Heavy tails : the effect of the service discipline

Heavy Tails: The Effect of the Service Discipline*

S.C. Borst[1,2,3], O.J. Boxma[1,2], and R. Núñez-Queija[1,2]
sem@cwi.nl, boxma@win.tue.nl, sindo@cwi.nl

[1] Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[2] CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[3] Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974, USA

Abstract. This paper considers the $M/G/1$ queue with regularly varying service requirement distribution. It studies the effect of the service discipline on the tail behavior of the waiting- or sojourn time distribution, demonstrating that different disciplines may lead to quite different tail behavior. The orientation of the paper is methodological: We outline three different methods of determining tail behavior, illustrating them for service disciplines like FCFS, Processor Sharing and LCFS.

This paper is dedicated to the memory of Vincent Dumas, a dear friend and gifted young mathematician.

1 Introduction

Measurements indicate that traffic in high-speed networks exhibits burstiness on a wide range of time scales, manifesting itself in long-range dependence and self-similarity, see for instance Leland *et al.* [29], Paxson & Floyd [38]. The occurrence of these phenomena is commonly attributed to extreme variability and long-tailed characteristics in the underlying activity patterns (connection times, file sizes, scene lengths), see for instance Beran *et al.* [5], Crovella & Bestavros [20], Willinger *et al.* [43]. This has triggered a lively interest in queueing models with long-tailed traffic characteristics.

Although the presence of long-tailed traffic characteristics is widely acknowledged, the practical implications for network performance and traffic engineering remain to be fully resolved. An interesting aspect is the role of scheduling and priority mechanisms in controlling the effect of long-tailed traffic characteristics on network performance. In a fundamental paper, Anantharam [3] considered a single-server queue fed by a Poisson arrival process of sessions, with each session lasting for an independent integer time T for which holds: $\mathbf{P}\{T = k\} \sim \alpha k^{-(\alpha+1)} L(k)$, where $1 < \alpha < 2$ and $L(\cdot)$ is a slowly varying function (see Definition 1). Each session brings in work at unit rate while it is active. Hence, the work brought in by each arrival is regularly varying and, because $1 < \alpha < 2$, the arrival process of work is long-range dependent, but $\mathbf{E}T < \infty$. Anantharam shows that, in the steady-state case, for *any* stationary Non-Preemptive service policy at the queue, the stationary sojourn time of a typical session

* Invited paper in the proceedings of Performance TOOLS 2002, 12th International Conference on Modelling Tools and Techniques for Computer and Communication System Performance Evaluation (Imperial College of Science Technology and Medicine, London, April 15-17, 2002)

must stochastically dominate a regularly varying random variable having infinite mean. Non-preemption means that once service on a session has begun, it is continued until all the work associated with it has been completed. Anantharam does not make any assumptions as to whether the service policy is work-conserving, or whether the length of a session is known at the time of arrival. In marked contrast to the above, Anantharam also shows that there exist causal stationary *preemptive* policies, which do not need information about the session durations at the time of their arrival, for which the stationary sojourn time of a session is stochastically dominated by a regularly varying random variable with finite mean.

The results of Anantharam raise several questions, like (i) are there (preemptive) service disciplines for which the tail of the sojourn time distribution is not heavier than the tail of the service requirement distribution, and (ii) what is the effect of various well-known scheduling disciplines on the tail behavior of the waiting- or sojourn time distribution?

A related issue arises when there are *several classes* of customers, which may be treated in different ways by the server (e.g., using fixed priorities, or according to a polling discipline). Then it is important to understand under what conditions, or to what extent, the tail behavior of the service requirements of one class affects the performance of other classes. The above issues have recently been investigated by the present authors and some of their colleagues. This paper summarizes the results. We focus on the classical $M/G/1$ queue and its multi-class generalizations (although some of the recently obtained results allow a general renewal arrival process, or a fluid input).

The orientation of the paper is methodological. After introducing the model and reviewing the main results for various basic disciplines in Section 2, we discuss three different methods of obtaining the tail behavior of waiting- or sojourn time distributions for $M/G/1$-type queues with regularly varying service requirement distribution(s): (i) an analytical one, which relies on Tauberian theorems relating the tail behavior of a probability distribution to the behavior of its Laplace-Stieltjes transform near the origin; (ii) a probabilistic one, which exploits a Markov-type inequality, relating an extremely large sojourn (or waiting) time with a single extremely large service requirement; and (iii) a probabilistic one, which is based on sample-path arguments which lead to lower and upper bounds for probability tails. These three approaches are the topics of Sections 3, 4 and 5, respectively. Sections 3 and 5 also discuss the multi-class case.

2 Model Description and Main Results

In this section, we formally describe the model, introduce some concepts and notation, and give an overview of the main results.

As stated before, we focus on the $M/G/1$ queue. In this system, customers arrive according to a Poisson process, with rate λ, at a single server who works at unit rate. Their service requirements $B_1, B_2, \ldots$ are independent and identically distributed, with distribution $B(\cdot)$ with mean β and Laplace-Stieltjes Transform (LST) $\beta\{\cdot\}$. A generic service requirement is denoted by B. There is no restriction on the number of customers in the system. We assume that the offered traffic load $\rho := \lambda\beta < 1$, so that the system reaches steady state. We study the stationary sojourn time S of a customer, and in some cases also the stationary waiting time W until service begins.

Before surveying the tail asymptotics of the waiting-time and sojourn time distributions for various service disciplines, we first introduce some useful notation and terminology. For any two real functions $g(\cdot)$ and $h(\cdot)$, we use the notational convention $g(x) \sim h(x)$ to denote

$\lim_{x\to\infty} g(x)/h(x) = 1$, or equivalently, $g(x) = h(x)(1 + o(1))$ as $x \to \infty$. For any stochastic variable X with distribution function $F(\cdot)$, with $\mathbf{E}X < \infty$, denote by $F^r(\cdot)$ the distribution function of the residual lifetime of X, i.e., $F^r(x) = \frac{1}{\mathbf{E}X}\int_0^x (1-F(y))\mathrm{d}y$, and by X^r a stochastic variable with distribution $F^r(\cdot)$.

Throughout this paper, we focus on the class $\mathcal{R}$ of *regularly-varying* distributions (which contains the Pareto distribution). This class is a subset of the class of subexponential distributions [27], that contains, a.o., the lognormal and Weibull distributions.

Definition 1. A distribution function $F(\cdot)$ on $[0, \infty)$ is called *regularly varying of index* $-\nu$ ($F(\cdot) \in \mathcal{R}_{-\nu}$) if

$$1 - F(x) = x^{-\nu}L(x), \quad \nu \geq 0,$$

where $L : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} L(\eta x)/L(x) = 1$, $\eta > 1$.

The class of regularly varying functions was introduced by Karamata [25], and its potential for probability theory was extensively discussed in Feller [23]. A key reference is Bingham *et al.* [7].

In the remainder of this section, we present an overview of the tail asymptotics of the waiting-time and sojourn time distributions in the $M/G/1$ queue for six key service disciplines: (i) First-Come-First-Served (FCFS); (ii) Processor Sharing (PS); (iii) Last-Come-First-Served Preemptive-Resume (LCFS-PR); (iv) Last-Come-First-Served Non-Preemptive Priority (LCFS-NP); (v) Foreground-Background Processor Sharing (FBPS); (vi) Shortest Remaining Processing Time First (SRPTF).

(i) The $M/G/1$ FCFS queue

The next theorem characterizes the tail asymptotics of the distribution of the steady-state waiting time W for the FCFS service discipline.

Theorem 1. *In the case of regular variation, i.e.,* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$,

$$\mathbf{P}\{W > x\} \sim \frac{\rho}{1-\rho}\mathbf{P}\{B^r > x\}, \quad x \to \infty. \tag{1}$$

Remark 1. The waiting-time tail in the $M/G/1$ FCFS queue appears to be one degree heavier than the service requirement tail, in the regularly varying case. This is explained by the fact that an arriving customer has a positive probability of arriving during a service time. His waiting time then is at least equal to the residual duration of the ongoing service – which is regularly varying of index $1 - \nu$ (cf. [7]).

Theorem 1 was first proved by Cohen [17] (who in fact considered the $GI/G/1$ case), and subsequently extended by several authors. In particular, Pakes [36] has proven that $\mathbf{P}\{W > x\} \sim \frac{\rho}{1-\rho}\mathbf{P}\{B^r > x\}$ even holds for the $GI/G/1$ queue, for the larger class of service requirement distributions for which the residual service requirement distribution is subexponential.

The fact that the sojourn time of a customer in the $M/G/1$ FCFS queue equals the sum of his waiting time and his lighter-tailed service requirement, the two quantities being independent, implies that the tail behavior of the sojourn time distribution is also given by the righthand side of (1).

(ii) The $M/G/1$ PS queue

The PS (processor sharing) service discipline operates as follows. If there are $n \geq 1$ customers present, then they are all served simultaneously, at a rate of $1/n$. At the ITC conference in 1997, J.W. Roberts raised the question whether the tail of the distribution of the steady-state sojourn time S_{PS} in the $M/G/1$ PS system might be just as heavy as the tail of the service requirement distribution. This question was motivated by the following observations: (i) in a PS system short jobs can overtake long jobs, so the influence of long jobs on the sojourn time of short jobs is limited, and (ii) the mean sojourn time in the $M/G/1$ PS queue only involves the *first* moment of the service requirement, whereas in the $M/G/1$ FCFS queue it involves the first moment of the *residual* service requirement (see also (8) below), and hence the *second* moment of the service requirement. In fact, if $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$ with $1 < \nu < 2$, then the second moment of the service requirement does not exist, and neither does the first moment of the waiting time in the $M/G/1$ FCFS case. Roberts' question can be answered affirmatively, as shown by the next theorem proven in [51].

Theorem 2. *If* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$,

$$\mathbf{P}\{S_{PS} > x\} \sim \mathbf{P}\{B > (1-\rho)x\}, \quad x \to \infty. \tag{2}$$

Apparently, in the $M/G/1$ PS queue the sojourn time tail is just as heavy as the service requirement tail.

Remark 2. Formula (2) allows the following interpretation. When a tagged customer has been in the system for a time period of length x, with x large, then the distribution of the number of other customers present has approximately become the steady-state distribution of the number of customers in the $M/G/1$ PS queue. Hence, on average, the server devotes ρ amount of time per time unit to those other customers. Therefore he will, on average, devote a fraction $1-\rho$ of the time to the tagged customer. Consequently, the amount of service received by that customer during x is approximately $(1-\rho)x$.

(iii) The $M/G/1$ LCFS-PR queue

In the LCFS Preemptive-Resume discipline, an arriving customer K is immediately taken into service. However, this service is interrupted when another customer arrives, and it is only resumed when all customers who have arrived after K have left the system.

