Queueing Syst (2006) 53: 7–18
DOI 10.1007/s11134-006-7583-y

Insensitive bounds for the moments of the sojourn time distribution in the M/G/1 processor-sharing queue

Sing-Kong Cheung · Hans van den Berg · Richard J. Boucherie

Abstract This paper studies the M/G/1 processor-sharing (PS) queue, in particular the sojourn time distribution conditioned on the initial job size. Although several expressions for the Laplace-Stieltjes transform (LST) are known, these expressions are not suitable for computational purposes. This paper derives readily applicable insensitive bounds for *all* moments of the conditional sojourn time distribution. The *instantaneous* sojourn time, i.e., the sojourn time of an infinitesimally small job, leads to insensitive upper bounds requiring only knowledge of the traffic intensity and the initial job size. Interestingly, the upper bounds involve polynomials with so-called Eulerian numbers as coefficients. In addition, stochastic ordering and moment ordering results for the sojourn time distribution are obtained.

Keywords M/G/1 PS · Conditional sojourn time · Moments · Insensitive bounds · Instantaneous sojourn time · Euler's number triangle · Moment ordering · Permanent customers

AMS Subject Classification: 60K25, 60E15

This work has been partially funded by the Dutch Ministry of Economic Affairs under the program 'Technologische Samenwerking ICT-doorbraakprojecten', project TSIT1025 BEYOND 3G.

S.-K. Cheung (✉) · R. J. Boucherie
University of Twente, Department of Applied Mathematics, Stochastic Operations Research Group, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: S.K.Cheung@utwente.nl; R.J.Boucherie@utwente.nl

H. van den Berg
TNO Information and Communication Technology, P.O. Box 5050, 2600 GB Delft, The Netherlands;
University of Twente, Department of Computer Science, Design and Analysis of Communication Systems, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: J.L.vandenBerg@telecom.tno.nl

1. Introduction

With the introduction of time-sharing computing in the 1960s, people became interested in the processor-sharing (PS) service discipline as the idealization of time-sharing queueing models, as initiated by Kleinrock [15, 16]. Nowadays, the PS discipline is of considerable interest in many application areas in which different users receive a share of a scarce common system resource. In particular, in the field of the performance evaluation of computer and communication systems, the PS discipline has been widely adopted as a convenient paradigm for modeling bandwidth sharing (e.g., see [2, 9]). For instance, one of the many applications of PS models is in the resource contention of the IEEE 802.11 Wireless Local Area Networks (WLANs), see [19].

In these types of communication networks, the most appropriate Quality-of-Service (QoS) measure from a user's perspective is the file transfer time $V(\tau)$ given that the user wanted to transfer a file of a given size $\tau > 0$. An important feature of *egalitarian* PS is that the conditional expected file transfer time $\mathbb{E}V(\tau)$ can be computed explicitly and grows linearly in τ, which reflects the 'fair' allocation of resources to the served flows. Moreover, it is insensitive to the flow (file) size distribution, depending on its mean only. Characterizing the distribution of $V(\tau)$ is an important problem.

The exact determination of the distribution of the conditional sojourn time $V(\tau)$ given its initial service requirement $\tau > 0$ (file transfer time given its initial file size) in the M/G/1 PS queue was an open problem for a long time. Several analytic solutions have been obtained; see Yashkov [31], Ott [21], Schassberger [23]. More recently, Zwart and Boxma [35] derived a new expression for the Laplace-Stieltjes transform (LST) of the sojourn time distribution, which avoids the complex contour integrals of the previous results. However these expressions are still fairly complex, and not readily applicable from a practical point-of-view. The known expressions lead to, at best, complicated recursive formulas for the moments which have mainly been examined only asymptotically, see e.g. [35].

In the present paper, we derive new results for the moments of the conditional sojourn time $V(\tau)$ in the M/G/1 PS queue, and we also study the sojourn time when the initial service requirement is arbitrarily small. We define $\widehat{V}(\tau)$ as the *instantaneous* sojourn time, i.e., the sojourn time of a customer with infinitesimally small service requirement. We show that the instantaneous sojourn time for arbitrarily small $\tau > 0$ leads to a moment ordering result between $\widehat{V}(\tau)$ and $V(\tau)$ for arbitrary $\tau > 0$. More specifically, our main result is that the moments of the instantaneous sojourn time provide upper bounds for all moments of the conditional sojourn time, which generalizes the upper bound for the second moment in Van den Berg [3]. Additionally, stochastic ordering results for the M/G/1 PS queue and also for PS models with a random number of permanent customers are obtained.

The upper bounds have the valuable characteristic of insensitivity requiring only knowledge of the traffic intensity ρ, and not of higher moments of the service requirement distribution. The upper bounds are also tight in a few appropriate senses, namely for all jobs with a small service requirement ($\tau \to 0$), and for all jobs in systems with heavy-traffic ($\rho \to 1$) or light-traffic ($\rho \to 0$). The latter valuable property follows from the fact that for $\rho \to 0$, the upper bounds coincide with the insensitive lower bounds given by the Jensen's inequality.

The paper is organized as follows. In Section 2 we give a short review of the M/G/1 PS queue. In Section 3 we establish the existence of insensitive upper (and lower) bounds for *all* conditional moments of the sojourn time, with a particular polynomial structure in ρ. Then, in Section 4, the instantaneous sojourn time $\widehat{V}(\tau)$ is introduced and we give readily applicable expressions using the so-called Eulerian numbers. In Section 5, we prove our main result that the moments of $\widehat{V}(\tau)$ provide upper bounds for the moments of $V(\tau)$, via stochastic comparison and moment ordering techniques. In addition, the stochastic ordering results proven in Section 5 provide simple characterizations of $V(\tau)$ under PS, which provide further support for the observation that the egalitarian PS service discipline is 'fair' from a tagged customer's perspective.

2. Preliminaries

In this section we introduce the notation used in the paper and give a short review of the M/G/1 PS queue. Customers arrive according to a Poisson process with rate $\lambda > 0$. Their service requirements are generally distributed with distribution function $B(x)$ and $B(0+) = 0$. Let β_k denote the k-th moment of the service requirement distribution. Every customer is being served with rate $1/n$, when $n > 0$ customers are present in the system. Assume that the workload is less than one, i.e., $\rho := \lambda\beta_1 < 1$, so the system allows a steady state.

The steady-state queue length distribution π_n is geometrically distributed and only depends on the workload (cf. [22]): $\pi_n = (1-\rho)\rho^n$, for $n \in \{0, 1, \dots\}$. We let $V(\tau)$ denote the (conditional) sojourn time of a customer entering the system in steady state having a service requirement of τ upon arrival. Define the k-th moment by $v_k(\tau) = \mathbb{E}V(\tau)^k$. The first moment of $V(\tau)$ is given by $v_1(\tau) = \tau/(1-\rho)$, see e.g. [8, 17].

For $\tau \geq 0$ define the Laplace-Stieltjes transform (LST) of $V(\tau)$ by $v(s,\tau) = \mathbb{E}\left[e^{-sV(\tau)}\right]$, $\operatorname{Re} s \geq 0$. Yashkov [31] derived an expression for $v(s,\tau)$ by writing the sojourn time as a functional on a branching process. Via different approaches, similar results for $v(s,\tau)$ were obtained by Ott [21], Schassberger [23], and Van den Berg and Boxma [4].

The expression for $v(s,\tau)$ obtained by Zwart and Boxma [35] which avoids the contour integrals in the expressions of [31, 21, 23], is the most suitable one for our purposes. They showed that $v(s,\tau)$ can be written as

$$v(s,\tau) = \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\alpha_n(\tau)\right)^{-1}, \tag{2.1}$$

where the coefficients $\alpha_n(\tau)$ are related to the waiting-time distribution in an equivalent M/G/1 queue with First Come First Served (FCFS) service discipline: $\alpha_0(\tau) := 1$, and for $n \geq 1$,

$$\alpha_n(\tau) = \frac{n}{(1-\rho)^n}\int_{x=0}^{\tau}(\tau - x)^{n-1}R^{(n-1)*}(x)dx, \tag{2.2}$$

with $R^{n*}(x)$ denoting the n-fold convolution of the waiting-time distribution $R(x)$ in the M/G/1 FCFS queue. It can be shown that for $n \geq 1$ (cf. [35]):

$$R^{n*}(x) = (1-\rho)^n \sum_{m=0}^{\infty}\binom{m+n-1}{n-1}\rho^m \widetilde{B}^{m*}(x), \quad \text{and}$$
$$R^{0*}(x) := 1, \tag{2.3}$$

where $\widetilde{B}^{m*}(x)$ is the m-fold convolution of the integrated tail or excess service requirement distribution $\widetilde{B}(x) = \frac{1}{\beta_1}\int_0^x(1 - B(u))du$.

As a consequence of the form of the LST in (2.1), it is shown in [35] that the moments $v_k(\tau)$ can be calculated recursively, as $v_0(\tau) := 1$ and for $k \geq 1$,

$$v_k(\tau) = -\sum_{j=1}^{k} \binom{k}{j} v_{k-j}(\tau)\alpha_j(\tau)(-1)^j. \tag{2.4}$$

In particular, it holds that $v_1(\tau) = \alpha_1(\tau) = \tau/(1-\rho)$, and $v_2(\tau) = 2\tau^2/(1-\rho)^2 - \alpha_2(\tau)$.

