classes, with scheduling disciplines that assign priorities to classes; see the contribution of Aalto et al. in this volume. Neither shall we explicitly discuss fairness of scheduling disciplines; see the contribution of Wierman in this volume.

The paper is organized as follows. Section 2 contains a description of the model under consideration, and a discussion of tail behavior with a distinction between light and heavy tails. We also present some definitions of optimality of scheduling disciplines, in the context of tail behavior. Sections 3 and 4 are devoted to tail asymptotics for two fundamental service disciplines: First Come First Served (FCFS) and Processor Sharing (PS). In the light-tailed case, FCFS turns out to be optimal in some asymptotic sense; in the heavy-tailed case, FCFS performs very poorly (since many customers may have to spend a long time in the system because of one very large customer ahead of them). In contrast, PS performs in some sense optimally in the heavy-tailed case. Several other disciplines are briefly touched upon in Section 5. Section 6 contains some concluding remarks. Technical details on a new optimality result for FCFS can be found in the appendix.

2. PRELIMINARIES

The goal of this section is to introduce some background on tail analysis and some notation. We introduce our model in Section 2.1, review the concepts of light tails and heavy tails in Section 2.2, and introduce new optimality notions in Section 2.3.

2.1 Model description

We consider a system where jobs, numbered by $i \geq 1$, arrive at a server one by one. The size (i.e. amount of work that needs to be processed) of job i is denoted by B_i. The sequence B_i, $i \geq 1$ is assumed to consist of independent and identically distributed (i.i.d.) random variables. The time between the arrivals of job $i-1$ and job i is given by A_i. The sequence A_i, $i \geq 1$ is i.i.d., independent of the sequence of job sizes. The mean amount of work offered per unit of time by this arrival stream is denoted by $\rho = E[B]/E[A]$, with B a generic job size, and A a generic inter-arrival time. We assume that this work is processed by a server, which works at speed 1 whenever there is work in the system; therefore we call ρ the server utilization. We assume that $\rho < 1$. Under this condition, the system reaches steady state under weak assumptions. By V, we denote the steady-state *response time* of a job, i.e. the time that elapses between arrival and departure of a job; we will also use the terms *system time* and *sojourn time*.

We allow for all non-anticipating scheduling disciplines (any information on jobs currently in the system may be used). In particular, we allow scheduling disciplines to be preemptive. We allow the server to work on several jobs simultaneously, and allow scheduling disciplines to be based on the sizes of jobs that are currently in the system. Let Π be the class of all such scheduling disciplines and write π for a particular scheduling discipline $\pi \in \Pi$. If we wish to express the dependence of V on a specific scheduling discipline π, then we write $V = V_\pi$.

2.2 Tails

We are interested in the tail behavior of V, i.e. in the behavior of $P(V > x)$ as x grows large. Examining such behavior is important in assessing how well a system is capable of preventing huge sojourn times. The concept of "huge" or "large" is not very well defined. One could think of the smallest value which is not acceptable in a particular application. In terms of probabilities, one can think of values of x for which $P(V > x)$ is of the order 10^{-3} or smaller.

There exist several techniques to obtain the tail behavior of V. These can be divided into analytic techniques and probabilistic techniques, and are documented elsewhere (see e.g. [9, 10, 26] and references therein). This set of techniques is often called *large deviations techniques* and the associated performance analysis is typically called *rare event analysis*. The goals of rare event analysis are twofold. The first goal is to obtain an accurate approximation of $P(V > x)$ which is valid for large x; this approximation gives insight into the frequency of long delays and is usually difficult to obtain by simulation. The second goal is to obtain insight into the way rare events occur, if they occur.

Since we only focus on results and insights, we merely point out that, given the context of the present paper, the solutions to these problems critically depend on the nature of the job size distribution. We distinguish between two different classes of job size distribution, namely light-tailed and heavy-tailed distributions. Both are defined below.

2.2.1 Light tails

Job sizes are said to have a light tail if there exists an $\epsilon > 0$ such that $E[e^{\epsilon B}] < \infty$. Important examples of light-tailed distributions are all distributions with bounded support and all phase-type distributions. Systems where all input distributions are light-tailed are typically well-behaved, since the convergence towards steady state is typically exponentially fast, as is the decay of correlations between successive response times. Moreover, generally all performance indicators, such as the response time, are light-tailed as well.

2.2.2 Heavy tails

We shall simply say that a random variable is heavy-tailed if it is not light-tailed, i.e. job sizes are heavy-tailed if $E[e^{\epsilon B}] = \infty$ for every $\epsilon > 0$. Quantities such as file sizes are typically heavy-tailed [17]. An important class of heavy-tailed distributions which is intuitively appealing is the class of *sub-exponential distributions*. We say that B has a sub-exponential distribution (or simply that B is sub-exponential) if

$$P(B_1 + \ldots + B_n > x) \sim P(\max_{i=1,\ldots,n} B_i > x),$$

where $f(x) \sim g(x)$ means that the ratio of f and g tends to 1 as $x \to \infty$. This class of distributions is appealing since it reflects that if the total amount of work delivered by a number of jobs is large, this is most likely due to a single large job. This intuition is completely different from the light-tailed case, where *all* job sizes are statistically larger than usual if the sum is larger than some large x, see for example [26]. Important examples of heavy-tailed distributions are the log-normal distribution, some Weibull distributions (with tails of the form e^{-x^β} for some $\beta \in (0, 1)$) and the following sub-classes of sub-exponential distributions:

- B is said to be *regularly varying* of index $\alpha > 0$ if $P(B > x) = L(x)x^{-\alpha}$, the function L being *slowly varying*, i.e. $L(ax)/L(x) \to 1$ for any $a > 0$. $L(x)$ can converge to a constant (leading to pure power tails) but can also be proportional to a power of $\log x$.

