Queues with Equally Heavy Sojourn Time and Service Requirement Distributions

Rudesindo Núñez-Queija
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Eindhoven University of Technology, Dept. Math. & Comp. Sc., The Netherlands

ABSTRACT
For the G/G/1 queue with First-Come First-Served, it is well known that the tail of the sojourn time distribution is heavier than the tail of the service requirement distribution when the latter has a regularly varying tail. In contrast, for the M/G/1 queue with Processor Sharing, Zwart and Boxma [26] showed that under the same assumptions on the service requirement distribution, the two tails are "equally heavy". By means of a probabilistic analysis we provide a new insightful proof of this result, allowing for the slightly weaker assumption of service requirement distributions with a tail of intermediate regular variation. The new approach allows us to also establish the "tail equivalence" for two other service disciplines: Foreground-Background Processor Sharing and Shortest Remaining Processing Time. The method can also be applied to more complicated models, for which no explicit formulas exist for (transforms of) the sojourn time distribution. One such model is the M/G/1 Processor Sharing queue with service that is subject to random interruptions. The latter model is of particular interest for the performance analysis of communication networks.

2000 Mathematics Subject Classification: 60K25, 68M20, 90B18, 90B22.
Keywords & Phrases: Processor Sharing; Foreground-Background Processor Sharing; Shortest Remaining Processing Time; Queues with Heavy Tails; (Intermediate) Regular Variation; Telecommunications.
Note: The work was carried out within PNA 2.

1 Introduction

Cohen [7] showed that in the G/G/1 queue with the *First-Come First-Served* (FCFS) discipline, the waiting-time distribution is regularly varying of index $1-\zeta$ if and only if the distribution of the service requirements is regularly varying of index $-\zeta$, where $\zeta > 1$. (A formal definition of regularly varying distributions is given in Appendix A.) Thus, the tail of the waiting-time distribution (and, hence, the sojourn time distribution) is "as heavy" as the *integrated tail* distribution of the service requirements and, therefore, heavier than the tail of the service requirement distribution itself. In particular, the m-th ($m > 0$) moment of the sojourn time distribution is finite if and only if the $(m+1)$-st moment of the service requirement distribution is finite.

Assuming Poisson arrivals and a regularly varying tail of the service requirement distribution with index $\zeta \in (1, 2)$, Anantharam [1] has shown that the mean of the sojourn time is infinite for any *non-preemptive* service discipline. In contrast, Anantharam [1] also showed that, under the same assumptions on the arrival process and the service requirements, there exist preemptive service disciplines for which the mean sojourn time is finite. More specifically, Zwart and Boxma [26] proved that in the M/G/1 queue with *Processor Sharing* (PS) the tail of the sojourn time distribution is *exactly as heavy as* that of the service requirement distribution when the latter has a regularly varying tail.

In this paper we provide an alternative *probabilistic* proof of Zwart and Boxma's "tail-equivalence" result (the original proof was based on Laplace transform techniques) and we extend it in several directions. Our approach allows for the slightly larger class of *intermediate* regularly varying service requirement distributions (see Assumption 2.1 for a definition). This shows that the tail equivalence still applies in cases where the tail of the service requirement distribution "fluctuates" between those of Pareto distributions with different indexes[1]. Besides PS, we show that the tail-equivalence also holds when the service discipline is either *Foreground-Background Processor Sharing* (FBPS) or *Shortest Remaining Processing Time* (SRPT). Our main interest being in PS, the latter two disciplines are only considered in the case that the variance of the service requirements is infinite. In the proofs, we use known expressions for the moments of the sojourn time distribution [21, 23, 24]. We emphasize that, although in the literature these expressions were derived from the Laplace transforms of the sojourn time distributions, they can alternatively be derived by solving simple ordinary differential equations. This allows us to apply the method to more complicated models where Laplace transform expressions are not available. In particular, the method's flexibility was illustrated in [17, Ch. 5] by establishing the tail equivalence for an M/G/1-PS model with random service interruptions.

Our method is based on establishing a relationship between a customer's sojourn time and its service requirement. Henceforth, the sojourn time of customers having a given service requirement will be referred to as the *conditional* sojourn time. We shall provide conditions on the moments of the conditional sojourn time distribution which imply that the tails of the sojourn time and the service requirement distributions are equally heavy when the latter is of intermediate regular variation (this will be made precise below).

Independent of the present analysis, Jelenković and Momčilović [12] used a related large deviations analysis to extend the tail equivalence result for the M/G/1-PS model to cases in which the tail of the service requirement distribution is "lighter" than that of Pareto distributions.

The analysis of queueing models with heavy tailed traffic characteristics is an important issue in the performance analysis of data communication systems. Heavy tails in the 'input' processes provide an explanation for phenomena such as *long-range dependence* and *self-similarity* that have been observed in data traffic measurements [14]. Moreover, in the context of telecommunications, the PS discipline is particularly relevant as it provides a justifiable modeling assumption for the way capacity is allocated to so-called *elastic traffic* flows [15, 17], which constitute the bulk of traffic in modern communication networks. The capacity available to elastic traffic is, however, not constant in time, but varies due to other applications that also share in the network's resources. PS systems in which the service capacity varies over time are, however, hard to analyze and closed-form expressions for performance measures such as (transforms of) the sojourn time distribution are only available in specific cases [16, 17]. For this reason it is important to develop methods that allow for analysis of such models, not relying on closed-form expressions. The flexibility of the technique presented in this paper is illustrated in [17, Ch. 5], where it is applied to establish the tail equivalence result in a modified M/G/1-PS model, with service that alternates between availability periods (exponentially distributed) and unavailability periods (generally distributed). For that model, even basic measures such as the mean number of customers in the system and the mean sojourn time are not known.

The structure of the remainder of the paper is as follows. Section 2 presents our approach in a general context. Intermediate results that are needed to apply the techniques of Section 2 to various

[1] An example of such a distribution is $H(x) = 1 - x^{\alpha_1 (\sin(\ln(\ln(x+1))) - \alpha_2)}$, $x \geq 0$, where $\alpha_1 > 0$ and $\alpha_2 \geq \sqrt{2}$. This distribution fluctuates between Pareto distributions with indexes $-\alpha_1(\alpha_2 - 1)$ and $-\alpha_1(\alpha_2 + 1)$

queueing models are gathered in Section 3. In Sections 4, 5 and 6 we show the tail equivalence for the M/G/1 queue with PS, FBPS and SRPT, respectively. The discussion in Section 7 provides an intuitive explanation of the results and Section 8 concludes the paper.

