Sojourn times in a processor sharing queue with multiple vacations

EURANDOM PREPRINT SERIES
2011-023

Sojourn times in a processor sharing queue with multiple vacations

U. Ayesta, O.J. Boxma, I.M. Verloop
ISSN 1389-2355

Sojourn times in a processor sharing queue with multiple vacations*

U. Ayesta[a,b], O.J. Boxma[c,d], I.M. Verloop[a]

[a]BCAM - Basque Center for Applied Mathematics,
Bizkaia Technology Park, Derio, Spain
[b]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[c]EURANDOM, Eindhoven, The Netherlands
[d]Department of Mathematics and Computer Science,
Technische Universiteit Eindhoven, The Netherlands

Abstract

We study an $M/G/1$ processor sharing queue with multiple vacations. The server only takes a vacation when the system has become empty. If he finds the system still empty upon return, he takes another vacation, and so on. Successive vacations are identically distributed, with a general distribution. When the service requirements are exponentially distributed we determine the sojourn time distribution of an arbitrary customer. We also show how the same approach can be used to determine the sojourn time distribution in an $M/M/1$-PS queue of a polling model, under the following constraints: the service discipline at that queue is exhaustive service, the service discipline at each of the other queues satisfies a so-called branching property, and the arrival processes at the various queues are independent Poisson processes. For a general service requirement distribution we investigate both the vacation queue and the polling model, restricting ourselves to the mean sojourn time.

1 Introduction

This study is devoted to the $M/G/1$-PS (Processor Sharing) system. In the egalitarian processor sharing discipline, when there are k customers present, they all are served simultaneously, receiving an equal share $1/k$ of the service capacity. Processor sharing was introduced by Kleinrock in the early sixties, as an idealised model of a time-sharing computer processor. In the last fifteen years it has gained renewed interest, partly because of its ability to represent 'fair' bandwidth sharing mechanisms like the Transmission Control Protocol (TCP) of the Internet.

The special feature of our study is that either (i) the server goes on a vacation after having emptied the $M/G/1$-PS system, or (ii) the $M/G/1$-PS system under consideration is just one out of several queues in a polling system, a single server visiting each of the queues in cyclic fashion. When service requirements are exponentially distributed, for both cases (i) and (ii), we determine the sojourn time distribution of customers from the $M/M/1$-PS system. For *general* service requirements we derive an integro-differential equation, the solution of which would immediately yield the mean conditional delay in the $M/G/1$-PS system with multiple vacations. Interestingly, that integro-differential equation is seen to coincide with an integro-differential equation that arises in a particular $M/G/1$ batch processor sharing queue [1]. For particular choices of the service requirement distribution, that integro-differential equation can be solved.

Motivation. Our motivation is twofold. On the one hand, it is theoretical: we wish to obtain a better insight into the effect of the PS service discipline on sojourn times, and we wish to develop

*The research of U. Ayesta and I.M. Verloop was partially supported by grant MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain) and grant PI2010-2 (Department of Education and Research, Basque Government). The research of O.J. Boxma was conducted within the framework of the European Network of Excellence Euro-NF.

probabilistic tools to accomplish this. On the other hand, we are motivated by the fact that vacation and polling systems arise very naturally in a host of application areas (in production systems, computer- and communication networks, traffic lights, maintenance, etc.). The literature on vacation and polling systems heavily concentrates on FCFS service per queue; however, in several of the above-mentioned application areas, scheduling customer service in a non-FCFS manner could be beneficial. E.g., polling models with non-FCFS service per queue arise in the IEEE 802-11 [19] and Bluetooth [20] communication protocols, in scheduling policies at routers and at I/O systems in web servers.

Well-known polling visit disciplines are the exhaustive discipline (the server serves the queue until it has become empty), the gated discipline (when the server arrives at a queue to find K customers, it serves exactly those K customers, and no more), and the 1-limited discipline (the server serves just one customer, assuming at least one is present). In Winands et al. [35] the mean delay in polling systems was already obtained under gated or exhaustive service and for various non-FCFS service disciplines per queue, but in the exhaustive case the PS discipline seemed to pose too hard mathematical problems. In [8] the LST (Laplace-Stieltjes Transform) of the sojourn time distribution was obtained for various service disciplines per queue like Last-Come-First-Served, Random Order of Service, Processor Sharing and Shortest Job First, under the gated visit discipline. Again, PS for the exhaustive discipline remained elusive. The present study aims to fill that gap.