- B is said to be of *intermediate regular variation* if $P(B > x + o(x)) \sim P(B > x)$ for any function $o(x)$ which is of the small order of x. This is not the standard definition, but a characterization due to D. Korshunov (personal communication). The standard definition is not very insightful and will therefore not be mentioned here.

Background on heavy tails, as well as their implications in areas like insurance and finance, can be found in [22].

2.3 Optimality

There is no well-established definition of optimality of a scheduling discipline in the context of tail behavior, and therefore we take the opportunity to propose one here.

Assume that the inter-arrival time and job size distribution are fixed. We say that a scheduling discipline π^* is *strongly tail optimal* (if the context is tails, we shall simply say strongly optimal), if

$$\limsup_{x\to\infty} \frac{P(V_{\pi^*} > x)}{P(V_\pi > x)} \leq 1, \qquad (1)$$

for any scheduling discipline $\pi \in \Pi$. As we shall see below, this is a rather strong property which is difficult to establish. Therefore, we propose two other definitions of optimality. We say that a scheduling discipline π^* is *weakly tail optimal* if there exists a finite constant M such that

$$\limsup_{x\to\infty} \frac{P(V_{\pi^*} > x)}{P(V_\pi > x)} \leq M, \qquad (2)$$

for any scheduling discipline π. Finally, we define an even weaker notion of optimality. We say that a scheduling discipline π^* is *logarithmically tail optimal* if

$$\liminf_{x\to\infty} \frac{\log P(V_{\pi^*} > x)}{\log P(V_\pi > x)} \geq 1. \qquad (3)$$

It is clear that any strongly optimal scheduling discipline is weakly optimal, and that every weakly optimal scheduling discipline is logarithmically optimal. Optimality of a scheduling discipline will generally depend on the distribution of the service times, and also on the load of the system; examples are given throughout the remainder of this paper.

3. FIRST COME FIRST SERVED

In this section we treat the most basic scheduling discipline, which is FCFS. We shall focus on both light-tailed and heavy-tailed service times. The qualitative tail behavior, as well as the intuition, completely differs between both cases. It turns out that both the tail estimate and the intuition are easier to describe in the heavy-tailed setting. This section is organized as follows. Light-tailed service times are discussed in Section 3.1, and heavy-tailed service times in Section 3.2. We close the section with a discussion of optimality properties of FCFS in Section 3.3.

To connect with existing literature, we present all results for FCFS for the waiting time W rather than the response time V. Since $V = W + B$ and W dominates B, it is easy to extend the results to response times.

3.1 Light-tailed service times

For light-tailed service times, the tail of the response time can be shown to be exponential under mild assumptions. Assume that there exists a constant $\gamma (:= \gamma_{FCFS}) > 0$ such that $E[e^{\gamma(B-A)}] = 1$ and $E[Be^{\gamma B}] < \infty$. Then there exists a constant C_{FCFS} such that

$$P(W > x) \sim C_{FCFS} e^{-\gamma x}. \tag{4}$$

The constant γ is known as the *decay rate*, or *adjustment coefficient*, and must be computed numerically in general. For the $M/M/1$ queue, $\gamma = \mu - \lambda$. The constant C_{FCFS} is explicit for $M/G/1$, but hard to compute in general. However, it has the appealing property that it is bounded from above by 1. In fact, a stronger, non-asymptotic result can be obtained:

$$P(W > x) \leq e^{-\gamma x} \text{ for all } x \geq 0. \tag{5}$$

Both results are attributed to Cramér & Lundberg, but the second result has independently been obtained by Kingman; see Chapter XIII of [4] for an overview of the literature. The approximation $P(W > x) \approx C_{FCFS} e^{-\gamma x}$ is in general excellent for moderate values of x; see [1] for an illustration.

Apart from obtaining insight into the order of magnitude of W, it is also useful to understand what the most likely cause of the event $\{W > x\}$ is, given that it happens. Anantharam [2] has investigated this problem for the workload process of the $GI/G/1$ queue and established that the workload most likely becomes large by *conspiracy*: for a long time, all inter-arrival times are statistically smaller than usual, and all service times are statistically larger than usual.

More precisely, the density of the inter-arrival time distribution, if it exists, will not equal $a(t)$, but (up to a normalizing constant) equals $e^{-\gamma t} a(t)$. Similarly, the density of the service time distribution changes from $b(t)$ to (a normalizing constant times) $e^{\gamma t} b(t)$. This makes the load of the system grow from ρ to a quantity which is bigger than 1. Consequently, the workload of the system grows at linear rate, leading to the event $\{W > x\}$. This type of change of density is called *exponential twisting* or *exponential tilting*. In the special case of the $M/M/1$ queue, this exponential twisting leads to the situation where the inter-arrival times are exponentially distributed with rate μ (instead of λ), and the service times are exponentially distributed with rate λ (instead of μ), resulting in a load of $1/\rho$ instead of ρ. This follows from the fact that $\gamma = \mu - \lambda$ in this special case. We refer to [4] and [26] for more background.

