Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Tail behavior of conditional sojourn times in
Processor-Sharing queues

R. Egorova, B. Zwart

REPORT PNA-R0607 APRIL 2006

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation 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)

Tail behavior of conditional sojourn times in Processor-Sharing queues

ABSTRACT

We investigate the tail behavior of the sojourn-time distribution for a request of a given length in an M/G/1 Processor-Sharing (PS) queue. An exponential asymptote is proven for general service times in two special cases: when the traffic load is sufficiently high and when the request length is sufficiently small. Furthermore, using the branching process technique we derive exact asymptotics of exponential type for the sojourn time in the M/M/1 queue. We obtain an equation for the asymptotic decay rate and an exact expression for the asymptotic constant. The decay rate is studied in detail and is compared to other service disciplines. Finally, using numerical methods, we investigate the accuracy of the exponential asymptote.

Tail behavior of conditional sojourn times in Processor-Sharing queues

Regina Egorova, Bert Zwart

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

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

E-mail: r.egorova@cwi.nl, zwart@win.tue.nl

Abstract

We investigate the tail behavior of the sojourn-time distribution for a request of a given length in an M/G/1 Processor-Sharing (PS) queue. An exponential asymptote is proven for general service times in two special cases: when the traffic load is sufficiently high and when the request length is sufficiently small. Furthermore, using the branching process technique we derive exact asymptotics of exponential type for the sojourn time in the M/M/1 queue. We obtain an equation for the asymptotic decay rate and an exact expression for the asymptotic constant. The decay rate is studied in detail and is compared to other service disciplines. Finally, using numerical methods, we investigate the accuracy of the exponential asymptote.

1 Introduction

The sojourn time of a customer, i.e. the time a customer spends in the system from its arrival until its service completion, is an important performance measure for queueing systems. In this paper, we investigate the tail behavior of the sojourn time distribution for a request of a given length in the stable M/G/1 Processor-Sharing (PS) queue. In the PS discipline all customers are served simultaneously: each customer in the system receives service at rate $1/N$, when the total number of customers present in the system is N. Queues with PS discipline became popular by the work of Kleinrock and were originally proposed as an idealization of time-sharing systems. The recent rise of interest in PS queues is related to their application in the performance analysis of bandwidth-sharing protocols in computer communication networks.
There exist a number of results on the complete distribution of the sojourn time in PS queues. Yashkov [22] found an analytic expression for the distribution function in terms of a double Laplace-Stieltjes transform (LST) based on the decomposition of the sojourn time into a set of independent branching processes. Schassberger [20] developed another

approach to derive the LST by considering PS as a limiting case of the round-robin discipline. Using methods similar to Yashkov's, the LST of the conditional sojourn time was also studied by Grishechkin [12], Ott [19] and Nunez-Queija [17]. Zwart and Boxma [24] derived a new, more explicit expression for the LST involving a series expansion.

The complexity of these results, and the renewed attention for the PS discipline as a flow-level model for the Internet, have led to an interest in the tail behavior of the sojourn time distribution. Although obtaining the tail behavior seems a more modest goal than obtaining the complete distribution, this task has still proven to be quite challenging and has recently been the subject of many papers. Several studies have focused on the analysis of the tail of the unconditional sojourn time distribution in the case when the service time distribution is heavy-tailed. A so-called reduced-load approximation was proven by Zwart and Boxma in [24] for the M/G/1 queue, which was extended by Núñez-Queija in [17], Jelenković and Momčilović [13]; see Borst *et al.* [5] for a survey.

For PS queues with light-tailed service time distributions only a few results are available. The tail asymptotics for the unconditional sojourn time in the M/M/1 PS queue are known, and are of quite remarkable form [4], [11]:

$$\mathbf{P}(V > x) \sim cx^{-5/6}\mathrm{e}^{-\alpha x^{1/3}}\mathrm{e}^{-\gamma_0 x}, \qquad x \to \infty, \tag{1.1}$$

for positive constants c, α, γ_0, and $f(x) \sim g(x)$ denoting $f(x)/g(x) \to 1$. Flatto [11] obtained this asymptotic tail behavior of the waiting time in the M/M/1 Random-Order-of-Service (ROS) queue. Subsequently, Borst *et al.* [4] showed that the waiting-time distribution in the M/M/1 ROS queue, conditioned to be positive, equals the sojourn time distribution in the M/M/1 PS queue.

Mandjes and Zwart [16] analyzed sojourn time asymptotics in the GI/GI/1 PS queue. Using large-deviation techniques, they derived logarithmic asymptotics for a broad class of light-tailed distributions. Recently, the exact asymptotics for the sojourn time in the M/D/1 PS queue were derived in [10].

The complexity of the asymptotics in the M/M/1 queue and the difficulty of extending Flatto's method to more general queues ([11], [4], [10]) motivate us to study the light-tailed case from a different perspective: in this paper we investigate the asymptotic behavior of the sojourn time distribution for a request of given length. Suppose the customer under consideration (the tagged customer) has a request of length τ (abbreviated as τ-request). Let $V(\tau)$ be its sojourn time. In order to emphasize this conditioning we will use the notation M/G(τ)/1 for the underlying queue, although we stress that all other customers still have generally distributed service times.

The analysis in this paper is based on two key ideas. The first cornerstone is the branching method introduced by Yashkov [22]. This approach enables us to represent the sojourn time in terms of a geometric random sum of "delay elements" and apply existing powerful asymptotic results for such random sums, which is the second cornerstone. Checking the assumptions under which these asymptotic results are valid, is still a challenging task, in particular for generally distributed service times. Assuming that either the traffic load is close to one, or that the request length is sufficiently small, we show in Section 2 that the asymptote

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{1.2}$$

is valid for general service time distributions.

Sections 3 and 4 are dedicated to the study of the system with exponential service time, for which no further assumptions are necessary. We obtain an equation for the asymptotic decay rate $\gamma(\tau)$ and an exact (though complicated) expression for the asymptotic constant

2

$\alpha(\tau)$. The decay rate equations are shown to be of quite an unusual form. The decay rate $\gamma(\tau)$ is the solution (in s) of

$$\tan\left(\frac{\tau}{2}\sqrt{-(\lambda+\mu-s)^2+4\lambda\mu}\right)=\frac{\sqrt{-(\lambda+\mu-s)^2+4\lambda\mu}}{\lambda-\mu+s\frac{1+\rho}{1-\rho}},\qquad \tau>\tau_0,\qquad(1.3)$$

$$\tanh\left(\frac{\tau}{2}\sqrt{(\lambda+\mu-s)^2-4\lambda\mu}\right)=\frac{\sqrt{(\lambda+\mu-s)^2-4\lambda\mu}}{\lambda-\mu+s\frac{1+\rho}{1-\rho}},\qquad \tau<\tau_0,\qquad(1.4)$$

where $\tau_0=\frac{1}{\sqrt{\lambda\mu}}\left(\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}\right)$.

In Section 3, we derive expressions for the delay elements of the sojourn time and in Section 4 we formulate the main asymptotic result for the M/M(τ)/1 queue. We show that the exponential asymptote (1.2) is valid if $\tau\neq\tau_0$ and is *not* valid if $\tau=\tau_0$.

Finally, in Sections 5 and 6 we give some numerical results. First, we analyze the behavior of the decay rate depending on the value of τ and compare it with decay rates for an M/M(τ)/1 system with a different service discipline such as Shortest Remaining Processing Time (SRPT), Foreground-Background (FB), First In First Out (FIFO) and Last In First Out (LIFO). In Section 6, we compare the asymptotic result to exact values of $\mathbf{P}(V(\tau)>x)$, obtained by numerical Laplace transform inversion. We also compare the accuracy of the asymptote and the heavy-traffic approximation. The results show that the exponential asymptote provides a good approximation to the tail probability.

2   General results

In this section we present some results for the sojourn time in a system with a general service time distribution. Under the condition that the traffic load is sufficiently high, we prove that the sojourn time tail behaves asymptotically as an exponential function. We also consider the situation when the service requirement of the given customer is close to zero.

In the following sections we will follow the approach presented in Yashkov [22], in which the general expression for the LST of the sojourn time of a τ-request in the M/G/1 queue is derived. The key idea is the decomposition of the sojourn time into a sum of certain functionals of independent branching processes. This approach enables one to reduce the problem to the computation of certain functionals of branching processes.

Customers arrive into the system according to a Poisson process with rate λ. Denote by B the generic service time. We assume that the queue is stable, i.e. that the traffic load in the system is less than one, $\rho=\lambda\mathbf{E}B<1$.

