UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Large deviations for sojourn times in processor sharing queues.

Mandjes, M.R.H.; Zwart, B.

DOI
10.1007/s11134-006-5567-6
Publication date
2006

Published in
Queueing Systems

Link to publication

Citation for published version (APA):
Mandjes, M. R. H., & Zwart, B. (2006). Large deviations for sojourn times in processor sharing queues. *Queueing Systems*, *52*(4), 237-250. https://doi.org/10.1007/s11134-006-5567-6

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Sep 2025

Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Large deviations of sojourn times in processor sharing queues

M.R.H. Mandjes, A.P. Zwart

REPORT PNA-E0410 JUNE 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2004, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Large deviations of sojourn times in processor sharing queues

ABSTRACT
This paper presents a large deviation analysis of the steady-state sojourn time distribution in the GI/G/1 PS queue. Logarithmic estimates are obtained under the assumption of the service time distribution having a light tail, thus supplementing recent results for the heavy-tailed setting. Our proof gives insight in the way a large sojourn time occurs, enabling the construction of an (asymptotically efficient) importance sampling algorithm. Finally our results for PS are compared to a number of other service disciplines, such as FCFS, LCFS, and SRPT.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: processor sharing queues, sojourn time, large deviations, change of measure, importance sampling

Large deviations of sojourn times in processor sharing queues

Michel Mandjes[*,†], Bert Zwart[*,‡]

[*]CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[†]Faculty of Mathematical Sciences
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

[‡]Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

michel@cwi.nl, zwart@win.tue.nl

May 27, 2004

Abstract

This paper presents a large deviation analysis of the steady-state sojourn time distribution in the $GI/G/1$ PS queue. Logarithmic estimates are obtained under the assumption of the service time distribution having a light tail, thus supplementing recent results for the heavy-tailed setting. Our proof gives insight in the way a large sojourn time occurs, enabling the construction of an (asymptotically efficient) importance sampling algorithm. Finally our results for PS are compared to a number of other service disciplines, such as FCFS, LCFS, and SRPT.

2000 Mathematics Subject Classification: 60K25.

Keywords & Phrases: processor sharing queues, sojourn time, large deviations, change of measure, importance sampling.

1 Introduction

There is a vast body of literature on waiting times in $GI/G/1$ queues operating under the First-Come-First-Served (FCFS) service discipline. Special emphasis was on the impact of the nature of the service-time distribution on the tail asymptotics, see for example Asmussen [3] and Glynn & Whitt [11]; notably there is a fundamental difference between waiting-time asymptotics under heavy-tailed and light-tailed service requirements. Perhaps the most prominent alternative service discipline is Processor Sharing (PS), in which the available capacity is shared equally among the active users. A number of recent papers were concerned with sojourn-time asymptotics in PS queues with heavy-tailed service times, see, e.g., [13, 16, 20, 31], but under light tails hardly any results are available.

To our best knowledge, only for the $M/M/1$ PS explicit asymptotics were known. It was shown that the probability that the sojourn-time V attains an extreme value obeys the following 'exact asymptotics':

$$\mathbb{P}\{V > x\} \sim cx^{-5/6}\mathrm{e}^{-\alpha x^{1/3}}\mathrm{e}^{-\gamma x}, \quad x \to \infty, \tag{1.1}$$

for positive constants c, α, γ, and $f(x) \sim g(x)$ denoting $f(x)/g(x) \to 1$. Flatto [10] proved the asymptotic relation (1.1) for the tail of the waiting-time distribution in the $M/M/1$ Random-Order-of-Service (ROS) queue; subsequently Borst *et al.* [5] showed that waiting times of the $M/M/1$ ROS queue can be directly related to sojourn times of the $M/M/1$ PS. In this context, also an early (1946) study by Pollaczek [23] is worth mentioning. We remark that the 'mixed polynomial-Weibullian-exponential asymptotics' appearing in (1.1) are quite uncommon in queueing theory.

The present paper departs from the exponentiality assumptions required in (1.1): we analyze sojourn-time asymptotics in the $GI/G/1$ PS queue, for a broad class of light-tailed service-time distributions. The nature of our asymptotics, however, is somewhat weaker than the type of asymptotics of (1.1): we settle for *logarithmic* (rather than exact) asymptotics. More specifically, the main result of our work is that

$$\log \mathbb{P}\{V > x\} \sim -\gamma x, \qquad x \to \infty, \tag{1.2}$$

for some $\gamma > 0$ that is determined by the distributions of the interarrival times and service times. Equivalently, one could say that $\mathbb{P}\{V > x\} = \mathrm{e}^{-\gamma x(1+\mathrm{o}(1))}$; the exact statement of our result is found in Theorem 3.1.

Contrasting the approach of Flatto [10], who relies on an explicit integral expression of the ROS waiting-time distribution for $M/M/1$, our proofs are of a probabilistic nature. As a consequence, they offer insight into *how* a large sojourn time is achieved. Evidently, a large sojourn time is caused by a combination of three effects:

(i) a large number of customers present at the arrival of the tagged customer (possibly with large residual service requirements);

(ii) the tagged customer having a large service requirement;

(iii) the work brought along by customers arriving during the sojourn time of the tagged customer.

Interestingly, our analysis indicates that for a broad class of light-tailed service times the logarithmic sojourn-time asymptotics are dominated by effect (iii). This is essentially different in the heavy-tailed case: if the service-time distribution is sufficiently heavy-tailed, then the large sojourn time of the tagged customer is predominantly due to effect (ii), i.e., its *own* large service time, see, e.g., [13, 16, 20, 31]. In other words: under heavy tails the service requirements of other customers have no significant impact.

Our proof is based on a change-of-measure argument: the interarrival times and services times are 'exponentially twisted' such that the tagged customer sees a system with load 1. This explains why our logarithmic asymptotics resemble those of the probability of a

long busy period. We prove (1.2) under two technical conditions which, loosely speaking, ensure that the service-time distribution is not too heavy (some finite exponential moments are required), and — perhaps surprising — *not too light*. For example, our conditions are satisfied by many distributions of practical interest, but *not* by distributions with bounded support, such as deterministic service times. In Section 4 we obtain a simple and elegant upper bound for $\mathbb{P}\{V > x\}$ in the $M/D/1$ PS queue which shows that the large-deviations behavior in this case is fundamentally different. It seems that the above-mentioned effect (i) plays an important role here.

We now turn to an overview of the content of our paper. Preliminary results are given in Section 2. Section 3 contains the main result and its proof, based on the change-of-measure mentioned above, and in addition a powerful fluid limit result for PS queues in overload, obtained recently by Puha *et al.* [24]. Section 4 treats a number of ramifications. An explicit exponential upper bound on $\mathbb{P}\{V > x\}$ is given for $M/G/1$ PS (Section 4.1), and an analysis of the decay rate γ under heavy traffic (Section 4.2). Section 4.3 indicates what happens if the tail of the service-time distribution is 'too light' (such as in $M/D/1$ PS), whereas Section 4.4 focuses on the role played by the assumption regarding exponential moments.

Section 5 applies our results to construct a rare-event simulation algorithm based on importance sampling. For the case of $M/G/1$ PS we prove that our choice of the importance-sampling parameters is the best in the class of exponential twists, in that the proposed change-of-measure is 'asymptotically optimal'. Our simulation approach outperforms straightforward Monte-Carlo simulation. A simulation study shows that, if the load of the PS queue is high, both the exact asymptotics (1.1) and the approximation based on the logarithmic asymptotics (1.2) are not very accurate, whereas for low load the fit is considerably better.

In Section 6, we compare our results for PS to results for other service disciplines, motivated by a recent interest in the impact of scheduling on system performance, see, e.g., [6, 30]. It turns out that the logarithmic asymptotics of the sojourn time distribution coincide for a large number of standard service disciplines, like Last-Come-First-Served (LCFS), Shortest-Remaining-Processing-Time (SRPT), Foreground-Background-PS (FBPS), and ROS. This section also mentions several open problems.

2 Model description and preliminaries

We consider a $GI/G/1$ queue operating under the PS service discipline; in PS, the server assigns each customer a service rate $1/n$ when there are n customers in the system. This policy is obviously work-conserving, and the $GI/G/1$ PS queue is positive recurrent when the load $\rho < 1$, which we assume throughout this paper. We consider a customer entering the system (say, at time 0) in steady state. Let V be the sojourn time of this tagged customer and B_0 its service requirement. Our main focus is on the asymptotic behavior of $\mathbb{P}\{V > x\}$ as $x \to \infty$, under light-tailed service times. Some specific assumptions will be imposed on the distribution of the service times.

To analyze the sojourn-time asymptotics, we will use an exponential change-of-measure, cf. Asmussen [3, Ch. XIII]. The proof of our main Theorem 3.1 requires that some random variables are exponentially twisted, whereas others remain unchanged. For the sake of clarity, we feel that it is appropriate to give a detailed description of the underlying filtered probability space.