Related work on the processor sharing queue. In the classical $M/G/1$-PS model, the steady-state queue length distribution is geometrically distributed with parameter ρ, the load of the system (arrival rate times mean service requirement). Next to this insensitivity property (the distribution of the actual service requirement B plays no role, only the mean), $M/G/1$-PS has another interesting property: the mean sojourn time of a customer, given its service requirement is τ, is linear in τ: $E(T|B=\tau) = \frac{\tau}{1-\rho}$. While these are quite simple results, the sojourn time *distribution* has turned out to be much more difficult to obtain. In 1970, Coffman, Muntz and Trotter [10] managed to derive the LST of the sojourn time distribution in the $M/M/1$-PS system. Sengupta and Jagerman [30] have obtained an expression for the same LST, conditioned on the number of customers seen upon arrival. Morrison [21] studied the sojourn time distribution itself (i.e., without LST). Almost simultaneously, Yashkov [37], Ott [27] and Schassberger [29] derived the LST of the sojourn time distribution in an $M/G/1$-PS; see [5] for an alternative derivation via an $M/M/1$-FCFS queue with feedback. Núñez-Queija [22] has derived the LST of the sojourn time distribution in an $M/M/1$-PS queue with service interruptions; notice that such interruptions occur randomly, whereas in our case vacations occur only when the system has become empty. In [18] Kleinrock et al. developed an integro-differential equation that characterizes the mean conditional sojourn time in a processor sharing queue with batch Poisson arrivals, and solved it when the service requirement distribution belongs to a particular class of distributions that includes the exponential distribution. More recently in [1, 4, 15, 26] this approach has been used to investigate general service requirement distributions.

Contributions. One of the main contributions of the present study is a derivation of the LST of the sojourn time distribution in the $M/M/1$-PS system with multiple vacations. If the system has become empty, the server takes a vacation. If, upon his return, the system is still empty, he takes another vacation, with the same distribution as the previous one; and so on. If, returning from a vacation, the system is not empty, then the server serves customers until the system has become empty once more. In our study of the sojourn time LST, we make use of an interesting intermediate result of [10]: an expression for the sojourn time LST in $M/M/1$-PS, conditional on his service requirement and on the number of customers found by a tagged customer upon arrival. In addition, we derive an expression for the sojourn time LST in an asymptotic regime when the length of the vacations grows large. This will be of particular interest in the context of polling systems.

Another main contribution pertains with the development of an integro-differential equation that characterizes the mean conditional delay in the $M/G/1$-PS system with multiple vacations. Using this approach we show that, as the service requirement τ grows to infinity, the mean conditional sojourn time has an asymptote of slope $\tau/(1-\rho)$ and we explicitly calculate the bias term.

The third main contribution of the paper concerns the application of the previous results to polling systems. We study the sojourn time in one queue Q_1 of an N-queue polling system. That queue receives exhaustive service and its service discipline is PS. In particular, for exponentially distributed service requirements we derive the LST of the sojourn time distribution. For the class of polling systems with so-called branching service disciplines at all queues (see, e.g., Resing [28]; exhaustive

and gated service are prominent examples), it is possible to derive the intervisit time distribution of Q_1 [8]. By first conditioning on the number of customers at Q_1 found by an arrival at Q_1, averaging the sojourn time LST over arrivals at Q_1 that take place while the server is at Q_1 and that take place during its intervisit time, we finally arrive at the unconditional sojourn time LST. In addition, we present results for two asymptotic regimes: a polling system having large switch-over times, and a polling system in a heavy-traffic regime.

Organization of the paper. Section 2 presents a model description of the $M/G/1$-PS queue with vacations, as well as results from [10] for the ordinary $M/M/1$-PS queue without vacations. The sojourn time LST in the $M/M/1$-PS queue with vacations is derived in Section 3. Section 4 considers the mean sojourn time in the case of the $M/G/1$-PS queue with vacations. We pay particular attention to the mean conditional sojourn time given the service requirement is τ, for $\tau \to \infty$. Finally, Section 5 is devoted to an $M/G/1$-PS queue in a polling system.

2 Model and preliminaries

We study a Processor Sharing (PS) queue with vacations. We assume that customers arrive according to a Poisson process with rate λ and have i.i.d. (independent, identically distributed) generally distributed service requirements; B denotes a generic service requirement, with mean $\mathbb{E}(B) = 1/\mu$ and distribution function $F(\cdot)$. We define $\rho = \lambda/\mu$. The scheduling policy applied in the queue is processor sharing. Once the queue empties, the server goes on vacation. Successive vacations are identically distributed; V denotes a generic vacation time. We let $\tilde{V}(s)$ be its LST, $F_V(\cdot)$ and $f_V(\cdot)$ the distribution function and density function of V, respectively. We denote by R_V (P_V) the length of a residual (past) vacation, hence $\mathbb{E}(R_V) = \mathbb{E}(P_V) = \frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)}$. We consider the system with multiple vacations, i.e., when the server returns from vacation but finds no customers in the system, it starts a new vacation. Throughout the paper we assume the system is stable, i.e., $\rho < 1$.

In the paper we will be interested in the sojourn time, denoted by T, as experienced by a customer. We further define W as the delay experienced by a customer, i.e., the sojourn time minus service requirement. Hence, $T \stackrel{d}{=} W + B$.

2.1 Preliminaries: ordinary $M/M/1$-PS queue

In Section 3 we make use of existing results for the sojourn time in the ordinary PS queue without vacations and *exponentially* distributed service requirements. These results are presented in what follows.

Let W_n be the delay (sojourn time minus service requirement) of the tagged customer in the $M/M/1$-PS queue without vacations, when he meets n customers at arrival. From [10] we have for $w_n(\tau, s) := \mathbb{E}(\mathrm{e}^{-sW_n}|B = \tau)$,

$$w_n(\tau, s) = \frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\beta(\tau, s)^n, \quad \tau \geq 0, \tag{1}$$

where

$$\beta(\tau, s) = \frac{r(1-\rho r)+(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}},$$

and r is the root (the one with minus the square-root) of $\lambda z^2 - (\lambda+\mu+s)z + \mu$. Note that $r(s)$ represents the LST of the length of a busy period in a standard $M/M/1$ queue [6].