For additional and related work on the M/G/1 PS queue, readers may refer to Asare and Foster [1], Yashkov [32, 33, 34], Grishechkin [11], Kitayev [12], Whitt [30], Ward and Whitt [29], and Núñez-Queija [20].

3. Upper and lower bounds for the conditional sojourn time

In this section, we establish insensitive bounds for all moments of the conditional sojourn time distribution, which have the form: $1 \leq (1-\rho)^k v_k(\tau)/\tau^k \leq \phi_{k-1}(\rho)$, where $\phi_{k-1}(\rho)$ is a polynomial in ρ of (at most) degree $k-1$ and with non-negative coefficients.

For the second moment of the conditional sojourn time in the M/G/1 PS queue, Van den Berg [3] obtained the following simple bounds:

$$\frac{1}{(1-\rho)^2}\tau^2 \leq v_2(\tau) \leq \frac{1+\rho}{(1-\rho)^2}\tau^2, \tag{3.1}$$

simply by using the fact that $R(0) = 1-\rho > 0$. We note that the upper bound for the second moment is $100\rho\%$ larger than the lower bound, and these bounds only depend on the mean service requirement and *not* on the second and higher moments. From (3.1) it is also interesting to note that $V(\tau)$ has a coefficient of variation less than $\sqrt{\rho}$.

By using the recursive formula (2.4) for $v_k(\tau)$ and 'ignoring' the alternating term $(-1)^j$, the following crude upper bound for all moments can be given: $v_k(\tau) \leq k!\,((e-1)\tau)^k/(1-\rho)^k$, see also Zwart [36]. As a consequence of this bound, the sojourn time $V(\tau)$ is always light-tailed conditional upon its service requirement. Intuitively it supports the conjecture that a large sojourn time is not due to excessive behavior of other customers present in the system.

The crude bound for the second moment is always worse than the upper bound given in (3.1), since $1+\rho < 2 < 2!(e-1)^2$. Furthermore, for $\rho \to 0$, we have the attractive property that the upper and lower bound in (3.1) coincide. Now we will generalize (3.1) for *all* moments, by using the recursive formula (2.4) and using the simple observation as for the second moment in [3], to obtain 'tight' bounds with a similar structure as (3.1).

Theorem 3.1. *For all $k \geq 2$, there exist non-negative constants $c_i^k \geq 0$, such that $v_k(\tau)$ is bounded by*

$$\frac{1}{(1-\rho)^k}\tau^k \leq v_k(\tau) \leq \frac{\phi_{k-1}(\rho)}{(1-\rho)^k}\tau^k, \tag{3.2}$$

where $\phi_{k-1}(\rho) = \sum_{i=0}^{k-1} c_i^k \rho^i$ is a polynomial in ρ of degree $k-1$ (if k even) or $k-2$ (if k odd) and $c_0^k = 1$.

Proof: The lower bound in (3.2) is straightforward by applying the Jensen's inequality. For the upper bound, we note that $1-\rho \leq R(x) \leq 1$ and hence $(1-\rho)^n \leq R^{n*}(x) \leq 1$. Therefore, by using (2.2) we obtain upper and lower bounds for $\alpha_n(\tau)$:

$$\frac{\tau^n}{1-\rho} \leq \alpha_n(\tau) \leq \frac{\tau^n}{(1-\rho)^n}, \quad \textit{for } n \geq 1, \textit{ and } \alpha_0(\tau) := 1. \tag{3.3}$$

Existence of the upper bound in (3.2) is obtained by induction. We rewrite the recursive formula (2.4) as $\overline{v}_k(\tau) = -\sum_{j=1}^{k}\binom{k}{j}\overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau)(-1)^j$, where $\overline{v}_k(\tau) := (1-\rho)^k v_k(\tau)/\tau^k$ and $\overline{\alpha}_j(\tau) := (1-\rho)^j\alpha_j(\tau)/\tau^j$. From (3.3), bounds for $\overline{\alpha}_j(\tau)$ are: $(1-\rho)^{j-1} \leq \overline{\alpha}_j(\tau) \leq 1$, for $j \geq 1$. Assume (induction hypothesis) that the following bounds hold for $\overline{v}_{k-1}(\tau), \overline{v}_{k-2}(\tau), \ldots$:

$$1 \leq \overline{v}_{k-j}(\tau) \leq \phi_{k-j-1}(\rho),$$

and where the bounds for $\overline{v}_0(\tau)$ and $\overline{v}_1(\tau)$ are satisfied by definition, since $\overline{v}_0(\tau) := 1$, and $\overline{v}_1(\tau) := 1$. Then, we have bounds for the product $\overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau)$, *for* $j = 1, \ldots, k$*:*

$$\begin{aligned} 1 &+ \sum_{i=1}^{j-1}(-\rho)^i\binom{j-1}{i} \\ &= (1-\rho)^{j-1} \leq \overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau) \leq \phi_{k-j-1}(\rho) \\ &= 1 + \sum_{i=1}^{k-j-1} c_i^{k-j}\rho^i. \end{aligned}$$

Now, apply induction and take into account the alternating term $(-1)^j$ (hence, we need both upper and lower bounds for $\overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau)$) to obtain the upper bound for $\overline{v}_k(\tau)$. Hence, straightforward term-by-term bounding (and 'splitting the positive and negative terms' in the recursive formula) gives:

$$\begin{aligned}\overline{v}_k(\tau) &= \sum_{\substack{j=1\\ j:\text{odd}}}^{k} \binom{k}{j} \overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau) - \sum_{\substack{j=2\\ j:\text{even}}}^{k}\binom{k}{j}\overline{v}_{k-j}(\tau)\overline{\alpha}_j(\tau)\\ &\le \sum_{\substack{j=1\\ j:\text{odd}}}^{k} \binom{k}{j}\left\{1+\sum_{i=1}^{k-j-1} c_i^{k-j}\rho^i\right\} - \sum_{\substack{j=2\\ j:\text{even}}}^{k}\binom{k}{j}\\ &\quad \left\{1+\sum_{i=1}^{j-1}(-\rho)^i\binom{j-1}{i}\right\} \equiv \sum_{i=0}^{k-1} c_i^k\rho^i. \qquad (3.4)\end{aligned}$$

By definition of the coefficients c_i^k in (3.4) and by comparing the terms, it is not difficult to see that: $c_0^k = 1$ for all k, $c_{k-1}^k = 1$ if k is even, and $c_{k-1}^k = 0$ if k is odd. Furthermore, it can be shown that $c_i^k \geq 0$, where the coefficients are constructed as in (3.4). However, for the existence of an upper bound of the described structure, it is not necessary to show that $c_i^k \geq 0$, since c_i^k can always be chosen sufficiently large (and finite). □

Remark 3.2. In principle, we can apply the 'alternating' procedure to obtain a lower bound as well. However, the resulting lower bound is always worse than Jensen's lower bound. Jensen's lower bound is obtained via the recursive formula if the coefficient $\alpha_n(\tau)$ is replaced by the upper bound $\tau^n/(1-\rho)^n$ for all $n \geq 1$. Hence, the procedure of recursively term-by-term bounding as in the proof of Theorem 3.1 for obtaining a lower bound (as well as for an upper bound) for $v_k(\tau)$, is too conservative. The latter fact can also be argued from the dependency of the coefficients $\{\alpha_n(\tau), \alpha_{n+1}(\tau), \dots\}$, $n \geq 2$. For example, if $\alpha_n(\tau)$ is 'close to its lower bound' $\tau^n/(1-\rho)$, then $\alpha_{n+1}(\tau)$ is generally 'not close to its upper bound' $\tau^{n+1}/(1-\rho)^{n+1}$. In fact, for a fixed $\tau > 0$, if it holds that $\alpha_n(\tau) = \tau^n/(1-\rho)$ for *some* $n \geq 2$, then necessarily $\alpha_n(\tau) = \tau^n/(1-\rho)$ for *all* $n \geq 2$. The latter observation will be important (see the similar observation in Lemma 5.8); the sequence $\alpha_n(\tau) = \tau^n/(1-\rho)$ for $n \geq 1$, also uniquely defines the so-called instantaneous sojourn time, which will be introduced in Section 4.

For the second moment, we obtain $c_1^2 = 1$ since $k = 2$ is even, and the upper bound is the same as in (3.1). As a direct consequence of Theorem 3.1 we have the following Corollary 3.3, which states that all conditional moments are finite in the stable M/G/1 PS system. This result is in sharp contrast with the stable M/G/1 FCFS queue, which provides further support for the observation that PS is a 'fair' service discipline.

Corollary 3.3. *If $\rho < 1$, then $v_k(\tau) < \infty$ for all $k \geq 1$.*

For the moments of the sojourn time in the stable M/G/1 FCFS queue, it is known that the k-th moment exists if and only if β_{k+1} is finite. For the PS case, the k-th moment of the (unconditional) sojourn time exists if and only if β_k is finite (see [35]).