2 Sufficient conditions for tail equivalence

We state the main result in a general setting. Let B be a non-negative random variable with distribution function $B(x)$, $x \geq 0$. For $\tau \geq 0$ let $V(\tau) \geq 0$ be a non-negative random process independent of B. We will be interested in $V(B)$, i.e., the value of the process $V(\cdot)$ at the stopping time B. The random variable $V(B)$ is well defined and its distribution function is given by

$$\mathbf{P}\{V(B) \leq t\} \quad = \quad \int_{\tau=0}^{\infty} \mathbf{P}\{V(\tau) \leq t\} \, \mathrm{d}B(\tau).$$

Remark 2.1 In the subsequent sections, $B(x)$ will represent the service requirement distribution and $V(\tau)$ will stand for the sojourn times of customers with service requirement τ. Consequently, the unconditional sojourn time of an arbitrary customer shall be distributed as $V(B)$.

Assumption 2.1 *The tail $\overline{B}(x) := 1 - B(x)$ of the distribution function $B(x)$ is of intermediate regular variation at infinity, i.e.,*

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

When this assumption is satisfied, we write $\overline{B}(x) \in \mathcal{IR}$. In particular, all functions with a regularly varying tail are of intermediate regular variation, see Cline [6] for a discussion. Assumption 2.1 implies that there exist numbers $\zeta \in (0, \infty)$, $x_0 \in (0, \infty)$, and $\eta \in (0, 1)$ such that, for all $x_2 \geq x_1 \geq x_0$,

$$\frac{\overline{B}(x_2)}{\overline{B}(x_1)} \quad \geq \quad \eta \left(\frac{x_2}{x_1} \right)^{-\zeta}, \tag{2.1}$$

see Appendix B.

Besides Assumption 2.1 on the distribution $B(x)$, we further impose on $V(\tau)$ the conditions in Assumption 2.2 below. In the subsequent sections we will show for several queueing models that Assumption 2.1 *implies* Assumption 2.2, where $B(x)$ and $V(\tau)$ have the interpretation given in Remark 2.1.

Assumption 2.2 *The following three conditions are satisfied:*

(i) For some $\overline{g} > 0$,

$$\lim_{\tau \to \infty} \frac{\mathbf{E}[V(\tau)]}{\tau} \quad = \quad \overline{g}. \tag{2.2}$$

(ii) For $\zeta \geq 0$ as in (2.1) there exist $\kappa > \zeta$ and $\delta > 0$ such that:

$$\lim_{\tau \to \infty} \tau^{-\kappa+\delta} \mathbf{E}\left[\left| \, V(\tau) - \mathbf{E}[V(\tau)] \, \right|^{\kappa} \right] \quad = \quad 0, \tag{2.3}$$

i.e.,

$$\mathbf{E}\left[\left| \, V(\tau) - \mathbf{E}[V(\tau)] \, \right|^{\kappa} \right] \quad = \quad \mathrm{o}(\tau^{\kappa-\delta}), \quad \tau \to \infty.$$

(iii) For all $t \geq 0$, the probability $\mathbf{P}\{V(\tau) > t\}$ is non-decreasing in $\tau \geq 0$. Hence, all moments $\mathbf{E}[V(\tau)^n]$, $n \in \mathbb{N}$, are non-decreasing in τ.

In the sequel, the constants $\overline{g}$, ζ, κ and δ will be as in Assumption 2.2. The main result is stated in Theorem 2.3. In its proof we use Lemmas 2.1 and 2.2 which rely on the following form of Markov's inequality (see Williams [22, Section 6.4]) for the tail distribution of $V(\tau)$:

$$\mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > t\} \leq \frac{\mathbf{E}\left[\left| V(\tau) - \mathbf{E}[V(\tau)] \right|^{\kappa}\right]}{t^{\kappa}}, \tag{2.4}$$

for all $\tau \geq 0$ and $t > 0$. When $\kappa = 2$, which is the case in two of the examples studied in the next sections, this reduces to Chebyshev's inequality:

$$\mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > t\} \leq \frac{\mathbf{Var}[V(\tau)]}{t^2}. \tag{2.5}$$

Our first lemma states that "when B is small, $V(B)$ can not be large".

Lemma 2.1 *Suppose Assumptions 2.1 and 2.2 are satisfied. Then, for fixed $\varepsilon \in (0,1)$,*

$$\lim_{x\to\infty} \frac{\mathbf{P}\{V(B) > \overline{g}\,x; B \leq x(1-\varepsilon)\}}{\mathbf{P}\{B > x(1-\varepsilon)\}} = 0.$$

Proof. For transparency of the presentation, the proof is deferred to Appendix C.

The following lemma complements the statements of Lemma 2.1 for the case that B is large.

Lemma 2.2 *If Assumption 2.2 is satisfied then, for all $\varepsilon > 0$,*

$$\lim_{x\to\infty} \frac{\mathbf{P}\{V(B) > \overline{g}\,x; B > x(1+\varepsilon)\}}{\mathbf{P}\{B > x(1+\varepsilon)\}} = 1.$$

Proof. Clearly, the lim sup of the above expression can not be larger than 1. Therefore, it suffices to show that the lim inf is at least 1. By Assumption 2.2 we have, for all $\tau \geq x(1+\varepsilon)$,

$$\mathbf{P}\{V(\tau) > \overline{g}\,x\} \geq \mathbf{P}\{V(x(1+\varepsilon)) > \overline{g}\,x\}.$$

Hence,

$$\begin{aligned}\mathbf{P}\{V(B) > \overline{g}\,x; B > x(1+\varepsilon)\} &= \int_{\tau=x(1+\varepsilon)}^{\infty} \mathbf{P}\{V(\tau) > \overline{g}\,x\}\,\mathrm{d}B(\tau)\\ &\geq \mathbf{P}\{V(x(1+\varepsilon)) > \overline{g}\,x\}\,\mathbf{P}\{B > x(1+\varepsilon)\}.\end{aligned}$$