3.2 Heavy-tailed service times

The first heavy-tailed result for the $GI/G/1$ queue with the FCFS discipline is due to Borovkov [7] and Cohen [15]. Cohen proved that the waiting time distribution is regularly varying of index $\alpha - 1$ if and only if the service time distribution is regularly varying of index α, $\alpha > 1$. The *if* part was already proven in [7]. The proof in [15] exploits a relation between the waiting time in the $GI/G/1$ queue and a renewal function. The renewals in this function represent idle periods in the "dual" unstable $GI/G/1$ queue (a queue with A and B interchanged). In addition the proof exploits a lemma on regular variation of a transient renewal function [14].

Pakes [42] extended this result to the case of sub-exponential *residual* service times; see also Veraverbeke [48] and the nice proof in [52]. We present the result in a slightly more general form due to Korshunov [33]. The following statements are equivalent: The waiting time tail in $GI/G/1$ FCFS is sub-exponential; the tail of the residual service time is sub-exponential; and

$$P(W > x) \sim \frac{\rho}{1-\rho} P(B^r > x). \tag{6}$$

Here B^r denotes a residual service time, with distribution $P(B^r > x) = \frac{1}{EB} \int_x^\infty P(B > u) \mathrm{d}u$. In the $M/G/1$ case (6) easily follows from the well-known representation of W as a geometric sum of residual service times. If, e.g., B has a Pareto distribution, asymptotically behaving like $x^{-\alpha}$, then the waiting time tail asymptotically behaves like $x^{1-\alpha}$.

It is interesting to observe that the tail of W is as heavy as the tail of the residual service time, which is the integrated tail of the service time, and hence heavier than the tail of the service time.

Formula (6) can be heuristically explained, cf. Figure 1. Consider for ease of exposition the case of Poisson arrivals. Suppose that the workload is larger than a large value x at time 0. We claim that this is most likely due to the arrival of a single customer with a large service time B at some time $-v \leq 0$ (one *catastrophe*, in stark contrast with the *conspiracy* in the light-tailed setting). Just before that time, the workload was $O(1)$, and after that time nothing exceptional happened, either; hence the workload drifts down to $B - (1-\rho)v$ at time 0. The intensity of the occurrence of the event that this exceeds x is $\lambda P(B > x + (1-\rho)v)$. Integrating over all v yields the righthand side of (6). These heuristics actually form the basis for a rigorous proof, cf. [52] and p. 30 of [54]. Related insights are discussed in [2] and [5]. The principle described here is sometimes called *the principle of a single big jump*, cf. [25].

3.3 Optimality properties of FCFS

A general folk theorem for FCFS scheduling is that it is an efficient scheduling discipline for job sizes with low variability. Since low variability is typically associated with light tails it may not come as a surprise that FCFS has some tail

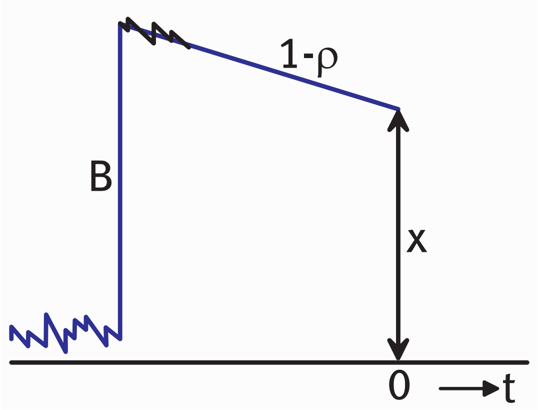

Figure 1: The principle of a single big jump

optimality properties for light tails. Stolyar & Ramanan [47] have shown that FCFS is logarithmically optimal for light-tailed job sizes. Their result applies in a more general setting than the $GI/G/1$ system. For the $GI/G/1$ system itself, it can actually be shown that FCFS is weakly optimal for light tails: See Appendix A. Whether FCFS is also *strongly* optimal is an open problem.

For heavy-tailed service times, FCFS is far from optimal: many jobs can have a huge response time by getting stuck behind a single large job. In a fundamental paper, Anantharam [3] shows that FCFS, as well as any other non-preemptive service discipline, is not effective in dealing with heavy-tailed job sizes. It can easily be shown that FCFS fails to be optimal for regularly varying job sizes. Even logarithmic optimality does not hold: in the next section we will identify a preemptive service discipline for which the response time distribution has tail behavior of the order $x^{-\alpha}$ rather than $x^{1-\alpha}$, implying that the limit in (3) equals $(\alpha-1)/\alpha$. This behavior is actually the worst one can get, since the time it takes to empty a system, if the system is in equilibrium at time 0, also has tail behavior of order $x^{1-\alpha}$ (for the case of regularly varying service times this follows from [55]).

4. PROCESSOR SHARING

Under the processor sharing (PS) service discipline, the total available service rate is equally shared among all users present: When $n \geq 1$ users are present, each of them receives service at rate $1/n$. PS was introduced to model round-robin scheduling mechanisms in time-shared computer systems [32], and is presently being used to model the flow level performance of bandwidth-sharing protocols [28, 37].

It is notoriously difficult to derive the distribution of the sojourn time under PS; see [51] for a survey concerning the $M/G/1$ case. Recent studies have focused on asymptotics. In our discussion we again distinguish between light-tailed and heavy-tailed service requirements. See [10] for a methodologically oriented survey on PS asymptotics.