The fact that no customer has to wait for the completion of a residual service requirement, suggests that the tail of the sojourn time distribution is just as heavy as the tail of the service requirement distribution. This was indeed proven in [13], using the following observation: The sojourn time of K has exactly the same distribution as the busy period of this $M/G/1$ queue. The busy period obviously has the same distribution for LCFS-PR as for FCFS. The tail behavior of the busy-period distribution in the $M/G/1$ queue has been studied by De Meyer and Teugels [30] for the case of a regularly varying service requirement distribution. This yields the next theorem (S_{LPR} denoting the steady-state sojourn time).

Theorem 3. *If* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$,

$$\mathbf{P}\{S_{LPR} > x\} \sim \frac{1}{1-\rho}\mathbf{P}\{B > (1-\rho)x\}, \quad x \to \infty. \tag{3}$$

(iv) The $M/G/1$ LCFS-NP queue
Let W_{LNP} denote the steady-state waiting time in the $M/G/1$ LCFS-NP queue. The possibility of preemption suggests that the tail of W_{LNP} will be determined by the tail of a residual service time. Indeed, in this paper we prove the following result, which in fact also holds for the sojourn time $S_{LNP} = W_{LNP} + B$, as is shown in Section 4.

Theorem 4. *If* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$,

$$\mathbf{P}\{W_{LNP} > x\} \sim \rho \mathbf{P}\{B^r > (1-\rho)x\}, \quad x \to \infty. \tag{4}$$

(v) The $M/G/1$ FBPS queue
The Foreground-Background Processor Sharing discipline allocates an equal share of the service capacity to the customers which so far have received the least amount of service, see Kleinrock [26] or Yashkov [46]. We will show (only for the case $1 < \nu < 2$) that the tail of the distribution of the sojourn time S_{FB} is the same as that for the ordinary PS discipline:

Theorem 5. *If* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$ *with* $1 < \nu < 2$,

$$\mathbf{P}\{S_{FB} > x\} \sim \mathbf{P}\{B > (1-\rho)x\}, \quad x \to \infty. \tag{5}$$

Although not proven here, it can be shown that the result remains true for $\nu \geq 2$.

(vi) The $M/G/1$ SRPTF queue
With this service discipline the total service capacity is always allocated to the customer(s) with the shortest remaining processing time (Shortest Remaining Processing Time First). Assuming that $B(x)$ is a continuous function, with probability 1, no two customers in the system have the same remaining service requirement [41]. The service of a customer is preempted 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 preempted is resumed as soon as there are no other customers with a smaller amount of work in the system. For the sojourn time S_{SR} we will prove the following theorem:

Theorem 6. *If* $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$ *with* $1 < \nu < 2$,

$$\mathbf{P}\{S_{SR} > x\} \sim \mathbf{P}\{B > (1-\rho)x\}, \quad x \to \infty. \tag{6}$$

Note that the tail of the service requirement distribution behaves as those of the PS and FBPS disciplines. Again, we remark (without proof) that the result is also valid for $\nu \geq 2$.

In the sequel we prove these theorems using different methods. This serves as a demonstration of the methods and allows us to compare them. Theorems 2 and 4 shall be proven in each of the Sections 3, 4 and 5, Theorems 1 and 3 are proven in Sections 3 and 5, and Theorems 5 and 6 are proven in Section 4.

3 Transform Approach

In this section we demonstrate an LST approach to the study of tails of waiting-time and sojourn time distributions in the $M/G/1$ queue and some of its generalizations. In Subsection 3.1 we consider the single-class $M/G/1$ queue, with as service discipline either FCFS, PS, LCFS-PR, or LCFS-NP. In Subsection 3.2 we consider the multi-class $M/G/1$ queue, in which the classes are served according to some scheduling mechanism.

For several of the above-mentioned cases, an expression for the LST of the waiting- and/or sojourn time distribution can be found in the literature. In exceptional cases, this expression is sufficiently simple to allow explicit inversion of the LST, thus possibly enabling one to determine the tail behavior of the corresponding distribution. An example is the $M/G/1$ queue with the FCFS discipline, which will first be considered in Subsection 3.1. But usually such an explicit inversion is not viable. Fortunately, there exists a very useful relation between the tail behavior of a regularly varying probability distribution and the behavior of its LST near the origin. That relation often enables one to conclude from the form of the LST of the waiting- and/or sojourn time distribution, that the distribution itself is regularly varying at infinity. We present this relation in Lemma 1 below.

Let $F(\cdot)$ be the distribution of a non-negative random variable, with LST $\phi\{s\}$ and finite first n moments $\mu_1, \dots, \mu_n$ (and $\mu_0 = 1$). Define

$$\phi_n\{s\} \ := \ (-1)^{n+1}[\phi\{s\} - \sum_{j=0}^{n} \mu_j \frac{(-s)^j}{j!}].$$

Lemma 1. *Let $n < \nu < n+1$, $C \geq 0$. The following statements are equivalent:*

$$\phi_n\{s\} = (C + o(1)) s^{\nu} L(1/s), \quad s \downarrow 0, \quad s \text{ real},$$

$$1 - F(x) = (C + o(1)) \frac{(-1)^n}{\Gamma(1-\nu)} x^{-\nu} L(x), \quad x \to \infty.$$

The case $C > 0$ is due to Bingham and Doney [6]. The case $C = 0$ was first obtained by Vincent Dumas, and is treated in [16], Lemma 2.2. The case of an integer ν is more complicated; see Theorem 8.1.6 and Chapter 3 of [7].

3.1 The Single-Class Case

(i) The $M/G/1$ FCFS queue
In the $M/G/1$ FCFS queue, the LST of the steady-state waiting-time distribution is given by the Pollaczek-Khintchine formula [18]:

$$\mathbf{E}[\mathrm{e}^{-sW}] \ = \ \frac{1-\rho}{1-\rho\beta^r\{s\}}, \quad \mathrm{Re}\ s \geq 0, \tag{7}$$

where $\beta^r\{s\} = (1 - \beta\{s\})/\beta s$ is the LST of the residual service requirement distribution $B^r(x)$. The geometric structure of this LST enables one to invert it, yielding (with B_i^r a random variable with distribution a residual service requirement distribution):

$$\mathbf{P}\{W < x\} \ = \ \sum_{n=0}^{\infty} (1-\rho)\rho^n \mathbf{P}\{B_1^r + \ldots + B_n^r < x\}, \quad x \geq 0. \tag{8}$$

A Karamata theorem (cf. Section 1.5 of [7]) implies that if $\mathbf{P}\{B > \cdot\} \in \mathcal{R}_{-\nu}$, then the integrated tail $\mathrm{P}(B^r > \cdot) \in \mathcal{R}_{1-\nu}$. More precisely, if

$$\mathbf{P}\{B > x\} \sim x^{-\nu} L(x), \quad \nu > 1, \quad x \to \infty, \tag{9}$$

then

$$\mathbf{P}\{B^r > x\} = \frac{1}{\beta}\int_x^\infty \mathbf{P}\{B > y\}\mathrm{d}y \sim \frac{1}{(\nu-1)\beta}x^{1-\nu}L(x), \quad x \to \infty. \tag{10}$$

A key property of regularly varying probability distributions (and of the larger class of subexponential distributions), specified to the sum of residual service requirements in (8), is that

$$\mathbf{P}\{B_1^r + \ldots + B_n^r > x\} \sim n\mathbf{P}\{B_1^r > x\}, \quad x \to \infty. \tag{11}$$

Put differently: the sum of n independent, identically distributed random variables with regularly varying tail exhibits the same tail behavior as the *maximum* of that sum. This implies that *when such a sum is large, that is most likely due to one of the terms being large.* This observation provides the key intuition to many of the results to be discussed below.

Combining (11) with (8) we obtain Theorem 1. We shall now demonstrate how the following statement, which implies Theorem 1, is easily obtained from the LST expression (7) and Lemma 1. For $\nu > 1$, $x \to \infty$,

$$\mathbf{P}\{B > x\} \sim x^{-\nu}L(x) \Longleftrightarrow \mathbf{P}\{W > x\} \sim \frac{\rho}{1-\rho}\frac{1}{(\nu-1)\beta}x^{1-\nu}L(x). \tag{12}$$

It follows from (9) and Lemma 1 that

$$1 - \beta^r\{s\} = 1 - \frac{1-\beta\{s\}}{\beta s} = -\left(\frac{\Gamma(1-\nu)}{\beta} + \mathrm{o}(1)\right)s^{\nu-1}L(1/s), \quad s \downarrow 0. \tag{13}$$

Combining this result with (7) yields:

$$1 - \mathbf{E}[\mathrm{e}^{-sW}] = \frac{\rho(1-\beta^r\{s\})}{1-\rho\beta^r\{s\}} \sim -\frac{\rho}{1-\rho}\frac{\Gamma(1-\nu)}{\beta}s^{\nu-1}L(1/s), \quad s \downarrow 0.$$

Another application of Lemma 1 gives the $\Longrightarrow$ part of (12). The reverse part is obtained in a similar way.

(ii) The $M/G/1$ PS queue

Theorem 2 indicates that, contrary to the FCFS case, the sojourn time tail in the $M/G/1$ PS queue is just as heavy as the service time tail. We now sketch the proof in [51], which is based on the application of Lemma 1 to an explicit expression of the sojourn time LST.

There are several expressions known for the LST of the sojourn time, cf. [35, 42, 45], but they contain contour integrals which are inversion formulas of Laplace transforms. Starting-point in [51] is an expression in [35] for the *conditional* LST of a customer's sojourn time $S_{PS}(\tau)$, given that his service requirement is τ: For Re $s \geq 0$, $\tau \geq 0$,

$$\mathbf{E}[\mathrm{e}^{-sS_{PS}(\tau)}] = \frac{1-\rho}{(1-\rho)H_1(s,\tau) + sH_2(s,\tau)},$$

where the functions $H_1(s,\tau)$ and $H_2(s,\tau)$ are given by their LST w.r.t. τ:

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}H_1(s,\tau) = \frac{x - \lambda(1-\beta\{x\})}{x - s - \lambda(1-\beta\{x\})}, \quad \mathrm{Re}\ x > 0,$$

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}H_2(s,\tau) = \frac{\rho x - \lambda(1-\beta\{x\})}{x(x - s - \lambda(1-\beta\{x\}))}, \quad \mathrm{Re}\ x > 0.$$

It follows from these relations that, for Re $s \geq 0$ and Re $x > 0$:

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}[\mathbf{E}[\mathrm{e}^{-sS_{PS}(\tau)}]]^{-1} = 1 + \frac{1}{1-\rho}\frac{s}{x}\frac{1}{1 - s\mathbf{E}[\mathrm{e}^{-xW}]/(x(1-\rho))}, \tag{14}$$

where W denotes the steady-state waiting time in the $M/G/1$ FCFS queue (we shall denote its distribution by $W(\cdot)$). Formula (14) implies (see [51]) that

$$\mathbf{E}[\mathrm{e}^{-sS_{PS}(\tau)}] = \left[\sum_{k=0}^{\infty} \frac{s^k}{k!}\alpha_k(\tau)\right]^{-1}, \tag{15}$$

with $\alpha_0(\tau) := 1$, $\alpha_1(\tau) := \tau/(1-\rho)$, and for $k \geq 2$,

$$\alpha_k(\tau) := \frac{k}{(1-\rho)^k}\int_{x=0}^{\tau} (\tau - x)^{k-1} W^{(k-1)*}(x)\mathrm{d}x.$$

In Corollary 3.2 of [51], Formula (15) is shown to imply that the kth moment of the sojourn time in the $M/G/1$ PS queue is finite iff the kth moment of the service requirement is finite. But Formula (15) is also suitable for applying Lemma 1. With S_{PS} the steady-state sojourn time in the $M/G/1$ PS queue, and using the fact that $\mathbf{E}[\mathrm{e}^{-sS_{PS}}] = \int_0^\infty \mathbf{E}[\mathrm{e}^{-sS(\tau)}]\mathrm{d}B(\tau)$, it can be shown [51] that, for $1 < \nu < 2$,

$$\mathbf{E}[\mathrm{e}^{-sS_{PS}}] - \beta\{\frac{s}{1-\rho}\} = \mathrm{o}(s^\nu L(1/s)), \quad s \downarrow 0,\ s \text{ real}.$$

One can now apply Lemma 1. Using the well-known fact that $\mathbf{E}S_{PS} = \beta/(1-\rho)$, it is seen that Theorem 2 holds for $1 < \nu < 2$ (and via a similar approach it is shown in [51] that this holds for all non-integer $\nu > 1$; we ignore the subtleties required in applying Lemma 1 for integer ν). In fact, a two-way application of Lemma 1 yields (cf. [51]): For $\nu > 1$, $x \to \infty$,

$$\mathbf{P}\{B > x\} \sim x^{-\nu}L(x) \Longleftrightarrow \mathbf{P}\{S_{PS} > x\} \sim \frac{1}{(1-\rho)^\nu}x^{-\nu}L(x). \tag{16}$$

(iii) The $M/G/1$ LCFS-PR queue

As observed in Section 2, the sojourn time in the $M/G/1$ LCFS-PR queue has the same distribution as the busy period in the $M/G/1$ queue. De Meyer and Teugels [30] have studied the tail of the latter distribution in the case of a regularly varying service requirement distribution. Their starting-point is the fact that the LST $\mu(s)$ of the steady-state busy period length P is the unique solution of the equation

$$\mu\{s\} = \beta\{s + \lambda(1 - \mu\{s\})\}, \tag{17}$$

with $|\mu\{s\}| \leq 1$ for Re $s \geq 0$. They apply Lemma 1 to show the following equivalence: for $\nu > 1$, $x \to \infty$,

$$\mathbf{P}\{B > x\} \sim x^{-\nu}L(x) \Longleftrightarrow \mathbf{P}\{P > x\} \sim \frac{1}{(1-\rho)^{\nu+1}}x^{-\nu}L(x). \tag{18}$$

Hence the tail of the busy period distribution is just as heavy as that of the service requirement distribution. Theorem 3 immediately follows from (18).

(iv) The $M/G/1$ LCFS-NP queue
Let W_{LNP} denote the steady-state waiting time in the $M/G/1$ LCFS-NP queue. The following is observed in [18], p. 431. If an arriving customer in the $M/G/1$ LCFS-NP queue meets a customer in service with a residual service requirement w, then his waiting-time distribution is that of a busy period with a special first service requirement w. That residual service requirement has distribution $B^r(\cdot)$ with LST $\beta^r\{s\}$ as introduced in the beginning of this section ([18], p. 432). It is now readily seen (cf., e.g., p. 299 of [19]) that

$$\mathbf{E}[\mathrm{e}^{-sW_{LNP}}] = 1 - \rho + \rho\beta^r\{\delta\{s\}\}, \quad \mathrm{Re}\ s \geq 0, \tag{19}$$

with $\delta\{s\}$ the unique zero in Re $s \geq 0$ of $\lambda(1 - \beta\{w\}) - w + s$, Re $w \geq 0$. In fact, cf. (17), $\delta\{s\} = s + \lambda(1 - \mu\{s\})$. In combination with (19), this gives another derivation of Formula (III.3.10) of [18]:

$$\mathbf{E}[\mathrm{e}^{-sW_{LNP}}] = 1 - \rho + \frac{\rho}{\beta}\frac{1-\mu\{s\}}{s + \lambda(1-\mu\{s\})}, \quad \mathrm{Re}\ s \geq 0. \tag{20}$$

Using Lemma 1, we can now easily verify that the tail of W_{LNP} is regularly varying of degree one heavier than the tail of the service requirement (as may be expected in view of the possibility of having to wait at least a residual service requirement). If (9) and hence also (13) hold, then it follows from (18) and Lemma 1 that

$$1 - \mu\{s\} - \frac{\beta}{1-\rho}s \sim -\frac{\Gamma(1-\nu)}{(1-\rho)^{\nu+1}}s^{\nu}L(1/s), \quad s \downarrow 0, \tag{21}$$

and therefore

$$1 - \mathbf{E}[\mathrm{e}^{-sW_{LNP}}] \sim -\frac{\lambda\Gamma(1-\nu)}{(1-\rho)^{\nu-1}}s^{\nu-1}L(1/s), \quad s \downarrow 0. \tag{22}$$

On the other hand, starting from (22) and using (20), one gets (21). Application of Lemma 1 and (18) now yields: For $\nu > 1$, $x \to \infty$,

$$\mathbf{P}\{B > x\} \sim x^{-\nu}L(x) \Longleftrightarrow \mathbf{P}\{W_{LNP} > x\} \sim \frac{\lambda}{(\nu-1)(1-\rho)^{\nu-1}}x^{1-\nu}L(x). \tag{23}$$

Both relations imply Theorem 4.

3.2 The Multi-Class Case

In this subsection we consider the $M/G/1$ queue with K classes of customers. We study several of the most important service scheduling disciplines, rules that specify at any time which class of customers is being served. We are interested in the question under what conditions, or to what extent, the tail behavior of the service requirements of one class affects the performance of other classes.

The notation is as introduced in Section 2, but quantities relating to class-i customers receive an index i. Hence, class-i customers arrive according to a Poisson process with rate λ_i, and their service requirements have distribution $B_i(\cdot)$ with mean β_i; $\rho_i := \lambda_i\beta_i$ and $\rho := \sum_{i=1}^{K}\rho_i$.

(i) Fixed priorities: Non-Preemptive priority
Assume that there are only two priority classes, class 1 having Non-Preemptive priority over

class 2. Cohen [18], Section III.3.8, gives the following expressions for the LST of the distribution of the steady-state waiting time W_1 of class-1 customers:

$$\mathbf{E}[\mathrm{e}^{-sW_1}] = \frac{1-\rho+\rho_2\beta_2^r\{s\}}{1-\rho_1\beta_1^r\{s\}}, \quad \mathrm{Re}\ s \geq 0, \quad \rho < 1, \tag{24}$$

$$\mathbf{E}[\mathrm{e}^{-sW_1}] = \frac{(1-\rho_1)\beta_2^r\{s\}}{1-\rho_1\beta_1^r\{s\}}, \quad \mathrm{Re}\ s \geq 0, \quad \rho_1 < 1, \quad \rho \geq 1. \tag{25}$$

In both cases, Lemma 1 can readily be applied to determine the tail behavior of the waiting-time distribution. Actually, this is one of the rare cases in which the LST can be easily inverted. For $\rho < 1$ this gives ($B_{1,i}^r$ has the residual service requirement distribution $B_1^r(\cdot)$, and B_2^r has the residual service requirement distribution $B_2^r(\cdot)$), with $\overset{d}{=}$ denoting equality in distribution,

$$W_1 \overset{d}{=} B_{1,1}^r + \ldots + B_{1,N}^r + Z,$$

where N is geometrically distributed with parameter ρ_1 while Z is zero with probability $(1-\rho)/(1-\rho_1)$ and $Z = B_2^r$ with probability $\rho_2/(1-\rho_1)$. For $\rho_1 < 1$ but $\rho \geq 1$, inversion of the LST in (25) yields:

$$W_1 \overset{d}{=} B_{1,1}^r + \ldots + B_{1,N}^r + B_2^r.$$

These results imply the following. If the service requirement distribution with the heaviest tail is regularly varying at infinity of index $-\nu$, then the waiting-time distribution of the high-priority customers is regularly varying at infinity of index $1-\nu$. More specifically: If the heaviest tail belongs to class 1, then the waiting-time tail of class-1 customers is as if no class 2 exists. If the heaviest tail belongs to class 2, then the waiting-time tail of class-1 customers behaves like the tail of a residual service requirement of class 2 if $\rho_1 < 1$ and $\rho \geq 1$, and like that tail multiplied by the factor $\rho_2/(1-\rho_1)$ if $\rho < 1$.

For class 2 the following result has been proven in [14]. If the service requirement distribution with the heaviest tail is regularly varying at infinity of index $-\nu$, then the waiting-time distribution of the low-priority customers is regularly varying at infinity of index $1-\nu$. This is proven by exploiting a representation for the LST of that waiting-time distribution, as given by Abate and Whitt [1], and then using Lemma 1. The result is not surprising, when one realizes that a low-priority customer may have to wait for a residual service requirement of either class. For the precise asymptotics we refer to [14].

(ii) Fixed priorities: Preemptive-Resume priority

First assume that there are only two priority classes, class 1 having Preemptive-Resume priority over class 2. As is well-known, class-1 customers are not affected by class-2 customers, so the results of Subsection 3.1 (for FCFS) apply to class 1. The waiting-time distribution of the low-priority customers *until the start of the – possibly interrupted – service* is the same as in the Non-Preemptive case. Those possible interruptions consist of full service requirements of high-priority customers, and in the regularly varying case these are less heavy than *residual* service requirements of those customers. Hence, in the scenario of regular variation, the tail behavior of low-priority customers is the same as in the Non-Preemptive case.

If there are $K > 2$ classes, then in studying class j one may aggregate classes $1, \ldots, j-1$ into one high-priority class w.r.t. class j, while the existence of classes $j+1, \ldots, K$ is irrelevant for class j.

(iii) Polling
Deng [21] has considered the extension of the two-class Non-Preemptive priority model to the case in which the server requires a switchover time to move from one class of customers to the other. She proves: If the service requirement distribution *or the switchover-time distribution* with the heaviest tail is regularly varying at infinity of index $-\nu$, then the waiting-time distributions of both classes are regularly varying at infinity of index $1 - \nu$. Again the key of the derivation is an explicit expression for the LST of the waiting-time distributions, in combination with Lemma 1.