From (2.2) it follows that $\mathbf{E}[V(x(1+\varepsilon))] - \overline{g}\,x > 0$, for x large enough. By Markov's inequality:

$$\begin{aligned}&\mathbf{P}\{V(x(1+\varepsilon)) \leq \overline{g}\,x\}\\ &= \mathbf{P}\{\mathbf{E}[V(x(1+\varepsilon))] - V(x(1+\varepsilon)) \geq \mathbf{E}[V(x(1+\varepsilon))] - \overline{g}\,x\}\\ &\leq \frac{\mathbf{E}\left[\left| V(x(1+\varepsilon)) - \mathbf{E}[V(x(1+\varepsilon))] \right|^{\kappa}\right]}{\left(\mathbf{E}[V(x(1+\varepsilon))] - \overline{g}x\right)^{\kappa}},\end{aligned}$$

and, by (2.2) and (2.3), this vanishes as $x \to \infty$. Therefore,

$$\lim_{x\to\infty} \mathbf{P}\{V(x(1+\varepsilon)) > \overline{g}\,x\} = 1,$$

and the proof is completed. □

Together, Lemmas 2.1 and 2.2 enable us to prove our main result which is stated in the following theorem.

Theorem 2.3 *Suppose Assumptions 2.1 and 2.2 are satisfied. Then the tail distributions of the random variables B and $V(B)$ are equally heavy in the sense that:*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \overline{g}\,x\right\}}{\mathbf{P}\left\{B > x\right\}} = 1.$$

Proof. The proof is given in two parts. First we write, for $\varepsilon > 0$,

$$\begin{aligned}\mathbf{P}\left\{V(B) > \overline{g}\,x\right\} &\leq \mathbf{P}\left\{V(B) > \overline{g}\,x; B \leq x(1-\varepsilon)\right\} \\ &\quad + \mathbf{P}\left\{B > x(1-\varepsilon)\right\}.\end{aligned}$$

By Lemma 2.1 and the fact that $\overline{B}(x) \in \mathcal{IR}$ we may neglect the first term on the right-hand side. Hence,

$$\limsup_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \overline{g}\,x\right\}}{\mathbf{P}\left\{B > x\right\}} \leq \limsup_{x\to\infty} \frac{\mathbf{P}\left\{B > x(1-\varepsilon)\right\}}{\mathbf{P}\left\{B > x\right\}}.$$

Letting $\varepsilon \downarrow 0$, the right-hand side tends to 1. For the second part of the proof we write, for $\varepsilon > 0$,

$$\mathbf{P}\left\{V(B) > \overline{g}\,x\right\} \geq \mathbf{P}\left\{V(B) > \overline{g}\,x; B > x(1+\varepsilon)\right\}.$$

Combining this with Lemma 2.2, we have

$$\liminf_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \overline{g}\,x\right\}}{\mathbf{P}\left\{B > x\right\}} \geq \liminf_{x\to\infty} \frac{\mathbf{P}\left\{B > x(1+\varepsilon)\right\}}{\mathbf{P}\left\{B > x\right\}}.$$

By Assumption 2.1 the right-hand side tends to 1 as $\varepsilon \downarrow 0$. □

3 Intermediate results

Before applying the results of the previous section to various models, we review some common notation and technical results that shall be used in the sequel. We start by considering an M/G/1 queue. Customers arrive according to a Poisson process with intensity λ and their service requirement distribution is $B(x)$ with mean $\beta_1 < \infty$ and k^{th} moment $\beta_k \leq \infty$, $k > 1$. Service is rendered at rate 1 whenever the system is not empty. The traffic load is denoted by $\rho = \lambda\beta_1$ and we assume that the system is stable: $\rho < 1$. The random variables B, $V(\tau)$ and $V(B)$ will generically represent the service requirement, the sojourn time conditional on $B = \tau$, $\tau \geq 0$, and the unconditional sojourn time, respectively.

When the second moment of the service requirement distribution (β_2) is infinite, we need to impose the following conditions:

Assumption 3.1 $\mathbf{E}\left[B^{\alpha}\right] < \infty$ *for some* $\alpha \in (1,2)$.

Assumption 3.2 $\mathbf{E}\left[B^{\zeta}\right] = \infty$ *for some* $\zeta \in (1,2)$.

It is straightforward to see that when Assumption 3.2 is satisfied and $\overline{B}(x) \in \mathcal{IR}$, then (2.1) holds for the same choice of ζ. Assumption 3.1 implies that the tail of the service requirement distribution is dominated by a Pareto tail, i.e., that for some $\theta > 0$ and all x large enough, $\overline{B}(x)$ is smaller than $x^{-\theta}$. We formalize the latter statement in a more general context in the next lemma.

Lemma 3.1 *If Z is a non-negative random variable with $\mathbf{E}\left[Z^{\theta}\right] < \infty$, for some $\theta \in \mathbb{R}$, then*

$$\mathbf{P}\{Z > u\} \quad = \quad \mathrm{o}(u^{-\theta}),$$

for $u \to \infty$. Hence, there exists a number $u_0 > 0$ such that $\mathbf{P}\{Z > u\} \leq u^{-\theta}$, for all $u \geq u_0$.
Conversely, if $\mathbf{P}\{Z > u\} = \mathrm{o}(u^{-\theta})$, for $u \to \infty$, then $\mathbf{E}\left[Z^{\theta-\varepsilon}\right] < \infty$, for all $\varepsilon \in (0, \theta)$.

Proof. The first statement follows from the fact that

$$\begin{aligned}
&\lim_{u\to\infty} \left(u^{\theta}\mathbf{P}\{Z > u\} \right) \\
&= \lim_{u\to\infty} \left(\theta \int_{x=0}^{u} x^{\theta-1}\mathbf{P}\{Z > x\}\,\mathrm{d}x - \int_{x=0}^{u} x^{\theta}\mathrm{d}\mathbf{P}\{Z \leq x\} \right) \\
&= \mathbf{E}\left[Z^{\theta}\right] - \mathbf{E}\left[Z^{\theta}\right] \; = \; 0.
\end{aligned}$$

The existence of the number u_0 is trivial and the last statement follows from:

$$\mathbf{E}\left[Z^{\theta-\varepsilon}\right] \; = \; (\theta - \varepsilon) \int_{u=0}^{\infty} u^{\theta-\varepsilon-1}\mathbf{P}\{Z > u\}\,\mathrm{d}u \; < \; \infty.$$

