Large Deviation Analysis of Subexponential Waiting Times in a Processor Sharing Queue

Predrag Jelenković and Petar Momčilović
Department of Electrical Engineering
Columbia University
New York, NY 10027
{predrag, petar}@ee.columbia.edu

October 2001, revised November 2002

Abstract

We investigate the distribution of the waiting time V in a stable M/G/1 processor sharing queue with traffic intensity $\rho < 1$. When the distribution of a customer service request B belongs to a large class of subexponential distributions with tails heavier than $e^{-\sqrt{x}}$, it is shown that

$$\mathbb{P}[V > x] = \mathbb{P}[B > (1-\rho)x](1 + o(1)) \quad \text{as} \quad x \to \infty.$$

Furthermore, we demonstrate that the preceding relationship does not hold if the service distribution has a lighter tail than $e^{-\sqrt{x}}$.

Keywords: M/G/1 queue, processor sharing, waiting time distribution, large deviations, subexponential distributions

This work is supported by the National Science Foundation PECASE Award No. 9875156.
`Mathematics of Operations Research, to appear`

1 Introduction

Processor sharing algorithms have been widely used in modeling computer and communication systems. These types of algorithms allow for efficient and fair distribution of resources. Early work on processor sharing [12] was motivated by the study of multi-user mainframe computer systems. Renewed interest in the processor sharing queues stems from their application in modeling of computer communication networks and Web servers. In particular, we would like to mention prospects of modeling congested links with TCP traffic as a processor sharing queue. More precisely, consider a number of independent TCP sessions that are running for an extended period of time. Then, by fairness of TCP (e.g., see [25, 19]), it follows that on a long run each session receives an equal share of bandwidth, which is exactly captured by processor sharing discipline. Similarly, majority of job schedulers in Web servers employ processor sharing based algorithms designed with the notion of fairness in mind.

In this paper we study the M/G/1 processor sharing (PS) queue that assumes Poisson arrivals of subexponential job sizes. It is widely accepted that, in the case of a large number of independent users, the job (session) arrival times are well modeled by a Poisson processes. On the other hand, heavy-tailed distributions are suitable for modeling job sizes; for discussion and references on traffic modeling with heavy-tailed/self-similar characteristics see [34]. The study of this paper is motivated by recent findings that server access patterns and file sizes may have moderately heavy tails, e.g. lognormal [39, 23, 24].

The literature on M/G/1 PS queue is extensive; a comprehensive survey with more than 200 references on mathematical problems of shared-processor systems can be found in [43]. Early investigations of processor sharing systems used Laplace transform technique, e.g. see [12, 42, 37, 32, 26]. In the case of the M/M/1 PS system the conditional Laplace transform of the waiting time was derived in [12]; further analysis of this system was carried out in [26]. Representative studies of the M/G/1 PS queue can be found in [42, 32, 37]. Recently, prediction methods for processor-sharing queues were developed in [40]. Waiting times in GI/G/1 PS queue were shown to be associated in [7]. The heavy-traffic and fluid approximations were studied in [38, 15, 44] and [10], respectively.

Empirical evidence of the presence of heavy tails in network traffic have stimulated the analysis of subexponential queueing systems [34]. The importance of scheduling in the presence of heavy tails was first recognized in [3]. Asymptotic behavior of the waiting time in the M/G/1 PS queue with polynomial-like tails has been derived in [45] and later generalized in [30]; these results will be specifically discussed in the following section.

In contrast to most of the preceding analyses, of both exponential (e.g., [12, 26]) and heavy-tailed [45] systems, that were based on the Laplace transform technique, in this paper we develop a novel sample path large deviation approach. Using this approach, we first provide a direct sample path proof of the result from [45, 30]. Then, we extend this result, in Theorem 3.1, to a large class of subexponential distributions with tails lighter than polynomial and heavier than $e^{-\sqrt{x}}$. Furthermore, in Proposition 3.1 we demonstrate that the result does not hold for service distributions with tails lighter than $e^{-\sqrt{x}}$. The uniform large deviation bounds for sums of subexponential random variables stated in Theorem 3.2 represent our main technical results that are of independent interest.

Our main result is related to the recent study of sampling at subexponential times that was investigated in [5, 13]; this relationship will be further discussed in Remark 3.4. It is interesting that, both in [5, 13] and this paper, the results require that the distributions have heavier tails than $e^{-\sqrt{x}}$. It is worth mentioning that the criticality of $e^{-\sqrt{x}}$ appeared in the early work of A.V. Nagaev [27].

The main result of the paper shows that the waiting time is asymptotically the same as if the customer were served in isolation at an equivalent rate $1-\rho$. A result of this type is usually referred to as reduce load equivalence, e.g. see Theorem 4.4 in [17], [2, 9, 22]. In view of our main result and [5, 13] it is tempting to infer that subexponential distributions with tails heavier than $e^{-\sqrt{x}}$ represent the right framework for which these results may hold.

The paper is organized as follows. In the next section we formally describe the model and discuss existing results on processor sharing with heavy tails. The main results are stated in Section 3. Concluding remarks can be found in Section 4. The proofs of the technical findings are postponed to Section 5.

2 M/G/1 processor sharing queue

In this subsection we present the basic theory of the M/G/1 processor sharing queue. Customers arrive to the queue of unit capacity according to a Poisson process with rate λ. Service requirements of customers are independent and identically distributed (i.i.d.) random variables (r.v.s) equal in distribution to B. Upon its arrival a customer joins the queue and starts receiving service immediately. The customers are served according to the processor sharing scheduling discipline. Namely, if there are n customers present in the queue, then each of the customers receives service at rate $1/n$. Once a customer receives service equal to its service requirement it departs from the system. The queue is assumed to be stable, i.e., the load $\rho \triangleq \lambda \mathbb{E}B$ of the system satisfies $\rho < 1$.

In analyzing renewal processes, excess random variables and distribution functions play an important role. For a nonnegative random variable X with distribution F and finite mean $\mathbb{E}X$, the excess distribution F_e is defined by

$$F_e(x) = \frac{1}{\mathbb{E}X} \int_0^x (1 - F(u))\, du, \qquad x \geqslant 0.$$

A random variable $X^{(e)}$ with distribution F_e is called the excess variable of X.

The distribution of the number of customers L in the queue in the stationary regime is known to be geometric [18, 36] and depends on B only through $\mathbb{E}B$ (insensitivity property), i.e.,

$$\mathbb{P}[L = n] = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

Furthermore, the stationary remaining service requirements of the customers present in the queue are i.i.d. random variables equal to $B^{(e)}$ in distribution [41, p. 387].

2.1 Existing results on processor sharing with heavy tails

In the case of regularly varying distributions Zwart and Boxma [45] established the asymptotic relationship between the tails of the waiting time in the M/G/1 PS queue and the customer service requirement. The main result from [45] is derived by means of Tauberian theorem that requires regularly varying service distribution with non-integer exponent. By using sample path arguments, Núñez-Queija [30] (see also [31]) generalized it to distributions with intermediately regularly varying tails. The result in [30] does not cover the technical case of $\mathbb{E}B^2 = \infty$ and $\mathbb{E}B^\zeta < \infty$ for all $\zeta \in (0, 1)$. In Section 5 we provide a proof, using a completely different approach, that does not require this minor condition.

Throughout the paper, for any two real functions $f(x)$ and $g(x)$, we use the standard notation $f(x) \sim g(x)$ as $x \to \infty$ to denote $\lim_{x\to\infty} f(x)/g(x) = 1$ or equivalently $f(x) = g(x)(1 + o(1))$ as $x \to \infty$. The class of intermediately regularly varying distributions $\mathcal{IR}$ is defined in the appendix.

Theorem 2.1 (Zwart & Boxma [45], Núñez-Queija [30]) *If* $B \in \mathcal{IR}$ *and* $\mathbb{E}B^\alpha < \infty$ *for some* $\alpha > 1$, *then* $\mathbb{P}[V > x] \sim \mathbb{P}[B > (1-\rho)x]$ *as* $x \to \infty$.

Proof: Presented in Section 5.6. □

In the remaining part of the paper we extend the preceding theorem to a class of subexponential distributions with lighter than polynomial tails, e.g. lognormal and Weibull.

3 Main results

This section contains the main results of this paper stated in Theorem 3.1 and Proposition 3.1. The proofs are based on an identity that will be described in the following paragraph.

Let B_i and V_i be the job size and waiting time of the customer arriving at time T_i, respectively. The sequence of arrival times $\{T_i\}_{i=1}^{\infty}$ is assumed to be Poisson. Without loss of generality, in view of PASTA property [41], we set $T_0 = 0$. Waiting time of a customer is defined as an amount of time between its arrival and departure, also referred to as sojourn time in the queueing literature. For the customer arriving at time T_0 define function $R_0(t) \equiv R_{B_0}(t)$ for $t \geqslant 0$ as the amount of work that remains to be completed at time t. The waiting time satisfies the following min-plus equality which stems from the features of processor sharing

$$V_0 = B_0 + \sum_{i=1}^{L} B_i^{(e)} \wedge B_0 + \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i), \tag{3.1}$$

where L is the number of customers in the system just before time $t = 0$, $N(t)$ denotes the number of Poisson arrivals in $(0, t)$ and $x \wedge y \equiv \min(x, y)$; the number of customers in the system L and their residual work $B_i^{(e)}$ are independent, e.g. see [41, p. 387]. The identity follows from the fact that in the PS queue any two customers present in the system for some interval of time receive equal amounts of service during that interval, irrespective of other departures and arrivals. In particular, a customer i, $1 \leqslant i \leqslant L$, present in the system just before $t = 0$ receives $B_0 \wedge B_i^{(e)}$ amount of service during $(0, V_0)$. Similarly, any customer arriving at $T_i \in (0, V_0)$ obtains $B_i \wedge R_0(T_i)$ service in interval $(0, V_0)$. Clearly, 0th customer receives its full requirement B_0 by time V_0. Therefore, by summing up the services that each customer present in the queue during $(0, V_0)$ receives, one derives (3.1). A related expression to (3.1) can be found in [42, eq. (3.4)] (see also Theorem 5.3.2 in [30]).

In this paper we focus on a class of distributions that belongs to a large set of subexponential distributions $\mathcal{S}^\star$ as defined in Appendix. This class of distribution functions (d.f.) was first introduced by A.V. Nagaev in [28].