The state of the PS system at time 0 is fully described by the queue length (i.e., number of customers) Q_0, and their residual service times $\bar{B}_1, ..., \bar{B}_{Q_0}$. Evidently, the workload at time 0 is given by $W = \bar{B}_1 + \ldots + \bar{B}_{Q_0}$. We define our filtration as follows. First let $\mathcal{F}_0 := \sigma\{B_0, Q_0, \bar{B}_1, ..., \bar{B}_Q\}$ denote the σ-algebra containing all events which are known at time 0 and relevant for $\{V > x\}$. Let $A_n, n \in \mathbb{N}$, be the time between the $(n-1)$st and nth arriving customer after time 0. Furthermore, let $B_n, n \in \mathbb{N}$ be the service time of the nth customer. We assume that $(A_n)_n$ and $(B_n)_n$ are mutually independent sequences, each consisting of i.i.d. random variables. Now define

$$\mathcal{F}_n := \mathcal{F}_0 \cup \sigma\{(A_i)_{i=1,...,n}, (B_i)_{i=1,...,n}\}$$

as the 'information available about the system' up to the nth arrival. We also introduce the random walks $S_n^A := A_1 + \ldots + A_n$, $S_n^B := B_1 + \ldots + B_n$, and $S_n := S_n^B - S_n^A$.

We define the 'generic' interarrival time and service time by A and B, respectively. As mentioned above, we assume the system to be stable: $\mathbb{E}\{B - A\} < 0$, or equivalently $\rho \equiv \mathbb{E}\{B\}/\mathbb{E}\{A\} < 1$. We also assume $\mathbb{E}\{A\} < \infty$. Define the moment generating functions $\Phi_B(s) = \mathbb{E}\{\mathrm{e}^{sB}\}$, $\Phi_A(s) = \mathbb{E}\{\mathrm{e}^{sA}\}$, $\Phi(s) = \mathbb{E}\{\mathrm{e}^{s(B-A)}\}$. Note that both Φ_A and Φ_B are strictly increasing and strictly convex functions in s, so that the inverse functions $\Phi_A^{\leftarrow}(s)$ and $\Phi_B^{\leftarrow}(s)$ are well defined.

Finally, we set $N(x) := \max\{n \in \mathbb{N} : S_n^A \leq x\}$ representing the number of arrivals between 0 and x. Furthermore, let $A(x) := S_{N(x)}^B$ be the total amount of work fed into the queue between time 0 and x. We recall that $f(x) \sim g(x)$ means $f(x)/g(x) \to 1$.

The following basic lemma is of crucial importance in the rest of the paper.

Lemma 2.1 *The asymptotic cumulant function of $A(x)$ is given by*

$$\Psi(s) := \lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}e^{sA(x)} = -\Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(s)}\right).$$

This result has been stated as a conjecture by Whitt [29], see in particular (1.15) of that reference. We have not been able to find the result in the present setting in later references, like [11]. Hence, for completeness, we include the proof. Note that no further assumptions are needed on, e.g., the right tail of the distribution of A.

Proof. Since

$$\mathbb{E}\{\mathrm{e}^{sA(x)}\} = \sum_{n=0}^{\infty} \mathbb{P}\left\{N(x) = n\right\} \Phi_B(s)^n = \mathbb{E}\{\Phi_B(s)^{N(x)}\}, \tag{2.1}$$

it suffices to determine the asymptotic behavior of $\mathbb{E}\{r^{N(x)}\}$. To do so, we consider its Laplace-Stieltjes transform w.r.t. x. Note that

$$\mathbb{E}\{r^{N(x)}\} = \sum_{n=0}^{\infty} r^n \left(\mathbb{P}\left\{S_n^A \leq x\right\} - \mathbb{P}\left\{S_{n+1}^A \leq x\right\}\right) = 1 + \left(1 - \frac{1}{r}\right) \sum_{n=1}^{\infty} r^n \mathbb{P}\left\{S_n^A \leq x\right\}.$$

Consequently,

$$\hat{N}(q) := \int_0^\infty \mathrm{e}^{-qx} \mathrm{d}\mathbb{E}\{r^{N(x)}\} = 1 + \left(1 - \frac{1}{r}\right) \frac{r\Phi_A(-q)}{1 - r\Phi_A(-q)}.$$

We see that the rightmost pole of $\hat{N}(q)$ is attained when $q = q(r)$ is such that $r\Phi_A(-q) = 1$, i.e. when $q(r) = -\Phi_A^{\leftarrow}(1/r)$. From this and the definition of $\hat{N}$ it immediately follows that

$$\log \mathbb{E}\{r^{N(x)}\} \sim -\Phi_A^{\leftarrow}(1/r)x. \tag{2.2}$$

Combining this result with (2.1) completes the proof. □

3 Main results

In this section we determine the exponential decay rate of $\mathbb{P}\{V > x\}$. Analogously to the proof of Cramér's theorem cf. [3], the proof consists of (i) an upper bound based on the classical Chernoff bound, and (ii) a lower bound based on a change-of-measure argument. The derivation of these bounds is carried out in the following two subsections.

3.1 Upper bound

We first derive the upper bound. The following assumption, which requires the existence of certain exponential moments, is imposed.

Assumption 3.1 *Set*

$$\omega \equiv \omega(\nu) := \Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(\nu)}\right) = -\Psi(\nu). \tag{3.1}$$

There exists a solution $\nu^* > 0$ to

$$\frac{\Phi_A'(\omega)}{\Phi_A(\omega)} = \frac{\Phi_B'(\nu)}{\Phi_B(\nu)}. \tag{3.2}$$

Furthermore, $\Phi_B(\nu) < \infty$ for ν in an environment of ν^*.

Define $\omega^* := \omega(\nu^*)$ and $\gamma := \nu^* + \omega^*$. Lemma 2.1 implies the following property.

Property 3.1 *ν^* is the optimizing argument of $\inf_{s\geq 0}[\Psi(s) - s]$.*

Proposition 3.1 *Under Assumption 3.1, we have*

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{V > x\} \leq -\gamma = -\omega^* - \nu^*.$$

Proof

We first consider the asymptotic behavior of $\mathbb{P}\{A(x) > x\}$. By the Chernoff bound we have, for all $s \geq 0$,

$$\mathbb{P}\{A(x) > x\} = \mathbb{P}\left\{e^{s[A(x)-x]} > 1\right\} \leq \mathbb{E}\{e^{s[A(x)-x]}\}$$

Hence, by virtue of Lemma 2.1, optimization over $s \geq 0$, and Property 3.1, we obtain

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{A(x) - x\} \leq \inf_{s\geq 0}[\Psi(s) - s] = -\omega^* - \nu^* = -\gamma.$$

Since $\Psi(s) - s$ equals 0 at $s = 0$, and has in addition a strictly negative derivative at $s = 0$, it follows that $-\gamma = -\omega^* - \nu^* < 0$, or equivalently, $\nu^* > -\omega^*$.

Next, we turn to $\mathbb{P}\{V > x\}$. Since the event $\{V > x\}$ implies that the queue does not empty before time x, we can write

$$\begin{aligned} \mathbb{P}\{V > x\} &\leq \mathbb{P}\{W + B_0 + A(x) - x > 0\} \\ &\leq \mathbb{E}\{\mathrm{e}^{\nu^* W}\}\mathbb{E}\{\mathrm{e}^{\nu^* B}\}\mathbb{E}\{\mathrm{e}^{-\nu^*(x-A(x))}\} \end{aligned} \tag{3.3}$$

Now note that $\mathbb{E}\{\mathrm{e}^{\nu^* B}\} < \infty$ in view of Assumption 3.1. Furthermore, we have

$$\begin{aligned} \mathbb{E}\{\mathrm{e}^{\nu^* W}\} &= \int_0^\infty \mathbb{P}\left\{\mathrm{e}^{\nu^* W} > x\right\} \mathrm{d}x = \int_0^\infty \mathbb{P}\left\{\sup_{n\geq 0} \mathrm{e}^{\nu^* S_n} > x\right\} \mathrm{d}x \\ &\leq \sum_{n=0}^\infty \int_0^\infty \mathbb{P}\left\{\mathrm{e}^{\nu^* S_n} > x\right\} \mathrm{d}x = \sum_{n=0}^\infty \Phi(\nu^*)^n < \infty; \end{aligned}$$

the finiteness of the last sum is due to $\Phi(\nu^*) = \Phi_A(-\nu^*)\Phi_B(\nu^*) < \Phi_A(\omega^*)\Phi_B(\nu^*) = 1$ (which in turn holds by (3.1) and the property $\nu^* > -\omega^*$). Taking logarithms, dividing by x, and letting $x \to \infty$ in (3.3) completes the proof. □

3.2 Lower bound

We now turn to the derivation of an asymptotic lower bound for $\mathbb{P}\{V > x\}$. For this, we need to make an additional assumption.