□

In the sequel, the random variable $W_{\lambda,B}$ is distributed as the (steady-state) waiting time in the M/G/1 FCFS queue with arrival rate λ and service time distribution $B(x)$, i.e.,

$$\mathbf{P}\{W_{\lambda,B} \leq t\} = (1-\rho)\left(1 + \sum_{n=1}^{\infty} \rho^n \mathbf{P}\{B_{\mathrm{res},1} + \cdots + B_{\mathrm{res},n} \leq t\}\right),$$

cf. Cohen [8, Part II, Expression (4.82)]. Here, $B_{\mathrm{res},1}$, $B_{\mathrm{res},2}, \ldots$ represents a sequence of i.i.d. random variables, drawn from the forward recurrence distribution of the service requirements, i.e., for all k,

$$\mathbf{P}\{B_{\mathrm{res},k} \leq t\} = \frac{1}{\beta_1} \int_{x=0}^{t} \mathbf{P}\{B > x\}\,\mathrm{d}x.$$

It will be convenient to write the distribution of $W_{\lambda,B}$ as follows,

$$\mathbf{P}\{W_{\lambda,B} \leq t\} = (1-\rho)\sum_{n=0}^{\infty} \rho^n \left[\frac{1}{\beta_1}\int_{x=0}^{t} \mathbf{P}\{B > x\}\,\mathrm{d}x\right]^{n\star}, \tag{3.1}$$

where the symbol * denotes the convolution operator for probability distributions, i.e., for a distribution function $H(x)$, $x \geq 0$, we define $H(x)^{0\star} := 1$, for all $x \geq 0$, and for $n \in \mathbb{N}_0$ and $x \geq 0$,

$$H(x)^{(n+1)\star} \quad := \quad \int_{u=0}^{x} H(x-u)^{n\star}\mathrm{d}H(u). \tag{3.2}$$

The next lemma states a direct implication of Assumption 3.1 for the distribution of $W_{\lambda,B}$. This relation will be useful in the analysis of sojourn times in the case that $\beta_2 = \infty$.

Lemma 3.2 *Let $\alpha > 1$. If $\mathbf{E}[B^{\alpha}] < \infty$ then $\mathbf{E}\left[(W_{\lambda,B})^{\alpha-1}\right] < \infty$.*

Proof. Asmussen [2, Thm. VIII.2.1]. An alternative proof can be found in [17, Ch. 5]. □

4 Processor Sharing

In the M/G/1 PS queue, at any point in time all customers in the system share equally in the service capacity. For an overview on the literature on PS queues we refer to Yashkov [24, 25]. More recent references can be found in [17]. Here we are interested in the tail of the sojourn time distribution. In order to prove that the tails of the sojourn time and service requirement distributions are equally heavy, in the sense of Theorem 2.3, we will need to verify Assumption 2.2. For this reason we shall first list some known results for the moments of $V(\tau)$. It is well known that the mean of the conditional sojourn time is given by:

$$\mathbf{E}\left[V(\tau)\right] = \frac{\tau}{1-\rho}, \tag{4.1}$$

see Sakata et al. [19, Eq. (10)], Sakata et al. [20, Eq. (49)], or Kleinrock [13, Eq. (4.17)]. The variance of $V(\tau)$ is given by

$$\mathbf{Var}\left[V(\tau)\right] = \frac{2}{(1-\rho)^2}\int_{u=0}^{\tau}(\tau-u)\mathbf{P}\left\{W_{\lambda,B} > u\right\}\mathrm{d}u, \tag{4.2}$$

cf. Yashkov [23, Eq. (3.20)] and Zwart and Boxma [26, Eq. (3.5), (3.10)]. $W_{\lambda,B}$ is distributed as in (3.1). When $\beta_2 < \infty$ we have for $k = 2, 3, \ldots$, and $\tau \to \infty$,

$$\mathbf{E}\left[V(\tau)^k\right] = \mathbf{E}\left[V(\tau)\right]^k + \frac{\beta_2}{2\beta_1}\frac{\rho}{1-\rho}\frac{k(k-1)}{(1-\rho)^k}\tau^{k-1} + \mathrm{o}(\tau^{k-1}), \tag{4.3}$$

cf. Zwart and Boxma [26, Rem. 3.3].

In the literature these results have mostly been obtained from expressions derived for the Laplace-Stieltjes transform of $V(\tau)$. However, (4.1)–(4.3) can be obtained from a set of simple (integro-)differential equations instead of deriving the Laplace-Stieltjes transform of $V(\tau)$, see Yashkov [23, Rem. 3] for an outline of how this can be done for (4.2).

The differential equations rely on a well known decomposition of the sojourn time of a customer with service requirement τ that arrives to the system when n customers are present with remaining service requirements $x_1, \ldots, x_n$, respectively. Denoting this conditional sojourn time by $V_n(\tau; x_1, \ldots, x_n)$ it holds that (cf. Yashkov[23, Eq. (3.4)])

$$V_n(\tau; x_1, \ldots, x_n) = D(\tau) + \sum_{i=1}^{n}\Phi(x_i, \tau), \tag{4.4}$$

where the terms on the right-hand side are independent random variables. The random variable $D(\tau)$ constitutes a “basic” component of the sojourn time: it has the distribution of the sojourn time of a customer with service requirement τ that enters into an empty system. When the system is not empty, the i-th customer present (with remaining service requirement x_i) “adds” a delay $\Phi(x_i, \tau)$ to the new customer's sojourn time. It is worth emphasizing that the delay components $\Phi(x_i, \tau)$, $i = 1, \ldots, n$, are independent of each other and independent of $D(\tau)$.

The decomposition in (4.4) has played a central role in the analysis of PS queues [9, 10, 18, 16]. It is also essential to the asymptotic analysis of the M/G/1 PS queue in Jelenković and Momčilović [12] and that of the already mentioned modified M/G/1 PS queue with random service interruptions in [17, Ch. 5]. It is beyond the scope of this paper to work out the details, instead, the reader interested in the derivation of the differential equations from which the moments can be found is referred to [16, 17].

The following result was previously proved by Zwart and Boxma [26, Thm. 4.1] for the case that the service requirement distribution has a regularly varying tail.

Theorem 4.1 *Consider the M/G/1 PS queue. If $\overline{B}(x) \in \mathcal{IR}$ and one of the following conditions is satisfied,*

(i) $\beta_2 < \infty$,

(ii) Assumptions 3.1 and 3.2 hold,

then

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\{B > x\}} = 1.$$

Remark 4.1 Note that we exclude the case that $\beta_2 = \infty$ and $\mathbf{E}\left[B^\zeta\right] < \infty$ for all $\zeta \in (1,2)$. This case can be included by studying the fourth moment of the sojourn time.