4. The instantaneous sojourn time

In this section we introduce the *instantaneous* sojourn time $\widehat{V}(\tau)$, defined as the sojourn time of an infinitesimally small job. The key idea is as follows. A customer with a (very small) initial service requirement $\tau > 0$ arrives at the system in steady state, say at time t_0. By the PASTA property, the tagged customer sees n other customers upon arrival with probability $\pi_n = (1-\rho)\rho^n$. If we denote the remaining service requirements of the n other customers at time t_0 by x_i, $i = 1, \dots, n$, then we may assume that $\tau << \min_{i=1,\dots,n} x_i$. Furthermore, we assume that τ is small enough such that no other customers arrive during the time interval $[t_0, t_0 + (n+1)\tau)$. Under these assumptions, it is as if the tagged customer arrived at a system with *n permanent* customers with probability π_n and with no other arriving customers. Hence, the sojourn time of the tagged customer is $(n+1)\tau$ with probability π_n. We define the *instantaneous* sojourn time as $\widehat{V}(\tau) = (N+1)\tau$, where N is distributed as $\mathbb{P}(N = n) = \pi_n$. The k-th moment of the true sojourn time $V(\tau)$ can be approximated with the k-th moment of the instantaneous sojourn time $\widehat{v}_k(\tau)$, as $\tau \to 0$,

$$v_k(\tau) \approx \widehat{v}_k(\tau) := \mathbb{E}\widehat{V}(\tau)^k = \sum_{n=0}^{\infty} \pi_n \left\{(n+1)\tau\right\}^k, \quad k \in \mathbb{N}.$$

Clearly, it holds that $\widehat{v}_1(\tau) = \mathbb{E}(N+1)\tau = \tau/(1-\rho)$, and thus the instantaneous sojourn time $\widehat{V}(\tau)$ as an approximation for $V(\tau)$, is exact for the first moment $v_1(\tau)$ and even for arbitrary $\tau > 0$. This will be an important fact for the rest of the paper. It can also be shown that $V(\tau)/\tau \overset{d}{\to} N+1$, as $\tau \to 0$; the convergence in distribution (denoted by $\overset{d}{\to}$) follows from:

$$\lim_{\tau \to 0} \frac{\alpha_n(\tau)}{\tau^n} = \frac{1}{1-\rho}, \quad \text{for } n \geq 1,$$

cf. (2.2) and (2.3), and hence $v(s/\tau;\tau)^{-1} = 1 + \sum_{n=1}^{\infty} \frac{s^n}{n!}\frac{\alpha_n(\tau)}{\tau^n}$ converges to $\frac{e^s-\rho}{1-\rho}$ (as $\tau \to 0$), which is the reciprocal of the LST of $(N+1)$, see also the last comment in Remark 3.2.

Interestingly, the moments $\widehat{v}_k(\tau)$ can be written explicitly by using the so-called Eulerian numbers. Eulerian numbers appear in many contexts in various fields of mathematics (number theory, combinatorics) and in various special functions (sinc functions, polylogarithms, etc.); we refer to [6, 10, 24, 25, 28] and references therein. An interpretation of the Eulerian number $\genfrac{\langle}{\rangle}{0pt}{}{k}{j}$ is that it counts the total number of permutations of the ordered set $\{1, \ldots, k\}$ that have j 'permutation ascents'. The first five rows of the Euler's number triangle are given by

$$\begin{array}{ccccccccc} & & & & \genfrac{\langle}{\rangle}{0pt}{}{1}{0} & & & & \\ & & & \genfrac{\langle}{\rangle}{0pt}{}{2}{0} & & \genfrac{\langle}{\rangle}{0pt}{}{2}{1} & & & \\ & & \genfrac{\langle}{\rangle}{0pt}{}{3}{0} & & \genfrac{\langle}{\rangle}{0pt}{}{3}{1} & & \genfrac{\langle}{\rangle}{0pt}{}{3}{2} & & \\ & \genfrac{\langle}{\rangle}{0pt}{}{4}{0} & & \genfrac{\langle}{\rangle}{0pt}{}{4}{1} & & \genfrac{\langle}{\rangle}{0pt}{}{4}{2} & & \genfrac{\langle}{\rangle}{0pt}{}{4}{3} & \\ \genfrac{\langle}{\rangle}{0pt}{}{5}{0} & & \genfrac{\langle}{\rangle}{0pt}{}{5}{1} & & \genfrac{\langle}{\rangle}{0pt}{}{5}{2} & & \genfrac{\langle}{\rangle}{0pt}{}{5}{3} & & \genfrac{\langle}{\rangle}{0pt}{}{5}{4} \end{array} \qquad \begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & & 1 & & & \\ & & 1 & & 4 & & 1 & & \\ & 1 & & 11 & & 11 & & 1 & \\ 1 & & 26 & & 66 & & 26 & & 1 \end{array}$$

Shenton and Bowman [24, 25] studied geometric distributions, and obtained an unusual recurrence relation for its cumulants, with 'cumulant components' that involve Eulerian numbers. To the best of our knowledge, the raw moments of a *shifted* geometric distribution on $\{1, 2, \ldots\}$, are not explicitly stated in the existing literature, in the form of Theorem 4.1.

Theorem 4.1. *The k-th moment of the instantaneous sojourn time is given by*

$$\widehat{v}_k(\tau) = \frac{\tau^k}{(1-\rho)^k} \sum_{j=0}^{k-1} \left\langle {k \atop j} \right\rangle \rho^j, \quad \textit{for } k \in \mathbb{N}.$$

Proof: Use the identity (e.g., see [28]):

$$\sum_{n=1}^{\infty} n^k r^n \equiv \frac{1}{(1-r)^{k+1}} \sum_{i=0}^{k} \left\langle {k \atop i} \right\rangle r^{k-i}$$
$$= \frac{r}{(1-r)^{k+1}} \sum_{i=0}^{k-1} \left\langle {k \atop i} \right\rangle r^{k-i-1},$$

where the last equality sign relies on the fact that $\genfrac{\langle}{\rangle}{0pt}{}{k}{k} = 0$. Then, we readily derive

$$\frac{\widehat{v}_k(\tau)}{\tau^k} = \sum_{n=0}^{\infty} \pi_n (n+1)^k = \frac{1-\rho}{\rho} \sum_{n=1}^{\infty} n^k \rho^n$$
$$= \frac{\sum_{i=0}^{k-1} \genfrac{\langle}{\rangle}{0pt}{}{k}{i} \rho^{k-i-1}}{(1-\rho)^k} = \frac{\sum_{j=0}^{k-1} \genfrac{\langle}{\rangle}{0pt}{}{k}{j} \rho^j}{(1-\rho)^k},$$

where $\genfrac{\langle}{\rangle}{0pt}{}{k}{i} = \genfrac{\langle}{\rangle}{0pt}{}{k}{k-i-1}$ is used in the last equality sign (symmetry of Euler's number triangle). □

In Figure 1, as an illustration of the instantaneous sojourn time, we have depicted $v_k(\tau)$ and $\widehat{v}_k(\tau)$ for the M/M/1 PS queue (together with the Jensen's lower bound $v_1(\tau)^k$), for $k = 2, 3, 4$, on a small and large scale for τ, respectively. Figure 2 depicts the moments, all properly scaled with $(1-\rho)^k/\tau^k$ for the large scale of τ. As expected, $\widehat{v}_k(\tau)$ is a good approximation for $v_k(\tau)$ when τ is small (and even for τ up to the mean β_1). The approximation is loose for large τ, since $V(\tau)/\tau \xrightarrow{\mathbb{P}} 1/(1-\rho)$ as $\tau \to \infty$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability (cf. [35]). In fact, for $k \geq 2$, we have an asymptotic estimate

$$v_k(\tau) = \left(\frac{\tau}{1-\rho}\right)^k + \frac{\lambda k(k-1)\beta_2}{2(1-\rho)^{k+1}} \tau^{k-1}$$
$$+ o(\tau^{k-1}), \quad \tau \to \infty,$$

whenever $\beta_2 < \infty$, cf. [35]. Note that $\widehat{v}_k(\tau)$ does not use knowledge of the higher moments of the service requirement distribution; $\widehat{v}_k(\tau)$ is also properly defined when $\beta_2 = \infty$.

Interestingly, the approximation for the second moment $v_2(\tau) \approx \widehat{v}_2(\tau) = \frac{1+\rho}{(1-\rho)^2} \tau^2$ for small $\tau > 0$, yields in fact $v_2(\tau) \leq \widehat{v}_2(\tau)$ for arbitrary $\tau > 0$, see (3.1). This might suggest that the moments of the instantaneous sojourn time $\widehat{v}_k(\tau)$ are upper bounds for $v_k(\tau)$, for all $k \geq 2$. In Section 5, we will prove our main result that $v_k(\tau) \leq \widehat{v}_k(\tau)$ for all $\tau \geq 0$ and $k \in \mathbb{N}$; see Theorem 5.11. The upper bounds hold under general service requirement distributions. An intuitive explanation is given in the next remark.