4.1 Light-tailed service requirements

Obtaining asymptotic estimates of the tail of the response time in PS systems with light-tailed job sizes is much more difficult than in the heavy-tailed case. The reason (as explained in [36, 10]) is that a large response time can be affected by any of the following three types of events: (i) the arrival has a large service requirement, (ii) the arrival sees a large number of customers present, and (iii) many new jobs arrive during the arrival's sojourn time. Usually, only a single event determines the tail behavior of the response time distribution: In Section 4.2 we shall see that for PS with heavy-tailed job sizes event (i) is the dominating cause of a large response time. In FCFS, event (ii) is the most likely cause of a large response time for both light-tailed and heavy-tailed job sizes (although the way this particular event occurs is completely different in these cases). Events of type (i) and (iii) can be ignored in FCFS. In PS with light-tailed job sizes, all three events can play a significant role - sometimes simultaneously. This leads to a more challenging analysis and more complicated results. For example, for the $M/M/1$ PS system, the following remarkable asymptotic expansion holds:

$$P(V > x) \sim C_1 x^{-5/6} \mathrm{e}^{-C_2 x^{1/3}} \mathrm{e}^{-\gamma_{BP} x}. \tag{7}$$

C_1 and C_2 are some complicated constants, and γ_{BP} is the decay rate of a busy period, and equals $(\sqrt{\mu}-\sqrt{\lambda})^2$ in this case (recall that λ is the arrival rate and μ the service rate). This result has been derived in [8], exploiting a link between $M/M/1$ PS and $M/M/1$ Random Order of Service; a detailed analytic study of the latter model can be found in [23].

This result, as well as its proof, does not offer much insight into *how* large response times occur. Motivated by this, [36] considers the more general setting of the $GI/G/1$ queue. The following estimate is the main result of that work:

$$\log P(V > x) \sim -\gamma_{BP} x. \tag{8}$$

The method used to prove this result is that of large deviations, and shows that the event $\{V > x\}$ is primarily caused by an event of type (iii): After time 0, the load of the system increases from ρ to 1. The system becomes critically loaded, causing the number of customers to grow, and the service rate per customer to drop. The number of customers at time 0 and the service requirement of the job itself are both of $o(x)$, and do not contribute to the logarithmic asymptotics (8). We conjecture that these quantities *do* play a (complicated) role when considering asymptotics of the type (7).

This intuition is only valid if the tail of the job size distribution is not too light, which is surprising since such a distinction was not necessary in the FCFS system. The $M/D/1$ PS queue has been considered in [21]. Using a relation with Yule processes and geometric random sums, it is shown there that

$$P(V > x) \sim C_3 \mathrm{e}^{-\gamma_{M/D/1-PS} x}, \tag{9}$$

for constants C_3 and $\gamma_{M/D/1-PS}$. It can be shown that $\gamma_{M/D/1-PS}$ is strictly larger than the decay rate $\gamma_{M/D/1-BP}$ of the busy period in the $M/D/1$ PS queue, and strictly smaller than $\gamma_{M/D/1-FCFS}$. It can also be shown that the way the event $\{V > x\}$ occurs is by a combination of events (ii) and (iii). Specifically, the number of customers at time 0 is proportional to x. The result (9) has been extended in [20] to the conditional response time in the $M/G/1$ PS queue. Another recent paper on light-tailed asymptotics in PS systems is [19], where the result (8) has been extended to a multiclass system with discriminatory processor sharing and a possibly fluctuating service rate.

4.2 Heavy-tailed service requirements

In the heavy-tailed case the following asymptotic equivalence has been established for various PS models:

$$P(V > x) \sim P(B > \kappa x), \tag{10}$$

for some constant κ, which equals $1 - \rho$ in work-conserving PS systems. Thus, if B has a power tail $x^{-\alpha}$, V also has a power tail of the form $x^{-\alpha}$ (with different constant). In simple terms: the tail of V is just as heavy as the tail of B.

For $M/G/1$ with regularly varying service requirements, [56] proves this result with $\kappa = 1 - \rho$; [38] extends it to the intermediately regularly varying case, and [31] to subexponential concave service distributions.

Van Ooteghem et al. [41] develop a probabilistic sample-path approach that avoids the explicit use of queue length information, and that yields (10) with $\kappa = 1 - \rho$ for the processor sharing queue with general renewal (non-Poisson) arrival process, and also for the queue with *discriminatory* processor sharing – at the price of assuming that the service times are regularly varying of index $\alpha > 2$ (instead of $\alpha > 1$). In [27], (10) is proven for model extensions which allow for admission control, impatience and multiple servers; κ may now differ from $1 - \rho$. See also the survey [10].

A heuristic explanation of (10) is the following. For the heavy-tailed service distributions which are considered in [31, 38, 56], the most likely way to have a very large sojourn time is that the tagged customer itself has a very large service requirement. Its sojourn time consists of that service requirement plus the amount of service provided to other customers during V. V is so long that "steady state is reached" in an early stage of the tagged customer's presence. Hence on average the server allocates a fraction ρ of its capacity to other work, so that $V \approx B + \rho V$, or $V \approx B/(1-\rho)$: the tagged customer on average experiences service at rate $1 - \rho$. Equation (10) is therefore sometimes called a *reduced service rate* approximation, cf. Figure 2. It also shows that the most likely way a job experiences a large response time, is that its own service time is large, which is event (i) mentioned above. An informal way to summarize this is: *If you stay in the system for a long time, it's your own fault.*

