Sojourn time tails in the M/D/1 processor sharing queue

TU/e technische universiteit eindhoven

SPOR-Report 2005-05

Sojourn Time Tails in the M/D/1 Processor Sharing Queue

R. Egorova
B. Zwart
O. Boxma

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, May 2005
The Netherlands

/department of mathematics and computing science

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, Statistics and Operations research
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Main Building 9.10
Telephone: + 31 40 247 3130
E-mail: wscosor@win.tue.nl
Internet: http://www.win.tue.nl/math/bs/spor

ISSN 1567-5211

Sojourn Time Tails in the M/D/1 Processor Sharing Queue

Regina Egorova, Bert Zwart, Onno Boxma
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

Abstract

We consider the sojourn time V in the M/D/1 processor sharing (PS) queue, and show that $\mathbf{P}(V > x)$ is of the form $C\mathrm{e}^{-\gamma x}$ as x becomes large. The proof involves a geometric random sum representation of V, and a connection with Yule processes, which also enables us to simplify Ott's (1984) derivation of the Laplace transform of V. Numerical experiments show that the approximation $\mathbf{P}(V > x) \approx C\mathrm{e}^{-\gamma x}$ is excellent even for moderate values of x.
2000 Mathematics Subject Classification: 60K25.
Keywords: random sum, Yule processes, branching processes, tail behavior, heavy traffic, transform inversion

1 Introduction

Queues with the PS service discipline became popular by the work of Kleinrock [16, 17, 18] and were originally proposed as an idealization of time-sharing (queueing) 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, see e.g. Núñez-Queija [20] and Roberts [24].
Several PS studies have focused on the analysis of the tail of the sojourn time distribution in a case when the service time distribution is heavy-tailed. The asymptotic tail behavior of the sojourn time in the M/G/1 PS queue with regularly varying service time distribution was derived by Zwart and Boxma in [29] and later generalized by Núñez-Queija in [20] for the case of distributions with intermediately regularly varying tails. They established the following asymptotic relationship between the distributions of the sojourn time V and the customer service time B (with ρ denoting the traffic load):

$$\mathbf{P}(V > x) \sim \mathbf{P}(B > (1-\rho)x), \tag{1.1}$$

as $x \to \infty$ (for any two real functions $f(\cdot)$ and $g(\cdot)$, $f(x) \sim g(x)$ as $x \to \infty$ denotes that $f(x)/g(x) \to 1$ as $x \to \infty$). This equivalence is often called a reduced-load approximation. In [13], Jelenković and Momčilović extended the equivalence result to the case when the service time belongs to a class of subexponential distributions with tails heavier than $\mathrm{e}^{-\sqrt{x}}$. The equivalence (1.1) was extended to other types of PS queues in [12] and [6]. Overall, the sojourn time asymptotics are well understood in the heavy-tailed case.

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

$$\mathbf{P}(V > x) \sim c x^{-5/6} \mathrm{e}^{-\alpha x^{1/3}} \mathrm{e}^{-\gamma_0 x}, \qquad x \to \infty, \tag{1.2}$$

for positive constants c, α, γ_0. 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.* [5] 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. We note that the proof of Flatto is purely analytical, and a probabilistic proof is still lacking.
Using large-deviation techniques, Mandjes and Zwart [19] analyzed sojourn time asymptotics in the GI/GI/1 PS queue. They derived logarithmic asymptotics for a broad class of light-tailed distributions:

$$\log \mathbf{P}(V > x) \sim -\gamma_0 x, \qquad x \to \infty. \tag{1.3}$$

Remark that in the M/M/1 case the decay rate γ_0 in (1.3) coincides with the exponential decay rate γ_0 in expression (1.2).
Apart from the shape of the asymptotics, it is of interest *how* large sojourn times take place. In a PS queue, three events may contribute to a large sojourn time: (i) a large service time of the tagged customer; (ii) a large number of customers present in the system upon arrival of the tagged customer; (iii) an unusually large number of arrivals after arrival of the tagged customer. When service times are heavy-tailed, event (i) is responsible for a large sojourn time. In [19], the authors show that for a broad class of light-tailed distributions, event (iii) determines the logarithmic asymptotics (1.3). Specifically, V becomes large if the traffic load ρ is increased to 1 during the sojourn time of the tagged customer. This intuition is valid under two technical conditions which ensure that the service-time distribution is not too heavy and not too light. The conditions in [19] are violated for any distribution with bounded support, such as deterministic service times.
This motivated us to take a closer look at the PS queue with deterministic service times. Specifically, we assume that customers arrive according to a Poisson process with rate λ at a single server. The server operates according to the PS discipline, i.e. when there are n customers in the system, each of them is served with rate $1/n$. The service time is constant for all customers, denoted by D. Let ρ be the traffic intensity, $\rho = \lambda D$. We assume that $\rho < 1$, so that the system reaches steady state. We investigate the asymptotic behavior of the sojourn time tail. Our main result is that the tail behavior of the steady-state sojourn time V is of the following form:

$$\mathbf{P}(V > x) \sim \alpha \mathrm{e}^{-\gamma x}, \qquad x \to \infty, \tag{1.4}$$

for some explicit constants α and γ. Observe that the asymptotic form is fundamentally different from the one for exponential service times. Moreover, from our analysis in Section 3, one can infer that the most likely way to cause the event $\{V > x\}$ not only involves more work feeding into the system between time 0 and x, but also an increased number of customers at time 0, i.e. the event $\{V > x\}$ occurs by a combination of the events (ii) and (iii) mentioned above.
To prove (1.4), we study the sojourn time of a customer by means of branching processes. The branching process representation and decomposition of the sojourn time into a sum of independent random variables (called *delay elements*), conditioned on the number of customers in the system, was established by Yashkov [27] for the M/G/1 PS queue and later extended by Ott [21]. This approach was also used by Rege and Sengupta [23] for

the M/G/1 queue with Discriminatory Processor Sharing and by Núñez-Queija [20] for M/M/1 PS queues with breakdowns. In this paper, we make the additional observation that the underlying branching process for the M/D/1 PS queue is a Yule process, which has been treated by Ross [25]. We use this connection to obtain a simplified derivation of the Laplace-Stieltjes Transform (LST) of the delay elements associated with V, which also leads to a relatively simple derivation of Ott's result ([21], formula (5.16)) for the LST of V. Since the number of customers in the system has a geometric distribution, we can apply existing theory for tail asymptotics of geometric random sums to obtain the tail behavior of V.