Remark 4.2. **(intuition for the upper bound)** In the instantaneous sojourn time analysis ($\tau \to 0$) we assumed that during a time interval of length $(n+1)\tau$, there is no other activity in the system. When n is large upon arrival, then this is not very likely: $\widehat{V}(\tau)$ overestimates the true sojourn time $V(\tau)$ when n is large upon arrival of the tagged customer, and underestimates $V(\tau)$ when n is small upon arrival. Apparently, for the first moment: over- and underestimation outweigh each other (weighted with probability π_n). For higher moments: overestimation is weighted more heavily than underestimation, since $\rho < 1$ and thus the queue length process shows a negative drift for a large initial value of the number of customers present in the system.

We note that $V(\tau)$ and $\widehat{V}(\tau)$ have a similar heavy-traffic behavior (when $\rho \to 1$). From the identity $\sum_{j=0}^{k-1} \genfrac{\langle}{\rangle}{0pt}{}{k}{j} = k!$ (i.e., the row sums of the Euler's number triangle), it follows that:

$$\lim_{\rho \to 1} (1-\rho)^k \widehat{v}_k(\tau) = k! \tau^k.$$

For $V(\tau)$ it is known that: $\lim_{\rho \to 1} (1-\rho)^k v_k(\tau) = k! \tau^k$ (cf. [35]), or in fact,

$$\mathbb{P}((1-\rho)V(\tau)/\tau \leq x) \to 1 - e^{-x}, \quad \text{as } \rho \to 1, \ x \geq 0,$$

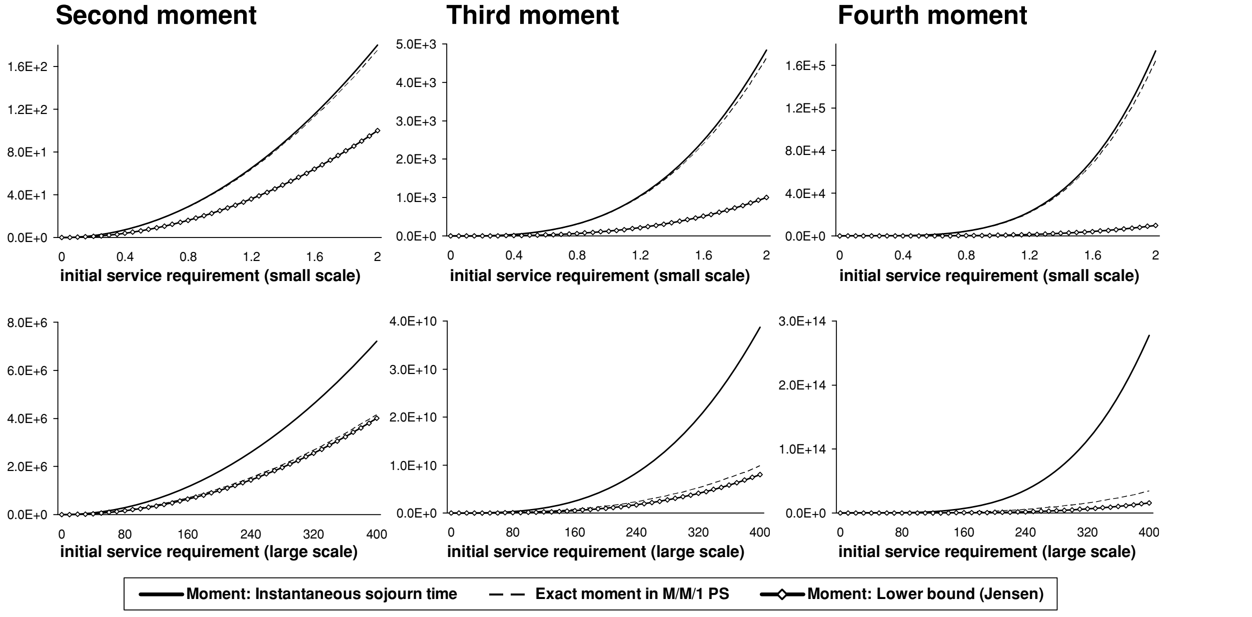

Fig. 1 The moments $v_k(\tau)$ in the M/M/1 PS queue with $\lambda = 0.4$, $\beta_1 = 2$, $\rho = 0.8$, and the instantaneous sojourn time moments $\widehat{v}_k(\tau)$, and Jensen's lower bound $v_1(\tau)^k$, for $k = 2, 3, 4$. Upper graphs: $\tau \in (0, \beta_1]$ (small scale for τ); and lower graphs: $\tau \in (0, 200\beta_1]$ (large scale).

i.e., $(1-\rho)V(\tau)/\tau$ converges in distribution to an exponential random variable with mean 1, when $\rho \to 1$ (cf. [34]). We also note that $\widehat{v}_k(\tau) = v_k(\tau) = \tau^k$ (deterministic) for $\rho \to 0$.

The above observations and Remark 4.2 suggest that $\widehat{v}_k(\tau)$ are (insensitive) upper bounds and tight in an appropriate sense. Furthermore, since $\{k!\tau^k/(1-\rho)^k\}_{k\geq 1}$ is the moment sequence of an exponentially distributed random variable $X(\tau)$ with mean $\tau/(1-\rho)$, it seems that $V(\tau)$ is 'less variable' than $X(\tau)$, in the *convex stochastic order* sense. In the next section we obtain more precise stochastic ordering results together with the formal proof that the instantaneous sojourn time moments are upper bounds for $v_k(\tau)$, for all $\tau \geq 0$ and $k \in \mathbb{N}$, with Eulerian numbers as coefficients for the polynomials in ρ.

5. Stochastic ordering

In this section we obtain some new results for the distribution of $V(\tau)$ in relation with stochastic ordering theory. For stochastic ordering theory we refer to Stoyan [27], and Shaked and Shanthikumar [26]. The main goal of this section is to prove that the moments of the instantaneous sojourn time serve as upper bounds for all moments of the conditional sojourn time.

In Section 5.1 we first establish a Laplace transform ordering between $V(\tau)$ and the instantaneous sojourn time $\widehat{V}(\tau)$. In addition, a characterization that $V(\tau)$ belongs to the so-called $\mathcal{L}$-class of life time distributions will be derived, which is related to the Laplace transform ordering. In Section 5.2 we prove the moment ordering result between $V(\tau)$ and $\widehat{V}(\tau)$,

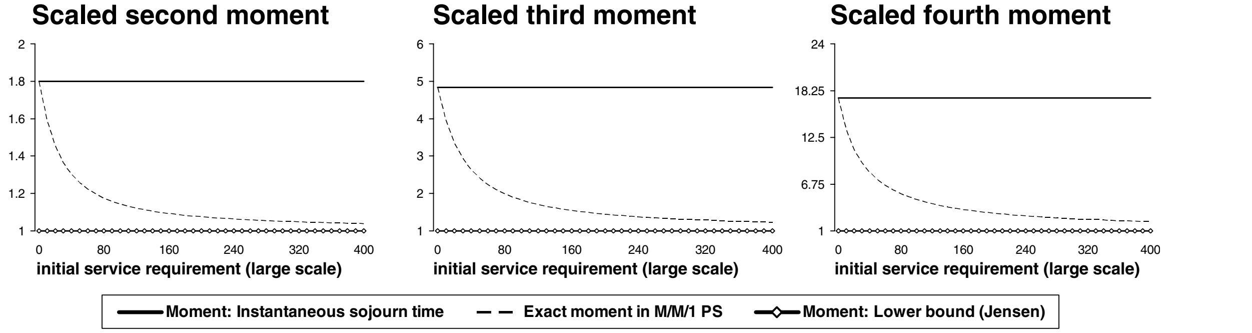

Fig. 2 The scaled moments $(1-\rho)^k v_k(\tau)/\tau^k$ in the M/M/1 PS queue ($\lambda = 0.4$, $\beta_1 = 2$), with $(1-\rho)^k \widehat{v}_k(\tau)/\tau^k$ as the scaled instantaneous sojourn time moments and the lower bound of 1.

i.e., $\mathbb{E}V(\tau)^k \le \mathbb{E}\widehat{V}(\tau)^k$ for all $\tau \ge 0$ and $k \in \mathbb{N}$, which follows from a more general moment ordering result between two PS queues, constructed with a random number of *permanent* customers.

5.1. Laplace transform ordering

The stochastic ordering in Laplace transforms denoted by $Y \ge_{Lt} X$, for any non-negative random variables X and Y, i.e., $v(s) = \mathbb{E}e^{-sY} \le \mathbb{E}e^{-sX} = w(s)$, $\mathrm{Re}\, s \ge 0$, is generally a weak ordering; it only implies $\mathbb{E}Y \ge \mathbb{E}X$. If in addition $\mathbb{E}Y = \mathbb{E}X$ is known besides the ordering $v(s) \le w(s)$, then it can be easily shown that $\mathbb{E}Y^2 \le \mathbb{E}X^2$, see Proposition 5.1. Implications for higher moments cannot be made in general. For $V(\tau)$ in the M/G/1 PS case we have a stronger Laplace transform ordering result; see Theorem 5.3.