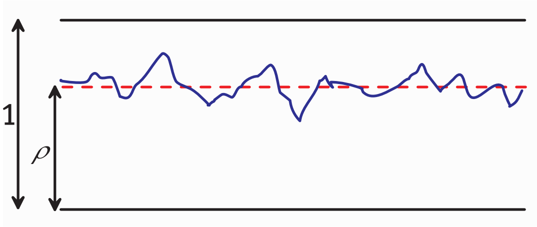

Figure 2: Reduced service rate approximation

A different way to arrive at (10) is as follows. Consider a customer with infinitely large job size (i.e. a permanent customer) that arrives in the system at time 0. Let $R(x)$ be the total amount of service obtained by that customer up to time x. We can write $P(V > x) = P(B > R(x))$, and the central limit theorem for renewal reward processes yields $R(x) \approx (1-\rho)x + O(\sqrt{x})$. Consequently, $P(V > x) = P(B > (1-\rho)x + O(\sqrt{x}))$. The equivalence $P(B > (1-\rho)x + O(\sqrt{x})) \sim P(B > (1-\rho)x)$, which is required to conclude (10), is called *square root insensitivity*. Regularly varying tails and log-normal tails are both square root insensitive. The Weibull tail e^{-x^β} is square root insensitive if and only if $\beta < 1/2$. The seminal work [31] makes the above heuristics rigorous, and also shows that the result (10) does not hold for sub-exponential distributions which are not square root insensitive.

In the case of an $M/G/1$ PS queue with *multiple* customer classes, (10) also holds for each individual class that has a regularly varying service requirement distribution – even if some other classes are "heavier". In more general models, the intuition behind (10) remains the same, but the results get harder to prove. We refer to [10] for (i) a discussion of various other extensions of (10) to multiclass generalizations of PS like DPS (Discriminatory Processor Sharing), (ii) an overview of the various methods via which results like (10) have been proven, and (iii) a discussion of the intimate relationship between (10) and a geometrically bounded queue length distribution. See also the paper by Aalto et al. in this volume.

4.3 Optimality properties of PS

Since the sojourn time for any service discipline is bounded from below by the job size, and since

$$\limsup_{x\to\infty} \frac{P(V_{PS} > x)}{P(B > x)} = \limsup_{x\to\infty} \frac{P(B > \kappa x)}{P(B > x)} < \infty,$$

if the tail of B is of (intermediate) regular variation, it follows that PS is weakly optimal for job size distributions of which the tail is (intermediate) regularly varying. Our conjecture is that PS is strongly optimal as well; it seems hard to improve upon $\kappa = 1 - \rho$ in the $GI/G/1$ queue. Proving this conjecture is a subject of current research.

For light-tailed service times, it can be shown from the above results that PS is not even logarithmically optimal. The reason is that the decay rate γ_{BP} is strictly smaller than the decay rate γ_{FCFS} for FCFS.

5. OTHER SERVICE DISCIPLINES

In this section we give an overview of existing results for other service disciplines. Size-based disciplines are considered in Section 5.1. Section 5.2 focuses on results for systems with multiple servers. Finally, we consider Random Order of Service and Last Come First Served in Sections 5.3 and 5.4. The results in the first two subsections are related to the subject of other surveys (on fairness and multi-server scheduling) in this special issue.

5.1 Size-based scheduling disciplines

Size-based scheduling disciplines make scheduling decisions based on the size of a job. A well-known example is Shortest Remaining Processing Time (SRPT). A related example is Foreground Background (FB) Processor Sharing, which serves (with preemption and according to PS) the jobs with the least attained service. The tail behavior of the response time for these scheduling disciplines has been derived for the $GI/G/1$ queue, both in the case of light-tailed and heavy-tailed job sizes [39]. The following two results which are proven in [39], covering light-tailed and heavy-tailed job sizes, are not only valid for FB and SRPT, but also for the class of SMART disciplines [50] discussed by Wierman in this volume.

Result 1. For light-tailed job sizes, the following holds if the distribution of B has no mass in its right endpoint (that is, if $P(B = x_B) = 0$, with $x_B = \sup\{x : P(B > x) > 0\}$):

$$\log P(V > x) \sim -\gamma_{BP} x. \qquad (11)$$

We see that the decay rate is the same as the decay rate for the busy period, suggesting that size-based scheduling policies are not effective in preventing large sojourn times when job sizes are light-tailed. The intuition behind this result is that the sojourn time of a job is upper bounded by a residual busy period, and is lower bounded by a residual busy period of jobs with size smaller than y, if the job size $B = y$. This also forms the basis of the proof.

If $P(B = x_B) > 0$, this result is not valid. An example is the $M/D/1$ queue, where SRPT coincides with FCFS. In [40] it is shown that the decay rate of V is strictly between γ_{BP} and γ_{FCFS} in this case.

These results can be refined to obtain the decay rate of $V(\tau)$, the sojourn time of a job with fixed service time τ. If $P(B = \tau) = 0$, then

$$\log P(V(\tau) > x) \sim -\gamma(\tau)x, \qquad (12)$$

for some function $\gamma(\tau)$. Under weak regularity assumptions, this function is continuous, strictly decreasing, converging to ∞ if $\tau \downarrow 0$ and converging to γ_{BP} if $\tau \to \infty$. There exists a critical τ^* for which $\gamma(\tau^*) = \gamma_{FCFS}$. The interpretation of τ^*, which is called the *critical job size*, is that jobs with size smaller than τ^* benefit from switching from FCFS to SRPT. For the $M/M/1$ queue, it is shown numerically in [40] that the percentage of such customers exceeds 85 for all values of the system load.