Suppose that a tagged customer with service request of length τ arrives at the epoch $t=0$. Every customer present in the system at the arrival of the tagged customer is called a *progenitor* while the new arrivals occurring after $t=0$ are assumed to be *descendants* of these progenitors. We take into account only those customers which arrive before the service of the tagged customer is completed. If n progenitors are present in the system then each new arrival is declared with probability $1/n$ to be a descendant of any of these progenitors. The tagged customer is also considered as a progenitor. Each branching process is formed by one progenitor and its descendants (for more details see [22]). The sojourn time of the tagged customer can be represented as

$$V(\tau)=V_0(\tau)+\sum_{i=1}^{Q}C_i(\tau),\qquad(2.1)$$

3

where $C_i(\tau)$ is the amount of service received by a certain progenitor and its direct descendants during the sojourn time of the tagged customer, $V_0(\tau)$ is equal to the amount of service received by the tagged customer and its direct descendants, and Q is the number of customers in the system. Notice that the random variable $V_0(\tau)$ is in fact the sojourn time of a customer which arrives into an empty system. In general we will call the variables $V_0(\tau)$ and $C_i(\tau)$ the *delay elements*. The essential observation here is that the elements $V_0(\tau)$ and $C_i(\tau)$, $i = 1, 2, ..., Q$, are independent of each other. This is due to the fact that the sizes of requests and arrivals are independent. The elements $C_i(\tau)$ are also identical in distribution.

For convenience, denote the sum $\sum_{i=1}^{Q} C_i(\tau)$ by $V_1(\tau)$. Using the well-known fact that $\mathbf{P}(Q = n) = (1 - \rho)\rho^n$, the probability distribution of V_1 can be written as

$$\mathbf{P}(V_1(\tau) > x) = \mathbf{P}(\sum_{i=0}^{Q} C_i(\tau) > x) = \sum_{n=0}^{\infty}(1 - \rho)\rho^n(1 - F_n(x)), \tag{2.2}$$

where F denotes the distribution of $C_i(\tau)$, and $F_n(x)$ is the n-fold convolution of F with itself. The random variable $V_1(\tau)$ is called a geometric random sum and such random sums arise in many applied probability settings. From the results in [14], it is well-known that if the Cramér condition holds, such a sum is asymptotically (as $x \to \infty$) equivalent to an exponential function. In particular, in relation to the sojourn time in the $\mathrm{M/G}(\tau)/1$ system, the following theorem holds.

Theorem 2.1 *Let the Cramér condition hold, i.e. suppose that there exists a $\gamma = \gamma(\tau) > 0$ such that*

$$\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}] = \frac{1}{\rho}. \tag{2.3}$$

(i) If $h(\tau) = \rho \int_0^{\infty} x\mathrm{e}^{\gamma(\tau)x}dF(x) = \rho\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=\gamma(\tau)} < \infty$, and $\mathbf{P}(B > \tau) > 0$, then the asymptotic relation

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{2.4}$$

holds with

$$\alpha(\tau) = \frac{1 - \rho}{h(\tau)\gamma(\tau)}\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]. \tag{2.5}$$

(ii) If $h(\tau) = \infty$, then

$$\lim_{x \to \infty} \mathbf{P}(V(\tau) > x)\mathrm{e}^{\gamma(\tau)x} = 0.$$

Proof.
The statement of the theorem follows from two known results. Since the sojourn time $V(\tau)$ can be represented as a sum of the r.v. $V_0(\tau)$ and the geometric random sum $V_1(\tau)$, we can apply the result of Kalashnikov and Tsitsiashvili [14] to $V_1(\tau)$. These authors have shown that if the Cramér condition holds, $h(\tau) < \infty$ and F is non-lattice, the sum $V_1(\tau)$ asymptotically behaves as

$$\mathbf{P}(V_1(\tau) > x) \sim k_C(\tau)\mathrm{e}^{-\gamma(\tau)x}, \tag{2.6}$$

where $k_C(\tau) = (1 - \rho)/(g(\tau)\,\gamma(\tau))$, and if $g(\tau) = \infty$,

$$\lim_{x \to \infty} \mathbf{P}(V_1(\tau) > x)\mathrm{e}^{\gamma(\tau)x} = 0. \tag{2.7}$$

The distribution function F of the delay element $C_i(\tau)$ is indeed non-lattice, since $\mathbf{P}(C_i(\tau) = B_i^r) > 0$, and the residual service time B_i^r has a density. The condition $\mathbf{P}(B > \tau) > 0$ implies that $\mathbf{P}(B^r > \tau) > 0$. Since we consider all elements only on the interval $[0, t]$, the elements $C_i(\tau)$ and $V_0(\tau)$ coincide (in distribution) if $B^r > \tau$, and

$$\infty > \frac{\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}]}{\mathbf{P}(B^r > \tau)} \geq \frac{\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}\mathbf{1}(B^r > \tau)]}{\mathbf{P}(B^r > \tau)} = \mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}|B^r > \tau] = \mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}].$$

Applying Breiman's theorem [6] under the weaker condition $\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}] < \infty$ (see [9]), we obtain that

$$\mathbf{P}(V(\tau) > x) = \mathbf{P}(\mathrm{e}^{V_0(\tau)}\mathrm{e}^{V_1(\tau)} > \mathrm{e}^x) \sim \mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]\mathbf{P}(V_1(\tau) > x), \qquad x \to \infty. \quad (2.8)$$

Substitution of (2.6) in (2.8) implies (2.4). Part (ii) of the theorem follows from (2.7) and the finiteness of $\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]$. $\qquad\qquad\square$

The above theorem provides an explicit expression for the tail behavior of the sojourn time. However, to verify the conditions of the approximation for the system with general service time appears to be a challenging task. In the following proposition, we prove the Cramér condition for the case when the traffic intensity is sufficiently large.

Proposition 2.1 *For any value of τ there exists a $\rho(\tau) < 1$ such that for all $\rho > \rho(\tau)$, there exists a solution $\gamma(\tau)$ of Equation (2.3) with $h(\tau) < \infty$.*

Proof. Due to convexity of the moment generating function (MGF), it suffices to show that for any fixed value of τ there exists a sufficiently large $\rho < 1$ such that there exists an $\bar{s}$ such that $\frac{1}{\rho} < \mathbf{E}[\mathrm{e}^{\bar{s}C_i(\tau)}] < \infty$. Observe that $C_i(\tau)$ is not greater than the busy period P_τ in a system with services defined as $\min(B, \tau)$ given that the first customer in the busy period has service request of length τ. Therefore, for the MGF's, the inequality holds: $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] \leq \mathbf{E}[\mathrm{e}^{sP_\tau}]$. Due to Theorem 7.1 in [1] it follows that P_τ has a decay rate $\hat{s}(\tau)$, defined as a solution of the equation $\lambda(d/ds)(\mathbf{E}[\mathrm{e}^{s\min(B,\tau)}]) = 1$, and since $\mathbf{P}(P_\tau > x) \sim const \cdot x^{-3/2}e^{-\hat{s}(\tau)x}$, we deduce $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)P_\tau}] < \infty$. Hence, $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i(\tau)}] < \infty$. Obviously, the decay rate is dependent on the value of ρ, or, having the service time fixed, on the arrival rate λ, $\hat{s}(\tau, \lambda)$.
To bound the MGF of $C_i(\tau)$ from below, notice that for any τ, $C_i(\tau) \geq \min(B^r, \tau)$, where B^r is the residual service time. Hence, $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)C_i(\tau)}] \geq \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)\min(B^r,\tau)}]$. If $\mathbf{P}(B > \tau) > 0$, then $\mathbf{E}[\min(B, \tau)] < \mathbf{E}B$ and the modified queue is still stable. Hence $\hat{s}(\tau, \frac{1}{\mathbf{E}B}) > 0$, and

$$\lim_{\lambda \to 1/\mathbf{E}B} \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)C_i(\tau)}] \geq \lim_{\lambda \to 1/\mathbf{E}B} \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)\min(B^r,\tau)}]$$

$$= \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,1/\mathbf{E}B)\min(B^r,\tau)}] > 1.$$

Thus, choosing $\rho > \frac{1}{\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\frac{1}{\mathbf{E}B})\min(B^r,\tau)}]}$, we can find the solution for Equation (2.3). $\qquad\square$

The straightforward consequence of this proposition is the following.

Theorem 2.2 *For any value of τ there exists a $\rho(\tau)$ such that for all $\rho > \rho(\tau)$ we have*

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{2.9}$$

where $\gamma(\tau)$ is the solution of Equation (2.3) and the constant $\alpha(\tau)$ is given by (2.5).

Using a similar approach, we can prove exponential asymptotics for the sojourn time of a customer with a very small service request.

Theorem 2.3 *For sufficiently small values of τ,*

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{2.10}$$

where $\gamma(\tau)$ is a solution of Equation (2.3) and the constant $\alpha(\tau)$ is given by (2.5).