We notice that

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{a^n}{n!}w_n(\tau, s) &= \frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\mathrm{e}^{a\beta(\tau,s)} \\ &= G(\tau, s)\mathrm{e}^{a\beta(\tau,s)}, \end{aligned} \tag{2}$$

with $G(\tau, s) := \frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}$.

Lemma 1 gives properties for the functions $G(\tau, s)$ and $\beta(\tau, s)$, which will be used later on. The proof of this Lemma is included in Appendix 1.

Lemma 1 *We have*

$$\begin{aligned}\frac{\partial G(\tau,s)}{\partial s}|_{s=0} &= -\frac{1}{\mu(1-\rho)}\left(\lambda\tau - \frac{\rho}{1-\rho}(1-\mathrm{e}^{-\mu\tau(1-\rho)})\right),\\ \frac{\partial \beta(\tau,s)}{\partial s}|_{s=0} &= -\frac{1}{\mu(1-\rho)}\left(1-\mathrm{e}^{-\mu\tau(1-\rho)}\right).\end{aligned}$$

3 $M/M/1$ processor sharing queue with multiple vacations

In this section we assume that customers have an exponentially distributed service requirement. In that case, we are able to obtain the LST of the delay of a customer with service requirement τ, see the Proposition below. In what follows we focus on one tagged customer (denoted as K in the ensuing), studying its delay W.

Proposition 1 *In an $M/M/1$-PS queue with multiple vacations,*

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sW}|B=\tau) &= \rho G(\tau,s)\frac{\beta(\tau,s)(1-\rho)}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}\\ &\quad +(1-\rho)G(\tau,s)\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s\mathbb{E}[V]},\end{aligned} \tag{3}$$

with $G(\tau,s)$ and $\beta(\tau,s)$ as defined in Section 2.1.

In particular, the first moment is given by

$$\mathbb{E}(W|B=\tau) = \frac{\rho\tau}{1-\rho} + \frac{\rho(2-\rho)\mathbb{E}(R_V)}{1-\rho}(1-\mathrm{e}^{-\mu\tau(1-\rho)}) + (1-\rho)\mathbb{E}(R_V). \tag{4}$$

Remark 1 (Unconditional delay) *The delay for an arbitrary customer can be obtained by unconditioning on the service requirement, i.e., $\mathbb{E}(\mathrm{e}^{-sW}) = \int_0^\infty \mathbb{E}(\mathrm{e}^{-sW}|B=\tau)\mu\mathrm{e}^{-\mu\tau}\mathrm{d}\tau$ and $\mathbb{E}(W) = \int_0^\infty \mathbb{E}(W|B=\tau)\mu\mathrm{e}^{-\mu\tau}\mathrm{d}\tau$. The mean unconditional delay is readily seen to equal $\mathbb{E}(W) = \frac{\rho/\mu}{1-\rho}+\mathbb{E}(R_V)$. This is in agreement with a well-known result [32] for the mean delay in an $M/M/1$-FCFS queue with exhaustive service and multiple vacations. That is no surprise; indeed, in the case of exponential service requirements, the queue length distributions for PS and FCFS are the same, hence the mean queue lengths are the same, and hence by Little's formula also the mean delays are the same.*

Remark 2 ($M/M/1$-PS without vacations) *For an $M/M/1$-PS queue without vacations, we have $\mathbb{E}(R_V)=0$, and we retrieve the known formula $\mathbb{E}(W|B=\tau) = \rho\tau/(1-\rho)$.*

Proof of Proposition 1: We have

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sW}|B=\tau) &= \sum_{n=1}^{\infty} p_n w_n(\tau,s)\\ &\quad +(1-\rho)\int_{u=0}^{\infty}\int_{v=0}^{\infty}\mathrm{e}^{-sv}\sum_{n=0}^{\infty}\mathrm{e}^{-\lambda(u+v)}\frac{(\lambda(u+v))^n}{n!}w_n(\tau,s)\mathrm{d}\mathbb{P}(P_V<u,R_V<v),\end{aligned}$$

with $p_n = \mathbb{P}(K$ arrives in a busy period and sees n customers upon arrival). The first term in the above equation corresponds to the case that the tagged customer arrives in a busy period and finds n customers upon arrival. The system behaves like an ordinary $M/M/1$-PS without vacations as far as his delay is concerned, hence the LST of the conditional delay of K is $w_n(\tau,s)$. The second term corresponds to the case that the tagged customer arrives during a vacation period. Given the length of the elapsed period of vacation u and the length of the residual vacation v, the probability of n customers arriving to the system during the vacation of length $u+v$ (excluding the tagged customer K) is given by $\mathrm{e}^{-\lambda(u+v)}(\lambda(u+v))^n/n!$. Since the policy is PS, after vacation, the tagged customer sees its delay as if it arrives at a PS queue where it meets n customers, i.e., $w_n(\tau,s)$.