Result 2. Another appealing property of the class of size-based policies considered in [39] is that, for heavy-tailed job sizes, the performance is as good as for PS. The following is valid for (intermediately) regularly varying job size distributions:

$$P(V > x) \sim P(B > (1 - \rho)x). \qquad (13)$$

The intuition behind this result is similar as for PS, with a small variation: The tagged job essentially receives the lowest priority, and only gets served when it is alone in the system, i.e. during the idle periods of the server. The fraction of time the server is idle in the interval $[0, x]$ is approximately $1 - \rho$, hence the event $\{V > x\}$ can be related to $\{B > (1 - \rho)x\}$.

5.2 Multiple servers

The $GI/G/s$ queue can be seen as a special case of the model described in Section 1, if we relax the assumption that work is processed at speed 1 whenever there is work in the system. The total capacity of "the server" is divided by s, and each server works at fixed rate $1/s$, independently of the other servers. Note however that this scheduling discipline is not work-conserving. For light-tailed job sizes, the tail behavior of the waiting time distribution is derived in [45].

For heavy-tailed job sizes, the tail behavior is only known for the case of two servers, see [24] for two identical servers and [11] for one general server and one exponential server. The results in those papers show that the tail of the waiting time distribution crucially depends on the value of the load ρ. If $\rho \in (1/2, 1)$, then a single large job is sufficient to destabilize the system. For regularly varying job sizes, it can be shown that there exists a constant C_4 such that

$$P(W > x) \sim C_4 P(B^r > x). \qquad (14)$$

In the case $\rho \in (0, 1/2)$, the behavior is different, since two huge jobs are necessary to cause a long delay: In that case, it follows for the $GI/G/2$ queue (see [24]) that there exists a constant C_5 such that

$$P(W > x) \sim C_5 P(B^r > x)^2. \qquad (15)$$

This set of results is consistent with finite moment conditions for W, which are derived in [46]. What is clear is the benefit of a *spare server*. With a spare server the impact of a single large job can be eliminated. If $\rho < 1/2$, the tail of W is even lighter than the tail of B. This implies that, if $\rho < 1/2$, $P(V > x) \sim P(B > x/2)$ (since each server works at rate $1/2$). Thus in this special case, two-server scheduling is weakly optimal. It is not strongly optimal, since the constant $1/2$ is not as good as the constant $1 - \rho$ appearing in the sojourn time asymptotics for PS. More results on multi-server scheduling can be found in the contribution of Squillante in this volume.

5.3 Random Order of Service

For exponential service time distributions, the results for Random Order of Service coincide with PS, as described in Section 4.1. This equivalence might be asymptotically true for more general light-tailed distributions, but we are not aware of any result in this domain. In [13] the waiting time tail is studied for the $GI/G/1$ queue with heavy-tailed job sizes. In the $M/G/1$ case, with regularly varying service times, the results take a simple form:

$$P(W > x) \sim \frac{\rho}{1 - \rho} h(\rho) P(B^r > x), \qquad (16)$$

with $h(\rho) \leq 1$; this should be compared with (6).

5.4 Last Come First Served

The preemptive version of Last Come First Served (LCFS) yields response times with distribution identical to that of the length of the $GI/G/1$ busy period. For light-tailed service times, it is shown in [43] that, under weak regularity conditions,

$$P(V > x) \sim C_6 x^{-1/2} e^{-\gamma_{BP} x}. \qquad (17)$$

The weaker result $\log P(V > x) \sim -\gamma_{BP} x$ has been proven (without regularity conditions) in [40]. The logarithmic tail estimate also holds for the preemptive version of LCFS.

For heavy-tailed service times, a similar tail equivalence of W and B^r as for Random Order of Service above is seen to hold for Last Come First Served Non-Preemptive, and in fact also for other disciplines for which an arriving customer may have to wait a residual service time before its service begins. For the preemptive version, the following result holds:

$$P(V > x) \sim E[N] P(B > x(1 - \rho)). \qquad (18)$$

The constant $E[N]$ is the expected number of customers served during a busy period. This result has been shown in [34] for regularly varying service times, and Poisson arrivals. In this special case, $E[N] = 1/(1 - \rho)$. This result has been extended in [55] to intermediate regular variation and renewal arrivals. In that paper, it is also shown that a large busy period is strongly related to a large cycle maximum. This has been generalized further to several classes of square root insensitive and sub-exponential service time distributions in [6, 30, 18].

6. OUTLOOK

In the case of a regularly varying service requirement distribution of index α, we have seen that some service disciplines give rise to waiting time tails that are regularly varying with the same index α, whereas for other service disciplines the index becomes $\alpha - 1$. We have also argued that it cannot get worse than $\alpha - 1$, as this is also the index for the residual busy period. This leaves the question whether any value between $\alpha - 1$ and α can be assumed. In [12] we show that this is indeed the case. E.g., one could split a job of size $x > 1$ into $x^{1-\beta}$ pieces of size x^{β}, $0 < \beta < 1$, and when one piece of a job is served, the remainder of the job moves back to the end of the queue. Any value ζ of the index of the waiting time tail between $\alpha - 1$ and α may be obtained by choosing $\beta = \alpha/(1+\zeta)$. The intuition behind this is the following. The most likely way to experience a long delay is to arrive during a long service piece. Each piece follows a power law with exponent $-\alpha/\beta$, and the residual of a piece follows a power law with exponent $-\zeta = 1 - \alpha/\beta$. Taking $\beta = 1$ yields $\zeta = \alpha - 1$, and $\beta = \alpha/(1+\alpha)$ would yield $\zeta = \alpha$, but for smaller values of β the length of the arriving job itself becomes the dominating tail factor.