Definition 3.1 *A nonnegative random variable* X *(or its d.f.) belongs to class* $\mathcal{SC}$ *(subexponential concave) if its hazard function* $Q(x) \triangleq -\log \mathbb{P}[X > x]$ *is* eventually concave, *such that, as* $x \to \infty$

$$Q(x)/\log x \to \infty \tag{3.2}$$

and for $x \geqslant x_0$, $\beta x \leqslant u \leqslant x$,

$$\frac{Q(x) - Q(u)}{Q(x)} \leqslant \alpha \frac{x-u}{x}, \tag{3.3}$$

where $0 < \alpha < 1$, $0 < \beta < 1$.

Examples of random variables in $\mathcal{SC}$ include random variables with hazard functions of type (i) $c(\log x)^\gamma$, $\gamma > 1$ and (ii) $c(\log x)^\gamma x^\alpha$, $\gamma > 0$, $0 < \alpha < 1$, i.e. widely used lognormal and Weibull distributions belong to $\mathcal{SC}$. Condition (3.2) implies that if $X \in \mathcal{SC}$ then X has all moments; this, in view of Theorem 2.1 that covers the hazard functions of type $c \log x$, is not restrictive.

The next theorem is the main result of the paper.

Theorem 3.1 *Let $B \in \mathcal{SC}$ with $\alpha < 1/2$ and*

$$\overline{\lim}_{x\to\infty} \frac{\mathbb{P}[B^{(e)} > x]}{x\mathbb{P}[B > x]} < \infty. \tag{3.4}$$

Then, as $x \to \infty$

$$\mathbb{P}[V > x] \sim \mathbb{P}[B > (1-\rho)x].$$

Remark 3.1 *(i)* The condition (3.4) is not very limiting since it is implied if $x^{1+\delta}\mathbb{P}[B > x]$ is eventually monotonically decreasing in x for some $\delta > 0$.
(ii) In the case when $Q(x)$ is absolutely continuous with hazard rate $q(x) \triangleq dQ(x)/dx$, the eventual concavity of $Q(x)$ is implied by $q(x)$ being eventually decreasing and the condition (3.3) is equivalent to

$$\frac{xq(x)}{Q(x)} \leqslant \alpha, \qquad \forall x \geqslant x_0.$$

These type of assumptions were used in [8].

The condition $\alpha < 1/2$ in Theorem 3.1 is crucial. As the proposition below shows, the result does not extend to the whole class of subexponential distributions. An intuitive explanation of this criticality arises from fluctuations induced by the CLT. Informally, the last sum in (3.1) is approximately equal to $\rho V_0 + O(\sqrt{V_0})$ for large V_0. Therefore, for the result to hold, the distribution of V_0, or equivalently B_0, has to be immune to these fluctuations, which translates to $\alpha < 1/2$.

Proposition 3.1 *If $\mathbb{P}[B > x] = e^{-x^\alpha}$, $\alpha > 1/2$, then $\mathbb{P}[B > x] = o(\mathbb{P}[V(1-\rho) > x])$ as $x \to \infty$.*

Remark 3.2 The proposition implies the result earlier obtained in [5], that the busy period P in the M/G/1 queue satisfies $\mathbb{P}[B > x] = o(\mathbb{P}[P(1-\rho) > x])$ as $x \to \infty$, when $\mathbb{P}[B > x] = e^{-x^\alpha}$, $\alpha > 1/2$.

The next lemma summarizes the basic properties of r.v.s in $\mathcal{SC}$. The proof is given in Section 5.

Lemma 3.1 *Let $X \in \mathcal{SC}$ and Q be its hazard function , then*

(i) $Q(x) \leqslant Q(u)\,(x/u)^\alpha$ *for all* $x_0 \leqslant u \leqslant x$.

(ii) *For any* $0 < \delta < 1-\alpha$, $\mathbb{P}[X > x - x^\delta] \sim \mathbb{P}[X > x]$ *as* $x \to \infty$.

(iii) $X \in \mathcal{S}^\star$.

(iv) *For any* $0 < \xi < 1$ *there is* $\delta > 0$ *such that for some* $\epsilon > 0$ *and sufficiently large* x

$$Q((\xi - \delta)x) + Q((1-\xi)x) \geqslant (1+\epsilon)Q(x).$$

In this paper C denotes a sufficiently large positive constant, while c denotes a sufficiently small positive constant. The values of C and c are generally different in different places. For example, $C/2 = C$, $C^2 = C$, $C + 1 = C$, etc.

The upper bound for Theorem 3.1 is derived by means of two bounds stated below. These large deviation bounds are our main technical results that are of independent interest. Related large deviation bounds can be found in [29].

Theorem 3.2 *Let $1 - e^{-Q(x)} \in \mathcal{SC}$ and $\mathbb{P}[X > x] \leqslant Cxe^{-Q(x)}$. Then*

(i) For all x and u

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i - \mathbb{E}X\mathbb{E}N(u) > x\right] \leqslant C\left(e^{-c\frac{x^2}{u}} + ue^{-\frac{1}{2}Q(x)}\right).$$

(ii) For any positive integer k there exists $1 > \gamma > 0$ such that for all $1 \leqslant n \leqslant Cx$

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \wedge \gamma x - n\mathbb{E}X > x\right] \leqslant Ce^{-kQ(x)}.$$

Remark 3.3 *(i)* A minor modification of the proof shows that the first part of the theorem holds when Poisson $N(u)$ is replaced by u.
(ii) When $\mathbb{P}[X > x] \leqslant Ce^{-Q(x)}$, the conditions on the hazard function $Q(x)$ can be relaxed. In particular, condition (3.2) can be replaced with

$$\underline{\lim}_{x\to\infty}\frac{Q(x)}{\log x} > 2,$$

which implies the existence of $(2+\epsilon)$ moment for some $\epsilon > 0$.
(iii) In view of this theorem, the condition $B \in \mathcal{SC}$ in Theorem 3.1 can be relaxed by B being asymptotically equivalent to a distribution in $\mathcal{SC}$.

Proof: See Section 5. □

For any sequence of i.i.d. r.v.s $\{X_i\}$ we use $W^{\phi}_{X\wedge Y}$ to denote the stationary workload in a queue with Poisson arrivals of rate λ, capacity ϕ and job sizes equal to $\{X_i \wedge Y\}$; let $W^{\phi}_X \equiv W^{\phi}_{X\wedge\infty}$. The following lemma estimates the stationary workload in a queue with truncated service requirements.

Lemma 3.2 *Assume $\phi > \rho$.*

(i) If $\mathbb{E}B^{1+\delta} < \infty$ for some $\delta > 0$, then for any α there exists $\epsilon > 0$ such that, as $x \to \infty$

$$\mathbb{P}\left[W^{\phi}_{B\wedge\epsilon x} > x\right] = o(x^{-\alpha}).$$

(ii) If $B \in \mathcal{SC}$ and

$$\overline{\lim}_{x\to\infty}\frac{\mathbb{P}[B^{(e)} > x]}{x\mathbb{P}[B > x]} < \infty,$$

then for any integer $k \geqslant 1$ there exists $1 > \epsilon > 0$ such that, as $x \to \infty$

$$\mathbb{P}[W^{\phi}_{B\wedge\epsilon x} > x] = o\left((\mathbb{P}[B > x])^k\right).$$

Proof: Given in Section 5. □

The next bound was derived in [9]. For completeness the proof is presented in Section 5.

Lemma 3.3 *If $B \in \mathcal{S}^\star$ then for any positive $\delta<1-\rho$ and all x the excess distribution of the residual busy period is bounded by*

$$\mathbb{P}[P^{(e)} > x] \leqslant C\mathbb{P}[B^{(e)} > (1-\rho-\delta)x].$$

The last preparatory result states that asymptotically the long waiting time of a customer cannot be caused by the customers present in the queue upon its arrival. Classes of heavy-tailed distributions $\mathcal{D}$, $\mathcal{L}$ and $\mathcal{S}^\star$ are defined in the appendix.

Proposition 3.2 *Let either (i) $X \in \mathcal{D} \cap \mathcal{L}$ and $\mathbb{E}X < \infty$ or (ii) $X \in \mathcal{S}^\star$ and*

$$\overline{\lim}_{x\to\infty} \frac{\mathbb{P}[X^{(e)} > x]}{x\mathbb{P}[X > x]} < \infty.$$

If $\{X_i^{(e)}\}_{i=1}^\infty$ are i.i.d. random variables equal in distribution to $X^{(e)}$ and independent of N with $\mathbb{E}\left[(1+\epsilon)^N\right]<\infty$ for some $\epsilon > 0$, then as $x\to\infty$

$$\mathbb{P}\left[X + \sum_{i=1}^{N} X_i^{(e)} \wedge X > x\right] \sim \mathbb{P}\left[X > x\right].$$

Proof: See Section 5. □

Proof of Theorem 3.1: Expression (3.1) for the waiting time of the 0th customer renders

$$V_0 - \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i) = B_0 + \sum_{i=1}^{L} B_0 \wedge B_i^{(e)} \triangleq \hat{B}_0, \tag{3.5}$$

where $\hat{B}_0$ is introduced for notational convenience. Next, for $\xi > \max\{1/2, \beta\}$ and $\delta > 0$, write $\mathbb{P}[V_0(1-\rho) > x] = f_1(x) + f_2(x) + f_3(x)$ where

$$\begin{aligned} f_1(x) &= \mathbb{P}[V_0(1-\rho) > x,\ \hat{B}_0 \leqslant \xi x], \\ f_2(x) &= \mathbb{P}\left[V_0(1-\rho) > x,\ \hat{B}_0 \in (\xi x, x - x^{1/2+\delta}]\right], \\ f_3(x) &= \mathbb{P}[V_0(1-\rho) > x,\ \hat{B}_0 > x - x^{1/2+\delta}]. \end{aligned}$$