The above 2-class model may also be viewed as a *polling* model with 2 queues Q_1, Q_2 and a server who alternatingly visits both queues, serving Q_1 exhaustively (i.e., until it is empty) and applying the 1-limited service discipline at Q_2 (i.e., serving one customer, if there is one, and then moving on to the other queue). In [15] a polling model with K queues has been studied, with the exhaustive or gated service discipline being employed at the various queues. In a similar way, the same conclusions as above have been obtained.

(iv) Processor sharing with several customer classes
In the multi-class disciplines that were discussed above, the worst tail behavior of any class determined the waiting-time tail behavior of all classes (except for high-priority customers in the case of Preemptive-Resume priority). Processor sharing turns out to be better able to protect customer classes from the bad behavior of other classes. Zwart [47] showed that the sojourn time distribution of a class-i customer is regularly varying of index $-\nu_i$ iff the service requirement distribution of that class is regularly varying of index $-\nu_i$, *regardless* of the service requirement distributions of the other classes. His method again relied on Lemma 1.

(v) Generalized processor sharing
The Generalized Processor Sharing (GPS) discipline operates as follows [37]. Customer class i is assigned a weight ϕ_i, $i = 1, \ldots, K$, with $\sum_{i=1}^{K} \phi_i = 1$. If customers of all classes are present, then one customer from each class is served simultaneously (processor sharing), a class-i customer receiving a fraction ϕ_i of the server capacity. If only some of the classes are present, then the service capacity is shared in proportion to the weights ϕ_i among the head-of-the-line customers of those classes.

GPS-based scheduling algorithms, such as Weighted Fair Queueing, play a major role in achieving differentiated quality-of-service in integrated-services networks. Hence, it is important to study the extent to which GPS manages to protect one class of customers from the adverse effects of bad traffic characteristics of other classes. Unfortunately, the queueing analysis of GPS is very difficult. A slightly more general model for $K = 2$ is the model with two parallel $M/G/1$ queues with service speeds depending on whether the other queue is empty or not. For general service requirement distributions, the joint distribution of the amounts of work of both classes has been obtained in [19] by solving a Wiener-Hopf problem (see [22] and [28] for the case of exponential service requirement distributions). The results of [19] have been exploited in [8, 10, 11]. In those papers, service requirements at Q_1 are either exponential or regularly varying; at Q_2 they are regularly varying. Whether the service requirement tail behavior at Q_2 affects the workload tail at Q_1 is shown to depend crucially on whether or not Q_1 is able to handle all its offered work by working at the low speed that occurs while Q_2 is non-empty (i.e., whether or not $\rho_1 < \phi_1$). The method employed in [8, 10, 11] starts from a – complicated – expression for the workload LST. In some cases Lemma 1 is applicable, but in other cases an extension of this lemma must be used. For $K \geq 3$ coupled queues, respectively for GPS with $K \geq 3$ classes, no explicit results are known. However, the sample-path techniques discussed in Section 5 have proven useful in obtaining tail asymptotics for an arbitrary number of classes [9].

4 Tail Equivalence via Conditional Moments

With heavy-tailed distributions, it is often the case that large occurrences of the variable of interest (e.g., a customer's waiting time or sojourn time) are essentially caused by a *single* large occurrence of one input variable (e.g., a service requirement). In this section we describe a generic approach that may be used to prove that the tails of the distributions of the *causal variable* and the *resultant* variable are *equally heavy*. Specifically, we say that two non-negative random variables X and Y have

equally-heavy tailed distributions if $\mathbf{P}\{Y > \overline{g}\,x\} \sim \mathbf{P}\{X > x\}$ for some constant $\overline{g} > 0$. In the examples below, a customer's own service requirement (denoted with B) or the *residual* service requirement of some other customer (denoted with B^r) will play the role of the causal variable X and the customer's sojourn time (denoted with S) that of the resultant Y. In order to explicitly express the dependence of the sojourn time on the (residual) service requirement, we shall use $S(\tau)$ to denote a customer's sojourn time *given* that the (residual) service requirement equals τ. Consequently, we may alternatively write $S(B)$ or $S(B^r)$ (depending on the causal variable) for the unconditional sojourn time S.

Theorem 7 below relates the tails of the distributions of S and the causal variable X (later replaced with either B or B^r). We shall make two assumptions: one regarding the distribution of the causal variable and one regarding $S(\tau)$. (The first assumption can be relaxed to distributions of *intermediate regular variation*, without invalidating Theorem 7, see [34].)

Assumption 1. $\mathbf{P}\{X > \cdot\} \in \mathcal{R}_{-\alpha}$ for some $\alpha > 0$.

Assumption 2. The following three conditions are satisfied:

(a) $\mathbf{E}S(\tau) \sim \overline{g}\,\tau$, for some $\overline{g} > 0$;

(b) With α as in Assumption 1, there exists $\kappa > \alpha$ such that

$$\mathbf{P}\{S(\tau) - \mathbf{E}S(\tau) > t\} \leq \frac{h(\tau)}{t^{\kappa}},$$

with $h(\tau) = \mathrm{o}(\tau^{\kappa-\delta})$, $\tau \to \infty$, for some $\delta > 0$;

(c) $S(\tau)$ is stochastically increasing in $\tau \geq 0$, i.e., for all $t \geq 0$, the probability $\mathbf{P}\{S(\tau) > t\}$ is non-decreasing in $\tau \geq 0$.

Theorem 7. *Suppose Assumptions 1 and 2 are satisfied. Then the tails of the distributions of the random variables X and $S(X)$ are equally heavy in the sense that:*

$$\mathbf{P}\{S(X) > \overline{g}\,x\} \sim \mathbf{P}\{X > x\}.$$

In particular, the distribution of $S(X)$ is also regularly varying with the same index $-\alpha$ as that of X.

Proof. We only give a sketch of the proof and refer to [34] for details. The proof consists of two parts. For the first part we write, with $\varepsilon > 0$,

$$\mathbf{P}\{S(X) > \overline{g}\,x\} \;\leq\; \mathbf{P}\{S(X) > \overline{g}\,x; X \leq x(1-\varepsilon)\} \,+\, \mathbf{P}\{X > x(1-\varepsilon)\}. \tag{26}$$

By conditioning on X and integrating over the distribution of X it can be shown (using Assumptions 1 and 2) that

$$\mathbf{P}\{S(X) > \overline{g}\,x; X \leq x(1-\varepsilon)\} \;=\; \mathrm{o}(\mathbf{P}\{X > x(1-\varepsilon)\}), \quad x \to \infty.$$

Hence, we may neglect the first term on the right-hand side of (26) and write

$$\limsup_{x\to\infty} \frac{\mathbf{P}\{S(X) > \overline{g}\,x\}}{\mathbf{P}\{X > x\}} \;\leq\; \limsup_{x\to\infty} \frac{\mathbf{P}\{X > x(1-\varepsilon)\}}{\mathbf{P}\{X > x\}} \;=\; (1-\varepsilon)^{-\alpha}.$$

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

$$\mathbf{P}\{S(X) > \overline{g}\,x\} \;\geq\; \mathbf{P}\{S(X) > \overline{g}\,x; X > x(1+\varepsilon)\}.$$

By conditioning again on X it can be shown that

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

Hence,

$$\liminf_{x\to\infty} \frac{\mathbf{P}\{S(X) > \overline{g}\,x\}}{\mathbf{P}\{X > x\}} \;\geq\; \liminf_{x\to\infty} \frac{\mathbf{P}\{X > x(1+\varepsilon)\}}{\mathbf{P}\{X > x\}} \;=\; (1+\varepsilon)^{-\alpha}.$$

Again, the right-hand side tends to 1 as $\varepsilon \downarrow 0$. □

We shall employ Theorem 7 to show for several queueing models that the tail of the sojourn time distribution is as heavy as that of the (residual) service requirement distribution. Assuming that the service requirement distribution is regularly varying, it suffices to verify that $S(\tau)$, the sojourn time conditioned on the (residual) service requirement, satisfies Assumption 2. Parts *(a)* and *(c)* of Assumption 2 are often not hard to verify. We shall use the following variant of Markov's inequality to verify part *(b)*:

$$\mathbf{P}\{S(\tau) > t\} \leq \frac{\mathbf{E}S(\tau)^{\kappa} - (\mathbf{E}S(\tau))^{\kappa}}{(t - \mathbf{E}S(\tau))^{\kappa}}, \qquad \tau \geq 0,\, t > \mathbf{E}S(\tau), \tag{27}$$

where $\kappa \geq 2$. In [34] Markov's inequality itself was used, but for the analysis of the $M/G/1$ LCFS-NP below, the form of (27) is more convenient. To see that this inequality holds, let $c(y)$, $y \geq 0$ be a convex function with a convex derivative $c'(y)$ and $c(0) = c'(0) = 0$. If Y is a non-negative random variable and $t \geq \mathbf{E}Y$,

$$c(Y) - c(\mathbf{E}Y) - c'(\mathbf{E}Y)\,(Y - \mathbf{E}Y) \geq \mathbf{1}_{\{Y>t\}} c(t - \mathbf{E}Y),$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. Taking expectations with respect to the distribution of Y, we obtain

$$\mathbf{E}c(Y) - c(\mathbf{E}Y) \geq \mathbf{P}\{Y > t\} c(t - \mathbf{E}Y).$$

Choosing $c(y) = y^\kappa$, with $\kappa \geq 2$, leads to the desired result.

The strength of the method described here is that it does not rely on the availability of the LST for the sojourn time distribution. In particular, the method's flexibility was demonstrated in [32] where it was employed in the analysis of an $M/G/1$ PS queue with random service interruptions (for that model even basic performance measures

such as mean queue length are not available). A limitation of the method is that it relies on the fact that an extremal occurrence of the performance measure of interest (e.g., the sojourn time) is essentially caused by the occurrence of *a single* extremal input variable.

(i) The $M/G/1$ PS queue
Consider again the $M/G/1$ PS queue described in Section 2. In this section $S_{PS}(\tau)$ will stand for the sojourn time of a customer

with service requirement $\tau \geq 0$, arriving when the system has reached stationarity. As before, the unconditional sojourn time will be denoted with S_{PS}, i.e., $S_{PS} = S_{PS}(B)$, where the random variable B stands for the customer's service requirement. We first list some known results for the moments of $S_{PS}(\tau)$. Then we shall use these to verify Assumption 2 and subsequently apply Theorem 7.