Assumption 3.2 *For each constant $c > 0$, we have*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{B > c \log x\} = 0.$$

Equivalently, Assumption 3.2 requires e^B to have an infinite moment generating function, in particular, e^B should be heavy-tailed. This assumption is satisfied by most distributions of interest (phase-type, Gamma, and Weibull distributions, etc.), but rules out distributions having extremely light tails. For instance, Assumption 3.2 is violated by service times for which $\mathbb{P}\{B > x\}$ is of the form $\exp(-\mathrm{e}^x)$, and also for any service time distribution with bounded support. Such distributions give rise to a fundamentally different behavior of the probability $\mathbb{P}\{V > x\}$, as will be demonstrated in Section 4.3.

Theorem 3.1 *If Assumptions 3.1 and 3.2 are valid, then*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{V > x\} = -\gamma = -\omega^* - \nu^*. \tag{3.4}$$

Evidently, the upper bound of Theorem 3.1 follows from Proposition 3.1. As announced above, our proof of the lower bound relies on a change-of-measure. In particular, we change the distributions of A_i and $B_i, i \in \mathbb{N}$, in such a way that the new load imposed on the system is *strictly larger* than 1, rather than $\rho < 1$. Under this new load, the event $\{V > x\}$ becomes considerably more likely than under the old load. More specifically, $\mathbb{P}\{V > x\}$ now decays at a *subexponential* rate, as shown by the following crucial lemma.

Lemma 3.1 *Let $\hat{\mathbb{P}}\{\cdot\}$ be a probability measure which is equal to $\mathbb{P}\{\cdot\}$ on $\mathcal{F}_0$ and which has the property that $\hat{\mathbb{E}}\{B_n - A_n\} = \hat{\rho} > 1$ for $n \geq 1$. Then*

$$\lim_{x\to\infty} \frac{1}{x} \log \hat{\mathbb{P}}\{V > x\} = 0.$$

By using this lemma, which is proven at the end of this subsection, we can give the following proof of Theorem 3.1.

Proof of Theorem 3.1. Define a probability measure $\mathbb{P}_\nu\{\cdot\}$ for $\nu \geq 0$ such that $\mathbb{P}_\nu\{E\} = \mathbb{P}\{E\}$ if $E \in \mathcal{F}_0$ and such that for $i \in \mathbb{N}$

$$\begin{aligned} \mathbb{P}_\nu\{A_i \in \mathrm{d}x\} &= \mathrm{e}^{\omega(\nu)x}\mathbb{P}\{A_i \in \mathrm{d}x\}/\Phi_A(\omega(\nu)), \\ \mathbb{P}_\nu\{B_i \in \mathrm{d}x\} &= \mathrm{e}^{\nu x}\mathbb{P}\{B_i \in \mathrm{d}x\}\ \Phi_B(\nu). \end{aligned}$$

Taking $\epsilon > 0$ sufficiently small, Assumption 3.1 entails that we can pick $\nu = \nu_\epsilon$ such that

$$\mathbb{E}_\nu\{B_i\}/\mathbb{E}_\nu\{A_i\} = \frac{\Phi'_B(\nu_\epsilon)}{\Phi_B(\nu_\epsilon)} \Big/ \frac{\Phi'_A(\omega(\nu_\epsilon))}{\Phi_A(\omega(\nu_\epsilon))} = 1 + \frac{\epsilon}{2}.$$

Denote the new probability measure by $\mathbb{P}_{\nu_\epsilon}\{\cdot\}$, and corresponding expectations by $\mathbb{E}_{\nu_\epsilon}\{\cdot\}$. Note that $\bar{N}(x) := N(x)+1$ is a stopping time w.r.t. the filtration $(\mathcal{F}_n)_{n\in\{0,1,\ldots\}}$. Furthermore, note that the event $\{V > x\}$ is $\mathcal{F}_{\bar{N}(x)}$-measurable. Finally, from Assumption 3.1 it follows that for every $\epsilon > 0$ small enough, the process $1/M_n^\epsilon, n \geq 1$, with

$$M_n^\epsilon = \mathrm{e}^{-\omega(\nu_\epsilon)S_n^A - \nu_\epsilon S_n^B},$$

is a martingale w.r.t $\mathcal{F}_n$ under $\mathbb{P}\{\cdot\}$, since $\Phi_A(\omega(\nu_\epsilon))\Phi_B(\nu_\epsilon) = 1$.
Thus, we have the following fundamental identity (see e.g. Theorem XIII.3.2 in [3]):

$$\mathbb{P}\{V > x\} = \mathbb{E}_{\nu_\epsilon}\{M^\epsilon_{\bar{N}(x)} I(V > x)\}. \tag{3.5}$$

Furthermore, we have for any event $S \subseteq \mathcal{F}_{\bar{N}(x)}$,

$$\mathbb{P}\{V > x\} \geq \mathbb{E}_{\nu_\epsilon}\{M^\epsilon_{\bar{N}(x)} I(V > x) I(\mathcal{S})\}. \tag{3.6}$$

Now choose, with $f_\pm(\epsilon) := (1 \pm \epsilon)/\mathbb{E}_{\nu_\epsilon}\{A\}$,

$$\mathcal{S} \equiv \mathcal{S}_\epsilon := \left\{ \frac{N(x)}{x} \in (f_-(\epsilon), f_+(\epsilon)); S^B_{N(x)} \leq (1+\epsilon)x \right\}.$$

Noting that, by definition, $S^A_{N(x)+1} > x$, and applying the definition of $\mathcal{S}_\epsilon$, we obtain the following lower bound for $\mathbb{P}\{V > x\}$ from (3.6):

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{V > x\} \geq -\nu_\epsilon - \omega(\nu_\epsilon)(1+\epsilon) + \liminf_{x\to\infty} \frac{1}{x} \log \mathbb{P}_{\nu_\epsilon}\{V > x; \mathcal{S}_\epsilon\}. \tag{3.7}$$

Consider the last term in (3.7); the goal is to prove that $\mathbb{P}_{\nu_\epsilon}\{V > x; S_\epsilon\}$ decays subexponentially, i.e., $\log \mathbb{P}_{\nu_\epsilon}\{V > x; S_\epsilon\} = \mathrm{o}(x)$ for any $\epsilon > 0$. We start by invoking the trivial lower bound

$$\mathbb{P}_{\nu_\epsilon}\{V > x; \mathcal{S}_\epsilon\} \geq \mathbb{P}_{\nu_\epsilon}\{V > x\} - \mathbb{P}_{\nu_\epsilon}\{\mathcal{S}_\epsilon^c\}.$$

Lemma 3.1 immediately implies that $\log \mathbb{P}_{\nu_\epsilon}\{V > x\} = \mathrm{o}(x)$. Therefore concentrate on $\mathbb{P}_{\nu_\epsilon}\{\mathcal{S}_\epsilon^c\}$, which is bounded from above by

$$\mathbb{P}_{\nu_\epsilon}\{\mathcal{S}_\epsilon^c\} \leq \mathbb{P}_{\nu_\epsilon}\left\{\frac{N(x)}{x} \leq f_-(\epsilon)\right\} + \mathbb{P}_{\nu_\epsilon}\left\{\frac{N(x)}{x} \geq f_+(\epsilon)\right\} + \mathbb{P}_{\nu_\epsilon}\left\{\frac{A(x)}{x} > 1+\epsilon\right\}. \quad (3.8)$$

We now show that the three probabilities in the r.h.s. of (3.8) decay to 0 exponentially fast in x. For the first two terms, note that

$$f_-(\epsilon) < \lim_{x\to\infty} \frac{1}{x}\mathbb{E}_{\nu_\epsilon}\{N(x)\} = \frac{1}{\mathbb{E}_{\nu_\epsilon}\{A_1\}} < f_+(\epsilon). \quad (3.9)$$

Furthermore, by using an argument similar to the one in Lemma 2.1, it can be shown that

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}_{\nu_\epsilon}\{r^{N(x)}\} = \Phi_A^{\leftarrow}(1/\Phi_B(\nu_\epsilon)) - \Phi_A^{\leftarrow}(1/(r\Phi(\nu_\epsilon))). \quad (3.10)$$

This function is finite for r in a neighborhood of $r = 1$. Combining (3.9), (3.10), and the Chernoff bound then directly leads to an exponential upper bound for the first two terms in (3.8).

The proof of the third term in (3.8) is similar; note that

$$\lim_{x\to\infty} \frac{1}{x}\mathbb{E}_{\nu_\epsilon}\{A(x)\} = 1 + \epsilon/2 < 1 + \epsilon,$$

and, as in Lemma 2.1,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}_{\nu_\epsilon}\{\mathrm{e}^{sA(x)}\} = \Psi(\nu_\epsilon + s) - \Psi(\nu_\epsilon),$$

which is finite around $s = 0$, in view of Assumption 3.1.