In what follows, we examine the asymptotic behavior of the three terms. We start with $f_1(x)$. Observe that, since $\hat{B}_0 \geqslant B_0 \geqslant R_0(t)$, (3.5) implies

$$\begin{aligned} V_0(1-\rho-\delta) &\leqslant \hat{B}_0 + \sup_{t\geqslant 0}\left\{\sum_{i=1}^{N(t)} B_i \wedge \hat{B}_0 - (\rho+\delta)t\right\} \\ &\stackrel{d}{=} \hat{B}_0 + W^{\rho+\delta}_{B\wedge\hat{B}_0}, \end{aligned}$$

where $\stackrel{d}{=}$ denotes the equality in distribution. Therefore, for $\delta_\rho \triangleq \delta/(1-\rho)$

$$\begin{aligned} f_1(x) &\leqslant \mathbb{P}\left[\hat{B}_0 + W^{\rho+\delta}_{B\wedge\hat{B}_0} > (1-\delta_\rho)x,\ \hat{B}_0 \leqslant \xi x\right] \\ &\leqslant \mathbb{P}\left[W^{\rho+\delta}_{B\wedge\delta_\rho x} > (1-2\delta_\rho)x\right] + \int_{\delta_\rho x}^{\xi x} \mathbb{P}\left[W^{\rho+\delta}_{B} > (1-\delta_\rho)x - u\right]\, d\mathbb{P}[\hat{B}_0 \leqslant u]. \end{aligned}$$

If δ_ρ (i.e. δ) is chosen small enough, the first term in the preceding sum, by Lemma 3.2 (ii), is upper bounded by $C(\mathbb{P}[B > (1-2\delta_\rho)x])^2 = o(\mathbb{P}[B > x])$, where the last equality follows from Lemma 3.1 (i). The bound for the second term is as follows. By Pakes' asymptotic result for the workload of a stable M/G/1 queue [33] and assumption (3.4)

$$f_1(x) \leqslant Cx \int_{\delta_\rho x}^{\xi x} e^{-Q((1-\delta_\rho)x-u)}\, d\mathbb{P}[\hat{B}_0 \leqslant u] + o(\mathbb{P}[B > x]).$$

Next, by discretizing the last integral for some $\Delta > 0$ one obtains

$$\begin{aligned} f_1(x) &\leqslant Cx \sum_{j=0}^{\left\lceil \frac{\xi-\delta_\rho}{\Delta} \right\rceil} \int_{(\delta_\rho+j\Delta)x}^{(\delta_\rho+(j+1)\Delta)x} e^{-Q((1-\delta_\rho)x-u)}\, d\mathbb{P}[\hat{B}_0 \leqslant u] + o(\mathbb{P}[B > x]) \\ &\leqslant Cx \sum_{j=0}^{\left\lceil \frac{\xi-\delta_\rho}{\Delta} \right\rceil} e^{-Q((1-2\delta_\rho-(j+1)\Delta)x)}\, \mathbb{P}[\hat{B}_0 > (\delta_\rho + j\Delta)x] + o(\mathbb{P}[B > x]) \\ &\leqslant Cx \sum_{j=0}^{\left\lceil \frac{\xi-\delta_\rho}{\Delta} \right\rceil} e^{-Q((1-2\delta_\rho-(j+1)\Delta)x)-Q((\delta_\rho+j\Delta)x)} + o(\mathbb{P}[B > x]), \end{aligned}$$

where the last inequality is due to Proposition 3.2. Next, Lemma 3.1 (iv) shows that each term in the last sum is $o(\mathbb{P}[B > x])$ for sufficiently small δ_ρ (i.e. δ) and Δ and, thus,

$$f_1(x) = o(\mathbb{P}[B > x]).$$

Bounding $f_2(x)$ requires the most work. Introduce a continuous function $H : \mathbb{R}_+ \to \mathbb{R}_+$ defined as

$$H(t) \triangleq \sup_{0 \leqslant u \leqslant t} \left\{ u - \sum_{i=1}^{N(u)} B_i \right\}, \quad t > 0. \tag{3.6}$$

The function H is nondecreasing and, hence, it is possible to define a right-continuous inverse $H^{\leftarrow}(x) = \inf\{t > 0 : H(t) > x\}$. From Figure 1, due to the memoryless property of exponential distribution, it is clear that $H(t)$ increases linearly at rate 1 over exponential intervals of parameter λ and then stays constant for the amounts of time that are equal in distribution to the busy period P of the original M/G/1 queue. Thus, $H^{\leftarrow}$ can be written in the following form

$$H^{\leftarrow}(t) = t + \sum_{i=1}^{N(t)} P_i, \tag{3.7}$$

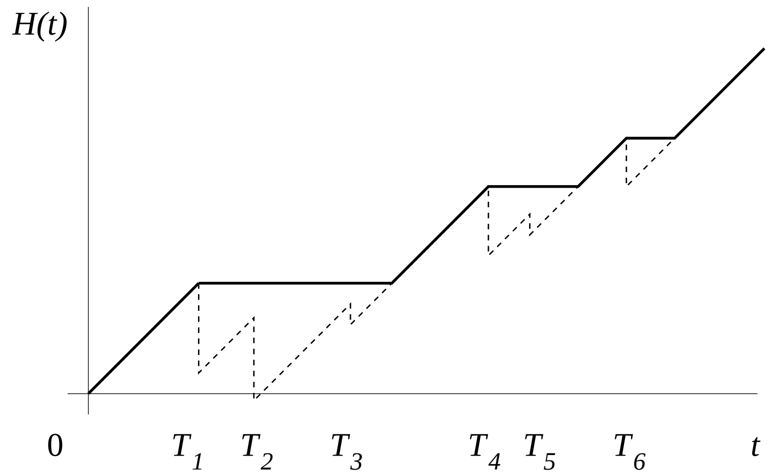

Figure 1: A typical sample path realization of function H; T_n-s are the Poisson arrival points.

where r.v.s $\{P_i\}_{i=1}^{\infty}$ are i.i.d. copies of P. Note that $H^{\leftarrow}(t)$ is the busy period of an M/G/1 queue with initial workload equal to t.

From (3.5), V_0 can be interpreted as the first time t that $t - \sum_{i=1}^{N(t)} B_0 \wedge R_0(t) = \hat{B}_0$, implying

$$\begin{aligned} V_0 &\leqslant \inf\left\{ t > 0 : t - \sum_{i=1}^{N(t)} B_i > \hat{B}_0 \right\} \\ &= \inf\{t > 0 : H(t) > \hat{B}_0\} = H^{\leftarrow}(\hat{B}_0). \end{aligned} \tag{3.8}$$

By using (3.7) and (3.8), $f_2(x)$ can be upper bounded by

$$\begin{aligned} f_2(x) &\leqslant \int_{\xi x}^{x-x^{1/2+\delta}} \mathbb{P}\left[\sum_{i=1}^{N(u)} P_i - \lambda u \mathbb{E}P > \frac{x-u}{1-\rho} \right] d\mathbb{P}[\hat{B}_0 \leqslant u] \\ &\leqslant C \int_{\xi x}^{x-x^{1/2+\delta}} \left(e^{-c\frac{(x-u)^2}{u}} + u e^{-\frac{1}{2}Q((1-\delta)(x-u))} \right) d\mathbb{P}[\hat{B}_0 \leqslant u] \\ &\triangleq f_{21}(x) + f_{22}(x), \end{aligned}$$

where the second inequality follows from Lemma 3.3, condition (3.4) and Theorem 3.2. Integration by parts and Proposition 3.2 give a bound for $f_{21}(x)$

$$\begin{aligned} f_{21}(x) &\leqslant C e^{-cx - Q(\xi x)} + C \int_{\xi x}^{x-x^{1/2+\delta}} \frac{x^2-u^2}{u^2} e^{-Q(u)} e^{-c\frac{(x-u)^2}{u}} \, du \\ &\leqslant C e^{-cx} + C e^{-Q(x)} \int_{\xi x}^{x-x^{1/2+\delta}} e^{Q(x)-Q(u)-c\frac{(x-u)^2}{x}} \, du, \end{aligned}$$

where in the second inequality we used that $(x^2-u^2)/u^2 = O(1)$ for all u in the interval of integration. To show that $f_{21}(x) = o(\mathbb{P}[B > x])$, it is enough to verify that the exponent in the last integral is upper bounded by $-cx^{2\delta}$ for the given interval of u. Thus, by assumption (3.3) and Lemma 3.1 (i), for all large x

$$\begin{aligned}
Q(x) - Q(u) - c\frac{(x-u)^2}{x} &\leqslant \alpha Q(x)\frac{x-u}{x} - c\frac{(x-u)^2}{x} \\
&\leqslant Cx^{\alpha}\frac{x-u}{x} - c\frac{(x-u)^2}{x} \\
&\leqslant Cx^{-(1/2-\alpha)+\delta} - cx^{2\delta};
\end{aligned} \tag{3.9}$$

since for all x large enough the right-hand side of the second inequality is increasing in u and $u \leqslant x - x^{1/2+\delta}$. Now, by choosing $\delta < 1/2 - \alpha$ it follows that (3.9) is upper bounded by $-cx^{2\delta}$. As far as $f_{22}(x)$ is concerned, by discretization of the integral below, we have

$$\begin{aligned}
f_{22}(x) &\leqslant Cx \int_{\xi x}^{x - x^{1/2+\delta}} e^{-\frac{1}{2}Q((1-\delta)(x-u))}\, d\mathbb{P}[\hat{B}_0 \leqslant u] \\
&\leqslant Cx \sum_{j=1}^{\lceil (1-\xi)x^{1/2-\delta}\rceil} e^{-\frac{1}{2}Q((1-\delta)jx^{1/2+\delta})} e^{-Q(x-(j+1)x^{1/2+\delta})} \\
&\leqslant C \max\left\{ x^{3/2} e^{-\frac{1}{2}Q((1-\delta)x^{1/2+\delta}) - Q(x-2x^{1/2+\delta})},\ x^{3/2} e^{-\frac{1}{2}Q((1-\xi)(1-\delta)x) - Q(\xi x - 2x^{1/2+\delta})} \right\},
\end{aligned} \tag{3.10}$$

where in the last inequality we used the concavity property of Q, i.e. the maximum of all the summands is equal to either the first or the last summand. Thus, Lemma 3.1 (i) and (ii) imply that the first term in the preceding maximum is $o(\mathbb{P}[B > x])$; the exponent of the second term is by Lemma 3.1 (i) bounded by

$$\frac{1}{2}Q((1-\delta)(1-\xi)x) + Q((\xi-\delta)x) - Q(x) \geqslant Q(x)\left(\frac{(1-\delta)^{\alpha}}{2}(1-\xi)^{\alpha} - \alpha(1-\xi+\delta)\right).$$

Therefore, for all δ sufficiently small, we obtain

$$f_2(x) = o(\mathbb{P}[B > x]).$$

The bound for $f_3(x)$ is straightforward by Lemma 3.1 and Proposition 3.2

$$\overline{\lim}_{x\to\infty} \frac{f_3(x)}{\mathbb{P}[B > x]} \leqslant \overline{\lim}_{x\to\infty} \frac{\mathbb{P}[B > x - x^{1/2+\delta}]}{\mathbb{P}[B > x]} = 1.$$

Combination of bounds for f_1, f_2 and f_3 yields the upper bound. The lower bound is a corollary of Lemma 3.4. □

Denote by $V^{(x)}$ the waiting time of a customer conditional on the fact that its service requirement is equal to x. In the same fashion, let $R^{(x)}(t)$ be the conditional amount of service to be completed at time t.