The remainder of the paper is organized as follows. In Section 2, we give a closed-form expression for the LST's of the distribution of the delay elements of the branching process decomposition. The main result is presented and proven in Section 3. In addition, the asymptotic behavior under heavy traffic is considered. It is shown that the limits with respect to time and traffic load are interchangeable. In the last section we present results from numerical experiments. We compute values of $\mathbf{P}(V > x)$ using transform inversion and compare them with values predicted by (1.4). These experiments demonstrate a remarkable accuracy of the obtained approximation (1.4).

2 Laplace-Stieltjes transform of the sojourn time distribution

In this section we derive the sojourn time LST in the M/D/1 PS queue. In fact, the explicit formula for the LST of V is well known. It was derived by Ott [21] as a special case of M/G/1/PS:

$$\mathbf{E}(\mathrm{e}^{-sV}) = \frac{(1-\rho)(\lambda+s)^2\mathrm{e}^{-(\lambda+s)D}}{s^2+\lambda(s+s(1-\rho)+\lambda(1-\rho))\mathrm{e}^{-(\lambda+s)D}}. \tag{2.1}$$

In this section we will give a new simplified proof of this formula using existing results for Yule processes. The first step in our proof is to represent the sojourn time as a function of a branching process. By conditioning on the number of customers in the system upon the arrival of the tagged customer, we will decompose its sojourn time into a number of independent random variables, called delay elements ([27]). In Section 3, we also use some intermediate results provided by this decomposition in the derivation of the tail asymptotics. Thus, our main focus in this section is on the LST of the delay elements.

To perform a branching process decomposition we consider the process on a transformed time scale. The time-change method is widely used in the analysis of PS queues, cf. [27], [20]. Throughout the paper, we perform all investigations depending on the amount of service t attained by the tagged customer, $t \in [0, D]$, and not the actual time scale. Moreover, we introduce the process $X(t)$ as the number of customers (including the tagged customer) at the server at the epoch when an amount of service t is received by the tagged customer. We study the sojourn time of the customer in terms of the process $X(t)$. A very useful observation is that on the time interval till the first departure (or time interval during which no departures occur) $X(t)$, $t \in [0, D]$, can be considered as a Yule process. Recall that a Yule process is a pure birth process in which each individual in the population independently gives birth at constant rate λ.

Let us now discuss the time change in more detail. Denote the number of customers in the system (including the tagged customer) at time x by $Q(x)$. The amount of service received by the tagged customer during the time interval $[0, x]$ is

$$t = T(x) = \int_0^x \frac{1}{Q(s)}\mathrm{d}s. \tag{2.2}$$

Then, the process $X(t)$ introduced above can be defined as $X(t) = Q(T^{-1}(t))$. Evidently, the sojourn time V_0 can be expressed in terms of the process $X(t)$ as

$$V_0 = \int_0^D X(t)\mathrm{d}t. \tag{2.3}$$

The remainder of this section is organized as follows. First we consider the situation when the tagged customer enters an empty system. We derive the LST of the sojourn time of this customer. Then we turn to the general case when there are a number of customers in the system upon arrival of the tagged customer. We give a detailed description of the sojourn time decomposition into delay elements and finally we prove Ott's formula (2.1).

2.1 Sojourn time of the first customer

In this subsection we derive the LST of the sojourn time of the first customer, i.e. the customer that enters an empty system. Notice that in this situation the above-defined process $\{X(t), t \in [0, D]\}$, where t is the amount of service received by the first customer, is a Yule process. In our model the births correspond to customer arrivals. Until the service requirement of the first customer is completed, a number of other customers may arrive but none leave the system before that time, since under the PS discipline with constant service requirements customers depart from the system in order of their arrival. The next proposition gives the LST of the first customer's sojourn time. From now on we will use variable t as time in the changed time scale.

Proposition 2.1

$$\mathbf{E}(\mathrm{e}^{-sV_0}) = \frac{\lambda + s}{\lambda + s\mathrm{e}^{(\lambda+s)D}}. \tag{2.4}$$

Proof:
The integral representation (2.3) of V_0 can be rewritten as follows:

$$V_0 = D + \sum_{k=1}^{X(D)-1} (D - t_k),$$

where $(t_k, k \geq 1)$ are the arrival times of customers that enter the system during the service of the first customer.
Since $\{X(t),\ t \geq 0\}$, is a Yule process, its marginal distribution is known (see e.g. [25], p. 236). At time t the population size is geometrically distributed with parameter $\mathrm{e}^{-\lambda t}$:

$$\mathbf{P}(X(t) = i) = (1 - \mathrm{e}^{-\lambda t})^{i-1}\mathrm{e}^{-\lambda t}, \quad t \in [0, D]. \tag{2.5}$$

Furthermore (see again [25]), the conditional joint probability density of the arrival times $t_1,\ t_2, \ldots,\ t_n$, given the number of customers, $X(t) = n + 1$, is given by

$$p(s_1, s_2, \ldots, s_n | X(t) = n + 1) = \prod_{i=1}^{n} f(s_i), \quad s_i \leq t, \tag{2.6}$$

where

$$f(x) = \frac{\lambda \mathrm{e}^{-\lambda(t-x)}}{1 - \mathrm{e}^{-\lambda t}}, \quad 0 \leq x \leq t. \tag{2.7}$$