Proof. The elements $C_i(\tau)$ can be bounded from above by the delay element $C_D(\tau)$ in the M/D/1 system with service requests of size τ. The results in [10] on the decay rate in the M/D/1 queue imply that there exists an $\hat{s}(\tau) > 0$ such that $\mathbf{E}\mathrm{e}^{\hat{s}(\tau)C_D(\tau)} = 1/\rho_D = 1/(\lambda\tau)$. Further, the same argument as in the proof of Proposition 2.1 is applicable. However in this case λ, $\mathbf{E}B$ and ρ are fixed and parameter τ is varying: $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i(\tau)}] > \mathbf{E}[\mathrm{e}^{\hat{s}(\tau)\min(B^r,\tau)}] > \mathrm{e}^{\hat{s}(\tau)\tau}\mathbf{P}(B^r > \tau)$. The equation for the decay rate $\hat{s}(\tau)$ (see Formula (3.2) in [10]) is

$$\frac{\lambda\tau(\lambda - s) + s - s\mathrm{e}^{(\lambda-s)\tau}}{(\lambda - s)(\lambda - s\mathrm{e}^{(\lambda-s)\tau})} = \frac{1}{\lambda}.$$

Taking $s = c\tau$ and letting $\tau \downarrow 0$ we see that $\liminf_{\tau\downarrow 0}\hat{s}(\tau)\tau \geq c$ for any c, and consequently, $\lim_{\tau\downarrow 0}\hat{s}(\tau)\tau = \infty$. Hence, the decay rate $\hat{s}(\tau)$ is increasing faster than linear in $1/\tau$ when τ becomes small. Thus, we can conclude that for any $\rho \in (0,1)$ there exists a τ_0 such that $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i(\tau)}] > 1/\rho$ holds for all $\tau < \tau_0$. $\qquad\square$

In the following sections, we focus on the behavior of the sojourn time in the M/M(τ)/1 queue.

3 The delay elements for exponential service times

The goal of this section is to derive the LST of the delay elements in the M/M(τ)/1 queue using the approach presented in Yashkov [22], in which the general expression for the LST of the sojourn time of a τ-request in the M/G/1 queue is derived. The LST of the sojourn time itself is a question of less importance for our tail behavior investigation; it has been derived in Coffman *et al.* [7].
Define $\varphi(s,\tau) = \mathbf{E}[\exp(-sC_i(\tau))]$ and $\delta(s,\tau) = \mathbf{E}[\exp(-sV_0(\tau))]$ the LST's of the random variables $C_i(\tau)$ and $V_0(\tau)$, respectively.

Theorem 3.1 *The delay elements of the sojourn time in the M/M/1 PS queue have LST's given by the expressions:*

$$\delta(s,\tau) = \frac{2g(s)\mathrm{e}^{-(\lambda+s-\mu)\frac{\tau}{2}}}{(\mu - \lambda + s)\left(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}\right) + g(s)\left(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)}\right)} \tag{3.1}$$

and

$$\varphi(s,\tau) = \frac{(\mu - \lambda - s)(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}) + g(s)(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)})}{(\mu - \lambda + s)(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}) + g(s)(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)})}, \tag{3.2}$$

where $g(s) = \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}$.

Proof. In order to derive the LST's of the delay elements we follow Yashkov [22]. Under the condition that the number of customers in the system upon arrival of the tagged customer is n and the remaining service of the ith progenitor at the epoch $t = 0$ is x_i, the sojourn time V can be represented as

6

$$V(\tau) = V_0(\tau) + \sum_{i=1}^{n} C_i(x_i, \tau). \tag{3.3}$$

Since the random variables $V_0(\tau)$ and $C_i(x_i, \tau)$ are independent, we can write

$$\mathbf{E}[\mathrm{e}^{-sV(\tau)}|n, x_1, ..., x_n] = \delta(s, \tau) \prod_{i=1}^{n} \varphi(s, x_i, \tau). \tag{3.4}$$

Unconditioning, we obtain that the LST of the sojourn time is

$$v(s, \tau) = (1 - \rho)\delta(s, \tau) \left[1 - \rho \int_{x=0}^{\infty} \varphi(s, x, \tau) \frac{(1 - B(x))}{\mathbf{E}B} dx \right]^{-1}$$

$$= (1 - \rho) \frac{\delta(s, \tau)}{1 - \rho\varphi(s, \tau)}, \tag{3.5}$$

where $\varphi(s, \tau)$ is the LST of the delay element $C_i(\tau)$ being the amount of service received by the ith progenitor and its direct descendants for the time interval during which the tagged customer is served until completion.

We now proceed to derive the expressions for $\delta(s, \tau)$ and $\varphi(s, \tau)$. Due to Formulas (3.9) and (3.14) in [22] we have

$$\varphi(s, x, \tau) = \begin{cases} \delta(s, \tau)/\delta(s, \tau - x), & x < \tau, \\ \delta(s, \tau), & x \geq \tau. \end{cases} \tag{3.6}$$

Using Formula (3.16) of [22] we obtain that

$$\delta(s, \tau) = \mathrm{e}^{-(s+\lambda)\tau}\psi(s, \tau)^{-1}, \tag{3.7}$$

where the LST $\tilde{\psi}(q, s)$ of the function $\psi(s, \tau)$ given by

$$\tilde{\psi}(q, s) = \int \mathrm{e}^{-q\tau}\psi(s, \tau)d\tau, \tag{3.8}$$

is a solution of the following equation (see Formulas (3.18)-(3.19) in [22])

$$q\tilde{\psi}(q, s) - 1 + \lambda\tilde{\psi}(q, s)\beta(q + s + \lambda) + \frac{\lambda(1 - \beta(q + s + \lambda))}{q + s + \lambda} = 0. \tag{3.9}$$

Substituting the LST of the service time $\beta(s) = \frac{\mu}{\mu+s}$ we obtain

$$\tilde{\psi}(q, s) = \frac{q + s + \mu}{q^2 + (\mu + \lambda + s)q + \lambda\mu}. \tag{3.10}$$

To derive an expression for $\psi(s, \tau)$ we must invert the LST $\tilde{\psi}(q, s)$ with respect to q. This can be easily done using partial-fraction decomposition of the latter expression. That will lead us to the LST of a sum of two exponential functions. As a result we get

$$\psi(s, \tau) = \frac{A\mathrm{e}^{-B\tau} + C\mathrm{e}^{-D\tau}}{g(s)}, \tag{3.11}$$

where $A = (\mu + s - \lambda + g(s))/2$, $B = (\mu + s + \lambda - g(s))/2$, $C = (-\mu - s + \lambda + g(s))/2$, $D = (\mu + s + \lambda + g(s))/2$, and $g(s) = \sqrt{(\mu + \lambda + s)^2 - 4\lambda\mu}$.

7

Knowing $\psi(s,\tau)$ we can determine the LSTs $\delta(s,\tau)$ and $\varphi(s,x,\tau)$:

$$\delta(s,\tau) = \mathrm{e}^{-(s+\lambda)\tau}\frac{g(s)}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}, \tag{3.12}$$

$$\varphi(s,x,\tau) = \begin{cases} \mathrm{e}^{-(s+\lambda)x}\frac{A\mathrm{e}^{-B(\tau-x)}+C\mathrm{e}^{-D(\tau-x)}}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}, & x < \tau, \\ \mathrm{e}^{-(s+\lambda)\tau}\frac{g(s)}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}, & x \geq \tau. \end{cases} \tag{3.13}$$

Expression (3.1) for the LST $\delta(s,\tau)$ follows in a straightforward manner. In order to derive the LST $\varphi(s,\tau)$ of the delay element $C_i(\tau)$, we integrate with respect to the residual service time x. After some simplifications we obtain Formula (3.2). □

In order to investigate the sojourn time tail behavior we will need the MGF's of the delay elements rather than the LST's. The results of the previous section yield that the MGF of the delay element $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ is

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}-\mathrm{e}^{-\frac{1}{2}\tau f(s)})+f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}+\mathrm{e}^{-\frac{1}{2}\tau f(s)})}{(\mu-\lambda-s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}-\mathrm{e}^{-\frac{1}{2}\tau f(s)})+f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}+\mathrm{e}^{-\frac{1}{2}\tau f(s)})}, \tag{3.14}$$

where $f(s) = g(-s)$ (Theorem 3.1),

$$f(s) = \sqrt{(\mu+\lambda-s)^2 - 4\lambda\mu}.$$

Let us study the function $f(s)$ in more detail.
The expression under the square root is a quadratic function with zeros at $s_l = \lambda + \mu - 2\sqrt{\lambda\mu} \equiv \mu(1-\sqrt{\rho})^2$ and $s_r = \lambda + \mu + 2\sqrt{\lambda\mu} \equiv \mu(1+\sqrt{\rho})^2$. The function is negative on the interval

$$s \in (\lambda + \mu - 2\sqrt{\lambda\mu}, \lambda + \mu + 2\sqrt{\lambda\mu})$$

and positive otherwise.
Taking into account the fact that the function $f(s)$ is purely imaginary inside the interval $[s_l, s_r]$, we can rewrite the MGF in two forms depending on the sign of the radicand.