As $\mathbb{P}_{\nu_\epsilon}\{V > x\}$ decays subexponentially and $\mathbb{P}_{\nu_\epsilon}\{\mathcal{S}_\epsilon\}$ exponentially in x, we conclude that, for x large, $\mathbb{P}_{\nu_\epsilon}\{V > x\} - \mathbb{P}_{\nu_\epsilon}\{\mathcal{S}_\epsilon\} \sim \mathbb{P}_{\nu_\epsilon}\{V > x\} = \mathrm{e}^{\mathrm{o}(x)}$ for any $\epsilon > 0$. Combining this with (3.7) we find that

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{V > x\} \geq -\nu_\epsilon - \omega(\nu_\epsilon)(1+\epsilon).$$

Now let $\epsilon \downarrow 0$. This yields $\nu_\epsilon \to \nu^*$ and $\omega(\nu_\epsilon) \to \omega^*$. We thus obtain

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbb{P}\{V > x\} \geq -\nu^* - \omega^*.$$

This result, in conjunction with Proposition 3.1, yields the stated. □

Proof of Lemma 3.1. Write

$$\hat{\mathbb{P}}\{V > x\} = \hat{\mathbb{P}}\left\{B_0 > \int_0^x \frac{1}{1+Q_p(u)}\mathrm{d}u\right\}$$

In this expression $Q_p(u), u \geq 0$, represents the evolution of the number of customers in a PS queue with a single permanent customer (with $Q_p(0) = Q_0$). Moreover, $(Q_p(u))_{u\in[0,x]}$ and B_0 are independent. We now invoke a crucial result about the fluid limit of the number of customers $Q(u)$ at time u for transient PS queues in overload, see Theorem 3.11 of Puha *et al.* [24] — for a similar result, see Jean-Marie & Robert [15]. It entails that, if $Q_0 = 0$, there exists a constant $\hat{\alpha}$ such that

$$Q(ux)/x \to \hat{\alpha}u,$$

$\hat{\mathbb{P}}$-a.s. in $D(0,\infty)$. This result implies

$$\hat{\mathbb{P}}\left\{Q(u) > \hat{\alpha}u/2;\ x/2 \le u \le x\right\} \to 1,$$

as $x \to \infty$. From Lemma 4 of Guillemin *et al.* [13] we know that the sample paths of $Q_p(u)$ and $Q(u)$ can be compared in such a way that $Q_p(u) \ge Q(u)$ a.s. for all $u \ge 0$. (This is intuitively obvious since a queue with one permanent customer is not working at full speed, resulting in a larger number of customers at any time.) Thus, we also have

$$\hat{\mathbb{P}}\left\{Q_p(u) > \hat{\alpha}u/2;\ x/2 \le u \le x\right\} \to 1, \tag{3.11}$$

as $x \to \infty$. Combining these results, we obtain

$$\begin{aligned}\hat{\mathbb{P}}\left\{V > x\right\} &\ge \hat{\mathbb{P}}\left\{B_0 > \int_{x/2}^{x} \frac{1}{1+Q_p(u)}\mathrm{d}u;\ Q_p(u) \ge \hat{\alpha}u/2;\ x/2 \le u \le x\right\} \\ &\ge \mathbb{P}\left\{B_0 > \hat{\alpha}\log x\right\}\hat{\mathbb{P}}\left\{Q_p(u) \ge \hat{\alpha}u/2;\ x/2 \le u \le x\right\}.\end{aligned}$$

The first probability behaves like $\mathrm{e}^{\mathrm{o}(1)}$ in view of Assumption 3.2. The second probability converges to 1, cf. (3.11). This completes the proof. □

3.3 Discussion

We close this section with some remarks on the proof and mention some related results which put Theorem 3.1 into perspective.

1. The proof of Lemma 3.1 clearly shows why we need Assumption 3.2. The fact that we twisted the load from ρ to $1+\epsilon/2 > 1$ (and not 1) in the proof of Theorem 3.1 is useful for two reasons. First, it allows us to apply general theorems for transient PS queues, as derived in [15, 24]. Secondly, we believe that directly twisting to a rate 1 leads to more restrictive assumptions. This can be seen as follows. Heavy-traffic limit theorems as in Gromoll [14] suggest that, in heavy traffic, $Q(u) = \mathrm{O}(\sqrt{u})$. Hence, when $\hat{\mathbb{P}}\{\cdot\}$ is chosen such that $\hat{\rho} = 1$, we have

$$\hat{\mathbb{P}}\left\{V > x\right\} \approx \mathbb{P}\left\{B_0 > \int_0^x \frac{1}{\mathrm{O}(\sqrt{u})}\mathrm{d}u\right\} \approx \mathbb{P}\left\{B_0 > \mathrm{O}(\sqrt{x})\right\}. \tag{3.12}$$

In this way one would rule out tails of the form e^{-x^p} with $p \ge 2$, which, as shown above, is not necessary.

2. We implicitly assumed that the tagged customer (with service time B_0) and customers arriving into the system after time 0 (with generic service time B) have the same service time distributions. This assumption is not necessary: if the distributions of B_0 and B are different, the conclusion of Theorem 3.1 still holds if B and B_0 both satisfy Assumption 3.1 and B_0 alone satisfies Assumption 3.2; it is not necessary that B satisfies the latter Assumption. This fact is exploited in Section 4.4.

3. An interesting implication of our results is the following. If we let P_r be the residual busy period of a $GI/G/1$ queue, we see that

$$\mathbb{P}\left\{V > x\right\} \le \mathbb{P}\left\{P_r > x\right\} \le \mathbb{P}\left\{W + B + A(x) - x > 0\right\}.$$

Thus, from the analysis in this section, it follows that the decay rates of P_r and V coincide under Assumptions 3.1 and 3.2. Using the connection between their Laplace-Stieltjes transforms, it can be shown that the decay rate of the $GI/G/1$ busy period itself coincides with the decay rate of the residual busy period. We thus draw the conclusion that, under Assumptions 3.1 and 3.2,

$$\log \mathbb{P}\{P > x\} \sim -(\nu^* + \omega^*)x, \tag{3.13}$$

a result which was, up to recently, only known for the $M/G/1$ queue. (During the preparation of this paper we became aware of the recent work [22] which derives exact asymptotics of $\mathbb{P}\{P > x\}$ in the $GI/G/1$ queue; see Section 6 for further details and references.)

4. Interestingly enough, our analysis shows an asymmetry between the twisting of the interarrival-time and service-time distributions: A is exponentially twisted by ω^*, and B by $\nu^* > -\omega^*$. When studying the large deviations of the *number of customers arriving during a busy period* (which we denote by N), however, there *is* such a symmetry: the interarrival times are twisted by $-\theta^*$ and the service times by θ^*. Here θ^* is determined by the so-called 'Petrov-equation' $\Phi'(s) = 0$; note that N is the first downward passage time of 0 of the random walk S_n, of which the tail asymptotics have been thoroughly investigated in Bertoin & Doney [4]. Indeed, when a solution θ^* to this equation exists, it follows that

$$\log \mathbb{P}\{N > x\} \sim -\theta^* x;$$

also exact asymptotics are known, see, e.g., [4].
Put differently: the logarithmic asymptotics for $\mathbb{P}\{N > x\}$ and $\mathbb{P}\{P > x\}$ do *not* coincide (except when interarrival times are deterministic), despite the fact that in both cases the twisted random walk has zero drift. The reason is that twisting the interarrival time distribution has a higher impact in the continuous-time case than in the discrete-time case, the additional effect being that if the mean interarrival time becomes smaller, it allows more steps of the random walk until time x. Bearing these considerations in mind, it is not surprising that the interarrival time distribution is twisted by $-\omega^*$ which is *smaller* than ν^*, the twist of the service times.

4 Special cases and complements

This section presents a number of ramifications of Theorem 3.1. In Section 4.1 we treat the special case of Poisson arrivals. Section 4.2 is devoted to the behavior of the decay rate γ in heavy traffic. The last two subsections show what happens when our Assumptions 3.1 and 3.2 are violated.

4.1 Poisson arrivals

In this section we consider the special case in which jobs arrive according to a Poisson process with rate λ. First of all, note that $\Phi_A(s) = \lambda/(\lambda - s)$, from which it readily follows

that, using Lemma 2.1, $\Psi(s) = \lambda(\Phi_B(s) - 1)$. This is not surprising since in this special case $A(s)$ is a Lévy process (even a compound Poisson process) which gives

$$\mathbb{E}\{\mathrm{e}^{sA(x)}\} = \mathrm{e}^{x\Psi(s)},$$

for *any* x. Assumption 3.1 simplifies in the Poisson case to the requirement that there exists a solution $\nu^* > 0$ to the equation $\lambda\Phi_B(\nu^*) = 1$, such that $\Phi_B(\nu)$ is finite for ν in a neighborhood of ν^*.
An explicit upper bound on $\mathbb{P}\{V > x\}$ can be given, uniformly in x.