Proposition 5.1. *For any non-negative random variables X and Y, with $\mathbb{E}Y = \mathbb{E}X$ and the LST ordering $v(s) = \mathbb{E}e^{-sY} \le \mathbb{E}e^{-sX} = w(s)$,* $\mathrm{Re}\, s \ge 0$*, it holds that: $\mathbb{E}Y^2 \le \mathbb{E}X^2$.*

Proof: By $\mathbb{E}Y = \mathbb{E}X$, the tangent line of $v(s)$ at $s = 0$ is equal to the tangent line of $w(s)$ at $s = 0$. Then, by convexity and analyticity of LSTs, and the ordering $v(s) \le w(s)$, it is readily seen that $\frac{d^2}{ds^2}v(s) \le \frac{d^2}{ds^2}w(s)$ for s in a neighborhood of 0. Hence, $\mathbb{E}Y^2 \le \mathbb{E}X^2$. □

Definition 5.2. **(Klefsjö [14])** It is said that $V(\tau)$ belongs to the $\mathcal{L}$-class of life time distributions if the LST ordering $v(s,\tau) \le x(s,\tau)$ holds, $\mathrm{Re}\, s \ge 0$, where $x(s,\tau)$ is the LST of an exponential distribution with mean $\tau/(1-\rho)$.

Theorem 5.3. *For the stable M/G/1 PS queue, the LST $v(s,\tau)$ of $V(\tau)$ is bounded by*

$$e^{-s\tau/(1-\rho)} \le v(s,\tau) \le \widehat{v}(s,\tau) = \frac{1-\rho}{e^{s\tau}-\rho} \le x(s;\tau)$$
$$= \frac{1}{1+s\tau/(1-\rho)}, \quad \mathrm{Re}\, s \ge 0,$$

where $\widehat{v}(s,\tau) := \mathbb{E}e^{-s\widehat{V}(\tau)}$ is the LST of $\widehat{V}(\tau)$; and $x(s;\tau)$ is the LST of an exponential random variable with mean $\tau/(1-\rho)$. In addition, $V(\tau) \in \mathcal{L}$, i.e., the conditional sojourn time belongs to the $\mathcal{L}$-class of life time distributions.

Proof: The bounds $e^{-s\tau/(1-\rho)} \le v(s,\tau) \le \frac{1-\rho}{e^{s\tau}-\rho}$ follow straightforwardly from (2.1) with the bounds (3.3), and it is also straightforwardly shown that $\frac{1-\rho}{e^{s\tau}-\rho}$ coincides with $\widehat{v}(s,\tau) := \mathbb{E}e^{-s\widehat{V}(\tau)}$, if $\rho < 1$. The inequality $\widehat{v}(s,\tau) \le x(s;\tau)$ follows from: $\frac{1-\rho}{e^{s\tau}-\rho} \le \frac{1-\rho}{1+s\tau-\rho} =: x(s,\tau)$, where $x(s,\tau)$ is clearly the LST of an exponential distribution with mean $\tau/(1-\rho)$. Hence, $V(\tau) \in \mathcal{L}$. □

Distributions belonging to the $\mathcal{L}$-class of life time distributions always have a finite second moment, and the coefficient of variation is not greater than one; see e.g. [5, 18]. More interestingly, although the $\mathcal{L}$-class is a wide class of distributions, Klar [13] obtained explicit and sharp 'reliability bounds' for any $\mathcal{L}$-class distribution. As an application of these reliability bounds (Theorem 4.1 from [13]), for the conditional sojourn time distribution in the M/G/1 PS queue, we obtain the next corollary.

Corollary 5.4. *For $x \le \tau/(1-\rho)$ we have the insensitive lower bound*

$$\mathbb{P}(V(\tau) > x) \ge 1 - \frac{1}{(x(1-\rho)/\tau)^2 - 2x(1-\rho)/\tau + 2}.$$

For $x > \tau/(1-\rho)$ we have the insensitive upper bound

$$\mathbb{P}(V(\tau) > x) \le \frac{1}{(x(1-\rho)/\tau)^2 - 2x(1-\rho)/\tau + 2}.$$

Remark 5.5. Stronger results for the reliability bounds exist for life time distributions belonging to subclasses of the $\mathcal{L}$-class.

5.2. Moment ordering

In this section, we will prove our main result that $v_k(\tau) \le (1 + \sum_{i=1}^{k-1} \left\langle {k \atop i} \right\rangle \rho^i)/\,[(1-\rho)/\tau]^k$, see Theorem 5.11. This moment ordering result follows from a more general moment ordering result between two PS queues with a random number of *permanent* customers, see Theorem 5.9. In Section 4, the instantaneous sojourn time $\widehat{V}(\tau)$ is defined as the sojourn time of an infinitesimally small job. Alternatively, $\widehat{V}(\tau)$ can also be viewed as the sojourn time of a customer (with an arbitrary service requirement τ) that enters a PS system with no other arriving customers but with a random number of permanent customers that is distributed as π_n. The latter viewpoint turns out to be convenient for proving our main result in the remainder of the paper.

We proceed with constructing two independent PS queues; both M/G/1 queues have the same service requirement distribution $B(x)$ with mean β_1 $(= \mathbb{E}X)$; they only have different Poisson arrival rates, λ_1 and λ_2 respectively. We let $V_i(\tau)$ denote the conditional sojourn time in the M/G/1 PS queue with arrival rate $\lambda_i, i = 1, 2$. Next, we define the random variable $V_i(\tau; n)$ as the conditional sojourn time in the M/G/1 PS queue with arrival rate λ_i, but now modified with n permanent customers in the system. The distribution of $V_i(\tau; n)$ is given by the $(n+1)$-fold convolution of the distribution of $V_i(\tau)$, see e.g. [3, 30]. Note that $V_i(\tau) \equiv V_i(\tau; 0)$.

We let $N^{(i)}$, $i = 1, 2$, be geometrically distributed with probability density function

$$\mathbb{P}(N^{(i)} = n) = \frac{1-\rho}{1-\rho_i}\left(\frac{\rho-\rho_i}{1-\rho_i}\right)^n, \quad n \in \mathbb{N} \cup \{0\}. \tag{5.1}$$

Hence the random variable $V_i(\tau; N^{(i)})$ can be interpreted as the conditional sojourn time in the M/G/1 PS queue with arrival rate λ_i and with a random number of permanent customers distributed as $N^{(i)}$, where $\rho_i = \lambda_i \mathbb{E}X$, and assume $\rho = \rho_1 + \rho_2 < 1$.

It is not difficult to show that $\mathbb{E}V_i(\tau; N^{(i)}) = \tau/(1-\rho)$, and $V_i(\tau; N^{(i)}) \in \mathcal{L}$, $i = 1, 2$. In general, we will prove that the following moment ordering holds: $\mathbb{E}V_1(\tau; N^{(1)})^k \leq \mathbb{E}V_2(\tau; N^{(2)})^k$, if the values for the arrival rates satisfy $\lambda_1 \geq \lambda_2$; see Theorem 5.9. First we derive the LST of $V_i(\tau; N^{(i)})$, for $i = 1, 2$.

Lemma 5.6. *For $i = 1, 2$, the LST defined by $\widehat{v}^{(i)}(s;\tau) = \mathbb{E}e^{-sV_i(\tau;N^{(i)})}$, $\mathrm{Re}\, s \geq 0$, for the random variable $V_i(\tau; N^{(i)})$ is expressed by*

$$\widehat{v}^{(i)}(s;\tau) = \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\widehat{\alpha}_n(\tau, \rho_i)\right)^{-1},$$

where $\widehat{\alpha}_n(\tau, \rho_i)$ is defined by: $\widehat{\alpha}_0(\tau, \rho_i) := 1, \widehat{\alpha}_1(\tau, \rho_i) := \tau/(1-\rho)$ and for $n \geq 2$:

$$\widehat{\alpha}_n(\tau, \rho_i) = \frac{n}{1-\rho}\sum_{m=0}^{\infty} \rho_i^m \binom{m+n-2}{n-2} \int_{x=0}^{\tau} (\tau - x)^{n-1} \widetilde{B}^{m*}(x)dx. \tag{5.2}$$

Proof: It holds that $\widehat{v}^{(i)}(s;\tau) = \sum_{n=0}^{\infty}\left(v^{(i)}(s;\tau)\right)^{n+1} \mathbb{P}(N^{(i)} = n)$ by definition of $V_i(\tau; N^{(i)})$ and conditioning on the event $\{N^{(i)} = n\}$, where $v^{(i)}(s;\tau) = \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\alpha_n^{(i)}(\tau)\right)^{-1}$ is the LST of $V_i(\tau; 0)$. Straightforward calculations give

$$\begin{aligned}
\widehat{v}^{(i)}(s;\tau) &= \frac{(1-\rho)v^{(i)}(s;\tau)}{1-\rho+(\rho-\rho_i)\left(1-v^{(i)}(s;\tau)\right)} \\
&= \left(\frac{1}{v^{(i)}(s;\tau)} + \frac{\rho-\rho_i}{1-\rho}\left(\frac{1}{v^{(i)}(s;\tau)} - 1\right)\right)^{-1} \\
&= \left(\sum_{n=0}^{\infty}\frac{s^n\alpha_n^{(i)}(\tau)}{n!} + \frac{\rho-\rho_i}{1-\rho}\sum_{n=1}^{\infty}\frac{s^n\alpha_n^{(i)}(\tau)}{n!}\right)^{-1} \\
&=: \left(\sum_{n=0}^{\infty}\frac{s^n}{n!}\widehat{\alpha}_n(\tau, \rho_i)\right)^{-1},
\end{aligned}$$

where $\widehat{\alpha}_n(\tau, \rho_i)$ is defined by: $\widehat{\alpha}_0(\tau, \rho_i) = \alpha_0^{(i)}(\tau) = 1$, and for $n \geq 1$:

$$\widehat{\alpha}_n(\tau, \rho_i) = \alpha_n^{(i)}(\tau)\left(1 + \frac{\rho-\rho_i}{1-\rho}\right) = \frac{1-\rho_i}{1-\rho}\alpha_n^{(i)}(\tau),$$

which leads to $\widehat{\alpha}_1(\tau, \rho_i) = (1-\rho_i)\alpha_1^{(i)}(\tau)/(1-\rho) = \tau/(1-\rho)$ and for $n \geq 2$, it is given by the expression (5.2); cf. (2.2) and (2.3) for the ordinary M/G/1 PS queue with workload ρ_i. □

As a direct consequence of Lemma 5.6, the moments of the random variables $V_i(\tau; N^{(i)})$, $i = 1, 2$, satisfy a similar recursion as for an ordinary M/G/1 PS queue.