In order to obtain the expression for the LST of V_0, we condition on the number of customers in the system upon departure of the first customer,

$$\begin{aligned}\mathbf{E}(\mathrm{e}^{-sV_0}) &= \mathbf{E}(\mathrm{e}^{-s\int_0^D X(t)\mathrm{d}t}) = \mathbf{E}(\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))}) \\ &= \sum_{n=0}^{\infty}\mathbf{E}(\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))}|X(D)=n+1)\mathbf{P}(X(D)=n+1), \quad (2.8)\end{aligned}$$

where, due to independence of the $t_k, k=1,\dots,n$, the conditional expectation is

$$\mathbf{E}(\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))}|X(D)=n+1) = \prod_{k=1}^{n}\mathbf{E}(\mathrm{e}^{-s(D-t_k)}|X(D)=n+1)\mathrm{e}^{-sD}. \quad (2.9)$$

Computing the inner term of the above product, we get

$$\begin{aligned}\mathbf{E}(\mathrm{e}^{-s(D-t_k)}|X(D)=n+1) &= \int_0^D \mathrm{e}^{-s(D-x)}\frac{\lambda\mathrm{e}^{-\lambda(D-x)}}{1-\mathrm{e}^{-\lambda D}}\mathrm{d}x \\ &= \frac{\lambda}{\lambda+s}\frac{1-\mathrm{e}^{-(\lambda+s)D}}{1-\mathrm{e}^{-\lambda D}}. \quad (2.10)\end{aligned}$$

Hence,

$$\mathbf{E}(\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))}|X(D)=n+1) = \left(\frac{\lambda}{\lambda+s}\right)^n\left(\frac{1-\mathrm{e}^{-(\lambda+s)D}}{1-\mathrm{e}^{-\lambda D}}\right)^n\mathrm{e}^{-sD}. \quad (2.11)$$

Substituting (2.11) into (2.8) we obtain the LST of the sojourn time,

$$\begin{aligned}\mathbf{E}(\mathrm{e}^{-sV_0}) &= \sum_{n=0}^{\infty}\left(\frac{\lambda}{\lambda+s}\right)^n\left(\frac{1-\mathrm{e}^{-(\lambda+s)D}}{1-\mathrm{e}^{-\lambda D}}\right)^n\mathrm{e}^{-sD}(1-\mathrm{e}^{-\lambda D})^n\mathrm{e}^{-\lambda D} \\ &= \frac{\mathrm{e}^{-(\lambda+s)D}}{1-\frac{\lambda}{\lambda+s}(1-\mathrm{e}^{-(\lambda+s)D})}. \quad (2.12)\end{aligned}$$

Rewriting this gives (2.4).

Remark 2.1 Interestingly, an analog of the above result is given in [15]. The authors consider an M/G/1 queue with any symmetric queueing discipline (processor sharing is a special case). Let B be a generic service time, D_1 the time till the first departure from the system. Assume that the system is empty at time 0. Then for any positive s,

$$\mathbf{E}(\mathrm{e}^{-sD_1}) = \frac{\lambda}{\lambda+s\mathbf{E}(\mathrm{e}^{(\lambda+s)B})}. \quad (2.13)$$

The expression is related to the LST of V_0 as

$$\mathbf{E}(\mathrm{e}^{-sD_1}) = \frac{\lambda}{\lambda+s}\mathbf{E}(\mathrm{e}^{-sV_0}),$$

which is a natural result, since $D_1 \stackrel{d}{=} A_1 + V_0$, where A_1 is the time till the first arrival.

2.2 Sojourn time of an arbitrary customer

Let us now turn to the derivation of the LST of the sojourn time of a customer who enters the system and sees a number of customers already in service upon its arrival. Denote its sojourn time by V. Suppose that the number of customers in the system upon its arrival is Q. As before, $X(t)$ is the number of customers at the epoch when an amount of service t is received by the tagged customer, $t \in [0, D]$. Then $X(0) = Q$, $X(0+) = Q + 1$.

Proof of Formula (2.1). Conditioning on the number of customers in the system upon arrival of the tagged customer we can write the LST as

$$\mathbf{E}(\mathrm{e}^{-sV}) = \sum_{n=0}^{\infty} \mathbf{E}(\mathrm{e}^{-sV}|Q = n)\mathbf{P}(Q = n). \tag{2.14}$$

Due to the PASTA property the probability that the tagged customer sees n other customers upon arrival is $\mathbf{P}(Q = n) = (1 - \rho)\rho^n$, where ρ is the traffic intensity.
Now we use a special decomposition of the sojourn time given the number of customers in the system upon arrival, first established by Yashkov [27]. Every customer being in the system at $t = 0$ is called a "progenitor" while the new arrivals occurring after $t = 0$ are assumed to be "descendants" of these progenitors. The tagged customer is considered as a progenitor. If n customers are present in the system then each new arrival is declared with probability $1/n$ as a descendant of any of these progenitors. Each branching process is formed by one progenitor and its descendants (for more details see [27]). Therefore, the sojourn time is decomposed into a sum of independent delay elements associated with $n + 1$ progenitors:

$$V|_{(Q=n)} = V_0 + \sum_{i=1}^{n} C_i, \tag{2.15}$$

where C_i, $i = 1, 2, ..., n$, are i.i.d. random variables equal to the sum of attained service of the ith progenitor (customer) and its direct descendants for the time interval during which the tagged customer will be served till completion (sum of all life times up to epoch D). Let R_i be the remaining service requirement of the ith progenitor at the moment of the tagged arrival. R_i is uniformly distributed on the interval $[0, D]$.
Using representation (2.15) the conditional expectation in (2.14) simplifies to

$$\mathbf{E}(\mathrm{e}^{-sV}|Q = n) = \mathbf{E}(\mathrm{e}^{-sV_0}) \left(\mathbf{E}(\mathrm{e}^{-sC_i})\right)^n. \tag{2.16}$$

Let us now derive the transform of the random variable C_i. Conditioning on R_i, we get

$$\mathbf{E}(\mathrm{e}^{-sC_i}) = \frac{1}{D}\int_0^D \mathbf{E}(\mathrm{e}^{-sC_i}|R_i = t)\mathrm{d}t. \tag{2.17}$$