Proof of Theorem 4.1. We show that Assumption 2.2 is satisfied. First we note that the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ, the last condition in Assumption 2.2, is easily seen using a sample-path argument: Comparing the sojourn times of two customers, for the same sequences of inter-arrival times and service requirements of other customers, it follows immediately that the one requiring the smaller amount of service leaves before the one with the larger service requirement. As a consequence of (4.1), we also have that (2.2) holds with $\overline{g} = 1/(1-\rho)$.

We now focus on (2.3) and first consider the case that $\beta_2 < \infty$. Equation (4.3) implies the following asymptotic result, for arbitrary $\varepsilon > 0$ and $k = 2, 3, \ldots$,

$$\mathbf{E}\left[(V(\tau) - \mathbf{E}[V(\tau)])^k\right] = \mathrm{o}(\tau^{k-1+\varepsilon}), \quad \tau \to \infty.$$

Thus, if $\overline{B}(x) \in \mathcal{IR}$ and ζ is as in (2.1), then let κ be an even integer which is larger than ζ. Then Assumption 2.2 is satisfied for any $\delta \in (0,1)$ with $\overline{g} = 1/(1-\rho)$, hence, Theorem 2.3 can be applied.

In the case that $\mathbf{E}[B^\alpha] < \infty$ and $\mathbf{E}\left[B^\zeta\right] = \infty$, for some $1 < \alpha < \zeta < 2$, we note that $\overline{B}(x)$ satisfies (2.1). Using Lemmas 3.1 and 3.2, we have $\mathbf{P}\{W_{\lambda,B} > u\} = \mathrm{o}(u^{1-\alpha})$ and, using (4.2), $\mathbf{Var}[V(\tau)] = \mathrm{o}(\tau^{3-\alpha+\varepsilon})$ for all $\varepsilon > 0$. Thus, Assumption 2.2 is satisfied with $\kappa = 2$ and $0 < \delta < \alpha - 1$. Now apply Theorem 2.3. □

5 Foreground-Background Processor Sharing

With the FBPS discipline, at all times, the service capacity is used to serve the customer(s) which so far have received the least amount of service, see Kleinrock [13] or Yashkov [24]. Note that more than one customer can have the (same) minimum amount of attained service. In that case the service capacity is shared equally among these customers, hence the term processor sharing.

Assuming $B(x)$ is absolutely continuous, the mean and variance of the sojourn time are given by:

$$\mathbf{E}[V(\tau)] = \frac{\tau}{1-\lambda h_1(\tau)} + \frac{\lambda h_2(\tau)}{2(1-\lambda h_1(\tau))^2}, \tag{5.1}$$

$$\mathbf{Var}[V(\tau)] = \frac{\lambda h_3(\tau)}{3(1-\lambda h_1(\tau))^3} + \frac{\lambda \tau h_2(\tau)}{(1-\lambda h_1(\tau))^3} + \frac{3(\lambda h_2(\tau))^2}{4(1-\lambda h_1(\tau))^4}, \tag{5.2}$$

cf. Yashkov [24, Form. (6.2) and (6.3)]. The functions $h_j(\tau)$, $j = 1, 2, 3$, are given by

$$h_j(\tau) = j \int_{x=0}^{\tau} x^{j-1} \overline{B}(x) \mathrm{d}x. \tag{5.3}$$

For FBPT the tail equivalence of sojourn time and service requirement distribution holds under the same assumptions as imposed on the M/G/1 queue with PS in the previous section. In this paper we shall only proof this for the case that $\beta_2 = \infty$ by using the above expressions. Similar as for the PS model, in the general case the proof can be given by first deriving differential equations for higher moments and, by means of these, showing that part (ii) of Assumption 2.2 is satisfied.

Theorem 5.1 *Consider the M/G/1 queue with FBPS. If $\overline{B}(x) \in \mathcal{IR}$ and Assumptions 3.1 and 3.2 are satisfied, then*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\left\{B > x\right\}} \;=\; 1.$$

Proof. First we remark that, as in the proof of Theorem 4.1, the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ, follows from a sample-path argument. Hence, it remains to be shown that (2.2) and (2.3) hold.

Note that the $h_j(\tau)$ defined in (5.3) are non-decreasing, positive functions, and that

$$\lim_{\tau\to\infty} h_1(\tau) \;=\; \beta_1 \;<\; \infty.$$

By Lemma 3.1 there is a number $x_0 > 0$ such that $\overline{B}(x) \leq x^{-\alpha}$, for all $x \geq x_0$. Using this in (5.3) for $j = 2, 3$, we have, for arbitrary $\varepsilon > 0$,

$$h_j(\tau) \;=\; \mathrm{o}(\tau^{j-\alpha+\varepsilon}), \quad \tau \to \infty.$$

Hence, by (5.1) and (5.2),

$$\lim_{\tau\to\infty} \frac{\mathbf{E}\left[V(\tau)\right]}{\tau} \;=\; \frac{1}{1-\rho}, \qquad \lim_{\tau\to\infty} \frac{\mathbf{Var}\left[V(\tau)\right]}{\tau^{3-\alpha+\varepsilon}} \;=\; 0.$$

Now, Assumption 2.2 is implied by Assumption 3.2 (with $\kappa = 2$ and $0 < \delta < \alpha - 1$). □

6 Shortest remaining processing time first

Now we consider an M/G/1 queue in which the total service capacity is always allocated to the customer with the shortest remaining processing time. The service of a customer is pre-empted when a new customer arrives with a service requirement smaller than the remaining service requirement of the customer being served. The service of the customer that is pre-empted is resumed as soon as there are no other customers with a smaller amount of work in the system. Currently, there is renewed interest in the SRPT discipline due to its relevance in Web server modeling [3, 11].

Remark 6.1 Note that if the service requirement distribution has discontinuity points, it may occur (with positive probability) that two customers have the same remaining service requirement, see Schrage and Miller [21]. Here we assume this is not the case, thus, $B(x)$ is a continuous function.