Corollary 3.1

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)\sin[\frac{1}{2}\tau f_2(s)]+f_2(s)\cos[\frac{1}{2}\tau f_2(s)]}{(\mu-\lambda-s)\sin[\frac{1}{2}\tau f_2(s)]+f_2(s)\cos[\frac{1}{2}\tau f_2(s)]} \quad \text{if } s \in [s_l, s_r], \tag{3.15}$$

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)\sinh[\frac{1}{2}\tau f_1(s)]+f_1(s)\cosh[\frac{1}{2}\tau f_1(s)]}{(\mu-\lambda-s)\sinh[\frac{1}{2}\tau f_1(s)]+f_1(s)\cosh[\frac{1}{2}\tau f_1(s)]} \quad \text{otherwise}, \tag{3.16}$$

where $f_1(s) = \sqrt{(\mu+\lambda-s)^2 - 4\lambda\mu}$ *and* $f_2(s) = \sqrt{-(\mu+\lambda-s)^2 + 4\lambda\mu}$.

4   Tail behavior in the M/M(τ)/1 queue

In this section we present our main result.

Theorem 4.1 *Define $\tau_0 = \frac{1}{\sqrt{\lambda\mu}}\left(\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}\right)$.*

(i) For all $\tau \neq \tau_0$,

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{4.1}$$

where $\gamma(\tau) > 0$ is the solution of Equation (2.3) and

$$\alpha(\tau) = \frac{2(1-\rho)}{\gamma(\tau)}\frac{[(\lambda+\mu-\gamma(\tau))^2 - 4\lambda\mu]\mathrm{e}^{-(-\gamma(\tau)+\lambda-\mu)\frac{\tau}{2}}}{K}, \tag{4.2}$$

with

$$K = (1+\rho)\left[f(\gamma(\tau))(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} - \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}}) + \gamma(\tau)\frac{\tau}{2}(\lambda+\mu-\gamma(\tau))(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} + \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})\right]$$

$$-(1-\rho)(\lambda+\mu-\gamma(\tau))\left[(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} + \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})(1 + \frac{(\mu-\lambda)\tau}{2}) + (\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} - \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})f(\gamma(\tau))\frac{\tau}{2}\right]$$

and $f(s) = \sqrt{(\mu+\lambda-s)^2 - 4\lambda\mu}$.

(ii) If $\tau = \tau_0$, then $\gamma(\tau) = (\sqrt{\mu} + \sqrt{\lambda})^2$ solves Equation (2.3) and

$$\lim_{x\to\infty} \mathbf{P}(V(\tau) > x)\mathrm{e}^{\gamma(\tau)x} = 0.$$

If the conditions stated in Theorem 2.1 hold in the case of exponential service times, the statement of the above theorem follows almost immediately. We will now show that the Cramér condition indeed holds, i.e. that there exists a positive solution of the equation $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{1}{\rho}$.

Let us first determine some useful thresholds that will play an essential role in our proof.

Proposition 4.1 *If $\tau < \tau_0 = \frac{1}{\sqrt{\lambda\mu}}(\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}})$, then the solution $\gamma(\tau)$ of Equation (2.3), if it exists, is larger than $s_r = (\sqrt{\mu}+\sqrt{\lambda})^2$, and if $\tau > \tau_0$, a solution must be inside the interval $[s_l, s_r] = [(\sqrt{\mu}-\sqrt{\lambda})^2, (\sqrt{\mu}+\sqrt{\lambda})^2]$.*

Proof. We claim that the solution $\gamma(\tau)$ of Equation (2.3), if it exists, is always larger than the threshold $s_l = \lambda+\mu-2\sqrt{\lambda\mu}$. Let γ_0 be the leftmost pole of the MGF $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$. Since the MGF is increasing in s on $[0, \gamma_0]$ we only need to show that

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_l} < \frac{1}{\rho}. \tag{4.3}$$

The value of the MGF at s_l is

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_l} = \frac{1 + \tau\mu - \tau\sqrt{\lambda\mu}}{1 - \tau\lambda + \tau\sqrt{\lambda\mu}}. \tag{4.4}$$

Thus, the inequality (4.3) simplifies to $\lambda + \tau\lambda(\mu - \sqrt{\lambda\mu}) < \mu + \tau\mu(\sqrt{\lambda\mu} - \lambda)$. Due to the stability assumption it is sufficient to show that $\lambda(\mu - \sqrt{\lambda\mu}) < \mu(\sqrt{\lambda\mu} - \lambda)$. Notice that this is equivalent to $\lambda + \mu - 2\sqrt{\lambda\mu} > 0$ and, hence, the claim is true.

Let us now check the behavior of the MGF at the right boundary $s_r = \lambda + \mu + 2\sqrt{\lambda\mu}$. We compare the value of the MGF with $1/\rho$:

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_r} = \frac{1 + \tau\mu + \tau\sqrt{\lambda\mu}}{1 - \tau\lambda - \tau\sqrt{\lambda\mu}} = \frac{1}{\rho}.$$
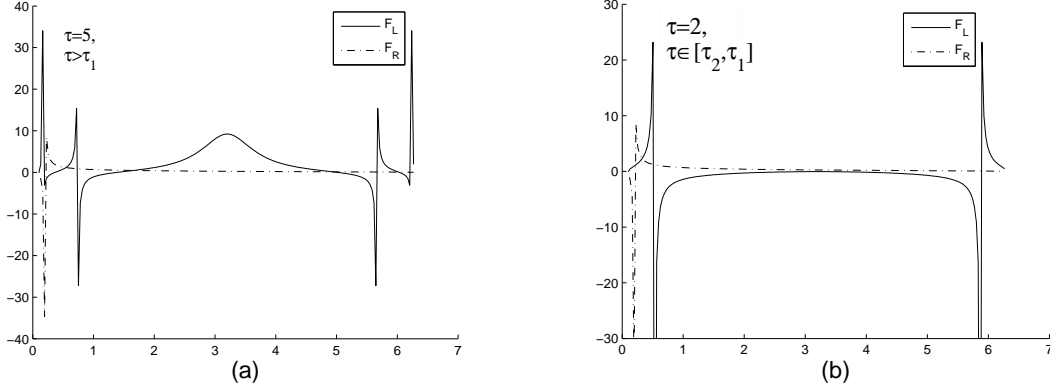
Figure 1: Functions F_L and F_R under different conditions on τ, $\lambda = 1.2$, $\mu = 2$.

This yields that the MGF at $s = s_r$ is equal to $1/\rho$ if

$$\tau_0 = \frac{\mu - \lambda}{\sqrt{\lambda\mu}(\lambda + \mu + 2\sqrt{\lambda\mu})} = \frac{1}{\sqrt{\lambda\mu}} \frac{(\sqrt{\mu} - \sqrt{\lambda})(\sqrt{\mu} + \sqrt{\lambda})}{(\sqrt{\mu} + \sqrt{\lambda})^2} = \frac{1}{\sqrt{\lambda\mu}}\left(\frac{1 - \sqrt{\rho}}{1 + \sqrt{\rho}}\right). \quad (4.5)$$

The statement of the proposition follows from the monotonicity of the MGF with respect to both s and τ. $\qquad \square$

In the next proposition we prove the existence of the decay rate $\gamma(\tau)$.

Proposition 4.2 *For any τ there exists a solution of Equation (2.3).*

Proof. Let us first assume that $\tau > \tau_0$. Hence, a solution of Equation (2.3) can only be inside the interval $[s_l, s_r]$. On this interval Equation (2.3) takes the following form

$$\frac{(\mu - \lambda + s) \sin[\frac{1}{2} f(s)\tau] + f(s) \cos[\frac{1}{2} f(s)\tau]}{(\mu - \lambda - s) \sin[\frac{1}{2} f(s)\tau] + f(s) \cos[\frac{1}{2} f(s)\tau]} = \frac{1}{\rho}, \quad (4.6)$$

where $f(s) = f_2(s) = \sqrt{-(\mu + \lambda - s)^2 + 4\lambda\mu}$.

After a simple computation we obtain that this equation is equivalent to

$$\tan\left(\frac{\tau}{2} f(s)\right) = \frac{f(s)}{\lambda - \mu + s\frac{1+\rho}{1-\rho}}. \quad (4.7)$$

Let us consider the left-hand side (denoted by F_L) and the right-hand side (denoted by F_R) of the latter equation in more detail. Depending on the value of τ, the behavior of F_L changes qualitatively. We will determine the intervals for τ on which F_L behaves differently and prove the Cramér condition on each interval.