Given $R_i = t$, we can express the conditional expectation $\mathbf{E}(\mathrm{e}^{-sC_i}|R_i = t)$ as in the previous section. However, in this situation we must distinguish between the intervals $[0, t]$ and $[t, D]$. Since no departures happen before t, on the interval $[0, t]$ we can apply ordinary Yule process properties as for V_0. On the interval $[t, D]$, we represent the number of customers in the system as a Yule process as well: the Yule process $Y(s)$, $s \in [0, D - t]$, that starts from a number of customers at the moment $s = 0$: $Y(0) = X(t) - 1$.
Rewriting the conditional expectation

$$\mathbf{E}(\mathrm{e}^{-sC_i}|R_i = t) = \mathbf{E}(\mathrm{e}^{-s\sum_{k=1}^{X(t)}(t-t_k)}\mathrm{e}^{-s\sum_{k=X(t)+1}^{X(D)}(t-t_k)}),$$

and using the memoryless property and (2.5), we have

$$\begin{aligned}
&\mathbf{E}(\mathrm{e}^{-sC_i}|R_i=t) = \mathbf{E}((\mathrm{e}^{-s\sum_{k=1}^{X(t)}(t-t_k)})(\mathrm{e}^{-s((X(t)-1)(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}))\\
= &\sum_{m=0}^{\infty}\mathbf{E}(\mathrm{e}^{-s\sum_{k=1}^{m+1}(t-t_k)}|X(t)=m+1)\\
&\qquad\times\mathbf{E}(\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}|X(t)=m+1)(1-\mathrm{e}^{-\lambda t})^m\mathrm{e}^{-\lambda t}.
\end{aligned}$$

Applying result (2.10) with D replaced by t, we can simplify this expression to obtain

$$\mathbf{E}(\mathrm{e}^{-sC_i}|R_i=t) = \sum_{m=0}^{\infty}\mathrm{e}^{-(\lambda+s)t}\left(\frac{\lambda(1-\mathrm{e}^{-(\lambda+s)t})}{\lambda+s}\right)^m\mathbf{E}(\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}).$$

For the expectation term in the right-hand side we perform a computation using the Yule process that starts from m individuals [25]. If the population starts from i individuals, the population size at epoch t is the sum of i i.i.d. geometric random variables with parameter $\mathrm{e}^{-\lambda t}$. Hence, the population size at epoch t has a negative binomial distribution with parameters i and $\mathrm{e}^{-\lambda t}$. As before the distribution of arrival times t_k is defined by (2.6). Using these facts, we obtain:

$$\begin{aligned}
&\mathbf{E}(\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}) =\\
&=\sum_{l=0}^{\infty}\mathbf{E}(\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{l}(D-t-t_k))}|Y(D-t)=l+m)\mathbf{P}(Y(D-t)=l+m)\\
&=\sum_{l=0}^{\infty}\mathrm{e}^{-(s+\lambda)m(D-t)}\left(\frac{\lambda}{\lambda+s}\right)^l(1-\mathrm{e}^{-(\lambda+s)(D-t)})^l\frac{(l+m-1)!}{(m-1)!l!}\\
&=\left(\frac{(\lambda+s)\mathrm{e}^{-(\lambda+s)(D-t)}}{s+\lambda\mathrm{e}^{-(\lambda+s)(D-t)}}\right)^m. \qquad (2.18)
\end{aligned}$$

Thus, substituting this into the expression for $\mathbf{E}(\mathrm{e}^{-sC_i}|R_i=t)$ we get

$$\begin{aligned}
\mathbf{E}(\mathrm{e}^{-sC_i}|R_i=t) &= \sum_{m=0}^{\infty}\mathrm{e}^{-(\lambda+s)t}\left(\frac{\lambda(1-\mathrm{e}^{-(\lambda+s)t})}{\lambda+s}\right)^m\left(\frac{(\lambda+s)\mathrm{e}^{-(\lambda+s)(D-t)}}{s+\lambda\mathrm{e}^{-(\lambda+s)(D-t)}}\right)^m\\
&= \frac{\lambda+s\mathrm{e}^{(\lambda+s)(D-t)}}{\lambda+s\mathrm{e}^{(\lambda+s)D}}, \qquad (2.19)
\end{aligned}$$

and

$$\mathbf{E}(\mathrm{e}^{-sC_i}) = \frac{1}{D}\int_0^D\mathbf{E}(\mathrm{e}^{-sC_i}|R_i=t)\mathrm{d}t = \frac{\rho(\lambda+s)-s+s\mathrm{e}^{(\lambda+s)D}}{D(\lambda+s)(\lambda+s\mathrm{e}^{(\lambda+s)D})}. \qquad (2.20)$$

Substituting (2.4) and (2.20) into (2.14), we obtain the sojourn time transform

$$\begin{aligned}
\mathbf{E}(\mathrm{e}^{-sV}) &= \frac{(\lambda+s)(1-\rho)}{\lambda+s\mathrm{e}^{(\lambda+s)D}}\sum_{n=0}^{\infty}\rho^n\left(\frac{\rho(\lambda+s)-s+s\mathrm{e}^{(\lambda+s)D}}{D(\lambda+s)(\lambda+s\mathrm{e}^{(\lambda+s)D})}\right)^n\\
&= \frac{(1-\rho)(\lambda+s)^2}{s^2\mathrm{e}^{(\lambda+s)D}+\lambda(s+s(1-\rho)+\lambda(1-\rho))}, \qquad (2.21)
\end{aligned}$$

which coincides with Ott's formula (2.1). □

In the next section, we shall use the equalities (2.19) and (2.20).

3 Tail behavior of the sojourn time

In this section we investigate the behavior of $\mathbf{P}(V > x)$ as $x \to \infty$. The following theorem is our main result.