Corollary 5.7. *For all $\tau \geq 0$, the moments defined by $\widehat{v}_k^{(i)}(\tau) = \mathbb{E}\left\{V_i(\tau; N^{(i)})\right\}^k$, are recursively given by $\widehat{v}_0^{(i)}(\tau) = 1$, and for $k \geq 1$:*

$$\widehat{v}_k^{(i)}(\tau) = -\sum_{j=1}^{k}\binom{k}{j}\widehat{v}_{k-j}^{(i)}(\tau)\widehat{\alpha}_j(\tau, \rho_i)(-1)^j.$$

In order to prove the general moment ordering: $\widehat{v}_k^{(1)}(\tau) \leq \widehat{v}_k^{(2)}(\tau)$, for $\lambda_1 \geq \lambda_2$ (Theorem 5.9), we first need the following Lemma 5.8, which implies that if we have that $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$ for some $k \geq 2$ and for some $\tau > 0$, then $V_1(\tau; N^{(1)})$ and $V_2(\tau; N^{(2)})$ are equally distributed. The converse statement is clearly true.

Lemma 5.8. *For any $\tau > 0$, and for any $k \geq 2$ (both fixed), the following equivalence holds (provided $\rho_1 + \rho_2 < 1$):*

$\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$ if and only if $\rho_1 = \rho_2$.

For $k = 1$, we have $\widehat{v}_1^{(1)}(\tau) = \widehat{v}_1^{(2)}(\tau) = \tau/(1-\rho)$, irrespective of the ordering between ρ_1 and ρ_2.

Proof: For $\tau > 0$ fixed, we will first prove the following equivalent statements:

(i) $\rho_1 = \rho_2$
(ii) *For all $n \geq 2$: $\widehat{\alpha}_n(\tau, \rho_1) = \widehat{\alpha}_n(\tau, \rho_2)$*
(iii) *For some $n \geq 2$: $\widehat{\alpha}_n(\tau, \rho_1) = \widehat{\alpha}_n(\tau, \rho_2)$*

Clearly, (i) implies (ii), which in turn implies (iii). Now we will show the non-trivial implication (iii) ⇒ (i). For $\tau > 0$, suppose that for some $n \geq 2$: $\widehat{\alpha}_n(\tau, \rho_1) = \widehat{\alpha}_n(\tau, \rho_2)$. Then, it follows from the structure of (5.2) that $\rho_1 = \rho_2$. To this end, note that $\widehat{\alpha}_n(\tau, \rho_i)$ is a real-valued polynomial in ρ_i with non-negative coefficients, and with positive coefficients if $\tau > 0$ (also observe that $\widetilde{B}(x)$ is a proper distribution function). Hence for $\tau > 0$, $\widehat{\alpha}_n(\tau, \rho_1)$ and $\widehat{\alpha}_n(\tau, \rho_2)$ are the same strictly increasing continuous functions (on $\mathbb{R}_+$), only evaluated at a different point, at ρ_1 and ρ_2 respectively. Hence, if $\widehat{\alpha}_n(\tau, \rho_1) = \widehat{\alpha}_n(\tau, \rho_2)$, then necessarily $\rho_1 = \rho_2$.

For any $k \geq 2$ and for any $\tau > 0$ (both fixed), we now proceed with proving the implication:

$$\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau) \Rightarrow \rho_1 = \rho_2. \tag{5.3}$$

The above implication (5.3) must hold for all $k \geq 2$ and all $\tau > 0$ (and it is *not* the same as the statement: {*For all* $k \geq 2$ *and all* $\tau > 0$: $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$} $\Rightarrow \{\rho_1 = \rho_2\}$*)*.

The implication (5.3) is straightforward for the second moment, since $\widehat{v}_2^{(1)}(\tau) = \widehat{v}_2^{(2)}(\tau)$ is equivalent to $\widehat{\alpha}_2(\tau, \rho_1) = \widehat{\alpha}_2(\tau, \rho_2)$, cf. the recursive formula in Corollary 5.7 with equal first moments, and thus $\rho_1 = \rho_2$ by the equivalent statements (i)-(ii)-(iii). For $k > 2$, the implication (5.3) is not trivial, mainly due to the presence of the alternating term $(-1)^j$ in the recursive formula. However, the statements (i)-(ii)-(iii) are equivalent in a strong sense. For example, if $\widehat{\alpha}_k(\tau, \rho_1) \neq \widehat{\alpha}_k(\tau, \rho_2)$ for *some* $k \geq 2$, then $\rho_1 \neq \rho_2$ and also $\widehat{\alpha}_k(\tau, \rho_1) \neq \widehat{\alpha}_k(\tau, \rho_2)$ for *all* $k \geq 2$; see also the similar observation in Remark 3.2.

Hence, if $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$, then we have two mutually exclusive possibilities:

(a) $\widehat{\alpha}_k(\tau, \rho_1) = \widehat{\alpha}_k(\tau, \rho_2)$
(b) $\widehat{\alpha}_k(\tau, \rho_1) \neq \widehat{\alpha}_k(\tau, \rho_2)$

The fact that $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$ implies either (a) or (b). By contradiction, we will now show that possibility (b) cannot occur. So, suppose $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau) \Rightarrow \widehat{\alpha}_k(\tau, \rho_1) \neq \widehat{\alpha}_k(\tau, \rho_2)$, but $\widehat{\alpha}_k(\tau, \rho_1) \neq \widehat{\alpha}_k(\tau, \rho_2)$ is equivalent to $\rho_1 \neq \rho_2$. Hence, assuming (b) true is the same as

$$\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau) \Rightarrow \rho_1 \neq \rho_2, \tag{5.4}$$

which is obviously false, since the negation of (5.4), i.e., $\rho_1 = \rho_2 \Rightarrow \widehat{v}_k^{(1)}(\tau) \neq \widehat{v}_k^{(2)}(\tau)$ is clearly false. Thus, the assumption that (b) holds is false, and hence $\widehat{v}_k^{(1)}(\tau) = \widehat{v}_k^{(2)}(\tau)$ implies possibility (a) which in turn is equivalent to $\rho_1 = \rho_2$ by the strong equivalences (i)-(ii)-(iii). □

Theorem 5.9. *For* $\tau > 0$ *and* $\rho = \rho_1 + \rho_2 < 1$, *if* $\rho_1 \geq \rho_2$, *then we have the moment ordering*

$$\widehat{v}_k^{(1)}(\tau) \leq \widehat{v}_k^{(2)}(\tau), \quad \textit{for all } k \in \mathbb{N}.$$

Proof: For the first moment we have $\widehat{v}_1^{(1)}(\tau) = \widehat{v}_1^{(2)}(\tau)$, irrespective of the ordering between ρ_1 and ρ_2. By Lemma 5.8, if $\rho_1 \neq \rho_2$, then it holds that $\widehat{v}_k^{(1)}(\tau) \neq \widehat{v}_k^{(2)}(\tau)$ for all $k \geq 2$ and all $\tau > 0$. Now consider the strict ordering $\rho_1 > \rho_2$. Lemma 5.8 guarantees for $\rho_1 > \rho_2$, that $\widehat{v}_k^{(1)}(\tau)$ and $\widehat{v}_k^{(2)}(\tau)$ cannot coincide for any $\tau > 0$ and $k \geq 2$. Then, continuity of $\widehat{v}_k^{(i)}(\tau)$ in τ implies for $\rho_1 > \rho_2$, that $\widehat{v}_k^{(1)}(\tau)$ and $\widehat{v}_k^{(2)}(\tau)$ cannot cross each other for any $k \geq 2$, as a function of $\tau > 0$. Hence, either $\{\widehat{v}_k^{(1)}(\tau) < \widehat{v}_k^{(2)}(\tau)$ for all $\tau > 0\}$, or $\{\widehat{v}_k^{(1)}(\tau) > \widehat{v}_k^{(2)}(\tau)$ for all $\tau > 0\}$ holds.