The function F_R is independent of τ. As a function of s, F_R has a pole at $s^* = (\mu - \lambda)\frac{1-\rho}{1+\rho}$. On the interval $[s_l, s^*]$, F_R is decreasing from 0 to $-\infty$, and on $[s^*, s_r]$ it is decreasing from $+\infty$ to 0.

Let us now study the behavior of F_L as a function of s and τ. The tangent has infinite jumps when its argument is equal to $\frac{\pi}{2} + \pi k$, $k \in \mathbb{N}$. We are only interested in the first jump, $k = 0$. Note that, due to symmetry of $f(s)$ around $s_0 = \lambda + \mu$, F_L is also symmetric as a function of s on the interval $[s_l, s_r]$ with respect to the center of the interval, $s_0 = \lambda + \mu$.

10

The first jump of the F_L occurs when

$$\frac{\tau}{2}f(s') = \frac{\pi}{2},$$

that is when

$$s' = \lambda + \mu - \sqrt{4\lambda\mu - \frac{\pi^2}{\tau^2}}.$$

We will consider two cases separately: (1) - when F_L has an infinite jump inside the interval $[s_l, s_r]$, (2) - when it does not have such a jump. We derive the conditions and values of τ these situations can occur.

(1-a) First suppose that F_L has an infinite jump before the infinite jump of F_R, that is $s' < s^*$. That is equivalent to

$$s' = \lambda + \mu - \sqrt{4\lambda\mu - \frac{\pi^2}{\tau^2}} < (\mu - \lambda)\frac{1-\rho}{1+\rho} = s^*,$$

and hence,

$$\tau > \frac{\pi}{2\sqrt{\lambda\mu}}\frac{\mu+\lambda}{\mu-\lambda} := \tau_1.$$

Thus, for any $\tau > \tau_1$ the function F_L jumps before F_R. Notice that F_R is negative up to s^* and F_L is positive up to s' and negative after s' increasing from $-\infty$. Hence we can conclude that under this condition on τ there is always a solution of the equation $F_L = F_R$. That means that there is a solution $\gamma(\tau)$ of the Equation (2.3) and it is located inside the interval $[s_l, s']$.

Consider now a different situation. Suppose that F_L has no infinite jumps inside the interval $[s_l, s_r]$. This is equivalent to the statement

$$\frac{\tau}{2}f(s) < \frac{\pi}{2},$$

for all $s \in [s_l, s_r]$, i.e.

$$\tau < \min_{s \in [s_l, s_r]} \frac{\pi}{f(s)} = \frac{\pi}{2\sqrt{\lambda\mu}} := \tau_2.$$

(1-b) Consequently, for any $\tau \in [\tau_2, \tau_1]$ (see Figure 1 (b)) there is a jump of F_L in the interval $[s^*, \lambda + \mu)$ (before $\lambda + \mu$ since F_L is symmetric). Due to the properties of both functions for these τ there is always a point $\gamma(\tau)$ at which F_L and F_R intersect, $\gamma(\tau) \in [s^*, \lambda + \mu)$.

(2) Thus, for any $\tau \in [\tau_0, \tau_2]$ the function F_L has no jumps in $[s_l, s_r]$. Comparing the values of F_L and F_R at the center of the interval there are two cases possible in this situation (see Figure 2 (a,b)): (a) $F_R|_{s=\lambda+\mu} < F_L|_{s=\lambda+\mu}$ and (b) $F_R|_{s=\lambda+\mu} > F_L|_{s=\lambda+\mu}$. The values of the functions at this point are:

$$F_L|_{s=\lambda+\mu} = \tan(\tau\sqrt{\lambda\mu}),$$

$$F_R|_{s=\lambda+\mu} = \frac{\mu-\lambda}{2\sqrt{\lambda\mu}}.$$

(2-a) Consider the first case. Let us derive conditions under which this event may occur. Due to the monotonicity of the tangent, the inequality

$$F_L|_{s=\lambda+\mu} = \tan(\tau\sqrt{\lambda\mu}) > \frac{\mu-\lambda}{2\sqrt{\lambda\mu}} = F_R|_{s=\lambda+\mu}$$
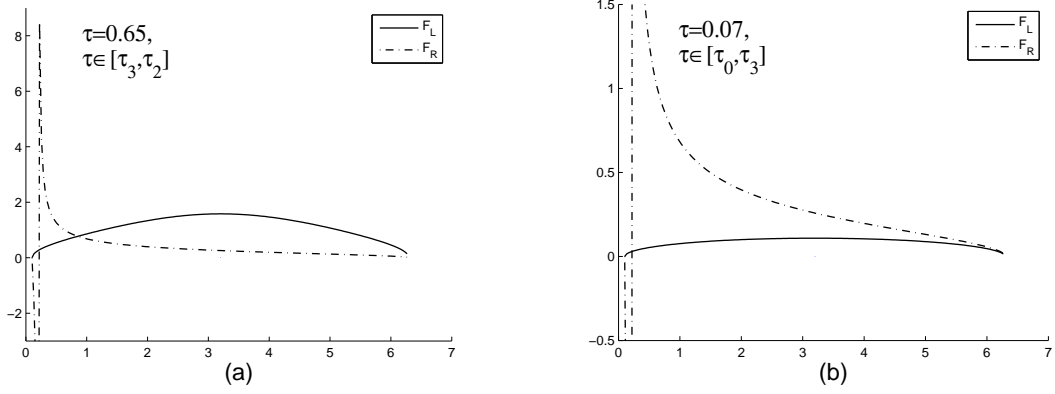
11

Figure 2: Functions F_L and F_R under different conditions on τ, $\lambda = 1.2$, $\mu = 2$.

reduces to

$$\tau > \frac{1}{\sqrt{\lambda\mu}} \arctan\left(\frac{\mu - \lambda}{2\sqrt{\lambda\mu}}\right) := \tau_3.$$

Hence, for all $\tau \in [\tau_3, \tau_2]$ the value of the F_R at the center point is lower than the value of F_L. Observe that F_R is decreasing on $[s^*, s_r]$ and the F_L is increasing on $[s_l, \lambda + \mu]$. Therefore, these two functions must intersect at the point $\gamma(\tau)$ on the interval $[s^*, \lambda + \mu]$. (2-b) Consider now the second case (Figure 2(b)): $F_L|_{s=\lambda+\mu} < F_R|_{s=\lambda+\mu}$. It is easy to check that the derivatives F_L' and F_R' are equal to infinity when $s = s_r$. For $\tau \in [\tau_0, \tau_3]$ it is impossible for F_L and F_R to intersect before the point $\lambda + \mu$. So we now consider $s \in [\lambda + \mu, s_r]$. For such s and τ both F_R and F_L are decreasing as functions of s and

$$F_R|_{s=s_r} = F_L|_{s=s_r} = 0,$$

$$F_R|_{s=\lambda+\mu} > F_L|_{s=\lambda+\mu}.$$

These functions can only intersect if and only if in some neighborhood of the point s_r the decrease of F_L is faster than the decrease of F_R, that is if and only if $F_L' < F_R'$.
The derivatives are given by

$$F_L' = \frac{\tau(\lambda + \mu - s)}{2f(s)\cos^2(\frac{\tau}{2}f(s))},$$

$$F_R' = \frac{4(\lambda - \mu)\lambda s\mu}{f(2\lambda\mu - \mu^2 + s\mu - \lambda^2 + \lambda s)^2}.$$

Thus, we have

$$F_L' = \frac{\tau(\lambda + \mu - s)}{2f(s)\cos^2(\frac{\tau}{2}f(s))} < \frac{4(\lambda - \mu)\lambda s\mu}{f(s)(2\lambda\mu - \mu^2 + s\mu - \lambda^2 + \lambda s)^2} = F_R',$$

$$\frac{\tau}{\cos^2(\frac{\tau}{2}f(s))} > \frac{8(\mu - \lambda)\lambda s\mu}{f(s)(s - \lambda - \mu)(2\lambda\mu - \mu^2 + s\mu - \lambda^2 + \lambda s)^2}.$$

When $s \to s_r$, $\cos(\frac{\tau}{2}f(s))$ converges to one from below. Hence, the right-hand side of the latter inequality is larger or equal to τ, while in this case $\tau > \tau_0 = \frac{\mu - \lambda}{\sqrt{\lambda\mu}(\lambda + \mu + 2\sqrt{\lambda\mu})}$. Notice that when $s \to s_r$ the left-hand side of the inequality converges to τ_0. Hence, the inequality holds for all s close enough to s_r.
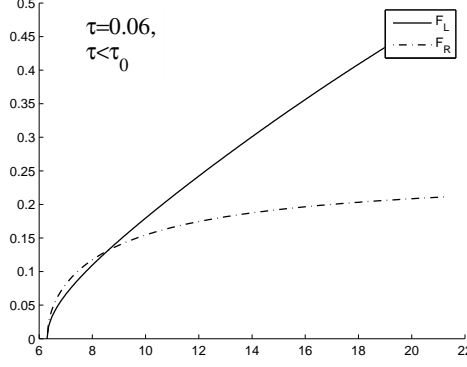
12

Figure 3: Functions F_L and F_R under conditions $\tau < \tau_0$, $\lambda = 1.2$, $\mu = 2$.