Following Schrage and Miller [21] we decompose the sojourn time into two different periods: The waiting time (the time until the customer is first served) and the residence time (the remainder of the sojourn time). For a customer with service requirement τ, we denote the waiting time by $W(\tau)$ and the residence time by $R(\tau)$. Thus, the sojourn time is given by $V(\tau) = W(\tau) + R(\tau)$. We emphasize that the residence time may contain service pre-emption periods caused by customers with a smaller service requirement. Schrage and Miller [21] obtained the LST of $W(\tau)$ and $R(\tau)$. For our purposes we only need the first two moments of these random variables. First we define $\rho(\tau)$ as the traffic load of customers with an amount of work less than or equal to τ,

$$\rho(\tau) := \lambda \int_{t=0}^{\tau} t \mathrm{d}B(t). \tag{6.1}$$

The first two moments of $W(\tau)$ are given by:

$$\mathbf{E}[W(\tau)] = \lambda \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 \overline{B}(\tau)}{2(1-\rho(\tau))^2}, \tag{6.2}$$

$$\begin{aligned} \mathbf{E}\left[W(\tau)^2\right] &= \lambda \frac{\int_{t=0}^{\tau} t^3 \mathrm{d}B(t) + \tau^3 \overline{B}(\tau)}{3(1-\rho(\tau))^3} \\ &\quad + \lambda^2 \int_{t=0}^{\tau} t^2 \mathrm{d}B(t) \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 \overline{B}(\tau)}{(1-\rho(\tau))^4}, \end{aligned} \tag{6.3}$$

and the mean and variance of $R(\tau)$ by

$$\mathbf{E}[R(\tau)] = \int_{t=0}^{\tau} \frac{1}{1-\rho(t)} \mathrm{d}t, \tag{6.4}$$

$$\mathbf{Var}[R(\tau)] = \lambda \int_{t=0}^{\tau} \frac{\int_{u=0}^{t} u^2 \mathrm{d}B(u)}{(1-\rho(t))^3} \mathrm{d}t. \tag{6.5}$$

These expressions enable us to apply Theorem 2.3, thus showing the tail equivalence in the case that $\beta_2 = \infty$. Although not shown here, the result is also true when $\beta_2 < \infty$.

Theorem 6.1 *Consider the M/G/1 queue with SRPT. If $\overline{B}(x) \in \mathcal{IR}$ and Assumptions 3.1 and 3.2 are satisfied, then*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\{B > x\}} = 1.$$

Proof. The proof proceeds along the same lines as those of Theorems 4.1 and 5.1. The monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ follows from a sample-path argument. Furthermore, note that $\rho(\tau)$ defined by (6.1) is a positive, non-decreasing function with $\rho(\tau) \longrightarrow \rho$, as $\tau \to \infty$. Using that the Césaro limit of a function is finite and equal to the ordinary limit when the latter exists, we have:

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[R(\tau)]}{\tau} = \lim_{t\to\infty} \frac{1}{1-\rho(t)} = \frac{1}{1-\rho}.$$

Now consider (6.5) and replace $\mathrm{d}B(u)$ by $-\mathrm{d}\overline{B}(u)$. By Lemma 3.1 there is a number $x_0 > 0$ such that $\overline{B}(x) \le x^{-\alpha}$, for all $x \ge x_0$. Using partial integration and the fact that $\rho(t) \le \rho$ for all $t \ge 0$, we have, for arbitrary $\varepsilon > 0$,

$$\begin{aligned} \mathbf{Var}[R(\tau)] &= -\lambda \int_{t=0}^{\tau} \frac{\int_{u=0}^{t} u^2 \mathrm{d}\overline{B}(u)}{(1-\rho(t))^3} \mathrm{d}t \\ &\le \frac{-\lambda}{(1-\rho)^3} \int_{t=0}^{\tau} \left(t^2 \overline{B}(t) - 2\int_{u=0}^{t} u\overline{B}(u)\mathrm{d}u \right) \mathrm{d}t \\ &= \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau \to \infty. \end{aligned}$$

In the same way, by partial integration we have for $\mathbf{E}[W(\tau)]$, using Formula (6.2),

$$\mathbf{E}[W(\tau)] = \lambda \frac{\int_{t=0}^{\tau} t\overline{B}(t)\mathrm{d}t}{(1-\rho(\tau))^2},$$

and similarly for $\mathbf{E}\left[W(\tau)^2\right]$. With the above bound for $\overline{B}(u)$, the following relations follow for all $\varepsilon > 0$:

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[W(\tau)]}{\tau^{2-\alpha+\varepsilon}} = \lim_{\tau\to\infty} \frac{\mathbf{E}\left[W(\tau)^2\right]}{\tau^{3-\alpha+\varepsilon}} = 0,$$

hence, since $3-\alpha > 2(2-\alpha)$,

$$\mathbf{Var}\left[W(\tau)\right] \;=\; \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau \to \infty.$$

Using the fact that the random variables $W(\tau)$ and $R(\tau)$ are independent for fixed $\tau > 0$, we have, for all $\varepsilon > 0$,

$$\lim_{\tau\to\infty} \frac{\mathbf{E}\left[V(\tau)\right]}{\tau} = \frac{1}{1-\rho}, \qquad \lim_{\tau\to\infty} \frac{\mathbf{Var}\left[V(\tau)\right]}{\tau^{3-\alpha+\varepsilon}} = 0.$$

Thus, Assumptions 2.1 and 2.2 are satisfied (for $\kappa = 2$ and $0 < \delta < \alpha - 1$) and we may apply Theorem 2.3. □

7 Discussion of the results

In all three models of Sections 4 – 6 we found that, when applying Theorem 2.3, the factor $\overline{g}$ is equal to $1/(1-\rho)$. An intuitive interpretation of this is as follows. Theorems 4.1, 5.1 and 6.1 state that the probability that a customer's sojourn time exceeds the value $x/(1-\rho)$ is asymptotically (for $x \to \infty$) equal to the probability that a customer's service requirement exceeds a value x. This property can be understood partly by noting that the three models share the property that if a customer with an infinite service requirement is placed in the queue, then the queue remains stable. Hence, after a very long period, say t time units with $t \to \infty$, the average capacity per unit of time devoted to the service of non-permanent customers is approximately equal to the average traffic load ρ (because the system is stable and, hence, all non-permanent customers eventually leave the system). The average total service capacity rendered by the system (per unit of time) is approximately 1. Thus, the average service capacity devoted to the permanent customer is approximately $1-\rho$. If the amount of service received by the permanent customer at time t is denoted by $S(t)$, we have that

$$\frac{S(t)}{t} \;\approx\; 1-\rho,$$

hence, the factor $\overline{g}$ above.