It is well known [26, 40] that the mean of the conditional sojourn time is proportional to the service requirement:

$$\mathbf{E}S_{PS}(\tau) = \frac{\tau}{1-\rho}. \tag{28}$$

The variance of $S_{PS}(\tau)$ is given by

$$\mathbf{Var}S_{PS}(\tau) = \frac{2}{(1-\rho)^2} \int_{u=0}^{\tau} (\tau - u)\mathbf{P}\{W > u\}\mathrm{d}u, \tag{29}$$

cf. Yashkov [45]. As before, W is distributed as the stationary waiting time in the $M/G/1$ queue. When the second moment of the service requirement distribution is finite we have for $k = 2, 3, \ldots$, cf. [51],

$$\mathbf{E}S_{PS}(\tau)^k = \left(\frac{\tau}{1-\rho}\right)^k + \frac{\lambda k(k-1)\mathbf{E}B^2}{2(1-\rho)^{k+1}}\tau^{k-1} + \mathrm{o}(\tau^{k-1}), \qquad \tau \to \infty. \tag{30}$$

In the literature these results have mostly been obtained from expressions for the LST of $S_{PS}(\tau)$. However, (28)–(30) can be obtained directly from a set of differential equations instead of deriving the LST of $S_{PS}(\tau)$, see Yashkov [45, Rem. 3] for an outline of how this can be done for the variance of $S_{PS}(\tau)$, and [32, 33] for higher moments. Later in this section we use similar ideas to derive differential equations for the moments of the conditional sojourn time in the $M/G/1$ LCFS-NP.

We shall now provide a new proof of Theorem 2. As before, we assume that $B(\cdot) \in \mathcal{R}_{-\nu}$. We further require that $\nu \neq 2$; in [34] it is indicated how this technical assumption can be overcome. Focusing on the sojourn time of a particular customer, its own service requirement B will act as the causal variable X, i.e., in the light of Assumption 1 we choose $\alpha = \nu$. We now verify that Assumption 2 is automatically satisfied. First we note that the monotonicity of $\mathbf{P}\{S_{PS}(\tau) > t\}$ in τ, the last condition in Assumption 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 (28), we also have that Condition *(a)* of Assumption 2 holds with $\overline{g} = 1/(1-\rho)$.

We now focus on Condition *(b)* and first consider the case that $\nu > 2$, ensuring that $\mathbf{E}B^2 < \infty$. Choose any integer $\kappa > \nu$ and use (30) to conclude that Condition *(b)* is satisfied for any $\delta \in (0,1)$. Hence, Theorem 7 can be applied.

In the case that $1 < \nu < 2$, it follows from (29) that $\mathbf{Var}S_{PS}(\tau) = \mathrm{o}(\tau^{3-\nu+\varepsilon})$ for all $\varepsilon > 0$. Thus, Assumption 2 is satisfied (with $\kappa = 2$ and $0 < \delta < \nu - 1$) and, hence, Theorem 7 can again be applied. In both cases we conclude that, cf. Theorem 2,

$$\mathbf{P}\{S_{PS} > \frac{x}{1-\rho}\} \sim \mathbf{P}\{B > x\},$$

(ii) The $M/G/1$ LCFS-NP queue

In the LCFS-NP case we focus on the sojourn time, $S_{LNP}(\tau)$, of a tagged customer that enters the system when the *remaining* service requirement of the customer in service equals τ. Because of the service discipline, if there are any customers in the queue, these are overtaken by the new customer and they will have no influence on $S_{LNP}(\tau)$, which we may write as

$$S_{LNP}(\tau) \;=\; \tau + \sum_{n=1}^{N(\tau)} P_n + B,$$

where, by convention, we set the empty sum equal to 0. $N(\tau)$ denotes the number of customers that enter the system during the remaining service requirement τ of the customer in service. P_n, $n = 1, 2, \ldots$, is an i.i.d. sequence having the distribution of the busy period in the $M/G/1$. Indeed, all customers that enter during the time τ overtake the tagged customer, and the same holds for the customers that arrive during their service time, and so on. Finally, B denotes the tagged customer's own service requirement. If there is no customer in service upon arrival, the sojourn time is just equal to the customer's own service requirement B. Note that the probability of arriving to a non-empty system is ρ. We thus have for S_{LNP}, the unconditional sojourn time of the tagged customer,

$$\mathbf{P}\{S_{LNP} \leq t\} \;=\; (1-\rho)\,\mathbf{P}\{B \leq t\} + \rho\,\mathbf{P}\{S_{LNP}(B^r) \leq t\}, \qquad t \geq 0. \tag{31}$$

Here B^r denotes the (unconditional) residual service requirement of the customer in service, i.e., B^r has distribution $B^r(x)$, $x \geq 0$, and, hence,

$$\mathbf{P}\{S_{LNP}(B^r) \leq t\} \;=\; \int_{\tau=0}^{\infty} \mathbf{P}\{S_{LNP}(\tau) \leq t\} \mathrm{d}B^r(\tau), \qquad t \geq 0.$$

We now prove Theorem 4 for *non-integer* $\nu > 2$ (see Remark 3 below), by showing that if $B(\cdot) \in \mathcal{R}_{-\nu}$

and, hence, $B^r(\cdot) \in \mathcal{R}_{-\alpha}$ with $\alpha := \nu - 1$, then $S_{LNP}(\tau)$ satisfies Assumption 2 and, by Theorem 7,

$$\mathbf{P}\{S_{LNP}(B^r) > \frac{x}{1-\rho}\} \;\sim\; \mathbf{P}\{B^r > x\}.$$

Since $1 - B(x) = \mathrm{o}(1 - B^r(x))$, $x \to \infty$, we have, from (31),

$$\mathbf{P}\{S_{LNP} > \frac{x}{1-\rho}\} \;\sim\; \rho\,\mathbf{P}\{B^r > x\}, \tag{32}$$

in accordance with Theorem 4.

Remark 3. The case $1 < \nu < 2$ needs special treatment. As we will see below, the approach aims at verification of Condition *(b)* of Assumption 2, taking κ equal to the nearest integer larger than $\nu - 1$, which in this case would be $\kappa = 1$. However, for $\kappa < 2$ we can not use (27). It is possible to verify Condition *(b)* using different probabilistic arguments which, however, we shall not pursue here.

A different problem occurs when ν is integer-valued. The approach would aim at choosing $\kappa = \nu$. This requires that $\mathbf{E}B^\nu < \infty$, which may not be the case. (Recall that in the analysis of the $M/G/1$ PS system we could choose κ equal to any integer larger than ν.)

In order to verify the conditions in Assumption 2 we first derive differential equations for the moments of $S_{LNP}(\tau)$. The random variable P shall have the distribution of the busy period in the $M/G/1$ queue. We assume that $m < \nu < m + 1$, for some integer $m \geq 2$, and therefore $\mathbf{E}B^m < \infty$ and $\mathbf{E}P^m < \infty$.

Remark 4. The fact that $\mathbf{E}P^m < \infty$ if and only if $\mathbf{E}B^m < \infty$ is a consequence of Theorem 3 (since $S_{LPR} \overset{d}{=} P$). Note that this result was proven in [30] using Laplace-Transform techniques (cf. (18)). The same can be achieved using (probabilistic) sample-path arguments [49].

Conditioning on whether or not an arrival occurs during a time interval of length $\Delta > 0$ we obtain, for $k = 1, 2, \ldots, m$,

$$\mathbf{E}\Big[S_{LNP}(\tau + \Delta)^k\Big] = (1 - \lambda\Delta)\mathbf{E}\Big[(\Delta + S_{LNP}(\tau))^k\Big]$$
$$+\lambda\Delta\mathbf{E}\Big[(\Delta + S_{LNP}(\tau) + P)^k\Big] + \mathrm{o}(\Delta), \qquad \Delta \to 0.$$

Re-arranging terms, dividing by Δ, passing $\Delta \to 0$ and using $\mathbf{E}P = \beta/(1 - \rho)$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{E}S_{LNP}(\tau)^k \;=\; \frac{k}{1-\rho}\mathbf{E}S_{LNP}(\tau)^{k-1} + \lambda \sum_{j=0}^{k-2} \binom{k}{j} \mathbf{E}S_{LNP}(\tau)^j \mathbf{E}P^{k-j}, \tag{33}$$

with initial condition $\mathbf{E}S(0)^k = \mathbf{E}B^k$. By solving (33) for $k = 1$ (setting the empty sum equal to 0), it can straightforwardly be verified that

$$\mathbf{E}S_{LNP}(\tau) \;=\; \beta + \frac{\tau}{1-\rho},$$

which verifies Condition *(a)* of Assumption 2. Recursively solving (33) for $k = 2, 3, \ldots, m$, leads to the general form

$$\mathbf{E}S_{LNP}(\tau)^k \;=\; \left(\frac{\tau}{1-\rho}\right)^k + p_{k-1}(\tau), \tag{34}$$

where $p_{k-1}(\tau)$ denotes a polynomial in τ of degree $k - 1$. The coefficients of this polynomial can be obtained recursively by substitution into (33); in particular, $p_{k-1}(0) = \mathbf{E}B^k$. Taking $\kappa = m$ we may use (27) and (34) to show that Condition *(b)* of Assumption 2 is satisfied (for any $0 < \delta < 1$). Finally, Condition *(c)* can again be verified by a sample-path argument similar to that in the analysis of the $M/G/1$ PS system.

(iii) The $M/G/1$ FBPS queue
With the FBPS discipline, the customers which so far have received the least

amount of service share equally in the total capacity. Using Theorem 7, we shall prove that the sojourn time tail is just as heavy as the service requirement tail if $B(\cdot) \in \mathcal{R}_{-\nu}$ with $1 < \nu < 2$. (For $\nu \geq 2$ we need to study higher moments of the conditional sojourn time.) $S_{FB}(\tau)$ denotes the sojourn time of a customer with service requirement τ. With a straightforward sample-path argument it can be shown that $S_{FB}(\tau)$ is stochastically non-decreasing.

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

$$\mathbf{E}S_{FB}(\tau) = \frac{\tau}{1-\lambda h_1(\tau)} + \frac{\lambda h_2(\tau)}{2(1-\lambda h_1(\tau))^2}, \tag{35}$$

$$\mathbf{Var}S_{FB}(\tau) = \frac{\lambda h_3(\tau)}{3(1-\lambda h_1(\tau))^3} + \frac{\lambda \tau h_2(\tau)}{(1-\lambda h_1(\tau))^3} + \frac{3(\lambda h_2(\tau))^2}{4(1-\lambda h_1(\tau))^4}, \tag{36}$$

cf. Yashkov [46, Form. (6.2) and (6.3)]. The functions $h_j(\tau)$, $j = 1, 2, 3$, are given by

$$h_j(\tau) = j \int_{x=0}^{\tau} x^{j-1} (1 - B(x))\, \mathrm{d}x. \tag{37}$$

These expressions can again be used [34] to prove that, for all $\varepsilon > 0$,

$$\mathbf{E}S_{FB}(\tau) \sim \frac{\tau}{1-\rho}, \qquad \mathbf{Var}S_{FB}(\tau) = \mathrm{o}(\tau^{3-\nu+\varepsilon}), \quad \tau \to \infty.$$

Consequently, Assumption 2 is implied by Assumption 1 (choosing $\kappa = 2$ and $0 < \delta < \nu - 1$) and we may again apply Theorem 7 to show that $\mathbf{P}\{S_{FB} > \frac{x}{1-\rho}\} \sim \mathbf{P}\{B > x\}$, cf. Theorem 5.