Thus, we have considered Equation (2.3) in four possible cases under the condition that $\tau > \tau_0$ and have shown that in all these cases there is a solution of Equation (2.3) and it lies inside the interval $[s_l, s_r]$.

(3) The only case left to consider is when $\tau < \tau_0$. For such values of τ, Equation (2.3) takes the form:

$$\mathbf{E}[e^{sC_i(\tau)}] = \frac{(\mu - \lambda + s)\sinh[\frac{1}{2}\tau f(s)] + f(s)\cosh[\frac{1}{2}\tau f(s)]}{(\mu - \lambda - s)\sinh[\frac{1}{2}\tau f(s)] + f(s)\cosh[\frac{1}{2}\tau f(s)]}, \tag{4.8}$$

or equivalently,

$$\tanh\left(\frac{\tau}{2}f(s)\right) = \frac{f(s)}{\lambda - \mu + s\frac{1+\rho}{1-\rho}}, \quad s \in [s_r, \infty), \tag{4.9}$$

where now $f(s) = f_1(s) = \sqrt{(\lambda + \mu - s)^2 - 4\lambda\mu}$.

A useful observation is that when $s \to \infty$, the left-hand side G_L converges to 1 and the right-hand side G_R converges to $\frac{1-\rho}{1+\rho}$, which is less than one for all $\rho > 0$. The derivatives of both functions are infinite at the point $s = s_r$ and both functions are strictly increasing for $s > s_r$ (see Figure 3). To prove the inequality we will use the same technique as in the previous case. We will show that there is a neighborhood of s_r in which the derivatives satisfy $G'_L < G'_R$, that is

$$\frac{\tau}{\cosh^2(\frac{\tau}{2}f(s))} < \frac{8(\mu - \lambda)\lambda s\mu}{f(s)(s - \lambda - \mu)(2\lambda\mu - \mu^2 + s\mu - \lambda^2 + \lambda s)^2}.$$

Notice that for $s \to s_r$ the function $\cosh(\frac{\tau}{2}f(s))$ converges to one from above, and so the left-hand side of the inequality is less or equal to τ, which is in this case less than τ_0. The inequality follows from the observation that the right-hand side converges to τ_0 when $s \to s_r$.

Thus, we have shown that for all $\tau > 0$ there exists a solution of Equation (2.3). $\qquad \square$

Now we are ready to complete the proof of Theorem 4.1.

Proof of Theorem 4.1. (i) Due to Propositions 4.1, 4.2 we know that the Cramér condition is satisfied. The decay rate $\gamma(\tau)$ is a solution of the following equation:

$$\frac{(\mu - \lambda + s)(e^{\frac{1}{2}\tau f(s)} - e^{-\frac{1}{2}\tau f(s)}) + f(s)(e^{\frac{1}{2}\tau f(s)} + e^{-\frac{1}{2}\tau f(s)})}{(\mu - \lambda - s)(e^{\frac{1}{2}\tau f(s)} - e^{-\frac{1}{2}\tau f(s)}) + f(s)(e^{\frac{1}{2}\tau f(s)} + e^{-\frac{1}{2}\tau f(s)})} = \frac{1}{\rho}, \tag{4.10}$$

13

where $f(s) = \sqrt{(\mu + \lambda - s)^2 - 4\lambda\mu}$.

Since the MGF $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ is differentiable in the point $s = \gamma(\tau)$, it follows that $h(\tau) < \infty$. The fact that the MGF $\mathbf{E}[\mathrm{e}^{sV_0(\tau)}]$ has the same abscissa of convergence as $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ (since B has unbounded support), implies that $\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]$ is finite for any τ. Thus, all conditions of Theorem 2.1 are satisfied and we can conclude that the asymptotic relationship (4.1) holds. The asymptotic constant $\alpha(\tau)$ is determined by Equation (2.5). We need to compute the derivative of the MGF of $C_i(\tau)$ at $\gamma(\tau)$. For compactness let us denote the denominator in Formula (3.14) by D and the numerator by N. The exponents e^{+} and e^{-} denote $\mathrm{e}^{f(s)\tau/2}$ and $\mathrm{e}^{-f(s)\tau/2}$ respectively. Then

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\bigg|_{s=\gamma(\tau)} = \left[\frac{N'}{D} - \frac{N \cdot D'}{D^2}\right]\bigg|_{s=\gamma(\tau)} = \left[\frac{N'}{D} - \frac{D'}{\rho \cdot D}\right]\bigg|_{s=\gamma(\tau)}$$

$$= \frac{\rho N' - D'}{\rho D}\bigg|_{s=\gamma(\tau)},$$

where

$$N' = (\mu - \lambda + s)(\mathrm{e}^{+} + \mathrm{e}^{-})\frac{\tau}{2}f'(s) + (\mathrm{e}^{+} - \mathrm{e}^{-}) + f'(s)\left((\mathrm{e}^{+} + \mathrm{e}^{-}) + f(s)(\mathrm{e}^{+} - \mathrm{e}^{-})\frac{\tau}{2}\right),$$

$$D' = (\mu - \lambda - s)(\mathrm{e}^{+} + \mathrm{e}^{-})\frac{\tau}{2}f'(s) - (\mathrm{e}^{+} - \mathrm{e}^{-}) + f'(s)\left((\mathrm{e}^{+} + \mathrm{e}^{-}) + f(s)(\mathrm{e}^{+} - \mathrm{e}^{-})\frac{\tau}{2}\right)$$

and $f'(s) = (\lambda + \mu - s)/f(s)$. Hence,

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\bigg|_{s=\gamma(\tau)} =$$

$$= \frac{1}{\rho D f(\gamma(\tau))}\left[(1 + \rho)\left[(f(\gamma(\tau))(\mathrm{e}^{+} - \mathrm{e}^{-}) + \gamma(\tau)\frac{\tau}{2}(\lambda + \mu - \gamma(\tau)))(\mathrm{e}^{+} + \mathrm{e}^{-})\right]\right.$$

$$\left. -(1 - \rho)(\lambda + \mu - \gamma(\tau))\left[(\mathrm{e}^{+} + \mathrm{e}^{-})(1 + (\mu - \lambda)\frac{\tau}{2}) + (\mathrm{e}^{+} - \mathrm{e}^{-})f(\gamma(\tau))\frac{\tau}{2}\right]\right].$$

Let us denote the last multiplier as K

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\bigg|_{s=\gamma(\tau)} = \frac{1}{\rho D f(\gamma(\tau))}K.$$

Since $\mathbf{E}[\mathrm{e}^{sV_0(\tau)}] \equiv \delta(-s, \tau) = \frac{2f(s)\mathrm{e}^{-(-s+\lambda-\mu)\tau/2}}{D}$ (Formula (3.1)) we obtain from Formula (2.5):

$$\alpha(\tau) = \frac{(1 - \rho)}{\gamma(\tau)K}2f^2(\gamma(\tau))\mathrm{e}^{-(-s+\lambda-\mu)\tau/2},$$

which gives formula (4.2).

(ii) When $\tau = \tau_0$ the decay rate follows immediately, $\gamma(\tau_0) = s_r$. However, in this case the function $h(\tau_0) = \rho\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau_0)}]$ is infinite. Hence, due to [14],

$$\lim_{x\to\infty}\mathbf{P}(V_1(\tau_0) > x)\mathrm{e}^{\gamma(\tau_0)x} = 0,$$

and consequently,

$$\lim_{x\to\infty}\mathbf{P}(V(\tau_0) > x)\mathrm{e}^{\gamma(\tau_0)x} = 0.$$

This completes the proof.

$\square$

5 The impact of the service discipline on the decay rate

In this section we investigate the behavior of the decay rate $\gamma(\tau)$ by solving Equation (2.3) numerically. Furthermore we perform a comparison of the PS decay rate with the decay rates in the $M/M(\tau)$ queue under different service disciplines: we consider the Shortest Remaining Processing Time (SRPT) and the Foreground-Background (FB) disciplines.