The above reasoning for the ratio $S(t)/t$ also holds when the service requirement distribution is not heavy tailed. However, it is known that the tail equivalence result does not hold for the M/M/1 PS queue [5]. The reason for this is that with a "light-tailed" service requirement distribution, Lemma 2.1 – which we need in the proof of Theorem 2.3 – does not hold: A large sojourn time may be due to the fact that many other customers are requesting service. Different from the heavy-tailed case, the probability of this happening is not negligible compared to that of a large sojourn time and a large service requirement occurring simultaneously.

8 Summary

We presented a new approach for the analysis of the tail of the sojourn time distribution when the service requirement distribution has a heavy tail of intermediate regular variation. We extended the "tail equivalence" of the sojourn time distribution and the service requirement distribution in the M/G/1 PS queue to distributions of this class. We also established the tail equivalence in the M/G/1 queues with FBPS or SRPT.

The strength of the approach outlined in this paper is its flexibility to be applied to models for which no closed-form expressions are available. An example of such a model is the M/G/1 PS queue with random service interruptions studied in [17, Ch. 5]. To establish the tail equivalence it suffices to verify the three asymptotic properties listed in Assumption 2.2. (This, in itself, is not a trivial exercise due to the fact that even basic performance measures are not known for this model. As for the ordinary M/G/1 PS queue it may be accomplished by means of a set of differential equations for the moments of the conditional sojourn times, which must be solved "asymptotically" [17].)

Acknowledgements

The author thanks Onno Boxma and Sem Borst for their valuable comments on earlier drafts of this paper.

A Definition of regularly varying distributions

A distribution function $H(x)$, $x \geq 0$, is said to have a regularly varying tail (at infinity) with index $\theta < 0$ if, for arbitrary $t > 0$,

$$\lim_{x \to \infty} \frac{1 - H(tx)}{1 - H(x)} = t^{\theta}.$$

The foremost important member of this class is the Pareto distribution: $H(x) = 1 - (1 + a\,x)^{\theta}$, $x \geq 0$, where $a > 0$ is an additional constant.

B Proof of Relation (2.1)

First we repeat the relation in the next lemma.

Lemma *Let $\overline{B}(x) \in \mathcal{IR}$. Then there exist numbers $\zeta \in (0, \infty)$, $x_0 \in (0, \infty)$, and $\eta \in (0, 1)$ such that, for all $x_2 \geq x_1 \geq x_0$,*

$$\frac{\overline{B}(x_2)}{\overline{B}(x_1)} \geq \eta \left(\frac{x_2}{x_1}\right)^{-\zeta}.$$

Proof. Let $\varepsilon > 0$. Because $\overline{B}(x) \in \mathcal{IR}$, there exists a $K = K(\varepsilon) \in (0, 1)$ and an $x_0 = x_0(\varepsilon, K)$ such that, for all $x \geq x_0$,

$$\frac{\overline{B}(x(1+\varepsilon))}{\overline{B}(x)} \geq K.$$

Let x_1 and x_2 be such that $x_2 \geq x_1 \geq x_0$, and let

$$n := \left\lceil \frac{\ln(x_2) - \ln(x_1)}{\ln(1+\varepsilon)} \right\rceil,$$

where $\lceil y \rceil$ is the smallest integer which is larger than or equal to $y \in \mathbb{R}$. Obviously, $n > 0$ and $x_2 \leq x_1(1+\varepsilon)^n$. We may write:

$$\begin{aligned} \overline{B}(x_1) &\leq K^{-1}\overline{B}(x_1(1+\varepsilon)) \leq \ldots \\ &\leq K^{-n}\overline{B}(x_1(1+\varepsilon)^n) \leq K^{-n}\overline{B}(x_2). \end{aligned}$$

Now the lemma is proved by setting

$$\zeta = \frac{-\ln(K)}{\ln(1+\varepsilon)} > 0,$$

and $\eta = (1+\varepsilon)^{-\zeta}$. $\square$

C Proof of Lemma 2.1

Lemma *Suppose Assumptions 2.1 and 2.2 are satisfied. Then, for fixed* $\varepsilon \in (0,1)$,

$$\lim_{x\to\infty} \frac{\mathbf{P}\{V(B) > \overline{g}\,x; B \le x(1-\varepsilon)\}}{\mathbf{P}\{B > x(1-\varepsilon)\}} \quad = \quad 0.$$

Proof. We prove the lemma using the following relations, which hold for x "large enough":

$$\begin{aligned}
\mathbf{P}\{V(B) > \overline{g}\,x; B \le x(1-\varepsilon)\} &= \int_{\tau=0}^{x(1-\varepsilon)} \mathbf{P}\{V(\tau) > \overline{g}\,x\}\,\mathrm{d}B(\tau) \\
&\le \int_{\tau=0}^{x(1-\varepsilon)} \mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > \overline{g}\,x - \mathbf{E}[V(x(1-\varepsilon))]\}\,\mathrm{d}B(\tau) \\
&\le \frac{\int_{\tau=0}^{x(1-\varepsilon)} \mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \mathrm{d}B(\tau)}{(\overline{g}\,x - \mathbf{E}[V(x(1-\varepsilon))])^{\kappa}}.
\end{aligned} \tag{C.1}$$

The first inequality is an immediate consequence of the monotonicity of $\mathbf{E}[V(\tau)]$ in τ, see Assumption 2.2. For the second inequality we use (2.4). Note that, indeed, for x large enough it must be that $\overline{g}\,x - \mathbf{E}[V(x(1-\varepsilon))]$ is positive, since by Assumption 2.2:

$$\frac{\overline{g}\,x}{\mathbf{E}[V(x(1-\varepsilon))]} \longrightarrow \frac{1}{1-\varepsilon} \quad > \quad 1, \quad x \to \infty.$$

Hence, for large x, the denominator of the right-hand side of (C.1) "behaves as" $(\overline{g}x\varepsilon)^{\kappa}$.