Theorem 3.1 *As* $x \to \infty$,

$$\mathbf{P}(V > x) \sim \alpha \mathrm{e}^{-\gamma x}, \tag{3.1}$$

where γ *is the real solution of the equation*

$$\frac{\lambda D(\lambda - s) + s - s\mathrm{e}^{(\lambda - s)D}}{D(\lambda - s)(\lambda - s\mathrm{e}^{(\lambda - s)D})} = \frac{1}{\rho}, \quad s \geq 0 \tag{3.2}$$

and

$$\alpha = \frac{(1-\rho)(\lambda - \gamma)}{2\lambda(1-\rho) - \gamma\rho(2-\rho)}. \tag{3.3}$$

Our derivation is based on the results obtained in the previous section. In particular, we will need the moment generating functions of the decomposition random variables V_0 and C_i, appearing in the representation (2.15) of V :

$$\mathbf{E}(\mathrm{e}^{sV_0}) = \frac{\lambda - s}{\lambda - s\mathrm{e}^{(\lambda - s)D}}, \tag{3.4}$$

$$\mathbf{E}(\mathrm{e}^{sC_i}) = \frac{\lambda D(\lambda - s) + s - s\mathrm{e}^{(\lambda - s)D}}{D(\lambda - s)(\lambda - s\mathrm{e}^{(\lambda - s)D})}. \tag{3.5}$$

This section is organized as follows. In Subsection 3.1 we analyze the singularities of the above moment generating functions with respect to s. This enables us to prove Theorem 3.1 with a version of the Cramér-Lundberg theorem for geometric random sums. The proof is given in Subsection 3.2. In Subsection 3.3 we show some implications of the obtained result.

3.1 Singularities of the delay element LST's

Let us consider singularities of the moment generating function $\mathbf{E}(\mathrm{e}^{sV_0})$. It is enough to consider only real values of s, since

$$|\mathbf{E}(\mathrm{e}^{sV_0})| \leq \mathbf{E}(\mathrm{e}^{Re(s)V_0}).$$

Let us rewrite the denominator of $\mathbf{E}(\mathrm{e}^{sV_0})$ as a function $f(s) = \lambda - s\mathrm{e}^{(\lambda - s)D}$. Obviously, singularities of the moment generating function might be only at zeros of the denominator. The trivial zero of $f(s)$ is $s = \lambda$. However, this is a removable singularity of $\mathbf{E}(\mathrm{e}^{sV_0})$: using L'Hospital we obtain that

$$\lim_{s \to \lambda} \mathbf{E}(\mathrm{e}^{sV_0}) = \frac{1}{1-\rho}. \tag{3.6}$$

However, there exists another zero of the function $f(s)$. The derivative of $f(s)$ is determined as $f'(s) = (Ds - 1)\mathrm{e}^{(\lambda - s)D}$ and $f'(s) = 0$ at $s = \frac{1}{D}$. Furthermore, $f(0) = \lambda$, $f(\infty) = \lambda$, $f(\lambda) = 0$ and by stability, $\lambda < \frac{1}{D}$. Since $f'(s) < 0$ for $s < 1/D$ and $f'(s) > 0$ for $s > 1/D$, we can conclude that there is a unique point $\gamma_0 > \frac{1}{D} > \lambda$ such that $f(\gamma_0) = 0$. An important issue is that this point is a pole of the moment generating function: $\mathbf{E}(\mathrm{e}^{\gamma_0 V_0}) = \infty$.

To analyze the behavior of $\mathbf{E}(\mathrm{e}^{sC_i})$, let us consider the conditional moment generating function $\mathbf{E}(\mathrm{e}^{sC_i}|R_i = t)$, $t \in [0, D]$:

$$\mathbf{E}(\mathrm{e}^{sC_i}|R_i = t) = \frac{\lambda - s\mathrm{e}^{(\lambda-s)(D-t)}}{\lambda - s\mathrm{e}^{(\lambda-s)D}}. \tag{3.7}$$

We already know the zeros of the denominator: λ and γ_0. Similarly, λ is a removable singularity, since

$$\lim_{s\to\lambda} \mathbf{E}(\mathrm{e}^{sC_i}|R_i = t) = \frac{1-\rho+\lambda t}{1-\rho}, \; t \in [0, D]. \tag{3.8}$$

However, we still have to check if $\mathbf{E}(\mathrm{e}^{sC_i}|R_i = t)$ has a singularity when $s = \gamma_0$. For this purpose we consider the numerator as a separate function, $f_t(s) = \lambda - s\mathrm{e}^{(\lambda-s)(D-t)}$. As a function of the parameter t, the numerator $f_t(s)$ increases for values $s < \lambda$ and decreases for $s > \lambda$. Since $f_0(\gamma_0) \equiv f(\gamma_0) = 0$ and $\gamma_0 > \lambda$, it follows that $f_t(\gamma_0)$ is strictly negative for any $t > 0$. Hence, γ_0 is a pole: $\mathbf{E}(\mathrm{e}^{\gamma_0 C_i}|R_i = t) = \infty$.
Summarizing this subsection we have

Proposition 3.1 *There exists a unique value $\gamma_0 > \lambda$ that satisfies the equation*

$$\lambda - s\mathrm{e}^{(\lambda-s)D} = 0, \tag{3.9}$$

and that is an abscissa of convergence of both $\mathbf{E}(\mathrm{e}^{sV_0})$ and $\mathbf{E}(\mathrm{e}^{sC_i}|R_i = t)$, $\forall t \in [0, D]$, and consequently, of $\mathbf{E}(\mathrm{e}^{sC_i})$.

We are now ready to give a proof of Theorem 3.1.