From (2) we obtain

$$\mathbb{E}(\mathrm{e}^{-sW}|B=\tau) = \sum_{n=1}^{\infty} p_n w_n(\tau,s) + (1-\rho)G(\tau,s)\int_{u=0}^{\infty}\int_{v=0}^{\infty}\mathrm{e}^{-sv}\mathrm{e}^{-\lambda(1-\beta(\tau,s))(u+v)}\mathrm{d}\mathbb{P}(P_V<u,R_V<v). \tag{5}$$

For general vacations we have (see [11, p. 113]; see also Remark 3 below)

$$\int_{u=0}^{\infty}\int_{v=0}^{\infty} \mathrm{e}^{-sv}\mathrm{e}^{-\lambda(1-\beta(\tau,s))(u+v)}\mathrm{d}\mathbb{P}(P_V<u,R_V<v) = \frac{\mathbb{E}(\mathrm{e}^{-\lambda(1-\beta(\tau,s))V}) - \mathbb{E}(\mathrm{e}^{-(\lambda(1-\beta(\tau,s))+s)V})}{s\mathbb{E}(V)}.$$

This follows from

$$\begin{aligned}
&\int_{u=0}^{\infty}\int_{v=0}^{\infty} \mathrm{e}^{-au}\mathrm{e}^{-bv}\mathrm{d}\mathbb{P}(P_V<u,R_V<v) = \frac{1}{\mathbb{E}(V)}\int_{u=0}^{\infty}\int_{v=0}^{\infty} \mathrm{e}^{-au-bv} f_V(u+v)\mathrm{d}v\mathrm{d}u \\
&= \frac{1}{\mathbb{E}(V)}\int_{z=0}^{\infty}\int_{v=0}^{z} \mathrm{e}^{-a(z-v)-bv} f_V(z)\mathrm{d}v\mathrm{d}z = \frac{1}{\mathbb{E}(V)}\int_{z=0}^{\infty} \mathrm{e}^{-az}\frac{1-\mathrm{e}^{-(b-a)z}}{b-a} f_V(z)\mathrm{d}z \\
&= \frac{1}{(b-a)\mathbb{E}(V)}(\mathbb{E}(\mathrm{e}^{-aV}) - \mathbb{E}(\mathrm{e}^{-bV})), \qquad (6)
\end{aligned}$$

where in the first step we used that $\mathbb{P}(P_V>u,R_V>v) = \frac{1}{\mathbb{E}(V)}\int_{u+v}^{\infty}(1-F_V(w))\mathrm{d}w$, see for example [3, p. 24].

We now consider p_n, which can be written as $p_n = \rho\mathbb{P}(K \text{ sees } n | K \text{ arrives in a busy period}) = \rho\mathbb{P}(N_{busy}=n)$, with N_{busy} the steady-state number of customers in a busy period. We have

$$\mathbb{E}(z^N) = (1-\rho)\mathbb{E}(z^{N_{vac}}) + \rho\mathbb{E}(z^{N_{busy}}), \qquad (7)$$

with N the steady-state number of customers, and N_{vac} the steady-state number of customers in the period of (subsequent multiple) vacations. Since the service requirements are exponentially distributed, the queue lengths are stochastically the same as the queue lengths in the $M/M/1$-FCFS queue with multiple vacations. From [7, Lemma 2.2.1] we get

$$\mathbb{E}(z^{N_{vac}}) = \frac{1-\mathbb{E}(z^{N_{end}})}{(1-z)\mathbb{E}(N_{end})}, \qquad (8)$$

with N_{end} the steady-state number of customers present in the system at the end of a vacation period, and (because of multiple vacations)

$$\mathbb{E}(z^{N_{end}}) = \frac{\tilde{V}(\lambda(1-z)) - \tilde{V}(\lambda)}{1-\tilde{V}(\lambda)}, \qquad (9)$$

(follows since $\mathbb{E}(z^{N_{end}}) = \mathbb{E}(z^{N_s}|N_s>0) = \frac{\mathbb{E}(z^{N_s};N_s>0)}{\mathbb{P}(N_s>0)} = \frac{\mathbb{E}(z^{N_s})-\mathbb{P}(N_s=0)}{1-\mathbb{P}(N_s=0)}$, with N_s the steady-state number of customers for a system with single vacations, so $\mathbb{E}(z^{N_s}) = \tilde{V}(\lambda(1-z)))$. Hence,

$$\mathbb{E}(z^{N_{vac}}) = \frac{1-\mathbb{E}(z^{N_{end}})}{(1-z)\mathbb{E}(N_{end})} = \frac{1-\tilde{V}(\lambda(1-z))}{\lambda(1-z)\mathbb{E}(V)}. \qquad (10)$$

Indeed, the last term is the PGF of the number of arrivals in P_V. Let $N_{M/M/1}$ be the number of customers present in steady state in a standard $M/M/1$ queue. Fuhrmann & Cooper [14] state that

$$N \stackrel{d}{=} N_{M/M/1} + N_{vac},$$

the latter two being independent. We thus have

$$\mathbb{E}(z^N) = \frac{1-\rho}{1-\rho z}\mathbb{E}(z^{N_{vac}}). \qquad (11)$$

Combining (7), (10) and (11) we obtain

$$\mathbb{E}(z^{N_{busy}}) \;=\; \frac{1}{\rho}\left(\frac{1-\rho}{1-\rho z}\mathbb{E}(z^{N_{vac}}) - (1-\rho)\mathbb{E}(z^{N_{vac}})\right) = z\frac{1-\rho}{1-\rho z}\frac{1-\tilde{V}(\lambda(1-z))}{\lambda(1-z)\mathbb{E}(V)},$$

which implies

$$N_{busy} \stackrel{d}{=} 1 + N_{M/M/1} + N_{vac}.$$

It then follows that $\mathbb{P}(N_{busy}=n) = \mathbb{P}(N_{M/M/1}+N_{vac}=n-1)$.