A topic for future research is to better understand optimality properties w.r.t. tail behavior of scheduling disciplines. Several open problems have been mentioned in this survey. Another open problem in this setting is the existence and construction of a "universally good" scheduling discipline: a scheduling discipline which is, in some sense, optimal for both light tails and heavy tails. Such a scheduling discipline may be implemented when no information about the exact tail behavior of the service distribution is available.

APPENDIX

A. WEAK OPTIMALITY OF FCFS

We give a proof sketch of the weak optimality property of FCFS, which is based on a result for the tail behavior of the maximum amount of work Q in the system during a busy cycle. This quantity is the same for all work-conserving disciplines. Under the assumption (4), it is shown in [29] that $P(Q > x) \sim C_Q e^{-\gamma_{FCFS} x}$ for a constant $C_Q > 0$. Let π be an arbitrary work-conserving discipline; N be the number of customers served in a busy cycle; $V_{\pi,i}, i \geq 1$ be the response time of the ith job in the system; and $I(A)$ be the indicator function of the event A. Observe that

$$\begin{aligned} P(V_\pi > x) &= \frac{1}{E[N]} E[\#\{i \leq N : V_i > x\}] \\ &\geq \frac{1}{E[N]} E[\#\{i \leq N : V_i > x\} I(Q > x)] \\ &\geq \frac{1}{E[N]} P(Q > x). \end{aligned}$$

The last inequality is valid since, if the amount of work in the system is larger than x, at least one of the customers in the system at that time will have a response time exceeding x, implying $\#\{i \leq N : V_i > x\} \geq 1$. Finally, we note that it can be shown that $P(V_{FCFS} > x) \sim E[e^{\gamma B}] P(W_{FCFS} > x)$. Putting all pieces together, we obtain

$$\begin{aligned} \limsup_{x\to\infty} \frac{P(V_{FCFS} > x)}{P(V_\pi > x)} &\leq E[N] \limsup_{x\to\infty} \frac{P(V_{FCFS} > x)}{P(Q > x)} \\ &= E[N] \frac{E[e^{\gamma B}] C_{FCFS}}{C_Q}. \end{aligned}$$

We conclude that the weak optimality criterion (2) is satisfied with $M = E[N]E[e^{\gamma B}]C_{FCFS}/C_Q$. It can be verified that this constant is strictly larger than 1, leaving strong optimality as an open problem.

B. REFERENCES

[1] Abate, J., Choudhury, G.L., Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems* **16**, 311–338.

[2] Anantharam, V. (1988). How large delays build up in a $GI/G/1$ queue. *Queueing Systems* **5**, 345–368.

[3] Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems* **33**, 73–89.

[4] Asmussen, S. (2003). *Applied Probability and Queues.* Springer-Verlag, New York.

[5] Baccelli, F., Foss, S. (2004). Moments and tails in monotone-separable stochastic networks. *Ann. Appl. Probab.* **14**, 612–650.

[6] Baltrunas, A., Daley, D., Klüppelberg, C. (2004). Tail behaviour of the busy period of a $GI/G/1$ queue with subexponential service times. *Stoch. Proc. Appl.* **111**, 237–258.

[7] Borovkov, A.A. (1970). Factorization identities and properties of the distribution of the supremum of sequential sums. *Theor. Probability Appl.* **15**, 359–402.

[8] Borst, S.C., Boxma, O.J., Morrison, J.A., Núñez-Queija, R. (2003). The equivalence between processor sharing and service in random order. *Oper. Res. Lett.* **31**, 254–262.

[9] Borst, S.C., Boxma, O.J., Núñez-Queija, R., Zwart, A.P. (2003). The impact of the service discipline on delay asymptotics. *Perf. Eval.* **54**, 175–206.

[10] Borst, S.C., Núñez-Queija, R., Zwart, A.P. (2006). Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* **53**, 31–51.

[11] Boxma, O.J., Deng, Q., Zwart, A.P. (2002). Waiting-time asymptotics for the $M/G/2$ queue with heterogeneous servers. *Queueing Systems* **40**, 5–31.

[12] Boxma, O.J., Denisov, D., Resnick, S.I. (2007). Paper in preparation.

[13] Boxma, O.J., Foss, S.G., Lasgouttes, J.-M., Núñez-Queija, R. (2004). Waiting time asymptotics in the single server queue with service in random order. *Queueing Systems* **46**, 35–73.

[14] Callaert, H., Cohen, J.W. (1972). A lemma on regular variation of a transient renewal function. *Z. Wahrscheinlichkeitstheorie u. Verw. Gebiete* **24**, 275–278.

[15] Cohen, J.W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* **10**, 343–353.

[16] Cohen, J.W. (1982). *The Single Server Queue.* North-Holland Publ. Cy., Amsterdam; revised edition.

[17] Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In: *Proc. ACM Sigmetrics '96*, 160–169.

[18] Denisov, D., Shneer, V. (2006). Asymptotics for first passage times of Lévy processes and random walks. Eurandom research report 2006-017.

[19] Egorova, R., Mandjes, M., Zwart, B. (2006). Sojourn

time asymptotics in processor sharing queues with varying service rate. Submitted for publication.

[20] Egorova, R., Zwart, B. (2007). Tail asymptotics for conditional sojourn times in Processor Sharing queues. *Queueing Systems* **55**, to appear.