The proof is completed, if we can find a $\tau^* > 0$ such that for all $k \geq 2$, if $\rho_1 > \rho_2$, then:

$$\widehat{v}_k^{(1)}(\tau^*) < \widehat{v}_k^{(2)}(\tau^*).$$

This can be done by choosing τ^* large enough, since $\frac{V_i(\tau)}{\tau} \xrightarrow{\mathbb{P}} \frac{1}{1-\rho_i}$, *as* $\tau \to \infty$, *and*

$$\frac{V_i(\tau; N^{(i)})}{\tau} \xrightarrow{d} \frac{N^{(i)} + 1}{1 - \rho_i}, \ \textit{as } \tau \to \infty.$$

It is readily verified that (cf. the proof of Theorem 4.1 and the geometric distribution (5.1)):

$$\mathbb{E}\left(\frac{N^{(i)} + 1}{1 - \rho_i}\right)^k = \frac{1 + \sum_{j=1}^{k-1} \left\langle {k \atop j} \right\rangle \left(\frac{\rho - \rho_i}{1 - \rho_i}\right)^j}{(1 - \rho)^k},$$

hence if $\rho_1 > \rho_2$, then $\frac{\rho_2}{1-\rho_1} = \frac{\rho-\rho_1}{1-\rho_1} < \frac{\rho-\rho_2}{1-\rho_2} = \frac{\rho_1}{1-\rho_2}$, and

$$\lim_{\tau \to \infty} \frac{\widehat{v}_k^{(1)}(\tau)}{\tau^k} = \mathbb{E}\left(\frac{N^{(1)} + 1}{1 - \rho_1}\right)^k$$
$$< \mathbb{E}\left(\frac{N^{(2)} + 1}{1 - \rho_2}\right)^k = \lim_{\tau \to \infty} \frac{\widehat{v}_k^{(2)}(\tau)}{\tau^k}, \quad \textit{for all } k \geq 2, \tag{5.5}$$

and with equality sign in (5.5) if and only if $\rho_1 = \rho_2$ (and for $k = 1$: $\widehat{v}_1^{(1)}(\tau) = \widehat{v}_1^{(2)}(\tau) = \frac{\tau}{1-\rho}$). Hence if $\rho_1 \geq \rho_2$, then $\widehat{v}_k^{(1)}(\tau) \leq \widehat{v}_k^{(2)}(\tau)$ for all $k \geq 1$ and all $\tau > 0$. □

Theorem 5.9 can be interpreted as follows. For a fixed $\tau > 0$, if the sojourn time $V_2(\tau; N^{(2)})$ is very large, then this is more likely due to the presence of many permanent customers in the system (large λ_1 implies that $N^{(2)}$ is 'stochastically' large) rather than a large arrival rate of non-permanent customers (large λ_2). By construction, the (random) number of permanent customers in system i is $N^{(i)}$ (denote system i as the model that corresponds to $V_i(\tau; N^{(i)})$, $i = 1, 2$). Interestingly, the number of non-permanent customers in system i is in distribution equal to $N^{(j)}$, $i \neq j$, for $i, j \in \{1, 2\}$; cf. Theorem 1 from [7] (where it is called a queue length decomposition result). Hence, if $\lambda_1 > \lambda_2$, then there are 'on average' more non-permanent and less permanent customers in system 1 compared to system 2. However, both systems have 'on average' an equal number of total customers (permanent plus non-permanent) regardless of the ordering between λ_1 and λ_2, which also explains the equality of the first moments: $\mathbb{E}V_1(\tau; N^{(1)}) = \mathbb{E}V_2(\tau; N^{(2)})$.

Remark 5.10. We conjecture that $V_1(\tau; N^{(1)}) \leq_{cx} V_2(\tau; N^{(2)})$ holds if $\lambda_1 \geq \lambda_2$, i.e., the random variables are ordered in the *convex stochastic order* sense (see [27, 26]). Then, it is said that the random variable $V_2(\tau; N^{(2)})$ is more variable (more likely to take extreme values) than the variable $V_1(\tau; N^{(1)})$. The first moments are necessarily equal. A sufficient condition for convex stochastic ordering is the so-called Karlin & Novikoff cut-criterion, cf. [27], which states that two random variables X and Y are convex stochastic ordered if the means are equal and the corresponding distribution functions cross each other once and exactly once. The difficulty to verify the cut-criterion is that we do not have the distribution functions explicitly. We note that the cut-criterion and the intuition for the conjecture given in the instantaneous sojourn time analysis, are similar (see Remark 4.2).

We arrive at our main result that the moments of the instantaneous sojourn time are upper bounds for the moments of the conditional sojourn time in the M/G/1 PS queue.

Theorem 5.11. *In the stable M/G/1 PS queue, we have the insensitive lower and upper bounds for all moments of the conditional sojourn time $V(\tau)$, for $\tau \geq 0$ and $k \in \mathbb{N}$:*

$$\frac{1}{(1-\rho)^k}\tau^k \leq v_k(\tau) \leq \frac{1 + \sum_{j=1}^{k-1} \left\langle {k \atop j} \right\rangle \rho^j}{(1-\rho)^k}\tau^k.$$

Proof: The result is trivial for the lower bound and for $\tau = 0$. For $\tau > 0$, we consider the special case of $\rho_2 = 0$ in Theorem 5.9. Then, for all $\rho \equiv \rho_1 \geq \rho_2 = 0$ and $\rho < 1$ it holds that

$$v_k(\tau) \equiv \widehat{v}_k^{(1)}(\tau) \leq \widehat{v}_k^{(2)}(\tau) = \tau^k \mathbb{E}(N+1)^k,$$

since 'with probability 1' we have: $N^{(1)} = 0$, $V(\tau) \equiv V_1(\tau; N^{(1)})$, $\mathbb{P}(N^{(2)} = n) = (1-\rho)\rho^n$, $V_2(\tau; 0) \equiv \tau$, and $V_2(\tau; N^{(2)}) \equiv V_2(\tau; N) \overset{d}{=} \tau(N+1) \equiv \widehat{V}(\tau)$, as in Section 4. □

Remark 5.12. The special choice of $\rho_2 = 0$ in Theorem 5.9 is essentially the same as the assumptions made in the instantaneous sojourn time analysis, as in Section 4. For $\rho_2 \to 0$: $V_2(\tau; N^{(2)}) \xrightarrow{d} \widehat{V}(\tau) = (N^{(2)} + 1)\tau$, as if the tagged customer arrived at a system with n permanent customers with probability $\mathbb{P}(N^{(2)} = n)$ and with no other arriving customers ($\rho_2 = 0$).

Remark 5.13. (**Theorem 5.11 in relation with the fluid and quasi-stationary regime**) Our main result can be related to the result obtained by Delcoigne, Proutière and Régnié [9]. They obtained the following (increased convex) stochastic ordering: $W^{fl} \leq_{icx} W \leq_{icx} W^{qs}$, for the stationary workload W in the M/G/1 PS queue with *time-varying* service capacity. Their bounds correspond to the workload in the so-called 'fluid' and 'quasi-stationary' regimes. As noted in [9], it proves much more difficult to derive similar results for the mean sojourn time.

In our paper, $\widehat{V}(\tau)$ can also be viewed as the sojourn time $V^{qs}(\tau)$ in a quasi-stationary regime, which can be obtained by considering a (modified) M/G/1 PS queue with fixed capacity, arrival rate λs and service requirement X/s. The (perturbation) parameter $s > 0$ represents the 'speed' of the queue length process, and it does not influence the queue length distribution. In the limit $s \to 0$, the queue length process freezes in some initial state, yielding the quasi-stationary regime, and it can be shown that $V^{qs}(\tau) \overset{d}{=} \widehat{V}(\tau)$. For the sojourn time $V^{fl}(\tau)$ in the fluid regime (i.e., for the limit $s \to \infty$), it can be shown that $V^{fl}(\tau)$ is constant and equal to $\tau/(1-\rho)$. Analogous to the insensitive bounds in [9], we conjecture that holds: $\tau/(1-\rho) \leq_{icx} V(\tau) \leq_{icx} \widehat{V}(\tau)$, where the $\leq_{icx}$-ordering could be replaced by the $\leq_{cx}$-ordering (since the means are equal).

Remark 5.14. With the moments of the instantaneous sojourn time as upper bounds, i.e., $v_k(\tau) \leq \widehat{v}_k(\tau)$, and the Chebyshev-Markov inequalities $\mathbb{P}(V(\tau) > x) \leq \frac{1}{x^k} v_k(\tau)$ for *all* $k \geq 1$, an insensitive upper bound for the tail probability $\mathbb{P}(V(\tau) > x)$ can be given

$$\mathbb{P}(V(\tau) > x) \leq \min_{k \geq 1} \frac{1 + \sum_{i=1}^{k-1} \left\langle {k \atop i} \right\rangle \rho^i}{(x(1-\rho)/\tau)^k}, \tag{5.6}$$

for $x \geq \tau > 0$, $\rho < 1$. The improvement upon Corollary 5.4 is considerable, particularly for large x. For $x \leq \tau/(1-\rho)$ the bound is not useful. However, if $x(1-\rho)/\tau > 1$, then both the numerator and denominator on the right-hand-side of (5.6) increase in k, but the denominator will be dominant for a certain k^*, defined by $k^* \equiv k^*(x; \tau) = \arg\min_{k \geq 1} \frac{\widehat{v}_k(\tau)}{x^k}$. We omit the details of the latter statement.