3.2 Proof of Theorem 3.1

For convenience, denote the sum $\sum_{i=0}^{N} C_i$ in representation (2.15) by V_1. As before, Q is the number of customers in the system upon arrival. The probability distribution of V_1 can be written as

$$\mathbf{P}(V_1 > x) = \mathbf{P}(\sum_{i=0}^{Q} C_i > x) = \sum_{n=0}^{\infty}(1-\rho)\rho^n(1 - F_n(x)), \tag{3.10}$$

where F denotes the distribution of C_i, and $F_n(x)$ is the n-fold convolution of F with itself. The random variable V_1 is called a geometric random sum and such random sums are used in many applied probability settings. In particular, it is well known ([10], [2], [14]) that $\mathbf{P}(V_1 > x)$ is asymptotically ($x \to \infty$) equivalent to the exponential function

$$\mathbf{P}(V_1 > x) \sim k_C \mathrm{e}^{-\gamma x}, \tag{3.11}$$

if there exists a $\gamma > 0$ (the Cramér exponent) such that it satisfies the Cramér condition

$$\rho \int_0^{\infty} \mathrm{e}^{\gamma x} \mathrm{d}F(x) = 1. \tag{3.12}$$

Note that the Cramér condition (3.12) can be presented as follows:

$$\mathbf{E}(\mathrm{e}^{\gamma C_i}) = \frac{1}{\rho}. \tag{3.13}$$

Since the function $\mathbf{E}(e^{sC_i})$ monotonically increases from 1 to ∞ on the interval $[0, \gamma_0)$ (by Proposition 3.1), for any nonzero value of ρ there exists a unique real solution γ of the preceding equation, $\gamma < \gamma_0$.
Substituting Expression (2.20) for $\mathbf{E}(e^{sC_i})$ into Equation (3.13) we get (cf. Equation (3.2)):

$$\mathbf{E}(e^{\gamma C_i}) = \frac{\lambda D(\lambda - \gamma) + \gamma - \gamma e^{(\lambda-\gamma)D}}{D(\lambda - \gamma)(\lambda - \gamma e^{(\lambda-\gamma)D})} = \frac{1}{\rho}.$$

So we have proved:

Proposition 3.2 *There exists a unique solution γ of Equation (3.2), $\gamma < \gamma_0$, that is an abscissa of convergence of $\mathbf{E}(e^{sV_1})$.*

To determine the coefficient k_C we apply the following theorem.

Theorem 3.2 ([2], [14]) *Let the Cramér condition (3.12) hold.*
If $g = \rho \int_0^\infty x e^{\gamma x} dF(x) = \rho \frac{d}{ds}\mathbf{E}(e^{sC_i})|_{s=\gamma} < \infty$, and F is non-lattice, then the asymptotic relation (3.11) holds with

$$k_C = \frac{1-\rho}{g\gamma}. \tag{3.14}$$

Since the moment generating function $\mathbf{E}(e^{sC_i})$ is differentiable in point $s = \gamma$, $\gamma < \gamma_0$, it follows that $g < \infty$ and we can determine the coefficient k_C and asymptotics for $\mathbf{P}(V_1 > x)$. Determining the derivative of the moment generating function, performing some simplification and substituting Condition (3.13) we obtain

$$\frac{d}{ds}\mathbf{E}(e^{sC_i})|_{s=\gamma} = \frac{\gamma\rho(\rho-2) + 2\lambda(1-\rho)}{\rho\gamma(\lambda - \gamma e^{(\lambda-\gamma)D})}. \tag{3.15}$$

Hence, by Theorem 3.2, the coefficient k_C is

$$k_C = \frac{1-\rho}{\gamma\rho\frac{d}{ds}\mathbf{E}e^{sC_i}|_{s=\gamma}} = \frac{(1-\rho)(\lambda - \gamma e^{(\lambda-\gamma)D})}{\gamma\rho(\rho-2) + 2\lambda(1-\rho)}. \tag{3.16}$$

Finally, F is non-lattice, since $\mathbf{P}(C_i = R_i) > 0$, and R_i has a density.
Summarizing the above, we obtain the asymptotic behavior of the random variable V_1.

Proposition 3.3 *Let γ be the solution of Equation (3.2). Then*

$$\mathbf{P}(V_1 > x) \sim \frac{(1-\rho)(\lambda - \gamma e^{(\lambda-\gamma)D})}{\gamma\rho(\rho-2) + 2\lambda(1-\rho)} e^{-\gamma x}, \qquad x \to \infty. \tag{3.17}$$

Knowing the LST of the first customer's sojourn time V_0 and the asymptotic tail behavior of V_1, we can derive an expression for the tail behavior of the sojourn time V.
Since V_1 has an asymptotically exponential tail and $\mathbf{E}(e^{(\gamma+\varepsilon)V_0}) < \infty$ for any $0 < \varepsilon < \gamma_0 - \gamma$ we can apply Breiman's theorem to e^{V_0}, e^{V_1} (see [7]):

$$\mathbf{P}(V > x) = \mathbf{P}(V_0 + V_1 > x) = \mathbf{P}(e^{V_0}e^{V_1} > e^x) \sim \mathbf{E}(e^{\gamma V_0})\mathbf{P}(V_1 > x), \ x \to \infty. \tag{3.18}$$

The above proposition and substitution of γ in (2.4) imply (3.1). $\square$

Remark 3.1 An interesting issue, raised in the introduction, is how the number of customers in the system is affecting a large sojourn time. Mandjes and Zwart [19] have shown that, in PS queues with for example phase-type service times, the initial number of customers is of $o(x)$ when $V > x$. In this remark, we show that this picture drastically changes when service times are deterministic.
The proof of Theorem 3.1 indicates that the realizations of the C_i's in the representation

$$V = V_0 + \sum_{i=1}^{N} C_i$$

are sampled from the exponentially tilted density $\mathrm{e}^{\gamma x}\mathrm{d}\mathbf{P}(C_i \leq x)/\mathbf{E}(\mathrm{e}^{\gamma C_i})$.
Under this density, the expected value of the C_i is

$$\mathbf{E}(C_i \mathrm{e}^{\gamma C_i})/\mathbf{E}(\mathrm{e}^{\gamma C_i}) = \rho\mathbf{E}(C_i \mathrm{e}^{\gamma C_i}) =: c(\gamma).$$

Thus, in order for V to be of size x, N should be around $x/c(\gamma)$.