Remark 3.4 Waiting time V_0 can be represented as sampling at subexponential time B_0 of the monotonically increasing process

$$V^{(x)} = x + \sum_{i=1}^{L} B_i^{(e)} \wedge x + \sum_{i=1}^{N(V^{(x)})} B_i \wedge R^{(x)}(T_i),$$

for which one could potentially use results obtained in [5, 13]. However, the major difficulty in carrying out this approach is that $V^{(x)}$ is implicitly defined, i.e. understanding $V^{(x)}$ requires the knowledge of $V^{(x)}$ and $R^{(x)}$. This is similar to the situation that arises in the analysis of the busy period, as pointed out in [5].

The following lemma provides a general lower bound on the waiting time. The proof is based on the Central Limit Theorem, and, therefore, the second moment of the service requirement is assumed.

Lemma 3.4 *If $\mathbb{E}B^2 < \infty$ and $\mathbb{P}[B > x] \sim \mathbb{P}[B > x + x^{1/2+\delta}]$ as $x \to \infty$ for some $\delta > 0$, then*

$$\underline{\lim}_{x\to\infty} \frac{\mathbb{P}[V > x]}{\mathbb{P}[B > (1-\rho)x]} \geqslant 1.$$

Proof: Consider the stationary M/G/1 PS queue at $t < 0$, as described in Section 2. Next, assume that at time $t = 0$ a special customer with infinite service requirement arrives. Let $Z(t)$, $t \geqslant 0$, be the total unfinished work in the new PS systems of all but the infinite customer. It is known that $Z(t)$ converges in distribution to an a.s. finite random variable Z [41, p. 339]. In what follows we exploit the fact that at time $t = V_0$ the remaining unfinished work in the original PS queue is equal to $Z(V_0)$. Thus, (3.1) and (3.6) render

$$\begin{aligned} H(V_0) \geqslant V_0 - \sum_{i=1}^{N(V_0)} B_i &\geqslant B_0 - \left(\sum_{i=1}^{N(V_0)} B_i - \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i) \right) \\ &\geqslant B_0 - \left(\sum_{i=1}^{N(V_0)} B_i - \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i) + \sum_{i=1}^{L} (B_i^{(e)} - B_i^{(e)} \wedge B_0) \right) \\ &= B_0 - Z(V_0). \end{aligned} \tag{3.11}$$

Next, from (3.11) one obtains

$$\begin{aligned} \mathbb{P}[(1-\rho)V_0 > x] &\geqslant \mathbb{P}\left[(1-\rho)H^{\leftarrow}(B_0 - Z(V_0)) > x,\ B_0 > x + k\sqrt{x},\ Z(V_0) \leqslant \sqrt{x}\right] \\ &\geqslant \mathbb{P}\left[(1-\rho)H^{\leftarrow}(x + (k-1)\sqrt{x}) > x,\ B_0 > x + k\sqrt{x},\ Z(V_0) \leqslant \sqrt{x}\right], \end{aligned}$$

where $k > 1$, and the last inequality is due to the monotonicity of $H^{\leftarrow}$. Then, by independence of $H^{\leftarrow}(x)$ and $Z(x)$ from B_0,

$$\begin{aligned} \mathbb{P}[(1-\rho)V_0 > x] &\geqslant \int_{x+k\sqrt{x}}^{\infty} \mathbb{P}\left[(1-\rho)H^{\leftarrow}(x + (k-1)\sqrt{x}) > x,\ Z(V^{(y)}) \leqslant \sqrt{x}\right] d\mathbb{P}[B \leqslant y] \\ &\geqslant \left(\mathbb{P}[(1-\rho)H^{\leftarrow}(x + (k-1)\sqrt{x}) > x] - \sup_{y \geqslant x + k\sqrt{x}} \mathbb{P}[Z(V^{(y)}) > \sqrt{x}] \right) \mathbb{P}[B > x + k\sqrt{x}]; \end{aligned}$$

the second inequality follows from the union bound. Observe that, since $Z(t)$ converges in distribution to a.s. finite Z, the supremum in the preceding inequality tends to 0 as $x \to \infty$. Therefore,

$$\underline{\lim}_{x\to\infty} \frac{\mathbb{P}[(1-\rho)V_0 > x]}{\mathbb{P}[B > x]} \geqslant \underline{\lim}_{x\to\infty} \mathbb{P}[(1-\rho)H^{\leftarrow}(x + (k-1)\sqrt{x}) > x].$$

Next, it is known that $\mathbb{E}P^2 < \infty$ if and only if $\mathbb{E}B^2 < \infty$ [1]. Thus, by (3.7), the process $H^{\leftarrow}(t)$ as a function of t satisfies the Central Limit Theorem, yielding

$$\underline{\lim}_{x\to\infty}\frac{\mathbb{P}[(1-\rho)V_0 > x]}{\mathbb{P}[B > x]} \geqslant \lim_{k\to\infty}\underline{\lim}_{x\to\infty}\mathbb{P}[(1-\rho)H^{\leftarrow}(x+(k-1)\sqrt{x}) > x] = 1,$$

which concludes the proof. □

Proof of Proposition 3.1: The proof is a minor modification of the proof of Lemma 3.4. Equation (3.11) leads to

$$\begin{aligned}\mathbb{P}[(1-\rho)V_0 > x] &\geqslant \mathbb{P}\left[(1-\rho)H^{\leftarrow}(B_0 - Z(V_0)) > x,\ B_0 > x-\sqrt{x},\ Z(V_0)\leqslant\sqrt{x}\right]\\ &\geqslant \mathbb{P}\left[(1-\rho)H^{\leftarrow}(x-2\sqrt{x}) > x,\ B_0 > x-\sqrt{x},\ Z(V_0)\leqslant\sqrt{x}\right]\\ &\geqslant \left(\mathbb{P}[(1-\rho)H^{\leftarrow}(x-2\sqrt{x}) > x] - \sup_{y\geqslant x-\sqrt{x}}\mathbb{P}[Z(V^{(y)}) > \sqrt{x}]\right)\mathbb{P}[B > x-\sqrt{x}].\end{aligned}$$

Thus,

$$\begin{aligned}\underline{\lim}_{x\to\infty}\frac{\mathbb{P}[(1-\rho)V > x]}{\mathbb{P}[B > x]} &\geqslant \underline{\lim}_{x\to\infty}\mathbb{P}[(1-\rho)H^{\leftarrow}(x-2\sqrt{x}) > x]\,\underline{\lim}_{x\to\infty}\frac{\mathbb{P}[B > x-\sqrt{x}]}{\mathbb{P}[B > x]}\\ &\geqslant c\lim_{x\to\infty}e^{x^\alpha-(x-\sqrt{x})^\alpha} = \infty,\end{aligned}$$

since $\alpha > 1/2$; the existence of $c > 0$ follows from the Central Limit Theorem. □

4 Concluding remarks

In this paper we study a processor sharing queue that represents a baseline model of efficient and fair network resource sharing algorithms, e.g. TCP flow control protocol and Web server job scheduling algorithms. Our main result extends the asymptotic reduced load equivalence relationship between the job sizes and their waiting times, derived in [45, 30] for polynomial tails, to a large class of subexponential distributions with tails heavier than $e^{-\sqrt{x}}$. In particular, this extension covers the practically important case of jobs with lognormal distributions that were recently empirically measured in [24, 23].

¿From a mathematical prospective, in contrast to the earlier work on this problem that utilized the Laplace transform technique, we derive a sample path method that could potentially be useful in analyzing more general models for which the Laplace transform solutions may not be available. Furthermore, we show that the derived relationship does not hold if the job sizes have lighter tails than $e^{-\sqrt{x}}$. The criticality of $e^{-\sqrt{x}}$ has appeared earlier in [27] and [5, 13]. In view of our analysis and [5, 13] one is tempted to conjecture that class of subexponential distributions with tails heavier than $e^{-\sqrt{x}}$ represents a natural framework for fully extending other reduced load equivalence results, e.g. Theorem 4.4 in [17] and the results of [2].

5 Proofs

This section contains the proofs of technical results: Lemmas 3.1, 3.2, 3.3, Proposition 3.2 and Theorems 3.2, 2.1. The large deviation bounds of Theorem 3.2 are our main technical results that are of independent interest.