The decay rate under the SRPT and FB disciplines has been studied in [18] and [15] respectively. For the SRPT discipline, Nuyens and Zwart [18] have shown that the decay rate of the conditional sojourn time $V_{SRPT}(\tau) = [V_{SRPT}|B = \tau]$ coincides with the decay rate of the residual busy period $\gamma^p_{SRPT}(\tau)$ in the queue with service time $B^\tau_{SRPT} = B\mathbf{1}(B < \tau)$. Mandjes and Nuyens [15] have derived similar result for the FB discipline. They proved that if the generic service time has an exponential moment then the sojourn time $V_{FB}(\tau)$ has the same decay rate $\gamma^p_{FB}(\tau)$ as the residual busy period in the queue with service time $B^\tau_{FB} = \min(B, \tau)$. It is known that the decay rate of the busy period can be determined as

$$\gamma^p(\tau) = -\kappa(\theta_0),$$

where $\kappa(s) = \lambda(\mathbf{E}[e^{sB^\tau}] - 1) - s$, and $\theta_0 > 0$ is a solution of the equation $\kappa'(\theta_0) = 0$ (or equivalently $\lambda(\mathbf{E}[e^{sB^\tau}])'_s = 1$).
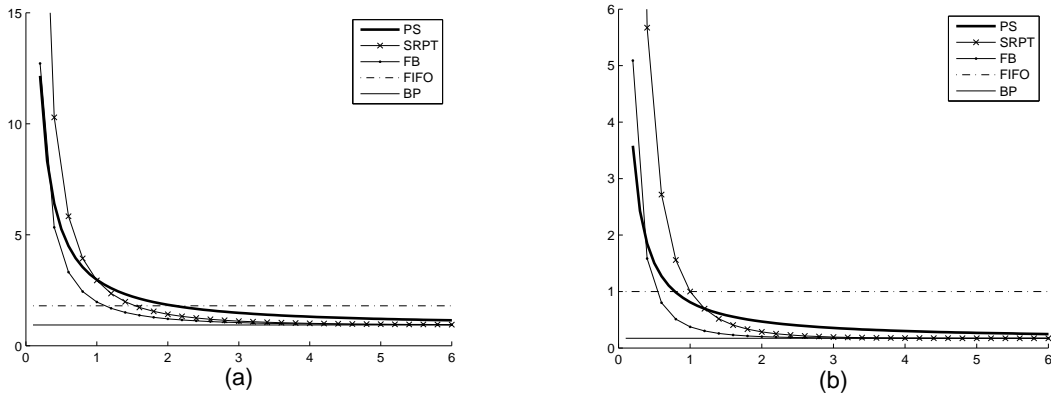

Figure 4: Decay rate as a function of τ in the $M/M(\tau)/1$ queue with PS, SRPT and FB service disciplines, $\mu = 2$: (a) $\lambda = 0.2$, (b) $\lambda = 1$.

Figure 4 presents the decay rate $\gamma(\tau)$ as a function of τ for the above mentioned disciplines. The generic service time is exponential, $\mu = 2$. Figure 4(a) shows the decay rates under very low traffic load, $\rho = 0.1$, and Figure 4(b) is for $\rho = 0.5$. In the figures, the horizontal lines shows the decay rate in the $M/M/1$ FIFO queue (dash-dotted line) and the decay rate of the busy period (solid line). The decay of FIFO queue is equal to $\gamma_{FIFO} = \mu - \lambda$ and the decay rate of the busy period is $\gamma^p = (\sqrt{\mu} - \sqrt{\lambda})^2$.

Figure 5 shows the decay rates when the traffic intensity is reasonably high, (a) $\rho = 0.9$, (b) $\rho = 0.95$. From the figure we clearly see that when the service request τ becomes larger, the decay rates for all disciplines decrease and converge to the decay rate of the busy period γ^p.

All graphs show that for moderate values of τ the decay rate of SRPT is the largest. For larger requests the FIFO discipline provides the largest decay rate. Our simulations and analytic results in [18] show that the majority of the customers (at least 85%) would prefer SRPT over FIFO. Interestingly, PS does not appear to be the best discipline (from the viewpoint of decay rates) for jobs of any size. If a customer has a large request, FIFO
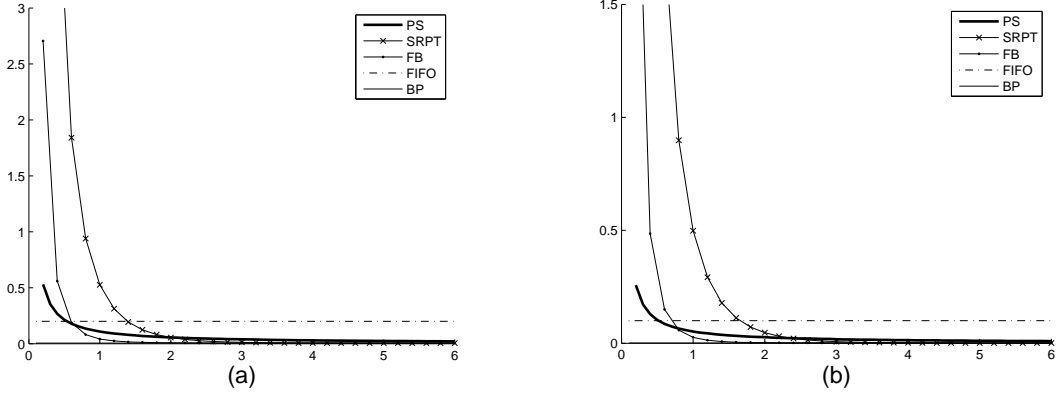
Figure 5: Decay rate as a function of τ in the M/M(τ)/1 queue with PS, SRPT and FB service disciplines, $\mu = 2$: (a) $\lambda = 1.8$, (b) $\lambda = 1.9$.

should be preferred, and if the request is small SRPT (and FB) provide shorter sojourn times, see Figures 4(b) and 5(a,b). Moreover, we conclude that the higher the traffic intensity, the less attractive is PS compared to the other disciplines.
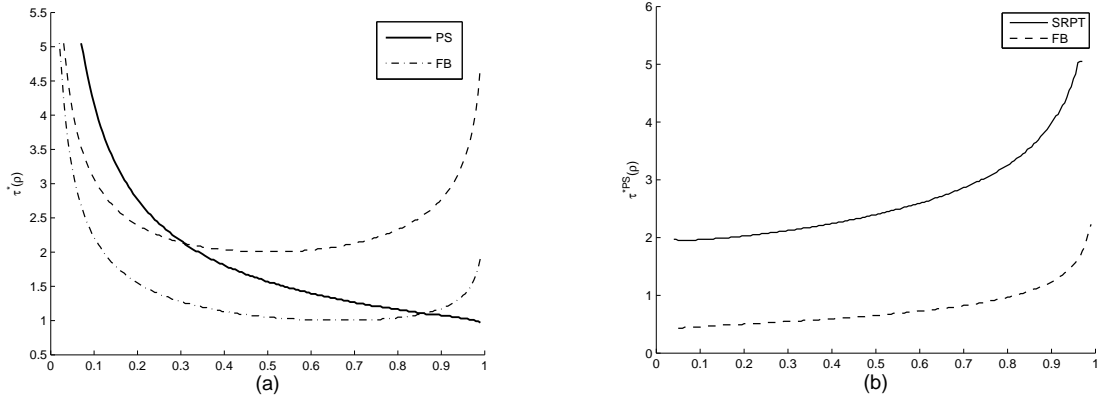

Figure 6: Decay rate τ^* as a function of ρ in the M/M(τ)/1 queue: (a) - intersection with FIFO, (b) - intersection with PS decay rate

However, in Figure 4(a) we see a somewhat different picture. In this case the traffic intensity is very low. Then for a certain range of requests, not too long and not too short, the decay rate for PS is the largest. Since there are not many customers in the system, the sojourn time is not affected by sharing capacity and there is no need to wait for other customers as in SRPT.

Let us introduce the variable τ^*_{PS} as the value of τ at which the PS decay rate $\gamma(\tau)$ is equal to the FIFO decay rate, i.e. the value at which $\gamma(\tau)$ crosses level $\mu - \lambda$. Define similarly variable τ^*_{SRPT} and τ^*_{SRPT}. Figure 6(a) shows the behavior of τ^*_{PS} and τ^*_{SRPT} as a function of the traffic load ρ. As we can see, for traffic load $\rho < 0.3$ the PS decay rate reaches the value $\mu - \lambda$ later than the SRPT decay rate. This means that for such ρ, there exists a positive ε_ρ, such that in the interval $[\tau^*_{PS} - \varepsilon_\rho, \tau^*_{PS}]$ the PS discipline has the largest decay rate (compared to FIFO and SRPT). In other words, for such ρ, a fraction $[e^{-\tau^*_{PS}}(e^{\varepsilon_\rho} - 1)]$ of customers would prefer the PS queue.