Since $p_n = \rho\mathbb{P}(N_{busy} = n)$, we have

$$\begin{aligned}
\sum_{n=1}^{\infty} p_n w_n(\tau,s) &= \sum_{n=1}^{\infty} \frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\beta(\tau,s)^n \rho\mathbb{P}(N_{M/M/1}+N_{vac}=n-1)\\
&= \beta(\tau,s)\rho\frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\sum_{n=0}^{\infty}\beta(\tau,s)^n\mathbb{P}(N_{M/M/1}+N_{vac}=n)\\
&= \beta(\tau,s)\rho\frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\mathbb{E}(\beta(\tau,s)^{N_{M/M/1}+N_{vac}})\\
&= \beta(\tau,s)\rho\frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}\frac{1-\rho}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}.
\end{aligned}$$

The latter is equal to $\beta(\tau,s)\rho G(\tau,s)\frac{1-\rho}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}$, which concludes the proof.

The derivation of the expression for the conditional mean delay as stated in Equation (4) can be found in Appendix 2. □

Remark 3 *The three key ingredients of our approach were knowledge of (i) the steady-state queue length distribution during busy periods and during vacations, (ii) the conditional sojourn time LST $w_n(\tau,s)$ in an $M/M/1$ PS system, and (iii) knowledge of the joint LST of past and residual vacation time, as given in (6). When the tagged customer arrives during a busy period, it will be served during that same busy period, and we can immediately use $w_n(\tau,s)$ from [10]. When the tagged customer arrives during a vacation, it will be served in the subsequent busy period. Because service is exhaustive, the system was empty at the beginning of the vacation. Hence we only need to know the number of other arrivals in the same vacation, before and after that of the tagged customer. So we only need to know the joint distribution of the past and residual length of one arbitrary vacation. Formula (6), a familiar result from renewal theory (hence with i.i.d. vacations) will still remain valid when successive vacations are dependent on each other and/or on previous busy periods. As mentioned in [8], this can be seen through the use of Palm theory, which can be employed to capture the biases that are mentioned above. The Palm framework allows one to work with the fact that, under the Palm measure induced by the point process consisting of the times at which a cycle begins, the sequence of cycle lengths formed in the stationary version of this polling system forms a stationary sequence, but does not form an i.i.d. sequence. For the use of Palm theory in queueing we refer to [3] and [31]; see also [33].*

3.1 Scaled vacations

In this subsection we are interested in the behavior of the system as the vacations grow to infinity. Scaling the length of the vacations will be of practical interest in the context of polling systems, as will be considered in Section 5.

We assume that the length of the vacation is a function of m, V_m, and grows with $1/g(m)$ as $m \to \infty$, where $g(m) \downarrow 0$. More precisely, we assume that the scaled vacation period $g(m)V_m$ converges in distribution to V^{Sc}, where V^{Sc} is non-defective. We denote the LST by $\tilde{V}^{Sc}(s) := \lim_{m\to\infty}\tilde{V}(g(m)s)$. In addition, we allow the traffic load to depend on m, having a limit $\hat{\rho}$ as $m \to \infty$, with $\hat{\rho} < 1$.

When the length of the vacation period grows to infinity, the delay a customer experiences will grow as well. It turns out that $g(m)$ is the appropriate scaling for the delay. The following proposition is a direct consequence of Proposition 1 and gives the LST of $\lim_{m\to\infty} g(m)W$.

Proposition 2 *Assume $g(m)V_m$ converges in distribution to V^{Sc}, where V^{Sc} is non-defective. We have*

$$\lim_{m\to\infty}\mathbb{E}(\mathrm{e}^{-sg(m)W}|B=\tau) = \hat{\rho}\frac{1-\tilde{V}^{Sc}(s\omega(\tau))}{s\omega(\tau)\mathbb{E}(V^{Sc})} + (1-\hat{\rho})\frac{\tilde{V}^{Sc}(s\omega(\tau)) - \tilde{V}^{Sc}(s(\omega(\tau)+1))}{s\mathbb{E}(V^{Sc})}, \qquad (12)$$

with $\omega(\tau) := \frac{\hat{\rho}}{1-\hat{\rho}}(1-\mathrm{e}^{-\mu\tau(1-\hat{\rho})})$.