Proposition 4.1 *Assume that Assumption 3.1 is satisfied. Then, for any* $x \geq 0$,

$$\mathbb{P}\{V > x\} \leq \frac{(1-\rho)\nu^*\Phi_B(\nu^*)}{\gamma}\,\mathrm{e}^{-\gamma x},$$

with $\gamma = \nu^* + \lambda - \lambda\Phi_B(\nu^*) > 0$. *In particular,*

$$\liminf_{x\to\infty} \frac{1}{x}\log\mathbb{P}\{V > x\} \leq -\gamma.$$

Note that ν^* is the maximizing argument of $s + \lambda - \lambda\Phi_B(s), s \geq 0$. It is easily verified that $\gamma > 0$ iff $\rho = \lambda\mathbb{E}\{B\} < 1$.

Proof. As in the previous section we obtain

$$\mathbb{P}\{V > x\} \leq \mathbb{E}\{\mathrm{e}^{sW}\}\mathbb{E}\{\mathrm{e}^{sB}\}\mathrm{e}^{-x(s+\lambda-\lambda\Phi_B(s))}. \tag{4.1}$$

From the Pollaczek-Khinchine formula, it readily follows that

$$\mathbb{E}\{\mathrm{e}^{\nu^*W}\} = \frac{(1-\rho)\nu^*}{\gamma} < \infty.$$

Combining this with (4.1) we obtain

$$\mathbb{P}\{V > x\} \leq \frac{(1-\rho)\nu^*\Phi_B(\nu^*)}{\gamma}\,\mathrm{e}^{-\gamma x}.$$

This immediately implies that the decay rate is bounded from above by $-\gamma$. □

The lower bound is also easier than in the case of $GI/G/1$ PS. One can work directly in continuous time, since the Poisson-arrival assumption enables us to use the Wald martingale associated with the compound Poisson process $A(x)$. Indeed, under $\mathbb{P}_{\nu_\epsilon}\{\cdot\}$, $A(x)$ is a compound Poisson process with arrival rate $\lambda_\epsilon = \lambda\Phi_B(\nu_\epsilon)$ and service times with moment generating function $\Phi_{B,\epsilon}(u) = \Phi_B(\nu_\epsilon + u)/\Phi_B(\nu_\epsilon)$. We have the following identity, see, e.g., Asmussen [3, Theorem XIII.3.2],

$$\mathbb{P}\{V > x\} = \mathbb{E}_{\nu_\epsilon}\{\mathrm{e}^{\Psi(\nu_\epsilon)x-\nu_\epsilon A(x)}I(V > x)\}. \tag{4.2}$$

This identity can be used as in the previous section to obtain an asymptotic lower bound for $\mathbb{P}\{V > x\}$.

4.2 A heavy traffic expansion of the decay rate

In this subsection we study the behavior of γ as $\rho \to 1$. A similar problem has been studied before by Abate & Whitt [1] (in the context of $M/G/1$ LCFS sojourn-time asymptotics).

Proposition 4.2 *Let $\gamma \equiv \gamma(\rho)$ be defined as in Theorem 3.1. Assume that $\mathbb{E}\{A^2\} < \infty$. Then, as $\rho \to 1$,*

$$\gamma(\rho) \sim -(1-\rho)^2 \frac{1}{2\beta}, \tag{4.3}$$

with

$$\beta = \frac{\mathbb{E}\{B^2\}}{\mathbb{E}\{A\}} - 2\mathbb{E}\{A\} + \frac{\mathbb{E}\{A^2\}}{\mathbb{E}\{A\}}.$$

Proof. From Proposition 3.1, we know that γ can be written as follows:

$$\gamma = \sup_\nu \left(\nu + \Phi_A^{-1}\left(\frac{1}{\Phi_B(\nu)} \right) \right) = \sup_\nu (\nu - (\rho\nu + \beta\nu^2) + \mathrm{o}(\nu^2)).$$

The latter representation (using a two-term Taylor expansion, the coefficients are derived below) is useful since the optimizing argument ν^* is converging to 0 as $\rho \to 1$. Using the strict concavity of $\nu + \Phi_A^{-1}(1/\Phi_B(\nu))$ one readily shows that

$$\nu^* \sim (1-\rho)/2\beta,$$

which is the optimizing argument of $\nu - (\rho\nu + \beta\nu^2)$. These considerations imply that

$$\gamma(\rho) \sim \nu^* - (\rho\nu^* + \beta(\nu^*)^2) + \mathrm{o}((\nu^*)^2),$$

from which the assertion follows. It remains to derive the coefficients α and β in the Taylor expansion given above. Clearly α is the mean arrival rate of traffic, i.e., $\rho = \mathbb{E}B/\mathbb{E}A$. Also,

$$\beta = \frac{\mathrm{d}^2}{\mathrm{d}\nu^2} \left(-\Phi_A^{-1}\left(\frac{1}{\Phi_B(\nu)} \right) \right) \Big|_{\nu=0} = \frac{\mathbb{E}(B^2)}{\mathbb{E}A} - 2\frac{(\mathbb{E}B)^2}{\mathbb{E}A} + \frac{\mathbb{E}(A^2)(\mathbb{E}B)^2}{(\mathbb{E}A)^3}.$$

[This is proven as follows. Define

$$\omega := \Phi_A^{-1}\left(\frac{1}{\Phi_B(\nu)} \right),$$

so that $\Phi_A(\omega)\Phi_B(\nu) = 1$. Differentiate this equation with respect to ω:

$$\Phi_A(\omega)\Phi_B'(\nu) + \Phi_A'(\omega)\Phi_B(\nu)\frac{\mathrm{d}\omega}{\mathrm{d}\nu} = 0;$$

from this $\mathrm{d}\omega/\mathrm{d}\nu$ can be solved. After inserting $\nu = \omega = 0$, it follows that $\alpha = -\mathrm{d}\omega/\mathrm{d}\nu = \mathbb{E}B/\mathbb{E}A$. A second differentiation yields:

$$\Phi_A(\omega)\Phi_B''(\nu) + 2\Phi_A'(\omega)\Phi_B'(\nu)\frac{\mathrm{d}\omega}{\mathrm{d}\nu} + \Phi_A''(\omega)\Phi_B(\nu)\left(\frac{\mathrm{d}\omega}{\mathrm{d}\nu}\right)^2 + \Phi_A'(\omega)\Phi_B(\nu)\frac{\mathrm{d}^2\omega}{\mathrm{d}\nu^2} = 0;$$

solving $\mathrm{d}^2\omega/\mathrm{d}\nu^2$ and inserting $\nu = \omega = 0$ yields the desired.] □

For the special $M/G/1$ case we recover the result $\gamma(\rho) \sim \frac{1}{2}(1-\rho)^2/\mathbb{E}(B^2)$ given in [1] for LCFS (note that in [1] the normalization $\mathbb{E}B = 1$ is used, such that heavy traffic corresponds to $\lambda \uparrow 1$). Interestingly, the asymptotics are of the order $(1-\rho)^2$, and hence intrinsically different from the $(1-\rho)$-behavior of FCFS.

4.3 Deterministic service times

Theorem 3.1 holds under Assumption 3.2, which rules out extremely light-tailed service-time distributions. In this section we show that imposing this assumption is not just a matter of mathematical convenience. By considering the asymptotic behavior of $\mathbb{P}\{V > x\}$ in the $M/D/1$ PS queue, we show that Assumption 3.2 is crucial.

First note that the decay rate in this case can be found by a careful inspection of the moment generating function of V that can be found in Ott [21]. However, we choose a more probabilistic approach. A simple and explicit upper bound for $\mathbb{P}\{V > x\}$ is derived, leading to decay rate $\hat{\gamma}$. The bound turns out to be sharp enough to imply that $\hat{\gamma} > \gamma$. For convenience, we take $D = 1$, which yields $\rho = \lambda$ and $\gamma = -\log\rho - (1-\rho)$.

As before, let Q_0 be the number of customers in the system upon arrival of the tagged customer; recall that $\mathbb{P}\{Q_0 = n\} = (1-\rho)\rho^n$. Further, let $\{Y_i(t), t \geq 0\}, i = 0, 1, \ldots$ be a collection of independent Yule processes with rate λ, that is, for each i, $Y_i(t)$ is a birth process with birth rate λn when $Y_i(t) = n$. Define $Z_i := \int_{t=0}^{1} Y_i(t)\mathrm{d}t$ and $Y_i := Y_i(1)$.