5.1 Proof of Lemma 3.1

(i) For any $x_0 \leqslant u \leqslant x$, we can choose $\beta < \gamma < 1$ and n such that $\gamma^{n+1}x \leqslant u \leqslant \gamma^n x$. Then (3.3) implies

$$Q(x)(1-\alpha+\alpha\gamma)^{n+1} \leqslant Q(\gamma^{n+1}x) \leqslant Q(u)$$

and, therefore,

$$Q(x)(1-\alpha+\alpha\gamma)^{1+\frac{\log x/u}{\log 1/\gamma}} \leqslant Q(u).$$

The last inequality can be restated in the following equivalent form

$$Q(x) \leqslant Q(u)\left(1-\alpha(1-\gamma)\right)^{-1}\left(\frac{x}{u}\right)^{\frac{\log(1-\alpha(1-\gamma))^{-1}}{\log\gamma^{-1}}}.$$

The statement (i) follows from the preceding inequality and the following limit

$$\lim_{\gamma\nearrow 1}\frac{\log(1-\alpha(1-\gamma))^{-1}}{\log\gamma^{-1}} = \alpha.$$

(ii) Set $u = x - x^\delta$ in (3.3) to obtain

$$Q(x-x^\delta) \geqslant Q(x)(1-\alpha x^{\delta-1}).$$

Part (i) implies $Q(x) \leqslant Cx^\alpha$ for $x \geqslant x_0$. Then for $x \geqslant x_0$

$$e^{-Q(x-x^\delta)+Q(x)} \leqslant e^{\alpha x^{\delta-1}Q(x)} \leqslant e^{C\alpha x^{\alpha+\delta-1}}$$

and the claim (ii) follows.
(iii) For any $\delta > 0$ and sufficiently large x the symmetry and concavity of $Q(x)$ yields

$$\begin{aligned}\int_0^x \mathbb{P}[X>u]\mathbb{P}[X>x-u]du &\leqslant 2\int_0^{x^\delta}\mathbb{P}[X>u]\mathbb{P}[X>x-u]\,du + \int_{x^\delta}^{x-x^\delta} e^{-Q(u)}e^{-Q(x-u)}\,du\\ &\leqslant 2\mathbb{E}X\ e^{-Q(x-x^\delta)} + xe^{-Q(x-x^\delta)}e^{-Q(x^\delta)}.\end{aligned}$$

Next, set $\delta < 1-\alpha$ and use (i) and (3.2) to obtain the upper bound. The lower bound follows from

$$\begin{aligned}\int_0^x \mathbb{P}[X>u]\mathbb{P}[X>x-u]\,du &\geqslant 2\int_0^{x^\delta}\mathbb{P}[X>u]\mathbb{P}[X>x-u]\,du\\ &\geqslant 2\mathbb{P}[X>x]\int_0^{x^\delta}\mathbb{P}[X>u]\,du.\end{aligned}$$

(iv) Directly from (i)

$$\begin{aligned}Q((\xi-\delta)x) &\geqslant Q(x)(\xi-\delta)^\alpha,\\ Q((1-\xi)x) &\geqslant Q(x)(1-\xi)^\alpha.\end{aligned}$$

Then, summing the last two inequalities results in

$$Q((\xi-\delta)x) + Q((1-\xi)x) \geqslant Q(x)\big((\xi-\delta)^\alpha + (1-\xi)^\alpha\big)$$

and the statement follows. $\square$

5.2 Proof of Lemma 3.2

The proof is based on the analysis of sums of truncated random variables. Denote by K the number of positive ladder heights in the Pollaczek-Khintchine representation of the stationary workload in an M/G/1 queue (see Ch. VII and IX in [4]).

(i) First observe that for all $y \leqslant \epsilon x$

$$\mathbb{P}[(B_i \wedge \epsilon x)^{(e)} > y] \leqslant \frac{\mathbb{E}B}{\mathbb{E}(B \wedge \epsilon x)} \mathbb{P}[B^{(e)} > y] \tag{5.1}$$

and introduce a new absolutely continuous random variable S defined by

$$\mathbb{P}[S > y] = \begin{cases} \Big((1+\delta)\mathbb{P}[B^{(e)} > y] \Big) \wedge 1 & y < y_0, \\ (1+\delta) y_0^{\beta} \mathbb{P}[B^{(e)} > y_0] y^{-\beta} & y \geqslant y_0, \end{cases}$$

where y_0 is finite. Then, for sufficiently large x and all y, $\mathbb{P}[S \wedge \epsilon x > y] \geqslant \mathbb{P}[(B \wedge \epsilon x)^{(e)} > y]$ and, thus,

$$\begin{aligned} \mathbb{P}\left[W^{\phi}_{B\wedge \epsilon x} > x\right] &= \mathbb{P}\left[\sum_{i=1}^{K} (B_i \wedge \epsilon x)^{(e)} > x\right] \\ &\leqslant \mathbb{P}\left[\sum_{i=1}^{\lfloor k \log \epsilon x \rfloor} S_i \wedge \epsilon x > x\right] + \mathbb{P}\left[K > k \log \epsilon x\right], \end{aligned} \tag{5.2}$$

where random variables $\{S_i\}$ are i.i.d. equal in distribution to S. Setting $s = \lfloor \epsilon^{-1} \rfloor$ and using an easy modification of the proof of Theorem 1 of [16] we derive

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{\lfloor k \log \epsilon x \rfloor} S_i \wedge \epsilon x > x\right] &\leqslant \binom{\lfloor k \log \epsilon x \rfloor}{s+1} C x^{-(s+1)\beta} \\ &\leqslant C (\log x)^{s+1} x^{-(s+1)\beta}. \end{aligned}$$

By an appropriate choice of s (and hence of ϵ) one can ensure that $\alpha < (s+1)\beta$ and, thus,

$$\mathbb{P}\left[\sum_{i=1}^{\lfloor k \log \epsilon x \rfloor} S_i \wedge \epsilon x > x\right] = o(x^{-\alpha}). \tag{5.3}$$

Furthermore, K is geometric and, therefore, a large enough k ensures $\mathbb{P}[K > k \log \epsilon x] = o(x^{-\alpha})$. Finally, this bound on K, (5.3) and (5.2) imply the proof of part (i) .

(ii) The proof is similar to the proof of part (i). The Pollaczek-Khintchine representation results in

$$\begin{aligned} \mathbb{P}[W^{\phi}_{B\wedge \epsilon x} > x] &= \mathbb{P}\left[\sum_{i=1}^{K} (B_i \wedge \epsilon x)^{(e)} > x\right] \\ &\leqslant \mathbb{P}\left[\sum_{i=1}^{\lfloor cx \rfloor} (B_i \wedge \epsilon x)^{(e)} > x\right] + \mathbb{P}[K > cx]. \end{aligned}$$

We point out that by (5.1) and the assumption of the lemma $\mathbb{P}[(B\wedge\epsilon x)^{(e)}>y]\leqslant Cye^{-Q(y)}$. Next, introduce a new random variable defined by $\mathbb{P}[S>y]=Cye^{-Q(y)}\wedge 1$ and note that for all $y\geqslant 0$

$$\mathbb{P}[(B\wedge\epsilon x)^{(e)}>y]\leqslant\mathbb{P}[S\wedge\epsilon x>y].$$

Thus, for any $1/k>\Delta>0$ we can choose $c<\Delta/\mathbb{E}S$, rendering for sufficiently small ϵ

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{\lfloor cx\rfloor}(B_i\wedge\epsilon x)^{(e)}>x\right]&\leqslant\mathbb{P}\left[\sum_{i=1}^{\lfloor cx\rfloor}S_i\wedge\epsilon x-\lfloor cx\rfloor\mathbb{E}S>(1-\Delta)x\right]\\&\leqslant Ce^{-(k+1)Q((1-\Delta)x)},\end{aligned}$$

where the last bound follows by Theorem 3.2. Assumption (3.3), for $\Delta<1-\beta$, yields $kQ(x)\leqslant(k+1)Q((1-\Delta)x)$ and, hence, part (ii) holds. □

5.3 Proof of Lemma 3.3

We start with constructing a new hybrid queue of unit capacity. It is fed by two arrival processes: (i) the arrival process of the original M/G/1 queue, and (ii) a fluid process of a constant rate $1-\rho-\delta$. The second process is served only if the workload of the first one is zero, i.e. the first process has the absolute priority. Let A be the stationary amount of work in the system that belongs to the second arrival process. Since the workload of the second process is not greater than the total workload in the system

$$\mathbb{P}[A>x]\leqslant\mathbb{P}[W_B^{\rho+\delta}>x]\sim\rho\delta^{-1}\mathbb{P}[B^{(e)}>x]\quad\text{as } x\to\infty,\tag{5.4}$$

where the asymptotics for the M/G/1 queue follows from Pakes' theorem (e.g. see [17]). Now, note that the workload of the second process evolves in the same way as the workload in a fluid queue loaded with an On-Off process. The On periods correspond to the busy periods in the original queue. Therefore, the probability that the process is in the active state is ρ and

$$\mathbb{P}[A>x]\geqslant\rho\mathbb{P}\left[P^{(e)}>\frac{x}{1-\rho-\delta}\right].$$

The last inequality and (5.4) yield the statement of the lemma. □

5.4 Proof of Proposition 3.2

The proof is an immediate consequence of the following four lemmas (5.1 - 5.4) and the dominant convergence.

Lemma 5.1 *Let $\{Y_i\}_{i=1}^n$ be independent, a.s. finite random variables and $X\in\mathcal{D}\cap\mathcal{L}$. Then, for any fixed n, as $x\to\infty$*

$$\mathbb{P}\left[X+\sum_{i=1}^{n}X\wedge Y_i>x\right]\sim\mathbb{P}[X>x].$$

Proof: Note that $\{X+\sum_{i=1}^n X\wedge Y_i > x\}$ only if $\{X > \frac{x}{n+1}\}$. Hence, for any $k>0$ the union bound yields

$$\mathbb{P}\left[X+\sum_{i=1}^n X\wedge Y_i > x\right] \leqslant \mathbb{P}[X > x-kn] + \mathbb{P}\left[X > \frac{x}{n+1}\right]\sum_{i=1}^n \mathbb{P}[Y_i > k].$$

Since X is both in $\mathcal{L}$ and $\mathcal{D}$ one easily obtains from the preceding inequality

$$\overline{\lim}_{x\to\infty}\frac{\mathbb{P}\left[X+\sum_{i=1}^n X\wedge Y_i > x\right]}{\mathbb{P}[X > x]} \leqslant 1 + C\sum_{i=1}^n \mathbb{P}[Y_i > k].$$

Setting $k\to\infty$, since $Y_i<\infty$ a.s., yields the statement of the lemma. □

Lemma 5.2 *Let $X\in\mathcal{D}\cap\mathcal{L}$, $\mathbb{E}X<\infty$ and $\{X_i^{(e)}\}_{i=1}^n$ be i.i.d. random variables equal in distribution to $X^{(e)}$. Then for any $\epsilon>0$ there exist C such that for all $x\geqslant 0$ and $n\geqslant 1$*

$$\mathbb{P}\left[X+\sum_{i=1}^n X\wedge X_i^{(e)} > x\right] \leqslant C(1+\epsilon)^n\mathbb{P}[X > x].$$

Proof: Define $S_n \triangleq \sum_{i=1}^n X_i^{(e)}$. Then

$$\begin{aligned}\mathbb{P}\left[X+\sum_{i=1}^n X\wedge X_i^{(e)} > x\right] &\leqslant \mathbb{P}\left[X > \frac{x}{n+1},\ X+S_n > x\right]\\ &\leqslant \mathbb{P}\left[X > \frac{x}{n+1}\right] + \mathbb{P}[X+S_n > x,\ S_n \leqslant x].\end{aligned} \tag{5.5}$$