Note that Equation (12) can be rewritten as

$$\begin{aligned}
\lim_{m\to\infty}\mathbb{E}(\mathrm{e}^{-sg(m)W}|B=\tau) &= \hat{\rho}\mathbb{E}(\mathrm{e}^{-sR_{\omega(\tau)V^{Sc}}}) + (1-\hat{\rho})\mathbb{E}(\mathrm{e}^{-s(\omega(\tau)P_{V^{Sc}}+(\omega(\tau)+1)R_{V^{Sc}})})\\
&= \hat{\rho}\mathbb{E}(\mathrm{e}^{-sR_{\omega(\tau)V^{Sc}}}) + (1-\hat{\rho})\mathbb{E}(\mathrm{e}^{-s(R_{V^{Sc}}+\omega(\tau)V^{Sc})}),
\end{aligned}$$

where we used that $\mathbb{E}(\mathrm{e}^{-(aP_{V^{Sc}}+bR_{V^{Sc}})}) = \frac{1}{(b-a)\mathbb{E}(V^{Sc})}(\tilde{V}^{Sc}(a)-\tilde{V}^{Sc}(b))$, see Equation (6), with $P_{V^{Sc}}$ and $R_{V^{Sc}}$ the past and residual length of V^{Sc}, respectively.

Proof of Proposition 2: We have that $\lim_{m\to\infty} G(\tau, g(m)s) = 1$ and $\lim_{m\to\infty}\beta(\tau, g(m)s) = 1$. Using Taylor expansion we obtain $\beta(\tau, g(m)s) = 1 + g(m)s\frac{\partial}{\partial s}\beta(\tau,s)|_{s=0} + O(g(m)^2)$ as $m\to\infty$. Hence, from Lemma 1 we obtain

$$\lim_{m\to\infty}\frac{\lambda(1-\beta(\tau,g(m)s))}{g(m)} = s\frac{\hat{\rho}}{1-\hat{\rho}}(1-\mathrm{e}^{-\mu\tau(1-\hat{\rho})}) = s\omega(\tau)$$

From Proposition 1 we directly have

$$\begin{aligned}
&\lim_{m\to\infty}\mathbb{E}(\mathrm{e}^{-sg(m)W}|B=\tau)\\
&= \hat{\rho}\frac{1-\tilde{V}^{Sc}(\frac{\lambda(1-\beta(\tau,g(m)s))}{g(m)})}{\frac{\lambda(1-\beta(\tau,g(m)s))}{g(m)}\mathbb{E}(V^{Sc})} + (1-\hat{\rho})\frac{\tilde{V}^{Sc}(\frac{\lambda(1-\beta(\tau,g(m)s))}{g(m)}) - \tilde{V}^{Sc}(\frac{\lambda(1-\beta(\tau,g(m)s))}{g(m)}+s)}{s\mathbb{E}(V^{Sc})}\\
&= \hat{\rho}\frac{1-\tilde{V}^{Sc}(s\omega(\tau))}{s\omega(\tau)\mathbb{E}(V^{Sc})} + (1-\hat{\rho})\frac{\tilde{V}^{Sc}(s\omega(\tau)) - \tilde{V}^{Sc}(s(\omega(\tau)+1))}{s\mathbb{E}(V^{Sc})},
\end{aligned}$$

which concludes the proof. □

4 $M/G/1$ processor sharing queue with multiple vacations

In this section we consider an $M/G/1$ processor sharing queue with multiple vacations. We will obtain an expression for the mean conditional delay of a tagged customer of size τ. This in contrast to Section 3 where we obtained the full distribution of the conditional delay, however, restricted to service requirements that are exponentially distributed.

The sojourn time T of the tagged customer of size τ is made up of two components, the queueing time Q (time between arrival and the beginning of service) and the time between the beginning of service and service completion, denoted by D. Since the scheduling discipline is PS, the queueing time of a customer is positive only if it arrives during a vacation period. The probability that the tagged customer finds the server on vacation is $1-\rho$. Conditioning on when the tagged customer arrives at the queue we obtain that $\mathbb{E}(Q) = 0\cdot\rho + (1-\rho)\mathbb{E}(R_V)$. Introducing $D(\tau) := \mathbb{E}(D|B=\tau)$ we have:

$$\mathbb{E}(W|B=\tau) = \mathbb{E}(Q) + D(\tau) - \tau = (1-\rho)\mathbb{E}(R_V) + D(\tau) - \tau. \tag{13}$$

In the following Proposition we will develop an integro-differential equation that $\frac{\mathrm{d}}{\mathrm{d}\tau}D(\tau)$ must satisfy.

Proposition 3 *The mean conditional delay in an $M/G/1$-PS queue with multiple vacations is given by (13), where $\frac{\mathrm{d}}{\mathrm{d}\tau}D(\tau)$ is the unique solution $z(\tau)$ of*

$$z(\tau) = 1 + (1-\rho)2\lambda\mathbb{E}(R_V)\overline{F}(\tau) + \lambda\int_0^\infty z(y)\overline{F}(\tau+y)\mathrm{d}y + \lambda\int_0^\tau z(y)\overline{F}(\tau-y)\mathrm{d}y. \tag{14}$$

The proof approach we follow was initiated by Kleinrock et al. [18] (see also [16]) who studied a processor sharing queue with batch Poisson arrivals. In [23] the author derived the conditional sojourn time for the foreground-background queue (also known as least-attained-service queue) using the tagged-customer approach. The same approach was used in the seminal paper [12] which studied the conditional delay in a discriminatory processor-sharing queue. More recently this approach has been used in [1, 4, 26].