To obtain an upper bound for V, assume that all customers in the system at time 0 have a remaining service requirement equal to 1. By a standard time-change argument (see e.g. Section 7.3 in Robert [26]), we obtain $V \leq \sum_{i=0}^{Q_0} Z_i$, which (since $Z_i \leq Y_i$) implies

$$V \leq \sum_{i=0}^{Q_0} Y_i =: \bar{V}. \tag{4.1}$$

We now derive the tail behavior of $\bar{V}$. From Ross [27], p. 236, we have

$$\mathbb{P}\{Y_i = n\} = \mathrm{e}^{-\rho}(1-\mathrm{e}^{-\rho})^{n-1},$$

i.e. Y_i has a geometric distribution with 'success probability' $\mathrm{e}^{-\rho}$. Since Q_0+1 satisfies the same property (with success probability $1-\rho$), $\bar{V}$ is a geometric sum of i.i.d. geometrically distributed random variables. This implies that $\bar{V}$ itself is geometrically distributed with success probability $(1-\rho)\mathrm{e}^{-\rho}$, and hence,

$$\mathbb{P}\{V > x\} \leq \mathbb{P}\left\{\bar{V} > x\right\} = (1-(1-\rho)\mathrm{e}^{-\rho})^{[x]}. \tag{4.2}$$

We now compare this bound with the upper bound provided by Proposition 2.1:

$$\mathbb{P}\{V > x\} \leq \mathrm{e}^{-\gamma x(1+\mathrm{o}(1))} = (\rho\mathrm{e}^{(1-\rho)})^{x(1+\mathrm{o}(1))}. \tag{4.3}$$

Now note that $(1-(1-\rho)\mathrm{e}^{-\rho})$ is strictly smaller than $\rho\mathrm{e}^{(1-\rho)}$. This is because the function e^x is strictly convex, and hence $(1-\rho)+\rho\mathrm{e} > \mathrm{e}^{\rho}$, implying that

$$\rho\mathrm{e}^{1-\rho} > 1-(1-\rho)\mathrm{e}^{-\rho}.$$

The above computations indicate that in the $M/D/1$ PS queue, large sojourn times are also due to a large number of customers present at time 0. Notice that this is markedly different than the behavior observed in the previous sections, where large sojourn times were predominantly due to the work brought along by users that arrived after the tagged customer.

4.4 On Assumption 3.1

In this section we indicate how the statement of Theorem 3.1 changes if Assumption 3.1 is violated. For convenience, focus on the $M/G/1$ queue. Define $\bar{\nu} := \sup\{\nu : \Phi_B(\nu) < \infty\}$; suppose $\bar{\nu} > 0$. Noting that $\Phi_B(\nu)$ is strictly convex, it follows that Assumption 3.1 is equivalent to $\bar{\rho} = \lambda\mathbb{E}\{B\mathrm{e}^{\bar{\nu}B}\} > 1$. Below we show that if this inequality is *not* satisfied (i.e., $\bar{\rho} \leq 1$) the exponential decay rate is $\bar{\gamma} = \bar{\nu} + \lambda - \lambda\Phi_B(\bar{\nu}) > 0$.
For the upper bound, we get as in the proof of Proposition 4.1 that

$$\mathbb{P}\{V > x\} \leq \frac{(1-\rho)\bar{\nu}\Phi_B(\bar{\nu})}{\bar{\gamma}}\mathrm{e}^{-\bar{\gamma}x}.$$

For the lower bound, we provide two different arguments. The first argument proceeds as follows: consider the twisted probability measure $\bar{\mathbb{P}}\{\cdot\}$ which coincides with $\mathbb{P}_\nu\{\cdot\}, \nu = \bar{\nu}$. Under this twisted probability measure, we still have an $M/G/1$ PS queue, but now service times after time 0 have become heavy-tailed rather than light-tailed.
This is enough to ensure that $\bar{\mathbb{P}}\{V > x\}$ is heavy-tailed. For example, let $\mathcal{H}_n$ be the event that n customers of size at least x enter the system between time 0 and 1. Then we have

$$\begin{aligned}\bar{\mathbb{P}}\{V > x\} &\geq \bar{\mathbb{P}}\{\mathcal{H}_n\}\,\bar{\mathbb{P}}\{V > x \mid \mathcal{H}_n\} \\ &\geq \bar{\mathbb{P}}\{N(1) \geq n\}\ (\bar{\mathbb{P}}\{B > x\})^n\,\mathbb{P}\left\{B_0 > 1 + \frac{x}{n+1}\right\},\end{aligned}$$

since after time 1 the service rate is at most $1/(n+1)$. From the above inequality and the tail behavior of B_0 we conclude that the decay rate of $\bar{\mathbb{P}}\{V > x\}$ is a most $\bar{\nu}/(n+1)$. Since this is valid for any n, we obtain

$$\frac{1}{x}\log\bar{\mathbb{P}}\{V > x\} \to 0.$$

From this result it is then possible to prove that

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbb{P}\{V > x\} = \bar{\gamma}. \tag{4.4}$$

A second argument to reach this conclusion (which gives some different insights) is to replace the service times $B_i, i \geq 1$, by $B_i^m = \min(B_i, m)$ (we do not change B_0 in order not to violate Assumption 3.2). Denote the corresponding sojourn time by V_m. Clearly, $\mathbb{P}\{V > x\} \geq \mathbb{P}\{V_m > x\}$, and the logarithmic tail behavior of $\mathbb{P}\{V_m > x\}$ follows from Theorem 3.1: with obvious notation we can write

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbb{P}\{V_m > x\} = \inf_{s\geq 0}[\Psi_m(s) - s] \to \inf_{s\geq 0}[\Psi(s) - s],$$

as $m \to \infty$ which follows from monotone convergence ($\Psi_m(s)$ is monotone in m). The latter expression is equal to $\bar{\gamma}$.

5 Importance sampling

In this section, we apply the analysis of the previous sections to construct an importance sampling algorithm which enables us to efficiently simulate $\mathbb{P}\{V > x\}$ in the $M/G/1$ PS

queue. The efficiency gain with respect to direct simulation methods is considerable, particularly for large values of x. We will develop the algorithm in Subsection 5.1 and show an asymptotic optimality property. The algorithm is then illustrated in Subsection 5.2 by means of a numerical example.

5.1 Asymptotically optimal algorithm

Importance sampling is a variance reduction technique in which the simulation is done by using a distribution under which the rare event occurs relatively frequently. The simulation output is weighed by the so-called likelihood ratio, keeping track of the difference between the original and new measures, thus obtaining unbiased estimates.
Let the alternative measure be denoted by $\mathcal{Q}$. Under $\mathcal{Q}$, say, n runs are performed, yielding indicator variables $I_1, \ldots, I_n$ (1 if $\{V > x\}$, and 0 else) and likelihoods $L_1, \ldots, L_n$. It is easily proven that $\mathbb{E}_{\mathcal{Q}}\{LI\} = \mathbb{P}\{V > x\}$, such that $n^{-1}\sum_{i=1}^{n} L_i I_i$ is an unbiased estimator. Due to Jensen's inequality,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}_{\mathcal{Q}}\{(LI)^2\} \geq 2 \lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}_{\mathcal{Q}}\{LI\},$$

where the latter expression is obviously twice the decay rate of $\mathbb{P}\{V > x\}$, i.e., $-2(\omega^*+\nu^*)$. We call an alternative measure $\mathcal{Q}$ *asymptotically optimal* if this lower bound $-2(\omega^* + \nu^*)$ is attained, i.e.,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{E}_{\mathcal{Q}}\{(LI)^2\} = -2(\omega^* + \nu^*). \tag{5.1}$$

In our importance-sampling setup we extensively use that in $M/G/1$ PS the steady state distribution of the number of customers in the system (including the distribution of their residual amount of work) is known.
The simulation of the event $\{V > x\}$ involves the three elements that were also mentioned in the Introduction:

(i) The number of customers present at time 0, denoted by Q_0. The distribution of Q_0 is geometric: $\mathbb{P}\{Q_0 = n\} = (1-\rho)\rho^n$. The amount of work left for each of them has the usual 'residual life distribution': $\mathbb{P}\left\{\bar{B} < x\right\} = (\mathbb{E}\{B\})^{-1} \int_0^x (1 - F_B(y))dy$; these are independent, and, in addition, also independent of the value of Q_0.

(ii) Our tagged customer B_0, arriving at time 0, with distribution B.

(iii) All customers arriving in $(0, x]$. The number of them has a Poisson distribution with mean λx, and each of them is distributed according to B.

Obviously, the simulation can be stopped at time x.
Our proof of the decay rate of $\mathbb{P}\{V > x\}$ suggests that – in the interval $(0, x]$ – the service times should be twisted by ν^*, i.e., we do not simulate under the original density $f_B(\cdot)$, but rather under density $g_B(\cdot)$ given by

$$g_B(x) = f_B(x)\frac{\mathrm{e}^{\nu^* x}}{\Phi_B(\nu^*)}, \quad x \geq 0.$$

Also, it suggests that the arrival rate λ be replaced by $\lambda^* := \lambda - \omega^* = \lambda\Phi_B(\nu^*)$. As the proofs of the decay rate do not involve a twisting of the number of jobs initially present Q_0, the residual job sizes $\bar{B}$, nor the tagged job B_0, we twist them by some θ_{Q_0}, $\theta_{\bar{B}}$, and θ_{B_0}, respectively; below we will select appropriate values for these twists. Let, as before, W denote the workload at time 0.