Next, by definition of the class of distributions $\mathcal{D}$ we have $s\triangleq \sup\frac{\mathbb{P}[X>x]}{\mathbb{P}[X>2x]}<\infty$ and, hence,

$$\mathbb{P}\left[X > \frac{x}{n+1}\right] \leqslant s^{\lceil \log_2(n+1)\rceil}\mathbb{P}[X > x]. \tag{5.6}$$

On the other hand, Theorem A.1 and Lemma A.6 (i) result in

$$\begin{aligned}\mathbb{P}[X+S_n > x,\ S_n\leqslant x] &= \int_0^x \mathbb{P}[X > x-y]\,d\mathbb{P}[S_n\leqslant y]\\ &\leqslant C(1+\epsilon)^n\int_0^x \mathbb{P}[X > x-y]\mathbb{P}[X>y]\,dy\\ &\leqslant C(1+\epsilon)^n\mathbb{P}[X > x],\end{aligned} \tag{5.7}$$

where in the last inequality we used Definition A.5 of $\mathcal{S}^\star$. Inequality (5.5) in conjunction with (5.6) and (5.7) yields the statement of the lemma. □

Lemma 5.3 *Let $X \in \mathcal{S}^\star$ and*

$$\overline{\lim}_{x\to\infty} \frac{\mathbb{P}[X^{(e)} > x]}{x\mathbb{P}[X > x]} < \infty.$$

If $\{X_i^{(e)}\}_{i=1}^n$ are i.i.d. random variables equal in distribution to $X^{(e)}$, then, for any fixed n, as $x \to \infty$

$$\mathbb{P}\left[X + \sum_{i=1}^n X \wedge X_i^{(e)} > x\right] \sim \mathbb{P}[X > x].$$

Proof: Define $S_n \triangleq \sum_{i=1}^n X_i^{(e)}$. Then,

$$\begin{aligned}\mathbb{P}\left[X + \sum_{i=1}^n X \wedge X_i^{(e)} > x\right] &= \mathbb{P}\left[X > \frac{x}{n+1},\ X + \sum_{i=1}^n X \wedge X_i^{(e)} > x\right] \\ &\leqslant \mathbb{P}\left[X > \frac{x}{n+1}\right]\mathbb{P}[S_n > x-k] + \mathbb{P}[X + S_n > x,\ S_n \leqslant x-k] \\ &\triangleq I_1(x) + I_2(x),\end{aligned} \tag{5.8}$$

where the second line follows from the independence of X and S_n. Next we examine the asymptotic behavior of $I_1(x)$ and $I_2(x)$. First, the assumption of the lemma and Theorem A.1 yield

$$\begin{aligned}\overline{\lim}_{x\to\infty}\frac{I_1(x)}{\mathbb{P}[X>x]} &\leqslant \overline{\lim}_{x\to\infty}\frac{\mathbb{P}[X > \frac{x}{n+1}]\mathbb{P}[X^{(e)} > x]}{\mathbb{P}[X>x]}\ \overline{\lim}_{x\to\infty}\frac{\mathbb{P}[S_n > x-k]}{\mathbb{P}[X^{(e)}>x]} \\ &\leqslant n\,\overline{\lim}_{x\to\infty}\frac{\mathbb{P}[X^{(e)}>x]}{x\mathbb{P}[X>x]}\,\overline{\lim}_{x\to\infty} x\mathbb{P}\left[X > \frac{x}{n+1}\right] = 0.\end{aligned} \tag{5.9}$$

By definition of the class of distributions $\mathcal{S}^\star$ (see the appendix)

$$\int_0^x \frac{\mathbb{P}[X > x-y]\mathbb{P}[X>y]}{\mathbb{P}[X>x]}\,dy \longrightarrow 2\mathbb{E}X \quad \text{ as } x\to\infty,$$

and, therefore, the following double limit holds

$$\lim_{k\to\infty}\overline{\lim}_{x\to\infty}\int_k^{x-k}\frac{\mathbb{P}[X>x-y]\mathbb{P}[X>y]}{\mathbb{P}[X>x]}\,dy = 0. \tag{5.10}$$

The quantity $I_2(x)$ can be upper bounded as

$$I_2(x) \leqslant \mathbb{P}[X > x-k] + \int_k^{x-k}\mathbb{P}[X>x-y]\,d\mathbb{P}[S_n \leqslant y]. \tag{5.11}$$

By Lemma A.6 (ii) for any $\epsilon > 0$ there exists k_0 such that for all $x > k > k_0$

$$\int_k^{x-k}\mathbb{P}[X>x-y]\,d\mathbb{P}[S_n\leqslant y] \leqslant \frac{(1+\epsilon)n}{\mathbb{E}X}\int_k^{x-k}\mathbb{P}[X>x-y]\mathbb{P}[X>y]\,dy. \tag{5.12}$$

Finally, substituting (5.12) in (5.11), using (5.10) and recalling $X \in \mathcal{S}^\star \subset \mathcal{L}$ results in

$$\overline{\lim}_{x\to\infty}\frac{I_2(x)}{\mathbb{P}[X>x]} \leqslant 1.$$

Combining (5.8) with (5.9) and the preceding limit concludes the proof. □

Lemma 5.4 *Let $X \in \mathcal{S}^\star$ and*

$$\overline{\lim}_{x\to\infty} \frac{\mathbb{P}[X^{(e)} > x]}{x\mathbb{P}[X > x]} < \infty.$$

If $\{X_i^{(e)}\}_{i=1}^n$ are i.i.d. random variables equal in distribution to $X^{(e)}$, then for any $\epsilon > 0$ there exist C such that for all $x \geqslant 0$ and $n \geqslant 1$

$$\mathbb{P}\left[X + \sum_{i=1}^n X \wedge X_i^{(e)} > x\right] \leqslant C(1+\epsilon)^n \mathbb{P}[X > x].$$

Proof: Define $S_n \triangleq \sum_{i=1}^n X_i^{(e)}$. Then

$$\begin{aligned}\mathbb{P}\left[X + \sum_{i=1}^n X \wedge X_i^{(e)} > x\right] &\leqslant \mathbb{P}\left[X > \frac{x}{n+1},\ X + S_n > x\right] \\ &\leqslant \mathbb{P}\left[X > \frac{x}{n+1}\right] \mathbb{P}[S_n > x] + \mathbb{P}[X + S_n > x,\ S_n \leqslant x]. \end{aligned} \tag{5.13}$$

Next, Theorem A.1 and Lemma A.6 yield

$$\begin{aligned}\mathbb{P}\left[X > \frac{x}{n+1}\right] \mathbb{P}[S_n > x] &\leqslant C(1+\epsilon)^n \mathbb{P}\left[X > \frac{x}{n+1}\right] \mathbb{P}[X^{(e)} > x] \\ &\leqslant C(1+\epsilon)^n (n+1) \sup_{z\geqslant 0} \{z\mathbb{P}[X > z]\}\, x^{-1}\mathbb{P}[X^{(e)} > x] \\ &\leqslant Cn(1+\epsilon)^n \mathbb{P}[X > x], \end{aligned} \tag{5.14}$$

where in the last inequality we used the assumption and the fact that $\mathbb{E}X < \infty$. The second term in (5.13) can be bounded in the same way as in (5.7) since $X \in \mathcal{S}^\star$

$$\mathbb{P}[X + S_n > x,\ S_n \leqslant x] \leqslant C(1+\epsilon)^n \mathbb{P}[X > x].$$

Combining (5.13) with (5.14) and the preceding inequality concludes the proof. □

5.5 Proof of Theorem 3.2

(i) For any $\max(1/2, \beta) < \gamma < 1$ the union bound gives

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i - \mathbb{E}X\mathbb{E}N(u) > x\right] \leqslant \mathbb{P}\left[\sum_{i=1}^{N(u)} X_i \mathbf{1}\{X_i \leqslant \gamma x\} - \mathbb{E}X\mathbb{E}N(u) > x\right] + \lambda u \mathbb{P}[X > \gamma x].$$

We point out that, by assumption $\mathbb{P}[X > x] \leqslant Cxe^{-Q(x)}$ and (3.2),

$$\lambda u \mathbb{P}[X > \gamma x] \leqslant Cuxe^{-\gamma Q(x)} \leqslant Cue^{-\frac{1}{2}Q(x)},$$

from which one concludes that the first statement of the lemma holds if for all x and u

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i \mathbf{1}\{X_i \leqslant \gamma x\} - \mathbb{E}X\mathbb{E}N(u) > x\right] \leqslant C\left(e^{-c\frac{x^2}{u}} + e^{-\frac{1}{2}Q(x)}\right). \tag{5.15}$$

In the proof of this statement we restrict our attention to $u \leqslant \eta x^2$ for some $\eta > 0$, since for any $\eta > 0$ and $u > \eta x^2$ the bound holds trivially if C is chosen large enough, i.e. $Ce^{-c/\eta} > 1$. Next, let

$$\frac{1}{\gamma x} \leqslant s \leqslant \frac{Q(x)}{x}. \tag{5.16}$$

Then, Markov's inequality yields

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i \mathbf{1}\{X_i \leqslant \gamma x\} - \lambda u \mathbb{E}X > x\right] \leqslant e^{-s(x+\lambda u \mathbb{E}X)} e^{\lambda u \left(\mathbb{E}e^{sX\mathbf{1}\{X \leqslant \gamma x\}} - 1\right)}. \tag{5.17}$$

We start with estimating the moment generating function of $X\mathbf{1}\{X \leqslant \gamma x\}$

$$\mathbb{E}e^{sX\mathbf{1}\{X\leqslant \gamma x\}} = \int_0^{1/s} e^{sy}\, d\mathbb{P}[X \leqslant y] + \int_{1/s}^{\gamma x} e^{sy}\, d\mathbb{P}[X \leqslant y] + \mathbb{P}[X > \gamma x]. \tag{5.18}$$