Comparing the PS decay rate to the FB one (see Figure 6(a)), the decay rate shows similar

16

behavior. In this case the threshold load is $\rho < 0.86$.

Figure 6(b) shows τ^{*PS}, the value of τ, at which the decay rate of the SRPT and FB disciplines is equal to the decay rate of PS. As we see, the higher the value of ρ, the smaller the group of customers preferring PS service over the other two disciplines.

Let us now summarize the results. In the PS queue, as well as in SRPT and FB, the decay rate decreases and converges to the decay rate of the busy period as $\tau \to \infty$. Interestingly, in most cases, except when the traffic load is quite low, the PS discipline is not a preferable discipline for any request length τ in the sense of reducing long sojourn times. For larger customers, FIFO has a higher decay rate than PS, and for smaller customers, SRPT performs the best. However, when ρ is not high, there is a certain interval for the length of requests for which PS has the largest decay rate. It would be an ultimate goal to design a more advanced scheduling discipline which give good performance for both small and large service times (here we consider it only from the decay rate point of view). We hope that these results can be potentially useful.

6   Accuracy of the asymptotics

Finally, we will study the accuracy of the exponential approximation (4.1) of the sojourn time in the M/M(τ)/1 queue:

$$\mathbf{P}(V(\tau) > x) \approx \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}.$$

The exponential asymptote is compared to exact values of $\mathbf{P}(V(\tau) > x)$ computed by numerical Laplace-Stieltjes transform inversion.

The inversion of the Laplace transform was considered to be numerically challenging for a long time. However, nowadays there is a number of reliable and effective inversion methods available. We will use the inversion algorithm of Abate and Whitt [2]. In this method the probability distribution function is presented as an infinite sum of complex-valued terms. For the summation of this infinite series the classical Euler summation method is applied. This method is known to provide high accuracy.

	$\tau = 0.8$			$\tau = 2$	
x	LST inversion	appr.(4.1)	x	LST inversion	appr.(4.1)
5	5.49E-01	5.77E-01	10	6.34E-01	7.25E-01
10	2.82E-01	2.96E-01	100	4.72E-03	5.41E-03
20	7.41E-02	7.79E-02	150	3.10E-04	3.56E-04
40	5.14E-03	5.39E-03	200	2.04E-05	2.34E-05
80	2.47E-05	2.59E-05	250	1.33E-06	1.54E-06
100	1.70E-06	1.79E-06	300	9.43E-08	1.01E-07
120	1.24E-07	1.24E-07	310	5.78E-08	5.89E-08

Table 1: Comparison of the exponential asymptote to results of numerical inversion.

Table 1 shows the numerical results for various request lengths τ. For simplicity we normalize the generic service time, $\mu = 1$, and take arrival rate $\lambda = 0.9$. For $\tau = 0.8$ and $\tau = 2$ the first column shows the probability $\mathbf{P}(V(\tau) > x)$ obtained by numerical inversion. The second column shows the exponential asymptotics derived in Theorem 4.1. The numbers show reasonably good accuracy of the asymptotic tail approximation. The relative error is on average about 5-10%. Due to the asymptotic constant $\alpha(\tau)$ in (4.2) which can take any positive value, the approximation (4.1) of $\mathbf{P}(V(\tau) > x)$ is not appropriate for smaller values of x.

	$\tau = 0.8$				$\tau = 2$		
x	LST inv	appr.(4.1)	HT	x	LST inv	appr.(4.1)	HT
10	5.42E-01	5.56E-01	5.35E-01	10	8.04E-01	8.59E-01	7.79E-01
20	2.84E-01	2.92E-01	2.87E-01	100	7.70E-02	8.20E-02	8.21E-02
50	4.10E-02	4.22E-02	4.39E-02	200	5.67E-03	6.03E-03	6.74E-03
100	1.63E-03	1.68E-03	1.93E-03	300	4.18E-04	4.44E-04	5.53E-04
150	6.45E-05	6.66E-05	8.48E-05	400	3.08E-05	3.26E-05	4.54E-05
200	2.54E-06	2.65E-06	3.73E-06	500	2.26E-06	2.40E-06	3.73E-06
240	1.96E-07	2.01E-07	3.06E-07	600	1.73E-07	1.76E-07	3.06E-07
250	1.05E-07	1.05E-07	1.64E-07	640	6.24E-08	6.21E-08	1.13E-07

Table 2: Comparison of the exponential asymptote to results of numerical inversion and heavy-traffic asymptotics.

Table 2 shows results for heavy traffic, in particular $\rho = 0.95$. In addition to the results from numerical inversion and asymptotics, the table presents results of the heavy-traffic approximation. From the results in [23], [21], it is known that under heavy traffic the sojourn time distribution in the M/G/1 PS queue behaves as

$$\mathbf{P}(V(\tau) > x) \approx e^{-\frac{(1-\rho)x}{\tau}}, \qquad x \to \infty.$$

These values are presented in the columns with headline HT.

The accuracy of the asymptotic approximation (4.1) is better for higher traffic load. It is also much more accurate than the heavy-traffic approximation for larger x, although for small x the heavy-traffic approximation performs better.

7   Conclusions

For the sojourn time in the M/G(τ)/1 queue, we established exponential asymptotics if either the load is sufficiently high or the service request is sufficiently small. In these cases the general formulas for the decay rate and the asymptotic constant are available, which allows to determine the asymptote numerically.

For the sojourn time in the M/M(τ)/1 queue we obtained the exponential asymptote for any traffic load and any request lengths. We derived an equation for the decay rate $\gamma(\tau)$, which turns out to be of quite remarkable form, and a complicated but exact expression for the constant $\alpha(\tau)$. Furthermore, we studied the behavior of the decay rate as a function of the request size. Comparison with other service disciplines shows an interesting result. It suggests that in order to have a shorter sojourn times in most of the cases it is not advisable to use the PS discipline. Most of the customers would prefer the SRPT service discipline, while the rest would benefit from FIFO. Finally, we investigated the accuracy of the asymptote by comparison with the tail probability obtained by LST inversion. The result showed that the asymptote provides a reasonably good approximation, especially under heavy traffic.

We finally suggest several possible extensions. A similar approach as in Sections 3 and 4 may be applied to a queue with phase-type distributed service times. In the phase-type case, we expect to have a finite number of special values of τ, at which the exponential asymptotic does not hold (recall that one such point τ_0 exists for exponential service times). It is quite probable that in a queue with general service time distribution, there could be found a whole spectrum of such points. This may necessitate an alternative approach to deal with the general queue.

References

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

[2] Abate, J., Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5–88.

[3] Van den Berg, J.L. (1990). Sojourn times in feedback and processor sharing queues. *PhD thesis,* Utrecht University.

[4] 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.

[5] Borst, S.C., Núñez-Queija,R., Zwart, A.P. (2006). Sojourn time asymptotics in Processor-Sharing queues. To appear in *Queueing Systems.*

[6] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications* **10**, 323–331.

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

[8] Cramér, H. (1930). On the mathematical theory of risk. *Skandia Jubilee Volume.*

[9] Denisov, D., Zwart, A.P. (2005). On a theorem of Breiman and a class of random difference equations. *EURANDOM report* **2005-039**, http://www.eurandom.nl/reports/2005/039DDreport.pdf.

[10] Egorova, R., Zwart, A.P., Boxma, O.J. (2006). Sojourn time tails in the M/D/1 Processor Sharing queue. *Probability in the Engineering and Informational Sciences* **20**, 427-444.

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

[12] Grishechkin, S. (1992). On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability* **24**, 653–698.

[13] Jelenković, P.R., Momčilović, P. (2004). Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operations Research* **28**, 587–608.

[14] Kalashnikov, V., Tsitsiashvili, G. (1999). Tail of waiting times and their bounds. *Queueing Systems* **32**, 257–283.

[15] Mandjes, M., Nuyens, M. (2004). Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the Engineering and Informational Sciences* **19**, 351–361.

[16] Mandjes, M., Zwart A.P. (2006). Large deviations for sojourn times in Processor Sharing queues. To appear in *Queueing Systems.*

[17] Núñez-Queija, R. (2000). Processor-sharing models for integrated-services networks. *PhD thesis,* Eindhoven University of Technology.

[18] Nuyens, M., Zwart, A.P. (2005). A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems*, submitted.

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

[20] Schassberger, R. (1984). A new approach to the M/G/1 processor sharing queue, *Advances in Applied Probability* **16**, 802–813.

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

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

[23] Yashkov, S.F. (1993). On a heavy-traffic limit theorem for the M/G/1 processor-sharing queue. *Stochastic Models* **9**, 467–471.

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