Remark 3.2 When $s = \lambda$, the denominator and the numerator of Expression (3.2) are both equal to zero. Using L'Hospital we get

$$\lim_{s\to\lambda} \mathbf{E}(\mathrm{e}^{sC_i}) = \frac{2-\rho}{2(1-\rho)}.$$

Solving the equation

$$\frac{2-\rho}{2(1-\rho)} = \frac{1}{\rho},$$

we obtain that, for $\rho = 2 - \sqrt{2}$, Equation (3.2) has a solution $\gamma = \lambda$.
Since the asymptotic constant α in (3.1) has a removable singularity at this value the tail behavior of V becomes:

$$\mathbf{P}(V > x) \sim \frac{1-\rho}{\rho(2-\rho)}\mathrm{e}^{-\lambda x} = \frac{1}{2}\mathrm{e}^{-\lambda x}, \qquad x \to \infty. \tag{3.19}$$

3.3 Implications of Theorem 3.1

In this subsection we treat a number of implications of our main result. First, we take a look at the relationship between the decay rate in the M/D/1 PS queue and decay rates in queues with FIFO and LIFO disciplines. Secondly, we consider the behavior of the decay rate γ in heavy traffic.

Other service disciplines
First we consider the FIFO service discipline. Due to a result of Ramanan and Stolyar [22] it follows that FIFO is optimal among all single-class work-conserving disciplines, i.e. it maximizes the decay rate. We can easily show that $\gamma_{FIFO} > \gamma_{PS}$. Recall ([4], Theorem XIII.5.2) that γ_{FIFO} is a solution of the equation $\rho\mathbf{E}(\mathrm{e}^{sB^{res}}) = 1$, where B^{res} is the remaining service time, which is R_i in the notation of Section 2. Using Equation (3.13) and the definition of C_i we get:

$$\mathbf{E}(\mathrm{e}^{\gamma_{FIFO} R_i}) = \mathbf{E}(\mathrm{e}^{\gamma_{PS} C_i}) > \mathbf{E}(\mathrm{e}^{\gamma_{PS} R_i}).$$

The decay rate inequality $\gamma_{FIFO} > \gamma_{PS}$ follows from monotonicity of the moment generating functions on the interval $[0, \gamma_0]$.

The strict inequality was also shown in [19] for the GI/G/1 PS queue for a class of light-tailed service time distributions excluding deterministic service time. Moreover, in [19] it was shown that in the M/D/1 queue the decay rate under the LIFO discipline does not exceed the decay rate in the PS case, $\gamma_{LIFO} < \gamma_{PS}$.
Table 1 shows decay rates for the M/D/1 queue with PS, FIFO and LIFO disciplines. For convenience, we take $D = 1$. A result in Cox and Smith [8] implies that in this case $\gamma_{LIFO} = -\log\rho - (1-\rho)$.

ρ	0.2	0.4	0.6	0.8
PS	1.9227	1.0462	0.5578	0.2331
FIFO	2.6604	1.6188	0.9474	0.4308
LIFO	0.8094	0.3163	0.1108	0.0231

Table 1: Asymptotic decay rates for the M/D/1 queue with PS, FIFO and LIFO disciplines.

The small value of γ_{LIFO} for $\rho = 0.8$ is related to the fact that $\gamma_{LIFO} = O((1-\rho)^2)$ as $\rho \to 1$, opposed to $\gamma_{FIFO} = O((1-\rho))$. In the next subsection, we show that in heavy traffic, γ_{PS} behaves like γ_{FIFO}.

Heavy traffic

Let us now study the sojourn time of a customer under heavy traffic, i.e. when the traffic intensity $\rho \to 1$.

Proposition 3.4 *Let γ and α be defined as in Theorem 3.1. Then, as $\rho \to 1$, the decay rate $\gamma \sim \lambda(1-\rho)$ and the coefficient $\alpha \to 1$.*

Proof: Obviously, when the traffic intensity $\rho \to 1$, the decay rate γ is converging to zero (see Equation (3.13)). Let us study the behavior of γ near zero in more detail. We expand the left-hand side of (3.13) into a two-term Taylor series: $\mathbf{E}(e^{\gamma C_i}) = \mathbf{E}(1 + \gamma C_i + O(\gamma^2))$. The second-order term is $O(\gamma^2)$ uniformly in ρ, since the second moment $\mathbf{E}(C_i^2)$ is finite if $\rho = 1$. The smoothness of the moment generating function $\mathbf{E}(e^{\gamma C_i})$ near zero implies that all moments of C_i are finite. To calculate the first moment $\mathbf{E}C_i$, let us take the derivative of the moment generating function at zero: $\mathbf{E}C_i = \frac{2-\rho}{\rho\lambda} \to \frac{1}{\lambda}$, and hence due to $\rho\mathbf{E}(e^{\gamma C_i}) = \rho + \gamma\mathbf{E}C_i + O(\gamma^2) = 1$, we get that

$$\gamma(1/\lambda + o(1)) = 1 - \rho,$$

and $\gamma \sim \lambda(1-\rho)$.
Substitution of the expression for γ into (3.3) gives the behavior of the asymptotic constant α :

$$\alpha = \frac{(1-\rho)(\lambda-\gamma)}{2\lambda(1-\rho) - \gamma\rho(2-\rho)} \sim \frac{(1-\rho)(\lambda - \lambda(1-\rho))}{2\lambda(1-\rho) - \lambda(1-\rho)\rho(2-\rho)} \to 1. \tag{3.20}$$

□

Remark 3.3 The above heavy-traffic behavior is related to a result of Yashkov in [28]. He derived a heavy-traffic limit result for the sojourn time in the M/G/1 PS queue conditioned on the service requirement. Replacing s in (2.1) by $(1-\rho)s$ and taking the limit when $\rho \to 1$ we have:

$$\lim_{\rho\to 1} \mathbf{E}(e^{-(1-\rho)sV}) = \frac{\lambda}{\lambda + s}. \tag{3.21}$$

Since the limiting value is the LST of the exponential distribution with parameter λ, we obtain the heavy-traffic approximation

$$\mathbf{P}(V > x) \approx \mathrm{e}^{-\lambda(1-\rho)x}. \tag{3.22}$$

Hence, summarizing Proposition 3.4 and Remark 3.3,

$$\lim_{\rho\to 1}\lim_{x\to\infty} \frac{\mathbf{P}((1-\rho)V > x)}{\alpha \mathrm{e}^{-\gamma x/(1-\rho)}} = \lim_{x\to\infty}\lim_{\rho\to 1} \frac{\mathbf{P}((1-\rho)V > x)}{\alpha \mathrm{e}^{-\gamma x/(1-\rho)}} = 1. \tag{3.23}$$

This suggests that the asymptotics given in Theorem 3.1 provide a good approximation of the sojourn time tail behavior if ρ is close to 1. The results in the next section confirm this.