The last term, by Markov's inequality, can be upper bounded as

$$\mathbb{P}[X > \gamma x] \leqslant \frac{\mathbb{E}X^2}{\gamma^2 s^2 x^2} s^2 \leqslant \mathbb{E}X^2 s^2; \tag{5.19}$$

recall that (3.2) implies $\mathbb{E}X^2 < \infty$. Inequality $e^x \leqslant 1 + x + x^2$ on $[0, 1]$, gives rise to

$$\begin{aligned} \int_0^{1/s} e^{sy}\, d\mathbb{P}[X \leqslant y] &\leqslant \int_0^{1/s} (1 + sy + s^2 y^2)\, d\mathbb{P}[X \leqslant y] \\ &\leqslant 1 + s\mathbb{E}X + s^2 \mathbb{E}X^2. \end{aligned} \tag{5.20}$$

Next, concavity of $Q(y)$ renders for $1/s \leqslant y \leqslant \gamma x$

$$sy - Q(y) \leqslant \max\{s\gamma x - Q(\gamma x),\ 1 - Q(1/s)\}$$

and, hence, integration by parts and Markov's inequality yield

$$\begin{aligned} \int_{1/s}^{\gamma x} e^{sy}\, d\mathbb{P}[X \leqslant y] &\leqslant e\mathbb{P}[X > 1/s] + Csx \int_{1/s}^{\gamma x} e^{sy - Q(y)} dy \\ &\leqslant s^2 e \mathbb{E}X^2 + Csx^2 \left(e^{s\gamma x - Q(\gamma x)} + e^{1 - Q(1/s)}\right) \\ &\leqslant Cs^2 \left(1 + x^3 \left(e^{s\gamma x - Q(\gamma x)} + e^{1 - Q(1/s)}\right)\right). \end{aligned} \tag{5.21}$$

The expression in brackets in (5.21) is an increasing function in s that achieves its maximum for $s = Q(x)/x$ (see (5.16)). Then, by (3.3), $\gamma Q(x) - Q(\gamma x) \leqslant -(1-\alpha)(1-\gamma)Q(x)$ and, by Lemma 3.1 (i) $Q(x/Q(x)) \geqslant Q(\sqrt{x})$; thus, the last two bounds, (3.2) and (5.21) yield

$$\int_{1/s}^{\gamma x} e^{sy}\, d\mathbb{P}[X \leqslant y] \leqslant Cs^2. \tag{5.22}$$

Hence, combining bounds (5.18), (5.19), (5.20) and (5.22) we derive

$$\mathbb{E}e^{sX\mathbf{1}\{X \leqslant \gamma x\}} \leqslant 1 + s\mathbb{E}X + C^* s^2, \tag{5.23}$$

where C^* is a constant. Substituting this estimate for $\mathbb{E}e^{sX\mathbf{1}\{X\leqslant\gamma x\}}$ in (5.17) yields

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i\mathbf{1}\{X_i\leqslant\gamma x\}-\lambda u\mathbb{E}X>x\right]\leqslant e^{-sx+\lambda uC^*s^2}.$$

Next, if $u\leqslant x^2/(2\lambda C^*Q(x))$, then by setting $s=Q(x)/x$ we derive

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i\mathbf{1}\{X_i\leqslant\gamma x\}-\lambda u\mathbb{E}X>x\right]\leqslant e^{-\frac{1}{2}Q(x)}. \tag{5.24}$$

On the other hand, if $u\geqslant x^2/(2\lambda C^*Q(x))$, $s=x/(2\lambda uC^*)$ $(\leqslant Q(x)/x)$ yields

$$\mathbb{P}\left[\sum_{i=1}^{N(u)} X_i\mathbf{1}\{X_i\leqslant\gamma x\}-\lambda u\mathbb{E}X>x\right]\leqslant e^{-\frac{x^2}{4\lambda C^*u}}. \tag{5.25}$$

Since for any value of u either (5.24) or (5.25) holds we conclude that (5.15) and, therefore, the first statement of the theorem holds.
(ii) The proof of the second statement proceeds along the similar lines. Choosing $\gamma<(k+1)^{-1/(1-\alpha)}$, $s=(k+1)Q(x)/x$ and using Lemma 3.1 (i) we verify

$$\begin{aligned}s\gamma x-Q(\gamma x)&\leqslant(k+1)\gamma Q(x)-Q(\gamma x)\\&\leqslant((k+1)\gamma^{1-\alpha}-1)\gamma^{\alpha}Q(x)<0.\end{aligned} \tag{5.26}$$

Next, Markov's inequality yields

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i\wedge\gamma x-n\mathbb{E}X>x\right]\leqslant e^{-s(x+n\mathbb{E}X)}\left(\mathbb{E}e^{s(X\wedge\gamma x)}\right)^n. \tag{5.27}$$

The moment generating function of $X\wedge\gamma x$ can be bounded as

$$\begin{aligned}\mathbb{E}e^{s(X\wedge\gamma x)}&=\int_0^{1/s}e^{sy}d\mathbb{P}[X\leqslant y]+\int_{1/s}^{\gamma x}e^{sy}d\mathbb{P}[X\leqslant y]+e^{\gamma sx-Q(\gamma x)}\\&\leqslant 1+s\mathbb{E}X+Cs^2,\end{aligned}$$

where the second inequality holds by the same arguments used in obtaining (5.23), and (5.26). Substituting the preceding bound in (5.27) results in

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n} X_i\wedge\gamma x-n\mathbb{E}X>x\right]&\leqslant e^{-s(x+n\mathbb{E}X)+n\log(1+s\mathbb{E}X+Cs^2)}\\&\leqslant e^{-s(x+n\mathbb{E}X)+n(s\mathbb{E}X+Cs^2)}\\&\leqslant e^{-(k+1)Q(x)\left(1-(k+1)CQ(x)/x\right)}\leqslant e^{-kQ(x)},\end{aligned}$$

for all x large enough, which renders the second part of Theorem 3.2. $\square$

5.6 Proof of Theorem 2.1

(Upper bound.) Note that $B \in \mathcal{IR}$ implies $B \in \mathcal{D} \cap \mathcal{L}$ and recall the definition of $\hat{B}_0$ from (3.5). Based on (3.1) and $\hat{B}_0 \geqslant B_0 \geqslant R_0(t)$ for all $t > 0$, the waiting time V_0 can be upper bounded as

$$\begin{aligned} V_0(1-\rho-\delta) &\leqslant \hat{B}_0 + \sum_{i=1}^{N(V_0)} B_i \wedge \hat{B}_0 - (\rho+\delta)V_0 \\ &\leqslant \hat{B}_0 + \sup_{t \geqslant 0} \left\{ \sum_{i=1}^{N(t)} B_i \wedge \hat{B}_0 - (\rho+\delta)t \right\} \end{aligned}$$

and, thus, for any positive $\delta < 1-\rho$

$$\mathbb{P}[V_0(1-\rho-\delta) > x] \leqslant \mathbb{P}[\hat{B}_0 > (1-\delta)x] + \mathbb{P}[W^{\rho+\delta}_{B \wedge \hat{B}_0} > \delta x]. \tag{5.28}$$

Next we examine the asymptotic behavior of the second term in (5.28)

$$\mathbb{P}[W^{\rho+\delta}_{B \wedge \hat{B}_0} > \delta x] \leqslant \mathbb{P}[\hat{B}_0 > \delta^2 x]\mathbb{P}[W^{\rho+\delta}_{B} > \delta x] + \mathbb{P}[W^{\rho+\delta}_{B \wedge \delta^2 x} > \delta x],$$

which by Lemmas 3.2, A.5 and Proposition 3.2 for δ sufficiently small results in

$$\mathbb{P}[W^{\rho+\delta}_{B \wedge \hat{B}_0} > \delta x] = o(\mathbb{P}[B > x]) \quad \text{as} \quad x \to \infty; \tag{5.29}$$

we use the fact that for any $B \in \mathcal{IR}$ there exists $\alpha > 0$, such that $\mathbb{P}[B > x] \geqslant C/x^\alpha$ (see (1.6) in [35]). Substituting (5.29) in (5.28) yields

$$\overline{\lim}_{x \to \infty} \frac{\mathbb{P}[V_0(1-\rho) > x]}{\mathbb{P}[B_0 > x]} \leqslant \overline{\lim}_{x \to \infty} \frac{\mathbb{P}\left[\hat{B}_0 > (1-\delta)x\right]}{\mathbb{P}[B_0 > \frac{1-\rho}{1-\rho-\delta}x]}.$$

Finally, recall Proposition 3.2 and then let $\delta \downarrow 0$ to complete the proof of the upper bound.
(Lower bound.) Recalling (3.11) for $\delta > 0$ results in

$$\mathbb{P}[(1-\rho)V_0 > x] \geqslant \mathbb{P}[(1-\rho)H^{\leftarrow}(x+\delta x) > x,\ B_0 > (1+2\delta)x,\ Z(V_0) \leqslant \delta x],$$

where Z is defined in the proof of Lemma 3.4. Next,

$$\mathbb{P}[(1-\rho)V_0 > x] \geqslant \left(\mathbb{P}[(1-\rho)H^{\leftarrow}(x+\delta x) > x] - \sup_{y \geqslant (1+2\delta)x} \mathbb{P}[Z(V^{(y)}) > \delta x]\right)\mathbb{P}[B > (1+2\delta)x]$$

and the result follows from the Law of Large Numbers, convergence of $Z(t)$ to a.s. finite Z and $B \in \mathcal{IR}$. □

Acknowledgment

The authors would like to thank Rudesindo Núñez-Queija and Bert Zwart for providing helpful comments.

Appendix: Heavy-tailed distributions

Here, we introduce some basic definitions and properties of heavy-tailed and subexponential distributions. First, we describe a family of long-tailed distribution functions. This is the largest operational class of heavy-tailed distributions. Let X be a random variable with distribution function (d.f.) F.

Definition A.1 *A nonnegative random variable X (or its d.f. F) is called long-tailed $X \in \mathcal{L}$ ($F \in \mathcal{L}$) if*

$$\lim_{x\to\infty} \frac{1-F(x-y)}{1-F(x)} = 1, \quad \forall y \in \mathbb{R}.$$

The following class of heavy-tailed distributions was introduced by Chistyakov [11].

Definition A.2 *A nonnegative random variable X (or its d.f. F) is called subexponential $X \in \mathcal{S}$ ($F \in \mathcal{S}$) if*

$$\lim_{x\to\infty} \frac{1-F^{2*}(x)}{1-F(x)} = 2,$$

where F^{2*} denotes the 2-fold convolution of F with itself, i.e., $F^{2*}(x) = \int_{[0,\infty)} F(x-y)F(dy)$.