(iv) The $M/G/1$ SRPTF queue

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 (Shortest Remaining Processing Time First). 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.

As in the $M/G/1$ FBPS queue, we restrict ourselves to the case $B(\cdot) \in \mathcal{R}_{-\nu}$ with $1 < \nu < 2$. We further assume that $B(x)$ is a continuous function, hence, with probability 1, no two customers in the system have the same remaining service requirement, see [41].

The sojourn time can be decomposed into two different periods: The waiting time (the time until the customer is first taken into service) and the residence time (the remainder of the sojourn time). The residence time may contain service pre-emption periods caused by customers with a smaller service requirement. 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 $S_{SR}(\tau) = W(\tau) + R(\tau)$. We define $\rho(\tau)$ as the traffic load of customers with an amount of work less than or equal to τ,

$$\rho(\tau) := \lambda \int_{t=0}^{\tau} t \mathrm{d}B(t). \tag{38}$$

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

$$\mathbf{E}W(\tau) = \lambda \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 (1 - B(\tau))}{2 (1 - \rho(\tau))^2}, \tag{39}$$

$$\mathbf{E}W(\tau)^2 = \lambda \frac{\int_{t=0}^{\tau} t^3 \mathrm{d}B(t) + \tau^3 \left(1 - B(\tau)\right)}{3 \left(1 - \rho(\tau)\right)^3} + \lambda^2 \int_{t=0}^{\tau} t^2 \mathrm{d}B(t) \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 \left(1 - B(\tau)\right)}{\left(1 - \rho(\tau)\right)^4}, \tag{40}$$

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

$$\mathbf{E}R(\tau) = \int_{t=0}^{\tau} \frac{1}{1 - \rho(t)} \mathrm{d}t, \tag{41}$$

$$\mathbf{Var}R(\tau) = \lambda \int_{t=0}^{\tau} \frac{\int_{u=0}^{t} u^2 \mathrm{d}B(u)}{\left(1 - \rho(t)\right)^3} \mathrm{d}t, \tag{42}$$

cf. [41]. These expressions may again be used [34] to verify that, when $1 < \nu < 2$, Assumption 1 implies Assumption 2 and, hence, $\mathbf{P}\{S_{SR} > \frac{x}{1-\rho}\} \sim \mathbf{P}\{B > x\}$, cf. Theorem 6.

5 Sample-Path Techniques

In the present section we describe how sample-path techniques may be used to determine the tail asymptotics of the delay distribution in the $M/G/1$ queue for various disciplines. By definition, the tail distribution of a random variable reflects the occurrence of rare events. Large-deviations theory suggests that, given that a rare event occurs, it happens with overwhelming probability in the most likely way. In case light-tailed processes are involved, the most likely path typically consists of an extremely long sequence of slightly unusual events, which conspire to make the rare event under consideration occur, see for instance Anantharam [2]. In contrast, for heavy-tailed characteristics, the most likely scenario usually involves just a single catastrophic event (or in general, a 'minimal combination' of disastrous events that is required to cause the event under consideration to happen). Typically, the scenario entails the arrival of a customer with an exceedingly large service requirement.

The fact that the most likely scenario usually involves just a single exceptional event, provides a heuristic method for obtaining the tail asymptotics by simply computing the probability of that scenario occurring. By way of illustration, we now sketch a heuristic derivation of the tail asymptotics of the workload V in the $M/G/1$ queue as described in Section 2.

Let us focus on the workload in the system at time $t = 0$. The assumption is that a large workload level is most likely due to the prior arrival of a customer with a large service requirement B, let us say at time $t = -y$. (Of course, this assumption is nothing but an educated guess at this stage. However, it turns out that this supposition leads to the correct result, and can actually be strengthened into a rigorous proof, as will be illustrated below.) Note that from time $t = -y$ onward, the workload decreases in a roughly linear fashion at rate $1 - \rho$. So in order for the workload at time $t = 0$ to exceed the level x, the service requirement B must be larger than $x + y(1-\rho)$. Observing that customers arrive as a Poisson process of rate λ, integrating w.r.t. y, and making the substitution $z = x + y(1-\rho)$, we obtain, for large x,

$$\mathbf{P}\{V > x\} \approx \int\limits_{y=0}^{\infty} \mathbf{P}\{B > x + y(1-\rho)\} \lambda \mathrm{d}y = \frac{\lambda}{1-\rho} \int\limits_{z=x}^{\infty} \mathbf{P}\{B > z\} \mathrm{d}z = \frac{\rho}{1-\rho} \mathbf{P}\{B^r > x\}. \tag{43}$$

With some additional effort, the heuristic derivation can often be strengthened into a rigorous proof. The typical approach consists of deriving lower and upper bounds which asymptotically coincide. It is often relatively straightforward to convert the heuristic arguments into a strict lower bound by calculating the probability of the most likely scenario occurring. The construction of a suitable upper bound tends to be more challenging. The upper bound usually consists of a dominant term which corresponds to the probability of the most likely scenario. The main difficulty lies in showing that this scenario is indeed the only plausible one, in the sense that all other possible sample paths do not significantly contribute. This is done by partitioning the 'irrelevant' sample paths into a few sets which must then all be shown to have an asymptotically negligible probability.

Although the above approach is fairly typical, it is hard to describe a universal method that can be mechanically executed. The identification of the most likely scenario is problem-specific and requires some sort of an educated guess. Categorizing the 'irrelevant' sample paths is not an automatic task either. The next lemma however characterizes the structure that typically emerges.

Lemma 2. *Suppose that for any $\delta > 0$, $\epsilon > 0$,*

$$\mathbf{P}\{X > x\} \geq F(-\delta)\mathbf{P}\{Y > G(\epsilon)x\} \prod_{i=1}^{K} \mathbf{P}\{D_i^{-\delta,\epsilon}(x)\}, \tag{44}$$

$$\mathbf{P}\{X > x\} \leq F(\delta)\mathbf{P}\{Y > G(-\epsilon)x\} + \sum_{j=1}^{L} \mathbf{P}\{E_j^{\delta,-\epsilon}(x)\}, \tag{45}$$

$\mathbf{P}\{Y > x\}$ is regularly varying of index $-\nu$, $\lim_{\delta \to 0} F(\delta) = F$, $\lim_{\epsilon \to 0} G(\epsilon) = G$, $\mathbf{P}\{D_i^{-\delta,\epsilon}(x)\} \to 1$ as $x \to \infty$, and $\mathbf{P}\{E_j^{\delta,-\epsilon}(x)\} = \mathrm{o}(x^{-\nu})$ as $x \to \infty$. Then

$$\mathbf{P}\{X > x\} \sim F\mathbf{P}\{Y > Gx\}.$$

Proof. The proof is straightforward. Relying on the lower bound (44) and the fact that $\mathbf{P}\{D_i^{-\delta,\epsilon}(x)\} \to 1$ as $x \to \infty$, we obtain

$$\liminf_{x \to \infty} \frac{\mathbf{P}\{X > x\}}{F\mathbf{P}\{Y > Gx\}} \geq \frac{F(-\delta)}{F} \liminf_{x \to \infty} \frac{\mathbf{P}\{Y > G(\epsilon)x\}}{\mathbf{P}\{Y > Gx\}}.$$

Letting $\delta, \epsilon \downarrow 0$, and recalling that $\mathbf{P}\{Y > x\}$ is regularly varying, we find

$$\liminf_{x \to \infty} \frac{\mathbf{P}\{X > x\}}{F\mathbf{P}\{Y > Gx\}} \geq 1.$$

Similarly, using the upper bound (45), the fact that $\mathbf{P}\{E_j^{\delta,-\epsilon}(x)\} = \mathrm{o}(x^{-\nu})$ as $x \to \infty$, and observing that $\mathbf{P}\{Y > x\}$ is regularly varying of index $-\nu$, we deduce

$$\limsup_{x \to \infty} \frac{\mathbf{P}\{X > x\}}{F\mathbf{P}\{Y > Gx\}} \leq \frac{F(\delta)}{F} \liminf_{x \to \infty} \frac{\mathbf{P}\{Y > G(-\epsilon)x\}}{\mathbf{P}\{Y > Gx\}}.$$

Letting $\delta, \epsilon \downarrow 0$, we conclude

$$\limsup_{x \to \infty} \frac{\mathbf{P}\{X > x\}}{F\mathbf{P}\{Y > Gx\}} \leq 1.$$

□

It is worth observing that the above proof technique in fact extends to the class of *intermediately* regularly varying distributions.

As a 'toy example', we now sketch how the above lemma may be used to strengthen the heuristic derivation of (43) into a rigorous proof. The approach is similar as outlined in Chapter 2 of Zwart [48]. We use the time-reversed sample-path representation

$$V \overset{d}{=} \sup_{t\geq 0}\{A(0,t)-t\} \tag{46}$$

with $A(0,t)$ denoting the amount of work arriving in the time interval $(0,t)$.

We first construct a lower bound of the form (44). Define $U^c := \sup_{t\geq 0}\{ct - A(0,t)\}$, with $c<\rho$. For any $\delta>0$, $\epsilon>0$,

$$\mathbf{P}\{V>x\} \geq \int\limits_{y=0}^{\infty} \mathbf{P}\{A(0,y)+B-y>x\}\lambda \mathrm{d}y$$

$$\geq \lambda \int\limits_{y=0}^{\infty} \mathbf{P}\{A(0,y)-y(\rho-\delta)\geq -\epsilon x\}\mathbf{P}\{B>x(1+\epsilon)+y(1-\rho+\delta)\}\mathrm{d}y$$

$$\geq \mathbf{P}\{\inf_{u\geq 0}\{A(0,u)-u(\rho-\delta)\}\geq -\epsilon x\}\lambda \int\limits_{y=0}^{\infty} \mathbf{P}\{B>x(1+\epsilon)+y(1-\rho+\delta)\}\mathrm{d}y$$

$$= \frac{\rho}{1-\rho+\delta}\mathbf{P}\{B^r>x(1+\epsilon)\}\mathbf{P}\{U^{\rho-\delta}\leq \epsilon x\}.$$

Note that $\mathbf{P}\{U^{\rho-\delta}\leq \epsilon x\}\to 1$ as $x\to\infty$ because of the law of large numbers.