4 Numerical experiments

In this section we present some numerical results. In particular, we compare the behavior of the sojourn time tail computed numerically from Ott's formula (2.1) with the asymptotics we have obtained. In Ott's formula the sojourn time distribution is expressed in terms of its LST.

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 that allow for computing probabilities and other quantities without any complication. In our study we will compute the sojourn time distribution using the inversion algorithm of Den Iseger [9] and will perform a cross-check with the algorithm proposed by Abate and Whitt [3]. Both methods are known to perform with high accuracy, and produced similar results. Since the sojourn time distribution has a jump at point D, we will apply the modified Den Iseger algorithm for functions with discontinuities.

Table 2 shows computational results for various arrival rates and service requirements normalized to $D = 1$. For each value of ρ, the first column shows, for different values of x, the approximation (3.1) for $\mathbf{P}(V > x)$ based on the asymptotic expansion. The second column presents the estimates derived with the Den Iseger inversion algorithm.

	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
x	asympt.	LST inv.	asympt.	LST inv.	asympt.	LST inv.
5	1,09356-02	1,09364-02	9,05070-02	9,05076-02	3,67385-01	3,67385-01
10	5,84744-05	5,84744-05	5,56564-03	5,56564-03	1,14522-01	1,14522-01
15	3,12671-07	3,12671-07	3,42253-04	3,42253-04	3,56993-02	3,56993-02
20	1,67189-09	1,67189-09	2,10465-05	2,10465-05	1,11283-02	1,11283-02
25	8,93986-12	8,93984-12	1,29423-06	1,29423-06	3,46896-03	3,46896-03
30	4,78027-14	4,80726-14	7,95876-08	7,95876-08	1,08135-03	1,08135-03
35	2,55608-16	2,11279-16	4,89416-09	4,89416-09	3,37084-04	3,37084-04
40	1,36677-18	1,21772-18	3,00961-10	3,00961-10	1,05077-04	1,05077-04

Table 2: Asymptotic approximation and numerical results.

The result shows a remarkable accuracy of the asymptotic tail approximation. The numbers obtained with LST inversion and the asymptotic formula differ sometimes within 10^{-16}, which is in fact the maximum accuracy of the inversion algorithm. Moreover, the asymptotics perform well even for relatively small values of x. Already for $x = 10$ the

x	LST inversion	asympt.(3.1)	heavy-traffic asympt.(3.22)
10	6,17856022E-01	6,17856022E-01	6,21885056E-01
30	2,19011860E-01	2,19011860E-01	2,40508463E-01
50	7,76332889E-02	7,76332889E-02	9,30144892E-02
70	2,75187267E-02	2,75187267E-02	3,59725188E-02
90	9,75458251E-03	9,75458251E-03	1,39120487E-02

Table 3: Asymptotic approximations and numerical results for $\rho = 0.95$.

error is of order 10^{-13}. Results with similar accuracy of exponential asymptotics in FIFO queues are presented in the paper of Abate *et al.* ([1], Table 1).
Table 3 presents results for the model with high load, $\rho = 0.95$. As before, the service requirement D is equal to 1. We consider two approximations: the asymptotic approximation from Theorem 3.1 (second column), and the heavy-traffic asymptotics (3.22) (third column). The first column shows the results from the numerical inversion. Remarkably, the heavy-traffic asymptotics perform less accurately than Approximation (3.1).

References

[1] Abate, J., Choudhury, G.L., Whitt, W. (1994). Waiting-time tail probabilities in queues with long tail service-time distributions. *Queueing Systems* **16**, 311–338.

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

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

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

[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., Van Ooteghem, D., Zwart B. (2005). Tail asymptotics for discriminatory processor sharing queues with heavy-tailed service requirements. To appear in *Performance Evaluation* (special issue on heavy tails and LRD).

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

[8] Cox, D.R., Smith, W.L. (1961). *Queues.* Methuen, London.

[9] Den Iseger, P. (2005). Numerical inversion of Laplace transforms using a Gaussian quadrature rule for the Poisson summation. To be submitted.

[10] Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Volume II.* Wiley, New York.

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

[12] Guillemin, F., Robert, P., Zwart, A.P. (2003). Tail asymptotics for processor-sharing queues. *Advances in Applied Probability* **36**, 525–543.

[13] Jelenković, P., 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] Kella, O., Zwart, A.P., Boxma, O.J. (2005). Some transient properties of symmetric M/G/1 queues. *Journal of Applied Probability* **42**, 223–234.

[16] Kleinrock, L. (1964). Analysis of a time-shared processor. *Naval Research Logistics Quarterly* **11**, 59–73.

[17] Kleinrock, L. (1976). *Queueing Systems.* John Wiley, New York.

[18] Kleinrock, L. (1976). Time-shared systems: A theoretical treatment. *Journal of the Association for Computing Machinery* **14**, 242–261.

[19] Mandjes, M., Zwart A.P. (2004). Large deviations for sojourn times in Processor Sharing queues. *SPOR-Report 2004-09,* Eindhoven University of Technology.

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

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

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

[23] Rege, K.M., Sengupta, B. (1994). A decomposition theorem and related results for the discriminatory processor sharing queue. *Queueing Systems* **18**, 333–351.

[24] Roberts, J.W. (2000). Engineering for quality of service. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 401–420.

[25] Ross, S.M. (1996). *Stochastic Processes.* John Wiley, New York.

[26] Tijms, H.C. (2003). *A First Course in Stochastic Models.* John Wiley, Chichester.

[27] 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.

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

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