Interestingly we observe that Equation (14) is related to the equation that characterizes the mean sojourn time in a processor sharing queue with batch arrivals (see Equation (1) in [1]). In fact, the integro-differential equation (14) coincides with that of a batch processor sharing queue where the batch arrival rate is λ, and the first and second moment of the batch size distribution are given by 1 and $(1-\rho)2\lambda\mathbb{E}(R_V)+1$, respectively. (This in particular means that batches of size 0 occur with strictly positive probability.) This integro-differential equation has been solved in [16, Section 4.7], [18] for exponential service requirements, see also Remark 4. The integro-differential equation has been

solved in [26] for hyper-exponential service requirements and in [4] for distributions having rational LST. In [1] the integro-differential equation has been studied for general service requirements, and properties of the solution have been obtained; see Appendix 3 for more details.

Proof of Proposition 3: Since the employed policy is PS, $D(\tau)$ can be interpreted as the average time needed for a customer in order to get τ units of service. Since at each moment in time all customers equally share the total amount of capacity available, for sufficiently small Δ we have

$$D(\tau + \Delta) = D(\tau) + \Delta + \Delta \mathbb{E}(L(\tau)) + o(\Delta),$$

where $L(\tau)$ is the number of customers in the system when the tagged customer is receiving service and has attained τ units of service. Here we used that when the tagged customer obtains Δ units of service, any other customer in the system also receives Δ units of service.

Taking the limit $\Delta \to 0$, it is readily seen that the derivative of the expected conditional sojourn time exists and is given by

$$D'(\tau) = 1 + \mathbb{E}(L(\tau)). \tag{15}$$

We now develop an expression for $L(\tau)$. Let us write $L(\tau) = L_1(\tau) + L_2(\tau)$, where:

- $L_1(\tau)$ is the number of customers that were in the system when the tagged customer started service, and are still present when the tagged customer has received τ units of service.
- $L_2(\tau)$ is the number of customers that arrive during the service of the tagged customer, and are still present when the tagged customer has received τ units of service.

Let us consider $\mathbb{E}(L_1(\tau))$. With probability $1-\rho$ the tagged customer finds the server idle. Hence we have

$$\begin{aligned} \mathbb{E}(L_1(\tau)) &= (1-\rho)\mathbb{E}(L_1(\tau)|K \text{ arrived in vacation period}) \\ &\quad +\rho\mathbb{E}(L_1(\tau)|K \text{ arrived in busy period}). \end{aligned} \tag{16}$$

At the start of the busy period there are on average $2\lambda\mathbb{E}(R_V)\overline{F}(\tau)$ customers present with service requirement larger than or equal to τ (the factor 2 comes from the fact that the expected remaining vacation is equal to the expected elapsed time of the vacation period). Hence,

$$\mathbb{E}(L_1(\tau)|K \text{ arrived in vacation period}) = 2\lambda\mathbb{E}(R_V)\overline{F}(\tau). \tag{17}$$

We will express $\mathbb{E}(L_1(\tau)|K$ arrived in busy period) as a function of $N(y)$, the number of customers in steady-state that have attained at most y units of service. Using Little-type arguments, it was shown in [24] (previously obtained by Kleinrock and Coffman in [17]) that in an arbitrary ergodic system and for an arbitrary scheduling discipline,

$$\mathrm{d}\mathbb{E}(N(y)) = \lambda D'(y)\overline{F}(y)\mathrm{d}y, \quad y > 0. \tag{18}$$

To explain (18) we interpret $\mathrm{d}\mathbb{E}(N(y)) = \mathbb{E}(N(y+\mathrm{d}y)) - \mathbb{E}(N(y)) + o(\mathrm{d}y)$ as the mean number of customers that have attained service in $[y, y+\mathrm{d}y)$ and apply Little's theorem to the black box formed by customers that have attained service in $[y, y+\mathrm{d}y)$. The arrival rate of such customers is $\lambda\overline{F}(y)$ and the mean amount of time that a customer spends in the black box, i.e., the expected amount of time a customer spends in the system in order for its attained service to pass from y to $y+\mathrm{d}y$, is $D'(y)\mathrm{d}y$ (follows since $D'(y) = \frac{D(y+\mathrm{d}y)-D(y)}{\mathrm{d}y} + o(1)$), and Equation (18) follows.

Using the PASTA property, $N(y)$ can be interpreted as the number of customers upon arrival of the tagged customer that have attained service less than or equal to y. Conditioning on the moment that the tagged customer arrives, we obtain

$$\mathrm{d}\mathbb{E}(N(y)) = (1-\rho)\cdot 0 + \rho\mathrm{d}\mathbb{E}(N(y)|K \text{ arrived in busy period}). \tag{19}$$