6. Conclusion

In this study, we have investigated the sojourn time $V(\tau)$ conditional on the initial service requirement $\tau > 0$ in the M/G/1 processor-sharing (PS) queue. In particular, we have studied all moments of $V(\tau)$ and we have obtained upper and lower bounds. Our main result (Theorems 5.11 and 5.9) is that there exist upper bounds, given by $\left(1+\sum_{i=1}^{k-1}\left\langle {k \atop i}\right\rangle\rho^{i}\right)\tau^{k}/(1-\rho)^{k}$, where $\left\langle {k \atop i}\right\rangle$ are Eulerian numbers, and they only depend on τ and the traffic intensity $\rho < 1$. A lower bound follows easily from Jensen's inequality. The main result has been proved via stochastic comparisons of two related PS models with random number of permanent customers.

An attractive feature of the upper bound of the above structure is that it is independent of second and higher moments of the service requirement distribution. Another attractive feature is that the upper bound coincides with Jensen's lower bound when $\rho \to 0$. Moreover, the k-th moment of $V(\tau)$ and the above upper bound, converge to the same expression, after proper scaling when $\rho \to 1$. The upper bounds of the above structure with Eulerian numbers are in fact the moments of the so-called *instantaneous* sojourn time $\widehat{V}(\tau)$, i.e., the sojourn time of a customer with an infinitesimally small initial service requirement ($\tau \to 0$). If the initial service requirement $\tau > 0$ is arbitrary (and not necessarily small), the instantaneous sojourn time also corresponds to the sojourn time of a tagged customer entering a PS system with no other arrivals but with a random number of permanent customers. The instantaneous sojourn time is also the sojourn time in a so-called quasi-stationary regime.

By studying the higher moments and providing insensitive upper bounds, we provide further support for the observation that PS is a 'fair' service discipline. In the stable M/G/1 PS system, excessive behavior of other customers in the system always has a limited influence on the sojourn time of the tagged customer. Intuitively, from a tagged customer point-of-view, the influence of the service requirements of other customers on the sojourn time of the tagged customer, is nearly insensitive. Even when there is a customer with infinite service requirement, the influence of this *permanent* customer on non-permanent customers is limited.

The influence of other customers is even more limited for jobs with a small initial service requirement; its sojourn time may be reasonably approximated by the instantaneous sojourn time. Moreover, it provides tight upper bounds for the higher moments of $V(\tau)$. However, for very large τ, these upper bounds are 'quite loose'. This can be seen as the price that must be paid for obtaining *insensitive* upper bounds. Nevertheless, the moments of $V(\tau)$ are always bounded from above by the moments of $\widehat{V}(\tau)$, which in turn are bounded from above by the moments of an exponential random variable with mean $\tau/(1-\rho)$, regardless of the service requirement distribution (even when the service requirement has an infinite second moment).

We conclude this paper with the remark that considerable attention has been paid in the literature to the exact analysis of the sojourn time in the M/G/1 PS queue. Relatively little work has been done on the investigation of the practical implications of the results. The discovery of simple bounds for *all* moments of $V(\tau)$ stimulates the investigation of simple but nevertheless good approximations for the distribution of $V(\tau)$, the moments and the tail probabilities. In addition, a logical next step is to investigate if similar results also hold for other PS queues. For extensions of PS service disciplines, such as the *discriminatory* PS, it is to be expected that the nice properties (regarding the instantaneous sojourn time) are lost. For the M/G/1 queue with the egalitarian PS discipline and queue-dependent service capacity [8] it may be worthwhile to investigate if the structures remain valid. This remains a topic for further research.

Acknowledgments

The authors would like to thank the referees for offering valuable and constructive comments on an earlier version of this paper.

References

1. B.K. Asare and F.G. Foster, Conditional response times in the M/G/1 processor-sharing system, Journal of Applied Probability 20 (1983) 910–915.
2. S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts, Statistical bandwidth sharing: A study of congestion at flow level, in: *Proceedings ACM SIGCOMM'01* (2001) pp. 111–122.
3. J.L. van den Berg, *Sojourn Times in Feedback and Processor Sharing Queues*, Ph.D. thesis, Utrecht University, The Netherlands (1990).
4. J.L. van den Berg and O.J. Boxma, The M/G/1 queue with processor sharing and its relation to a feedback queue, Queueing Systems 9 (1991) 365–402.
5. A. Bhattacharjee and D. Sengupta, On the coefficient of variation of the $\mathcal{L}$- and $\overline{\mathcal{L}}$-classes, Statistics Probability Letters 27 (1996) 177–180.
6. L. Carlitz, Eulerian numbers and polynomials of higher order, Duke Mathematical Journal 27 (1960) 401–423.
7. S.-K. Cheung, J.L. van den Berg, and R.J. Boucherie, Decomposing the queue length distribution of processor-sharing models into queue lengths of permanent customer queues, Performance Evaluation 62(1–4) (2005) 100–116.
8. J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.
9. F. Delcoigne, A. Proutière, and G. Régnié, Modeling integration of streaming and data traffic, Performance Evaluation 55(3–4) (2004) 185–209.
10. R.L. Graham, D.E. Knuth, and O. Patashnik, *Eulerian Numbers, §6.2 in Concrete Mathematics: A Foundation for Computer Science*, 2nd ed. (Reading, MA, Addison-Wesley, 1994) pp. 267–272.
11. S. Grishechkin, On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes, Advances in Applied Probability 24 (1992) 653–698.

12. Y. Kitayev, The M/G/1 processor-sharing model: Transient behavior, Queueing Systems 14 (1993) 239–273.
13. B. Klar, A note on the $\mathcal{L}$-class of life distributions, Journal of Applied Probability 39 (2002), no. 1: 11–19.
14. B. Klefsjö, A useful ageing property based on Laplace transform, Journal of Applied Probability 20 (1983) 615–626.
15. L. Kleinrock, Analysis of a time-shared processor, Naval Research Logistics Quarterly 11 (1964) 59–73.
16. L. Kleinrock, Time-shared systems: A theoretical treatment, Journal of the Association for Computing Machinery 14 (1967) 242–261.
17. L. Kleinrock, *Queueing Systems, Vol II: Computer Applications* (Wiley, New York, 1976).
18. G. Lin, Characterizations of the $\mathcal{L}$-class of life distributions, Statistics Probability Letters 40 (1998) 259–266.
19. R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie, and M. Fleuren, Analysis of flow transfer times, in IEEE 802.11 wireless LANs, Annals of Telecommunications 59(11–12) (2004) 1407–1432.
20. R. Núñez-Queija, *Processor sharing models for integrated services networks*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands (2000).
21. T.J. Ott, The sojourn time distribution in the M/G/1 queue with processor sharing, Journal of Applied Probability 21 (1984) 360–378.
22. M. Sakata, S. Noguchi, and J. Oizumi, Analysis of a processor-shared queueing model for time-sharing systems, in: *Proceedings 2nd Hawaii International Conference on System Sciences*, Jan. (1969), pp. 625–628.
23. R. Schassberger, A new approach to the M/G/1 processor sharing queue, Advances in Applied Probability 16 (1984) 202–213.
24. L.R. Shenton and K.O. Bowman, The geometric distribution, cumulants, Eulerian numbers, and the logarithmic function, Far East Journal of Theoretical Statistics 5 (2001) 113–142.
25. L.R. Shenton and K.O. Bowman, The geometric distribution's central moments and Eulerian numbers of the second kind, Far East Journal of Theoretical Statistics 7 (2002) 1–17.
26. M. Shaked and J.G. Shanthikumar, *Stochastic Orders and Their Applications* (Academic Press Inc., 1994).
27. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, Chichester, 1983).
28. Weisstein, W. Eric, Eulerian Number, *From MathWorld–A Wolfram Web Resource*. http://mathworld.wolfram.com/EulerianNumber.html
29. A. Ward and W. Whitt, Predicting response times in processor-sharing queues, in: P.W. Glynn, D.J. MacDonald and S.J. Turner (eds.), *Proceedings of the Fields Institute Conference on Communication Networks*, 2000.
30. W. Whitt, The M/G/1 processor-sharing queue with long and short jobs. Unpublished manuscript (1998).
31. S.F. Yashkov, A derivation of response time distribution for a M/G/1 processor-sharing queue, Problems of Control and Information Theory 12 (1983) 133–148.
32. S.F. Yashkov, Processor-sharing queues: Some progress in analysis, Queueing Systems 2 (1987) 1–17.
33. S.F. Yashkov, Mathematical problems in the theory of processor-sharing queueing systems, Journal of Soviet Mathematics 58 (1992) 101–147.
34. S.F. Yashkov, On a heavy traffic limit theorem for the M/G/1 processor sharing queue, Communications in Statistics—Stochastic Models 9(3) (1993) 467–471.
35. A.P. Zwart and O.J. Boxma, Sojourn time asymptotics in the M/G/1 processor sharing queue, Queueing Systems 35 (2000) 141–166.
36. A.P. Zwart, *Queueing systems with heavy tails*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands (2001).