The likelihood ratio is composed from contributions of the items (i), (ii), and (iii) above. First focus on (i): the customers already present at time 0. We again denote by $\bar{B}_i$ the residual service requirement of the ith customer, where $i = 1, \ldots, Q_0$; clearly $W = \sum_{i=1}^{Q_0} \bar{B}_i$. The likelihood induced by these customers is

$$\left(\mathrm{e}^{-\theta_{Q_0} Q_0}\mathbb{E}\{\mathrm{e}^{\theta_{Q_0}}\}^{Q_0}\right)\left(\mathrm{e}^{-\theta_{\bar{B}} W}\mathbb{E}\{\mathrm{e}^{\theta_{\bar{B}}\bar{B}}\}^{Q_0}\right). \tag{5.2}$$

Considering (ii), the contribution from the tagged customer is clearly

$$\mathrm{e}^{-\theta_{B_0} B_0}\Phi_B(\theta_{B_0}). \tag{5.3}$$

Finally, the contribution from (iii), i.e, the customers arriving in $(0, x]$, is

$$\left(\mathrm{e}^{(\lambda^*-\lambda)x}\left(\frac{\lambda}{\lambda^*}\right)^{N(x)}\right)\left(\mathrm{e}^{-\nu^* A(x)}\Phi_B(\nu^*)^{N(x)}\right) = \mathrm{e}^{-\omega^* x - \nu^* A(x)}, \tag{5.4}$$

where the first term in the left hand side is due to the arrivals (from a Poisson distribution with mean $\lambda^* t$ rather than λt), and the second term due to the amount of work brought along by these arrivals. The likelihood L of a simulation run is clearly the product of (5.2), (5.3), and (5.4).

As mentioned above, we still have the freedom to choose θ_{Q_0}, $\theta_{\bar{B}}$, θ_{B_0}. Suppose we set $\theta_{\bar{B}} = \theta_{B_0} = \nu^*$ and $\theta_{Q_0} = \log\mathbb{E}\{\mathrm{e}^{\nu^*\bar{B}}\}$. Noticing that, as before, $\Phi_B(\nu^*) < \infty$ and

$$\mathbb{E}\{\mathrm{e}^{\theta_{Q_0} Q_0}\} = \mathbb{E}\{\mathbb{E}\{\mathrm{e}^{\nu^*\bar{B}}\}^{Q_0}\} = \mathbb{E}\{\mathrm{e}^{\nu^* W}\} < \infty,$$

this implies that, for some finite κ, the likelihood L is majorized by $\kappa\,\mathrm{e}^{-\omega^* x - \nu^*(A(x)+B+W)}$. As an aside we mention that we easily retrieve the exponential upper bound on $\mathbb{P}\{V > x\}$:

$$\mathbb{P}\{V > x\} = \mathbb{E}_{\mathcal{Q}}\{LI(V > x)\} \le \mathbb{E}_{\mathcal{Q}}\{LI(A(x) + B + W > x)\} \le \kappa\,\mathrm{e}^{-\omega^* x - \nu^* x}.$$

However, we also get an upper bound on the second moment of the estimator:

$$\mathbb{E}_{\mathcal{Q}}\{L^2 I(V > x)\} \le \mathbb{E}_{\mathcal{Q}}\{L^2 I(A(x) + B + W > x)\} \le \kappa^2\,\mathrm{e}^{-2\omega^* x - 2\nu^* x},$$

such that the change of measure is asymptotically optimal, as desired, see (5.1).

Our proof of the logarithmic asymptotics suggests that the twist $\theta_{Q_0} = \theta_{\bar{B}} = \theta_B = 0$ might also work well. It turns out that this is not necessarily the case. Under this twist we have that $L = \exp(-\omega^* x - \nu^* A(x))$, and hence

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\{L^2 I(V > x)\} &\le \mathrm{e}^{-2\omega^* x - 2\nu^* A(x)}\mathbb{E}_{\mathcal{Q}}\{\mathrm{e}^{2\nu^*(B+W)}\} \\ &\overset{!}{=} \mathrm{e}^{-2\omega^* x - 2\nu^* x}\Phi_B(2\nu^*)\mathbb{E}\{\mathrm{e}^{2\nu^* W}\},\end{aligned}$$

but the last two factors (i.e., $\Phi_B(2\nu^*)$ and $\mathbb{E}\{\mathrm{e}^{2\nu^* W}\}$) are not necessarily finite. Consequently, we do not have straightforward bounds on the likelihood. For this reason it remains unclear whether this twist (suggested by the results in Section 3) would yield asymptotic optimality.

x	sim.	appr. (1.1)	appr. (5.5)
20	$4.3 \cdot 10^{-2}$	$3.4 \cdot 10^{5}$	$5.0 \cdot 10^{-2}$
40	$7.9 \cdot 10^{-3}$	$9.8 \cdot 10^{3}$	$1.1 \cdot 10^{-2}$
60	$2.0 \cdot 10^{-3}$	$8.1 \cdot 10^{2}$	$3.4 \cdot 10^{-3}$
80	$6.0 \cdot 10^{-4}$	$1.1 \cdot 10^{2}$	$1.2 \cdot 10^{-3}$
100	$2.2 \cdot 10^{-4}$	$2.0 \cdot 10^{1}$	$5.1 \cdot 10^{-4}$
120	$7.2 \cdot 10^{-5}$	$4.4 \cdot 10^{0}$	$2.3 \cdot 10^{-4}$

Table 1: Simulation results and approximations; $\rho = 0.8$.

5.2 Numerical analysis

In this section we demonstrate the importance sampling algorithm proposed in Section 5.1 assuming exponential service times.

We first focus on a case with relatively *high load*: $\rho = 0.8$. Normalize the service rate μ to 1 (and hence $\lambda = 0.8$). The first column of Table 1 shows, for various values of the threshold x, estimates based on importance sampling simulations. The simulation was repeated until the width of the confidence interval was below 5% of the estimated value; the simulation time needed on a Pentium PC ranged from a few milliseconds to a few seconds. We chose the threshold values $x = 20, 40, \ldots, 120$ to compare with the exact values in Figure 2.a of [12]. The second column displays the approximation based on the asymptotically exact expansion (1.1). In the regime with relatively high load, one could expect that heavy-traffic approximations could be accurate. The third column presents results based on the heavy-traffic expansion suggested in e.g. Zwart & Boxma [31], which indicates that

$$\mathbb{P}\{V > x\} \approx \mathbb{P}\{E\,B > (1-\rho)x\}\,, \tag{5.5}$$

with E distributed exponentially with mean 1, independently of B. We remark that for the regime of relatively high load an accurate approximation can be found in Guillemin & Boyer [12].

Strikingly, the exact expansion (1.1) performs badly for these parameters; further numerical experiments for larger x show that the convergence of $-\log \mathbb{P}\{V > x\}/x$ to its theoretical limit $(\sqrt{\lambda} - \sqrt{\mu})^2$ is extremely slow.

Table 2 concentrates on a *low-load* scenario: $\lambda = 0.3$, $\mu = 1$. For such load values the heavy-traffic approximation obviously does not apply. The first column is the estimate based on importance sampling, the second the asymptotic exact approximation, and the third the decay rate $-\log \mathbb{P}\{V > x\}/x$ (where $\mathbb{P}\{V > x\}$ was approximated by the simulated value).

For this low-load scenario, the asymptotic exact expansion (1.1) works relatively well, but should clearly be treated with care. Note that the convergence of the decay rate to its theoretical limiting value 0.205 is still rather slow. Even for the extremely small probabilities of Table 2, the importance sampling took just 5-10 minutes; again the simulation was

x	sim.	appr. (1.1)	decay rate
20	$8.3 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	0.47
40	$1.2 \cdot 10^{-7}$	$9.3 \cdot 10^{-8}$	0.40
60	$3.8 \cdot 10^{-10}$	$2.1 \cdot 10^{-10}$	0.36
80	$1.5 \cdot 10^{-12}$	$7.6 \cdot 10^{-13}$	0.34
100	$7.6 \cdot 10^{-15}$	$3.5 \cdot 10^{-15}$	0.32
120	$4.9 \cdot 10^{-17}$	$1.9 \cdot 10^{-17}$	0.31

Table 2: Simulation results and approximations; $\rho = 0.3$.

terminated when the width of the confidence interval was below 5% of the estimated value. We also estimated the probability of $\{V > 40\}$ by using direct Monte Carlo simulation; it took 14 hours to obtain an estimate with 5% precision. The use of importance sampling is emphasized by the observation that, extrapolating, it would take about a factor 10^{10} longer to estimate $\mathbb{P}\{V > 120\}$ by the direct method — here we use the rule of thumb that the number of runs required in the direct simulation is inversely proportional to the probability to be estimated.

6 Other service disciplines

In this section we investigate the impact of the service discipline on the decay rate of the sojourn-time distribution. In particular, we find the decay rate for various other service disciplines.

Note that PS minimizes the decay rate: It is shown in Section 3 that the crude upper bound $V_{\rm PS} \leq P_r$ (with P_r the time to empty the queue, i.e. the residual busy period) is attained, in the sense that both random variables have the same decay rate. As we will show in this section, this worst-case property is shared by many other service disciplines. We now give an overview of various service disciplines.