Next we study the numerator. We can choose $\delta > 0$ small enough such that $\kappa - \delta > \zeta$. Let x_0 be as in (2.1), and $\tau_0 \ge x_0$ such that, for all $\tau \ge \tau_0$:

$$\mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \quad \le \quad \tau^{\kappa-\delta}.$$

Such a τ_0 exists by Assumption 2.2. If x is such that $x(1-\varepsilon) > \tau_0$ then (C.1) leads to:

$$\begin{aligned}
&\int_{\tau=\tau_0}^{x(1-\varepsilon)} \mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \mathrm{d}B(\tau) \\
&\le\; -\int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta}\mathrm{d}\overline{B}(\tau) \\
&\overset{\text{p.i.}}{=}\; \tau_0^{\kappa-\delta}\overline{B}(\tau_0) - (x(1-\varepsilon))^{\kappa-\delta}\,\overline{B}(x(1-\varepsilon)) \\
&\qquad +(\kappa-\delta)\int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta-1}\overline{B}(\tau)\mathrm{d}\tau \\
&\le\; \tau_0^{\kappa-\delta}\overline{B}(\tau_0) + (\kappa-\delta)\overline{B}(x(1-\varepsilon)) \int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta-1}\left(\frac{\tau}{x(1-\varepsilon)}\right)^{-\zeta} \mathrm{d}\tau \\
&\le\; \tau_0^{\kappa-\delta}\overline{B}(\tau_0) + \frac{\kappa-\delta}{\kappa-\delta-\zeta}\overline{B}(x(1-\varepsilon))\,(x(1-\varepsilon))^{\kappa-\delta},
\end{aligned}$$

where "p.i." indicates the use of partial integration. In the second inequality we used (2.1) and the fact that $(x(1-\varepsilon))^{\kappa-\delta}\,\overline{B}(x(1-\varepsilon)) \;\ge\; 0$. Since

$$\int_{\tau=0}^{\tau_0} \mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \mathrm{d}B(\tau)$$

is independent of x and $\kappa - \delta > \zeta$, the numerator of the right-hand side of (C.1) is bounded from above by a function that tends to infinity as $\overline{B}(x(1-\varepsilon))\,(x(1-\varepsilon))^{\kappa-\delta}$. Recall that the denominator "behaves as" $(\overline{g}x\varepsilon)^{\kappa}$. Therefore, dividing the right-hand side of (C.1) by $\overline{B}(x(1-\varepsilon))$, and letting $x \to \infty$, proves the lemma. □

References

1. V. Anantharam. Scheduling strategies and long-range dependence. Technical Report, University of California, Berkeley (1997).
2. S. Asmussen. *Applied Probability and Queues.* Wiley, Chichester (1987).
3. N. Bansal, and M. Harchol-Balter. Analysis of SRPT Scheduling: Investigating Unfairness. *Performance Evaluation Review* 29 (2001), Special issue – Proc. Sigmetrics/Performance 2001 (Cambridge, MA), 279–290.
4. N.H. Bingham, C.M. Goldie, J.L. Teugels. *Regular Variation.* University Press, Cambridge (1987).
5. S.C. Borst, O.J. Boxma, J.A. Morrison, R. Núñez-Queija. The equivalence between processor sharing and service in random order. SPOR Report 2002-01, Eindhoven University of Technology (2002).
6. D.B.H. Cline. Intermediate regular and Π variation. *Proc. London Mathematical Society (*3^{rd} *series)* 68 (1994), 594–616.
7. J.W. Cohen. Some results on regular variation in queueing and fluctuation theory. *Journal of Applied Probability* 10 (1973), 343–353.
8. J.W. Cohen. *The Single Server Queue.* (2nd ed.) North-Holland, Amsterdam (1982).
9. S.A. Grishechkin. Crump-Mode-Jagers branching processes as a method of investigating M/G/1 systems with processor sharing. *Theory of Probability and its Applications* 36 (1991), 19–35; translated from *Teoriya Veroyatnostei i ee Primeneniya* 36 (1991), 16–33 (in Russian).
10. S.A. Grishechkin. On a relationship between processor sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability* 24 (1992), 653–698.
11. M. Harchol-Balter, N. Bansal, B. Schroeder. Implementation of SRPT Scheduling in Web Servers. Technical Report CMU-CS-00-170, Carnegie Mellon University (2001).
12. P. Jelenković, P. Momcilović. Resource Sharing with Subexponential Distributions. Proc. IEEE Infocom 2002 (New York).
13. L. Kleinrock. *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York (1976).
14. W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* 2 (1994), 1–15.

15. L. MASSOULIÉ, J.W. ROBERTS. Arguments in favour of admission control for TCP flows. *In: Teletraffic Engineering in a Competitive World – Proc. ITC 16, Edinburgh.* Eds. D. Smith and P. Key. Elsevier, Amsterdam (1999), 33–44.

16. R. NÚÑEZ-QUEIJA. Sojourn times in a processor-sharing queue with service interruptions. *Queueing Systems* 34 (2000), 351–386.

17. R. NÚÑEZ-QUEIJA *Processor-Sharing Models for Integrated-Services Networks.* Ph.D. thesis, Eindhoven University of Technology (2000), ISBN 90-646-4667-8 (also available from the author upon request).

18. K.M. REGE, B. SENGUPTA. A decomposition theorem and related results for the discriminatory processor sharing queue. *Queueing Systems* 18 (1994), 333–351.

19. M. SAKATA, S. NOGUCHI, J. OIZUMI. Analysis of a processor-shared queueing model for time-sharing systems. *In: Proc. 2nd Hawaii International Conference on System Sciences* (1969), 625–628.

20. M. SAKATA, S. NOGUCHI, J. OIZUMI. An analysis of the M/G/1 queue under round-robin scheduling. *Operations Research* 19 (1971), 371–385.

21. L.E. SCHRAGE, L.W. MILLER. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research* 14 (1966), 670–684.

22. D. WILLIAMS. *Probability with Martingales.* University Press, Cambridge (1991).

23. S.F. YASHKOV. A derivation of response time distribution for a $M/G/1$ processor-sharing queue. *Problems of Control and Information Theory* 12 (1983), 133–148.

24. S.F. YASHKOV. Processor-sharing queues: Some progress in analysis. *Queueing Systems* 2 (1987), 1–17.

25. S.F. YASHKOV. Mathematical problems in the theory of processor-sharing queueing systems. *Journal of Soviet Mathematics* 58 (1992), 101–147.

26. A.P. ZWART, O.J. BOXMA. Sojourn time asymptotics in the M/G/1 processor-sharing queue. *Queueing Systems* 35 (2000), 141–166.