We now proceed to derive an upper bound of the form (45). For any interval $I\subseteq \mathbb{R}^+$, define $V(I) := \sup_{t\in I}\{A(0,t)-t\}$. For any $y\geq 0$, let $N_y(I)$ be the number of customers arriving during the time interval I whose service requirement exceeds the value y. Then, for all $M\geq 0$,

$$\begin{aligned}
&\mathbf{P}\{V>x\} \leq \mathbf{P}\{V([0,Mx])>x\} + \mathbf{P}\{V((Mx,\infty))>x\}\\
&= \mathbf{P}\{V([0,Mx])>x; N_{\epsilon x}([0,Mx])=0\} + \mathbf{P}\{V([0,Mx])>x; N_{\epsilon x}([0,Mx])=1\}\\
&+ \mathbf{P}\{V([0,Mx])>x; N_{\epsilon x}([0,Mx])\geq 2\} + \mathbf{P}\{V((Mx,\infty))>x\}.
\end{aligned}$$

The second term corresponds to the only plausible scenario and is dominant. As in [50], it may be shown that for any $\delta>0$, $\epsilon>0$, as $x\to\infty$,

$$\begin{aligned}
&\mathbf{P}\{V([0,Mx])>x; N_{\epsilon x}([0,Mx])=1\}\\
&\leq \int_{y=0}^{\infty} \mathbf{P}\{B>x(1-\epsilon)+y(1-\rho-\delta)\}\lambda \mathrm{d}y + \mathrm{o}(x^{1-\nu})\\
&= \frac{\rho}{1-\rho-\delta}\mathbf{P}\{B^r>x(1-\epsilon)\} + \mathrm{o}(x^{1-\nu}).
\end{aligned}$$

Applying Lemma 2 completes the proof, once we have shown that each of the other three terms can asymptotically be neglected.

For the first term, one may exploit a powerful lemma of Resnick & Samorodnitsky [39] to show that for any $\mu>0$ there exists an $\epsilon>0$ such that

$$\mathbf{P}\{V([0,Mx])>x; N_{\epsilon x}([0,Mx])=0\} = \mathrm{o}(x^{-\mu}), \qquad x\to\infty.$$

The idea is that when there are no large service requirements, the process $\{A(0,t)-t\}$ cannot significantly deviate from its normal drift over long intervals of the order x, so that the workload cannot reach a large level.

In order to control the third term, it may be checked using elementary arguments that $\mathbf{P}\{N_{\epsilon x}([0,Mx]) \geq 2\} = \mathrm{o}(x^{1-\nu})$ as $x \to \infty$. Note that the probability of two large service requirements occurring in a time interval of order x asymptotically vanishes compared to that of just one large service requirement.

Finally, in order to restrain the last term, it may be shown using results from Mikosch [31] that $\lim_{M\to\infty} \limsup_{x\to\infty} \mathbf{P}\{V((Mx,\infty)) > x\}/\mathbf{P}\{V > x\} = 0$. The idea is that a large workload level of order x must build up 'in linear time', since otherwise the process $\{A(0,t)-t\}$ must deviate from its normal drift for a prohibitively long period of time.

The above proof exploits and confirms the large-deviations notion that a large workload level is typically due to a single large service requirement by implicitly characterizing the most likely sample path. We refer to Baccelli and Foss [4] for yet stronger characterizations.

It is worth emphasizing that we used the above proof technique for illustration purposes only. The machinery is unnecessarily heavy for determining the workload asymptotics in the ordinary $M/G/1$ queue, given that a convenient expression for the LST is available, which can be explicitly inverted as shown in Section 3. The true merits of the methodology become manifest in more complicated systems, such as fluid queues or GPS models, where typically no useful expression for the LST is available [9, 12, 50].

5.1 The Single-Class Case

We now turn the attention to the tail asymptotics of the delay distribution in the $M/G/1$ queue. In contrast to the workload distribution, the delay distribution *does* strongly depend on the service discipline that is used.

(i) The $M/G/1$ FCFS queue
For FCFS, the waiting time is simply equal to the workload at the time of arrival. Because of the PASTA property, it then follows from (43) that

$$\mathbf{P}\{W_{FCFS} > x\} \sim \frac{\rho}{1-\rho}\mathbf{P}\{B^r > x\},$$

which agrees with Theorem 1.

(ii) The $M/G/1$ LCFS-NP queue
For LCFS Non-Preemptive priority, the waiting time is equal to 0 with probability $1-\rho$, and with probability ρ it is equal to a busy period starting with a residual service requirement, which gives

$$\mathbf{P}\{W_{LNP} > x\} \sim \rho\mathbf{P}\{B^r > x(1-\rho)\}, \tag{47}$$

as asserted in Theorem 4.

A heuristic derivation of the above formula proceeds as follows. Consider a tagged customer arriving at time $t = 0$. The premise is that a long waiting time is most likely due to a large service requirement B of the customer in service, if any. The waiting time W of the tagged customer then consists of the remaining service requirement, $B - y$, plus the amount of work arriving during its own waiting time, which is approximately ρW, so that $W \approx B - y + \rho W$, or equivalently, $W \approx (B-y)/(1-\rho)$. So in order for the waiting time to exceed the value x, the service requirement B must be larger than $y + x(1-\rho)$. Thus, observing that arrivals occur as a Poisson process of rate λ, and integrating w.r.t. y, we find, for large x,

$$\mathbf{P}\{W_{LNP} > x\} \approx \int\limits_{y=0}^{\infty} \mathbf{P}\{B > x(1-\rho) + y\}\lambda \mathrm{d}y = \rho \mathbf{P}\{B^r > x(1-\rho)\},$$

which is in agreement with (47). The above heuristic derivation may be translated into a rigorous proof in a similar fashion as indicated for the workload asymptotics.

(iii) The $M/G/1$ LCFS-PR queue
For LCFS Preemptive Resume, the sojourn time is simply equal to the busy period, yielding

$$\mathbf{P}\{S_{LPR} > x\} \sim \frac{1}{1-\rho}\mathbf{P}\{B > x(1-\rho)\},$$

as stated in Theorem 3. A sample-path proof of the tail asymptotics of the busy-period distribution may be found in Zwart [49].

(iv) The $M/G/1$ PS queue
We now turn to the tail asymptotics of the sojourn time for the Processor-Sharing discipline. Consider a tagged customer arriving at time $t = 0$. The sojourn time S of the tagged customer consists of its own service requirement B plus the amount of service provided to other customers during its sojourn time. In case of a long sojourn time, the amount of service received by other customers will be approximately ρS, so that $S \approx B + \rho S$, or equivalently, $S \approx B/(1-\rho)$. The assumption is thus that a long sojourn time is most likely due to a large service requirement of the tagged customer itself, suggesting that, for large x,

$$\mathbf{P}\{S_{PS} > x\} \sim \mathbf{P}\{B > x(1-\rho)\}, \tag{48}$$

which corroborates with Theorem 2.

We now show how the above rough derivation may be used as the basis for a rigorous proof of (48) using lower and upper bounds along the lines of Lemma 2. The proof is similar to that in Jelenković & Momcilović [24]. Let B_0 and S_0 be the service requirement and the sojourn time, respectively, of a tagged customer arriving at time $t = 0$. Let B_i and T_i denote the service requirement and the arrival time of the i-th customer arriving after time $t = 0$. Let $L(0)$ be the number of customers in the system just before time $t = 0$, and let B_l^r denote the remaining service requirement of the l-th customer. We use the sample-path representation

$$S_0 = B_0 + \sum_{l=1}^{L(0)} \min\{B_l^r, B_0\} + \sum_{i=1}^{N((0,S_0))} \min\{B_i, R_0(T_i)\}, \tag{49}$$

with $N((0,t))$ denoting the number of customers arriving during the time interval $(0,t)$ and $R_0(t)$ the remaining service requirement of the tagged customer at time t, see for instance Wolff [44].

The next lemma presents a lower bound for the sojourn time of the tagged customer. Denote $Z(t) := \sum\limits_{i=1}^{N((0,t))} \max\{B_i - R_0(T_i), 0\}$.

Lemma 3. *For any $\delta > 0$,*

$$S_0(1-\rho+\delta) \geq B_0 - U^{\rho-\delta} - Z(S_0).$$

Proof. Using the representation (49), we have

$$\begin{aligned}
S_0(1-\rho+\delta) &= B_0 + \sum_{l=1}^{L(0)} \min\{B_l^r, B_0\} + \sum_{i=1}^{N((0,S_0))} \min\{B_i, R_0(T_i)\} - (\rho-\delta)S_0 \\
&\geq B_0 + \sum_{i=1}^{N((0,S_0))} B_i - (\rho-\delta)S_0 + \sum_{i=1}^{N((0,S_0))} \min\{B_i, R_0(T_i)\} - \sum_{i=1}^{N((0,S_0))} B_i \\
&= B_0 + A(0,S_0) - (\rho-\delta)S_0 + \sum_{i=1}^{N((0,S_0))} \min\{R_0(T_i) - B_i, 0\} \\
&\geq B_0 + \inf_{t\geq 0}\{A(0,t) - (\rho-\delta)t\} - \sum_{i=1}^{N((0,S_0))} \max\{B_i - R_0(T_i), 0\} \\
&= B_0 - U^{\rho-\delta} - Z(S_0).
\end{aligned}$$

□

The next lemma provides an upper bound for the sojourn time of the tagged customer. For any $y > 0$, let $A_y(0,t)$ be a version of the process $A(0,t)$ where all service requirements of arriving customers are truncated at the level y. For any $c > \rho$, define $V_y^c := \sup_{t\geq 0}\{A_y(0,t) - ct\}$.

Lemma 4. *For any $\delta > 0$,*

$$(1-\rho-\delta)S_0 \leq \hat{B}_0 + V_{B_0}^{\rho+\delta},$$

with $\hat{B}_0 := B_0 + \sum_{l=1}^{L(0)} \min\{B_l^r, B_0\}$.

Proof. Using the representation (49),

$$\begin{aligned}
S_0(1-\rho-\delta) &= B_0 + \sum_{l=1}^{L(0)} \min\{B_l^r, B_0\} + \sum_{i=1}^{N((0,S_0))} \min\{B_i, R_0(T_i)\} - (\rho+\delta)S_0 \\
&\leq \hat{B}_0 + \sum_{i=1}^{N((0,S_0))} \min\{B_i, B_0\} - (\rho+\delta)S_0 \\
&= \hat{B}_0 + A_{B_0}(0,S_0) - (\rho+\delta)S_0 \\
&\leq \hat{B}_0 + \sup_{t\geq 0}\{A_{B_0}(0,t) - (\rho+\delta)t\} \\
&= \hat{B}_0 + V_{B_0}^{\rho+\delta}.
\end{aligned}$$

□

The above two lemmas provide the necessary ingredients for the proof of (48) along the lines of Lemma 2.