- *First-come first-served (FCFS).*
  First we consider the FCFS service discipline. This is, of course, well studied and the asymptotic behavior of $\log \mathbb{P}\{V_{\rm FCFS} > x\}$ is known under very general assumptions, see, e.g., Glynn & Whitt [11] and references therein. For the $GI/G/1$ queue, it can be shown that

  $$-\log \mathbb{P}\{V_{\rm FCFS} > x\} \sim \gamma_{\rm FCFS} x, \tag{6.1}$$

  with $\gamma_{\rm FCFS}$ the unique positive solution of the equation $\Phi_A(-s)\Phi_B(s) = 1$; for the situation in which this solution does not exist, see, e.g., Embrechts & Veraverbeke [9] and Korshunov [17]. Invoking a powerful result by Ramanan & Stolyar [25], it follows that FCFS is optimal among all non-anticipating work-conserving disciplines, in that it *minimizes the decay rate.* Furthermore, it is easy to see that $\gamma_{\rm FCFS} > \gamma_{\rm PS} = \gamma$ if

Assumptions 3.1 and 3.2 hold (where γ is defined in Section 3). From the result (see Section 3.1)

$$\mathbb{E}\{\mathrm{e}^{\nu^* V_{\mathrm{FCFS}}}\} = \mathbb{E}\{\mathrm{e}^{\nu^*(W+B)}\} < \infty,$$

it follows that $\gamma_{\mathrm{FCFS}} \geq \nu^*$. On the other hand we have $\gamma_{\mathrm{PS}} = \nu^* + \omega^* < \nu^* \leq \gamma_{\mathrm{FCFS}}$.

- *Last-come first-served (LCFS).*
  There are two types of LCFS disciplines, namely Preemptive Resume (PR) and Non-preemptive Resume (NPR). Exact asymptotics in the PR case has recently been obtained by Palmowski & Rolski [22]; their results imply that the logarithmic tail asymptotics are the same as those of PS, since the LCFS-PR sojourn-time distribution coincides with that of the busy-period distribution (see Remark 3 in Section 3.3). For earlier treatments of the $M/G/1$ LCFS-PR queue we refer to Abate & Whitt [2] and references therein. For the nonpreemptive case, note that $\mathbb{P}\{V_{\mathrm{LCFS-NPR}} > x\}$ can be lower bounded by

  $$\mathbb{P}\{W > 0\}\,\mathbb{P}\{N(B^r) \geq 1\}\,\mathbb{P}\{P > x\},$$

  which has the same decay rate as the upper bound $\mathbb{P}\{P^r > x\}$. Thus, the sojourn time distributions of the $GI/G/1$ LCFS-PR and LCFS-NPR queues both have the same decay rate.

- *Shortest remaining processing time (SRPT).*
  The shortest remaining processing time (see e.g. Schrage & Miller [28]) discipline has both a preemptive and non-preemptive variant. For both cases, it is easy to see that the decay rate of the sojourn time coincides with that of γ_{PS}. We give an outline of the argument for the case that the service time has an unbounded support. Note that, for every $y > 0$, we have

  $$\mathbb{P}\{V_{\mathrm{SRPT}} > x\} \geq \mathbb{P}\{B > y\}\,\mathbb{P}\{P(y) > x\},$$

  with $P(y)$ a busy period of a $GI/G/1$ queue in which customers only enter if their service time is less than y. This lower bound holds since all customers with service time less than y have priority over a customer with service time bigger than y. The decay rate $\gamma(y)$ of $P(y)$ can easily be shown to converge to γ_{PS} as $y \to \infty$. For a detailed proof in the case of Poisson arrivals, we refer to Mandjes & Nuijens [18], who consider the strongly related *Foreground-background* PS discipline.

- *Random Order of Service (ROS).*
  For exponential service times, the waiting-time distribution (when positive) can be identified with the sojourn-time distribution in the PS queue [5], which clearly indicates that also $\gamma_{\mathrm{ROS}} = \gamma_{\mathrm{PS}}$. A simple argument for general service times has escaped us.

It is interesting to compare the above findings with the situation of heavy-tailed service times, in particular regularly varying service times. For light tails we saw that FCFS was performing best, and PS (and some other disciplines) worst. In the case of heavy tails, however, the opposite applies. PS can be shown to be optimal in the sense that $\mathbb{P}\{V > x\}$ has the same tail behavior as $\mathbb{P}\{B > x\}$ (up to a constant), while FCFS has worst-case behavior ($x\mathbb{P}\{B > x\}$, up to a constant), see Borst *et al.* [6] for an overview.
Service disciplines which can be shown to perform bad in both the light-tailed and heavy-tailed case are ROS [7] and LCFS-NPR. It would be interesting to find a service discipline which is both optimal for light-tailed and heavy-tailed service times, although it could be the case that such a service discipline does not exist.
For any work-conserving light-tailed $GI/G/1$ queue, the decay rate of the sojourn time distribution is, whenever it exists, lower bounded by $\gamma_{\rm PS}$ and upper bounded by $\gamma_{\rm FCFS}$. In this section, we considered several other service discplines resulting in decay rates which equaled either $\gamma_{\rm PS}$ or $\gamma_{\rm FCFS}$; it would be interesting to find service disciplines having decay rates in between these two values.

References

[1] Abate, J., Whitt, W. (1994). A heavy-traffic expansion for the asymptotic decay rates of tail probabilities in multi-channel queues. *Operations Research Letters* **15**, 1994, 223–230.

[2] Abate, J., Whitt, W. (1997). Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. *Queueing Systems* **25**, 173–233.

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

[4] Bertoin, J., Doney, R. (1996). Some asymptotic results for transient random walks. *Advances in Applied Probability* **28**, 207–226.

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

[6] Borst, S.C., Boxma, O.J., Nunez-Queija, R., Zwart, A.P. (2003). The impact of the service discipline on delay asymptotics. *Performance Evaluation* **54**, 177–206.

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

[8] Coffman, jr., E.G., Muntz, R.R., Trotter, H. (1970). Waiting time distributions for processor-sharing systems. *Journal of the ACM* **17**, 123–130.

[9] Embrechts, P., Veraverbeke, N. (1982). Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics* **1**, 55–72.

[10] Flatto, L. (1997). The waiting time distribution for the random order service $M/M/1$ queue. *Annals of Applied Probability* **7**, 382–409.

[11] Glynn, P., Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* **31A**, 131–156.

[12] Guillemin, F., Boyer, J. (2002). Analysis of $M/M/1$ queue with processor sharing via spectral theory. *Queueing Systems* **39**, 377–397.

[13] Guillemin, F., Robert, P., Zwart, A.P. (2004). Asymptotic results for processor sharing queues. *Advances in Applied Probability*, to appear.

[14] Gromoll, C. (2004). Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability* **14**, 555–611.

[15] Jean-Marie, A., Robert, P. (1994). On the transient behavior of the processor sharing queue. *Queueing Systems* **17**, 129–136.

[16] Jelenković, P., Momčilović, P. (2003). Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operations Research* **28**, 587–608.

[17] Korshunov, D.A. (1997). On distribution tail of the maximum of a random walk. *Stochastic Processes and their Applications* **72**, 97–103.

[18] Mandjes, M., Nuijens, M. (2004). Sojourn times in the $M/G/1$ FB queue with light-tailed service times. Submitted for publication.

[19] Morrison, J.A. (1985). Response-time distribution for a processor-sharing system. *SIAM Journal of Applied Mathematics* **45**, 152–167.

[20] Núñez-Queija, R. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research* **113**, 101–117.

[21] Ott, T. (1984). The sojourn-time distribution in the $M/G/1$ queue with processor sharing. *Journal of Applied Probability* **21**, 360–378.

[22] Palmowksi, Z., Rolski, T. (2004). Markov Processes conditioned to never exit a subspace of the state space with application to the single server queue. Submitted for publication, available at `http://www.math.uni.wroc.pl/~zpalma/publication.html`

[23] Pollaczek, F. (1946). La loi d'attente des appels téléphoniques. *Comptes Rendus Acad. Sci. Paris* **222**, 353–355.

[24] Puha, A., Stolyar, A., Williams, R. (2004). The fluid limit of an overloaded processor sharing queue. Preprint.

[25] Ramanan, K., Stolyar, A. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability* **11**, 1–48.

[26] Robert, P. (2003). *Stochastic Networks and Queues - A probabilistic approach*, Springer, New York.

[27] Ross, S. (1996). *Stochastic Processes*. Wiley, New York.

[28] Schrage, L.E., Miller, L.W. (1966). The queue $M/G/1$ with the shortest remaining processing time discipline. *Operations Research* **14**, 670–684.

[29] Whitt, W. (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems* **2**, 71–107.

[30] Wierman, A., Harchol-Balter, M. (2003). Classifying scheduling policies with respect to unfairness in an $M/GI/1$. *Proceedings of ACM Sigmetrics*, San Diego, CA.

[31] Zwart, A.P., Boxma, O.J. (1999). Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* **35** 141–166.