[21] Egorova, R., Zwart, A.P., Boxma, O.J. (2006). Sojourn time tails in the $M/D/1$ processor sharing queue. *Prob. Eng. Inf. Sciences* **20**, 429–446.

[22] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modeling Extremal Events.* Springer, New York.

[23] Flatto, L. (1997). The waiting time distribution for the random order service $M/M/1$ queue. *Ann. Appl. Prob.* **7**, 382–409.

[24] Foss, S., Korshunov, D. (2006). Heavy tails in multi-server queue. *Queueing Systems* **52**, 31–48.

[25] Foss, S., Konstantopoulos, T., Zachary, S. (2007). The principle of a single big jump: discrete and continuous time modulated random walks with heavy-tailed increments. *J. Theor. Prob.*, to appear.

[26] Ganesh, A., O'Connell, N., Wischik, D. (2003). *Big Queues.* Springer, New York.

[27] Guillemin, F., Robert, Ph., Zwart, A.P. (2004). Tail asymptotics for processor sharing queues. *Adv. Appl. Prob.* **36**, 525–543.

[28] Heyman, D.P., Lakshman, T.V., Neidhardt, A.L. (1997). A new method for analysing feedback-based protocols with applications to engineering Web traffic over the Internet. In: *Proc. ACM Sigmetrics '97*, 24–38.

[29] Iglehart, D. (1972). Extreme values in the $GI/G/1$ queue. *Ann. Math. Statist.* **43**, 627–635.

[30] Jelenković, P.R., Momčilović, P. (2004). Large deviations of square-root insensitive random sums. *Math. Oper. Res.* **29**, 398–406.

[31] Jelenković, P.R., Momčilović, P. (2003). Large deviation analysis of subexponential waiting times in a processor sharing queue. *Math. Oper. Res.* **28**, 587–608.

[32] Kleinrock, L. (1964). Analysis of a time-shared processor. *Nav. Res. Log. Quarterly* **11**, 59–73.

[33] Korshunov, D.A. (1997). On distribution tail of the maximum of a random walk. *Stoch. Proc. Appl.* **72**, 97–103.

[34] De Meyer, A., Teugels, J.L. (1980). On the asymptotic behaviour of the distribution and the service time in $M/G/1$. *J. Appl. Prob.* **17**, 802–813.

[35] Mandjes, M., Nuyens, M. (2005). Sojourn times in the $M/G/1$ FB queue with light-tailed service times. *Prob. Eng. Inf. Sciences* **19**, 351–361.

[36] Mandjes, M., Zwart, B. (2006). Large deviations for sojourn times in processor sharing queues. *Queueing Systems* **52**, 237–250.

[37] Massoulié, L., Roberts, J.W. (1999). Bandwidth sharing: Objectives and algorithms. In: *Proc. IEEE Infocom '99*, 1395–1403.

[38] Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. *Ann. Oper. Res.* **113**, 101–117.

[39] Nuyens, M., Wierman, A., Zwart, A.P. (2007). Preventing large sojourn times with `SMART` scheduling. *Oper. Res.*, to appear.

[40] Nuyens, M., Zwart, A.P. (2006). A large deviations analysis of the $GI/G/1$ SRPT queue. *Queueing Systems* **54**, 85–97.

[41] Van Ooteghem, D.T.M.B., Zwart, A.P., Borst, S.C. (2005). Tail asymptotics for discriminatory processor-sharing queues with heavy-tailed service requirements. *Perf. Eval.* **61**, 281–298.

[42] Pakes, A.G. (1975). On the tails of waiting-time distributions. *J. Appl. Prob.* **12**, 555–564.

[43] Palmowksi, Z., Rolski, T. (2006). On busy period asymptotics in the $GI/G/1$ queue. *Adv. Appl. Prob.* **83**, 92–103.

[44] Park, K., Willinger, W. (eds.) (2000). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York.

[45] Sadowsky, J.S., Szpankowski, W. (1990). On the analysis of the tail queue length and waiting time distributions of a $GI/G/c$ queue. In: *Proc. Performance '90*, 93–107.

[46] Scheller-Wolf, A., Vesilo, R. (2006). Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FCFS multiserver queues. *Queueing Systems* **54**, 221–232.

[47] Stolyar, A., Ramanan, K. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Ann. Appl. Probab.* **11**, 1–48.

[48] Veraverbeke, N. (1977). Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stoch. Proc. Appl.* **5**, 27–37.

[49] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM *Trans. Netw.* **5**, 71–86.

[50] Wierman, A., Harchol-Balter, M., Ogasami, T. Nearly insensitive bounds on SMART scheduling. In: *Proc. ACM Sigmetrics'05*, 205–216.

[51] Yashkov, S.F. (1987). Processor-sharing queues: Some progress in analysis. *Queueing Systems* **2**, 1–17.

[52] Zachary, S. (2004). A note on Veraverbeke's theorem. *Queueing Systems* **46**, 9–14.

[53] Zwart, A.P. (1999). Sojourn times in a multiclass processor sharing queue. In: *Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 335–344.

[54] Zwart, A.P. (2001). *Queueing Systems with Heavy Tails.* Ph.D. thesis, Eindhoven University of Technology. Available at http://alexandria.tue.nl/extra2/200112999.pdf

[55] Zwart, A.P. (2001). Tail asymptotics for the busy period in the $GI/G/1$ queue. *Math. Oper. Res.* **26**, 485–493.

[56] Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* **35**, 141–166.