Proof of Theorem 2 (Lower bound) Using Lemma 3, noting that $S_0 \geq B_0$, we obtain

$$\begin{aligned}\mathbf{P}\{S_{PS} > x\} &\geq \mathbf{P}\{B_0 - U^{\rho-\delta} - Z(S_0) > (1-\rho+\delta)x\}\\ &\geq \mathbf{P}\{B_0 > (1-\rho+\delta+2\epsilon)x\}\mathbf{P}\{U^{\rho-\delta} \leq \epsilon x\} \inf_{y \geq (1-\rho+\delta+2\epsilon)x} \mathbf{P}\{Z(y) > \epsilon x\}.\end{aligned}$$

Because of the law of large numbers, $\mathbf{P}\{U^{\rho-\delta} \leq \epsilon x\} \to 1$ as $x \to \infty$. As observed in [24], $\inf\limits_{y \geq (1-\rho+\delta+2\epsilon)x} \mathbf{P}\{Z(y) > \epsilon x\} \to 1$ as $x \to \infty$.
(Upper bound) Using Lemma 4, we find

$$\begin{aligned}\mathbf{P}\{S_{PS} > x\} &\leq \mathbf{P}\{\hat{B}_0 + V_{B_0}^{\rho+\delta} > (1-\rho-\delta)x\}\\ &\leq \mathbf{P}\{\hat{B}_0 > (1-\rho-\delta-\epsilon)x\} + \mathbf{P}\{V_{B_0}^{\rho+\delta} > \epsilon x\}.\end{aligned}$$

As demonstrated in [24], $\mathbf{P}\{B_0 > x\} \sim \mathbf{P}\{\hat{B}_0 > x\}$ and $\mathbf{P}\{V_{B_0}^{\rho+\delta} > \epsilon x\} = \mathrm{o}(x^{1-\nu})$ as $x \to \infty$. Invoking Lemma 2 then completes the proof. □

5.2 The Multi-Class Case

We finally consider the tail asymptotics of the waiting time in the multi-class $M/G/1$ queue with priorities as described in Subsection 3.2. We focus on the case of a Non-Preemptive priority discipline. As mentioned in Subsection 3.2, the tail asymptotics in the case of a Preemptive-Resume policy immediately follow from the results for a high-priority class in isolation and a low-priority class in the Non-Preemptive priority scenario. We assume that the service requirement distribution of at least one of the classes has a regularly varying tail. Let $\mathcal{M}$ be the index set of the classes with the 'heaviest' tail.

Consider a tagged class-k customer arriving at time $t = 0$. The premise is that a long waiting time is most likely due to the prior arrival of a customer with a large service requirement B, let us say at time $t = -y$, which may belong to any of the classes $m \in \mathcal{M}$. Of course, how likely it is for the culprit customer to belong to a given class $m \in \mathcal{M}$ depends on the arrival rates and mean service requirements of the various classes. Due to the Non-Preemptive priority policy, the identity of the culprit customer is not of any relevance for the impact on the tagged customer. However, the effect does strongly depend on the identity of the tagged customer itself. For compactness, denote $\sigma_k = \sum\limits_{l=1}^{k} \rho_l$. Note that from time $t = -y$ onward, the amount of work in the system that has precedence over the service of the tagged customer decreases in a roughly linear fashion at rate $1 - \sigma_k$. In addition, the tagged customer must wait for the amount of work arriving during its waiting time W_k from higher-priority classes at rate σ_{k-1}. Thus, $W_k \approx B - y(1-\sigma_k) + W_k\sigma_{k-1}$, or equivalently, $W_k \approx (B - y(1-\sigma_k))/(1-\sigma_{k-1})$. So in order for the waiting time of the tagged customer to exceed the value x, the service requirement B must be larger than $x(1-\sigma_{k-1}) + y(1-\sigma_k)$. Observing that class-$m$ customers arrive as a Poisson process of rate λ_m, integrating w.r.t. y, we obtain, for large x,

$$\begin{aligned}\mathbf{P}\{W_k > x\} &\approx \sum_{m \in \mathcal{M}} \int_{y=0}^{\infty} \mathbf{P}\{B_m > x(1-\sigma_{k-1}) + y(1-\sigma_k)\}\lambda_m \mathrm{d}y\\ &= \sum_{m \in \mathcal{M}} \frac{\rho_m}{1-\sigma_k} \mathbf{P}\{B_m^r > x(1-\sigma_{k-1})\}.\end{aligned}$$

References

1. Abate, J., Whitt, W. (1997). Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. *Queueing Systems* **25**, 173–233.
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. Baccelli, F., Foss, S. (2001). Moments and tails in monotone-separable stochastic networks. Research Report RR 4197, INRIA Rocquencourt.
5. Beran, J., Sherman, R., Taqqu, M.S., Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. IEEE *Trans. Commun.* **43**, 1566–1579.
6. Bingham, N.H., Doney, R.A. (1974). Asymptotic properties of super-critical branching processes. I: The Galton-Watson process. *Adv. Appl. Prob.* **6**, 711–731.
7. Bingham, N.H., Goldie, C.M., Teugels, J.L. (1987). *Regular Variation* (Cambridge University Press, Cambridge, UK).
8. Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Coupled processors with regularly varying service times. In: *Proc. Infocom 2000 Conference*, Tel-Aviv, Israel, 157–164.
9. Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Reduced-load equivalence and induced burstiness in GPS queues. *Queueing Systems*, to appear.
10. Borst, S.C., Boxma, O.J., Van Uitert, M.G.J. (2001). Two coupled queues with heterogeneous traffic. In: *Teletraffic Engineering in the Internet Era, Proc. ITC-17*, Salvador da Bahia, Brazil, eds. J. Moreira de Souza, N.L.S. da Fonseca, E.A. de Souza e Silva (North-Holland, Amsterdam), 1003–1014.
11. Borst, S.C., Boxma, O.J., Van Uitert, M.G.J. (2002). The asymptotic workload behavior of two coupled queues. CWI Report PNA-R0202. Submitted for publication.
12. Borst, S.C., Zwart, A.P. (2001). Fluid queues with heavy-tailed $M/G/\infty$ input. SPOR-Report 2001-02, Department of Mathematics and Computer Science, Eindhoven University of Technology. Submitted for publication.
13. Boxma, O.J., Cohen, J.W. (2000). The single server queue: Heavy tails and heavy traffic. In: *Self-similar Network Traffic and Performance Evaluation*, eds. K. Park, W. Willinger (Wiley, New York), 143–169.
14. Boxma, O.J., Cohen, J.W., Deng, Q. (1999). Heavy-traffic analysis of the M/G/1 queue with priority classes. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 1157–1167.
15. Boxma, O.J., Deng, Q., Resing, J.A.C. (2000). Polling systems with regularly varying service and/or switchover times. *Adv. Perf. Anal.* **3**, 71–107.
16. Boxma, O.J. Dumas, V. (1998). The busy period in the fluid queue. *Perf. Eval. Review* **26**, 100–110.
17. Cohen, J.W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* **10**, 343–353.
18. Cohen, J.W. (1982). *The Single Server Queue* (North-Holland Publ. Cy., Amsterdam; revised edition).
19. Cohen, J.W., Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis* (North-Holland Publ. Cy., Amsterdam).
20. Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web Traffic: evidence and possible causes. In: *Proc. ACM Sigmetrics '96*, 160–169.
21. Deng, Q. (2001). The two-queue $E/1-L$ polling model with regularly varying service and/or switchover times. SPOR-Report 2001-09, Department of Mathematics and Computer Science, Eindhoven University of Technology.
22. Fayolle, G., Iasnogorodski, R. (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Z. Wahrsch. Verw. Gebiete* **47**, 325–351.
23. Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Vol. II* (Wiley, New York).
24. Jelenković, P.R., Momcilović, P. (2002). Resource sharing with subexponential distributions. In: *Proc. IEEE Infocom 2002*, New York NY, USA, to appear.
25. Karamata, J. (1930). Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)* **4**, 38–53.
26. Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications* (Wiley, New York).
27. Klüppelberg, C. (1988). Subexponential distributions and integrated tails. *J. Appl. Prob.* **25**, 132–141.
28. Konheim, A.G., Meilijson, I., Melkman, A. (1981). Processor sharing of two parallel lines. *J. Appl. Prob.* **18**, 952–956.
29. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V. (1994). On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM *Trans. Netw.* **2**, 1–15.
30. 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.

31. Mikosch, T. (1999). Regular variation, subexponentiality and their applications in probability theory. EURANDOM Report 99-013.
32. Núñez-Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks.* Ph.D. thesis, Eindhoven University of Technology, ISBN 90-646-4667-8 (also available from the author upon request).
33. Núñez-Queija, R. (2000). Sojourn times in a processor-sharing queue with service interruptions. *Queueing Systems* **34**, 351–386.
34. Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. CWI Report PNA-R0201. Submitted for publication.
35. Ott, T.J. (1984). The sojourn-time distribution in the $M/G/1$ queue with processor sharing. *J. Appl. Prob.* **21**, 360–378.
36. Pakes, A.G. (1975). On the tails of waiting-time distributions. *J. Appl. Prob.* **12**, 555–564.
37. Parekh, A.K., Gallager, R.G. (1993). A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM *Trans. Netw.* **1**, 344–357.
38. Paxson, A., Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. IEEE/ACM *Trans. Netw.* **3**, 226–244.
39. Resnick, S., Samorodnitsky, G. (1999). Activity periods of an infinite server queue and performance of certain heavy-tailed fluid queues. *Queueing Systems* **33**, 43–71.
40. Sakata, M., Noguchi, S., Oizumi, J. (1971). An analysis of the $M/G/1$ queue under round-robin scheduling. *Oper. Res.* **19**, 371–385.
41. Schrage, L.E., Miller, L.W. (1966). The queue $M/G/1$ with the shortest remaining processing time discipline. *Oper. Res.* **14**, 670–684.
42. Schassberger, R. (1984). A new approach to the $M/G/1$ processor sharing queue. *Adv. Appl. Prob.* **16**, 802–813.
43. 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.
44. Wolff, R.W. (1989). *Stochastic Modeling and the Theory of Queues* (Prentice Hall, Englewood Cliffs).
45. Yashkov, S.F. (1983). A derivation of response time distribution for a $M/G/1$ processor-sharing queue. *Prob. Control Inf. Theory* **12**, 133–148.
46. Yashkov, S.F. (1987). Processor-sharing queues: Some progress in analysis. *Queueing Systems* **2**, 1–17.
47. Zwart, A.P. (1999). Sojourn times in a multiclass processor sharing queue. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 335–344.
48. Zwart, A.P. (2001). *Queueing Systems with Heavy Tails.* Ph.D. thesis, Eindhoven University of Technology.
49. Zwart, A.P. (2001). Tail asymptotics for the busy period in the $GI/G/1$ queue. *Math. Oper. Res.* **26**, 485–493.
50. Zwart, A.P., Borst, S.C., Mandjes, M. (2001). Exact asymptotics for fluid queues fed by multiple heavy-tailed On-Off flows. *Ann. Appl. Prob.*, to appear. Shortened version in: *Proc. Infocom 2001*, Anchorage AK, USA, 279–288.
51. Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* **35**, 141–166.