It is well known that $\mathcal{S} \subset \mathcal{L}$ [6]. A survey on subexponential distributions can be found in [14]. The class of intermediately regularly varying distributions $\mathcal{IR}$ is a subclass of $\mathcal{S}$.

Definition A.3 *A nonnegative random variable X (or its d.f. F) is called intermediately regularly varying $X \in \mathcal{IR}$ ($F \in \mathcal{IR}$) if*

$$\lim_{\eta\uparrow 1} \overline{\lim}_{x\to\infty} \frac{1-F(\eta x)}{1-F(x)} = 1.$$

Regularly varying distributions $\mathcal{R}_\alpha$ are the best known examples from $\mathcal{IR}$ ($\mathcal{R}_\alpha \subset \mathcal{IR}$).

Definition A.4 *A nonnegative random variable X (or its d.f. F) is called regularly varying with index α, $X \in \mathcal{R}_\alpha$ ($F \in \mathcal{R}_\alpha$) if*

$$F(x) = 1 - \frac{l(x)}{x^\alpha}, \quad \alpha \geqslant 0,$$

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

Lemma A.5 *Let $F \in \mathcal{IR}$, $\eta \in (0,1)$, then*

$$\sup_{x\in[0,\infty)} \frac{1-F(\eta x)}{1-F(x)} < \infty.$$

Proof: Follows immediately from the definition. □

Definition A.5 *A nonnegative random variable X (or its d.f. F) belongs to the class $\mathcal{S}^\star$ if X has finite expectation and*

$$\lim_{x\to\infty} \int_0^x \frac{1-F(x-y)}{1-F(x)} (1-F(y))\, dy = 2\mathbb{E}X.$$

Definition A.6 *A nonnegative random variable X (or its d.f. F) belongs to the class $\mathcal{D}$ of dominated-variation distributions if*

$$\overline{\lim}_{x\to\infty} \frac{1-F(x)}{1-F(2x)} < \infty.$$

Theorem A.1 (Klüppelberg [20]) *(a) If $F \in \mathcal{D} \cap \mathcal{L}$ has finite expectation, then $F \in \mathcal{S}^\star$. (b) If $F \in \mathcal{S}^\star$, then $F \in \mathcal{S}$ and $F^{(e)} \in \mathcal{S}$.*

Lemma A.6 (Klüppelberg [21]) *Let $\{X_i\}_{i=0}^{\infty}$ be i.i.d. random variables. If $X_0 \in \mathcal{S}^\star$, then*

(i) for each $\epsilon > 0$ there exists a constant $K(\epsilon) > 0$ such that

$$\frac{d\mathbb{P}\left[\sum_{i=1}^{n} X_i^{(e)} \leqslant x\right]}{dx} \leqslant K(\epsilon)(1+\epsilon)^n \mathbb{P}[X_0 > x], \qquad x \geqslant 0,\ n \geqslant 1.$$

(ii) for any fixed n, as $x \to \infty$

$$\frac{d\mathbb{P}[\sum_{i=1}^{n} X_i^{(e)} \leqslant x]}{dx} \sim n\frac{\mathbb{P}[X_0 > x]}{\mathbb{E}X_0}.$$

References

[1] J. Abate and W. Whitt. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Syst. Theory Appl.*, 25(1/4):173–223, 1997.

[2] R. Agrawal, A. Makowski, and Ph. Nain. On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Syst. Theory Appl.*, 33(1/3):5–41, 1999.

[3] V. Anantharam. Scheduling strategies and long-range dependence. *Queueing Syst. Theory Appl.*, 33(1/3), 1999.

[4] S. Asmussen. *Applied Probability and Queues.* Wiley, 1987.

[5] S. Asmussen, C. Klüppelberg, and K. Sigman. Sampling at subexponential times, with queueing applications. *Stochastic Process. Appl.*, 79:265–286, 1999.

[6] K. B. Athreya and P. E. Ney. *Branching Processes.* Springer-Verlag, 1972.

[7] F. Baccelli and D. Towsley. The customer response times in the processor sharing queue are associated. *Queueing Syst. Theory Appl.*, 7:269–282, 1990.

[8] A. Baltrunas. On the asymptotics of one-sided large deviation probabilities. *Lith. Math. J.*, 35(1):11–17, 1995.

[9] S. Borst, O. Boxma, and P. Jelenković. Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. *Queueing Syst. Theory Appl.*, 2001, to apper.

[10] H. Chen, O. Kella, and G. Weiss. Fluid approximations for a processor sharing queue. *Queueing Syst. Theory Appl.*, 27(1/2):99–125, 1997.

[11] V. P. Chistyakov. A theorem on sums of independent positive random variables and its application to branching random processes. *Theory Probab. Appl.*, 9:640–648, 1964.

[12] E.G. Coffman, R.R. Muntz, and H. Trotter. Waiting time distributions for processor-sharing systems. *J. ACM*, 17(1):123–130, 1970.

[13] S. Foss and D. Korshunov. Sampling at random time with a heavy-tailed distribution. *Markov Process. Related Fileds*, 6:543–568, 2000.

[14] C. M. Goldie and C. Klüppelberg. Subexponential distributions. In R. Adler, R. Feldman, and M.S. Taqqu, editors, *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*, pages 435–459. Birkhäuser, Boston, 1998.

[15] S. Grishechkin. GI/G/1 processor sharing queue in heavy traffic. *Adv. Appl. Probab.*, 26:539–555, 1994.

[16] P. Jelenković. Network multiplexer with truncated heavy-tailed arrival streams. In *Proc. IEEE Infocom*, New York, NY, March 1999.

[17] P. Jelenković and A. Lazar. Asymptotic results for multiplexing subexponential on-off processes. *Adv. Appl. Probab.*, 31(2):394–421, 1999.

[18] F. Kelly. *Reversibility and Stochastic Networks.* Wiley, 1979.

[19] F. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.*, 48, 1998.

[20] C. Klüppelberg. Subexponential distributions and integrated tails. *J. Appl. Probab.*, 25:132–141, 1988.

[21] C. Klüppelberg. Subexponential distributions and characterizations of related classes. *Probability Theory and Related Fields*, 82:259–269, 1989.

[22] C. Kotopoulos, N. Likhanov, and R. Mazumdar. Asymptotic analysis of the GPS system fed by heterogeneous long-tailed sources. In *Proc. IEEE Infocom*, Anchorage, AK, April 2001.

[23] Z. Liu, N. Niclausse, and C. Jalpa-Villanueva. Web traffic modeling and performance comparison between HTTP 1.0 and HTTP 1.1. In E. Gelenbe, editor, *Systems Performance Evaluation: Methodologies and Applications*, pages 177–189. CRC Press, 1999.

[24] Z. Liu, N. Niclausse, and C. Jalpa-Villanueva. Traffic model and performance evaluation of Web servers. *Performance Evaluation*, 46(2-3):77–100, 2001.

[25] L. Massoulié and J. Roberts. Bandwidth sharing: Objectives and algorithms. In *Proc. IEEE Infocom*, New York, NY, March 1999.

[26] J.A. Morrison. Response-time distribution for a processor sharing system. *SIAM J. Appl. Math.*, 45(1):152–167, 1985.

[27] A. V. Nagaev. Integral limit theorems taking large deviations into account when Cramér's condition does not hold I, II. *Theory Probab. Appl.*, 14:51–64, 193–208, 1969.

[28] A.V. Nagaev. On a propety of sums of independent random variables. *Theory Probab. Appl.*, 22(2):326–338, 1977.

[29] S. V. Nagaev. Large deviations of sums of independent random variables. *Ann. Probab.*, 7(5):745–789, 1979.

[30] R. Núnez-Queija. *Processor-Sharing Models for Integrated-Services Networks.* PhD thesis, Eindhoven University of Technology, January 2000.

[31] R. Núnez-Queija. Queues with equally heavy sojourn time and service requiremet distributions. CWI Research Report PNA-R0201. Submitted for publication, 2002.

[32] T.J. Ott. The sojourn-time distribution in the M/G/1 queue with processor sharing. *J. Appl. Probab.*, 21:360–378, 1984.

[33] A.G. Pakes. On the tails of waiting-time distribution. *J. Appl. Probab.*, 12:555–564, 1975.

[34] K. Park and W. Willinger, editors. *Self-similar Network Traffic and Performance Evaluation.* Wiley, 2000.

[35] S. Resnick and G. Samorodnitsky. Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Syst. Theory Appl.*, 33(1/3):43–71, 1999.

[36] M. Sakata, S. Noguchi, and J. Oizumi. Analysis of a processor shared queueing model for time sharing systems. In *Proc. of 2nd Hawaii Internat. Conf. on System Sciences*, pages 625–628, 1969.

[37] R. Schassberger. A new approach to the M/G/1 processor sharing queue. *Adv. Appl. Probab.*, 16:202–213, 1984.

[38] B. Sengupta. An approximation for the sojourn-time distribution for the GI/G/1 processor sharing queue. *Comm. Statist. Stochastic Models*, 8:35–57, 1992.

[39] M. Squillante, D. Yao, and L. Zhang. Web traffic modeling and web server performance analysis. In *Proc. IEEE 38th Conf. Decision and Control*, pages 4432–4437, Phoenix, AZ, 1999.

[40] A. Ward and W. Whitt. Predicting response times in processor-sharing queues. In D.J. MacDonald P.W. Glynn and S.J. Turner, editors, *Proc. of the Fields Institute Conference on Communication Networks*, 2000.

[41] R. W. Wolff. *Stochastic Modeling and Theory of Queues.* Prentice Hall, 1989.

[42] S.F. Yashkov. A derivation of response time distribution for a M/G/1 processor sharing queue. *Problems Control Inform. Theory*, 12:133–148, 1983.

[43] S.F. Yashkov. Mathematical problems in the theory of shared-processor systems. *J. Soviet Math.*, 58:101–147, 1992.

[44] S.F. Yashkov. On heavy traffic limit theorem for the M/G/1 processor-sharing queue. *Comm. Statist. Stochastic Models*, 9:467–471, 1993.

[45] A.P. Zwart and O.J. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Syst. Theory Appl.*, 35(1/4):141–166, 2000.