Using that a customer that has received y units of service when the tagged customer has arrived, is with probability $\frac{\overline{F}(\tau+y)}{\overline{F}(y)}$ present in the system when the tagged customer has received τ units of

service, together with the fact that $N(y)$ is the number of customers with attained service less than or equal to y that the tagged customer finds upon arrival, we obtain that

$$\begin{aligned}
&\mathbb{E}(L_1(\tau)|K \text{ arrived in busy period}) \\
&= \int_0^\infty \frac{\overline{F}(\tau+y)}{\overline{F}(y)} \mathrm{d}\mathbb{E}(N(y)|K \text{ arrived in busy period}) \\
&= \int_0^\infty \frac{1}{\rho}\frac{\overline{F}(\tau+y)}{\overline{F}(y)} \mathrm{d}\mathbb{E}(N(y)) = \lambda \int_0^\infty \frac{D'(y)}{\rho}\overline{F}(\tau+y)\mathrm{d}y, \qquad (20)
\end{aligned}$$

where the last steps follow from (18) and (19). Substituting (17) and (20) in (16) we get

$$\mathbb{E}(L_1(\tau)) = (1-\rho)\lambda 2\mathbb{E}(R_V)\overline{F}(\tau) + \lambda \int_0^\infty D'(y)\overline{F}(\tau+y)\mathrm{d}y. \qquad (21)$$

We now focus on $\mathbb{E}(L_2(\tau))$. The tagged customer needs $D'(y)\mathrm{d}y$ units of time in order for its attained service to pass from y to $y+\mathrm{d}y$. The mean number of arrivals during this time is thus $\lambda D'(y)\mathrm{d}y$. A customer that arrives at the system when the tagged customer has received y units of service is with probability $\overline{F}(\tau-y)$ present in the system when the tagged customer has received τ units of service. Now integrating the attained service of the tagged customer, y, from 0 to τ, we get

$$\mathbb{E}(L_2(\tau)) = \lambda \int_0^\tau D'(y)\overline{F}(\tau-y)dy. \qquad (22)$$

Combining (15), (21) and (22) we obtain

$$D'(\tau) = 1 + (1-\rho)\lambda 2\mathbb{E}(R_V)\overline{F}(\tau) + \lambda \int_0^\infty D'(y)\overline{F}(\tau+y)\mathrm{d}y + \lambda \int_0^\tau D'(y)\overline{F}(\tau-y)\mathrm{d}y,$$

completing the proof of the proposition.

The existence and uniqueness of the solution of Equation (14) is proved in Theorem 1 and Lemma 3 of [1], respectively, see Appendix 3 for more details. □

Remark 4 ($M/M/1$-PS with multiple vacations) *In the case of exponentially distributed service requirements, Equation (14) can be solved analytically and we can thus verify that Equation (13) gives the same result as the conditional expectation obtained in Proposition 1.*

We thus need to solve Equation (14) under the assumption of exponential service requirements. For simplicity of notation let $k = (1-\rho)2\lambda\mathbb{E}(R_V)$. Taking the derivative of (14) we get

$$\begin{aligned}
z'(\tau) &= -k\mu e^{-\mu\tau} - \mu\lambda \int_0^\infty z(y)e^{-\mu(\tau+y)}dy + \lambda z(\tau) - \mu\lambda \int_0^\tau z(y)e^{-\mu(\tau-y)}dy \\
&= -\mu z(\tau) + \mu + \lambda z(\tau) = z(\tau)(\lambda-\mu) + \mu.
\end{aligned}$$

The solution to this differential equation is

$$z(\tau) = Re^{(\lambda-\mu)\tau} - \frac{\mu}{\lambda-\mu}, \qquad (23)$$

where R is some arbitrary constant.

Taking $\tau = 0$ in Equations (14) and (23) we get $R - \frac{\mu}{\lambda-\mu} = 1 + k - \lambda\frac{\mu}{(\lambda-\mu)\mu} + \lambda R\int_0^\infty e^{(\lambda-2\mu)y}dy$ and solving for R we get

$$R = \frac{k}{1-\frac{\lambda}{2\mu-\lambda}} = \frac{k(2-\rho)}{2(1-\rho)}.$$

Recall that $z(\tau)$ represents $\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}(D|B=\tau)$, hence from (23) we obtain $\mathbb{E}(D|B=\tau) = \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{E}(D|B=y)\mathrm{d}y = -\frac{k(2-\rho)}{2\mu(1-\rho)^2}e^{-\mu(1-\rho)\tau} + \frac{\tau}{1-\rho} + C$, for some constant C. Since the discipline is PS, we have $\mathbb{E}(D|B=0) = 0$, which implies $C = \frac{k(2-\rho)}{2\mu(1-\rho)^2}$. Hence,

$$\mathbb{E}(D|B=\tau) = \frac{k(2-\rho)}{2\mu(1-\rho)^2}\left(1-e^{-\mu(1-\rho)\tau}\right) + \frac{\tau}{1-\rho} = \frac{\rho(2-\rho)\mathbb{E}(R_V)}{(1-\rho)}\left(1-e^{-\mu(1-\rho)\tau}\right) + \frac{\tau}{1-\rho}.$$

We conclude that the expression in Equation (13) indeed coincides with the mean delay as obtained in Proposition 1.

Remark 5 ($M/G/1$-PS) *For the ordinary $M/G/1$-PS queue, the mean conditional delay is known to be $\frac{\rho\tau}{1-\rho}$, in agreement with Equation (13) when setting $\mathbb{E}(R_V) = 0$. That expression for the mean delay follows, since the unique solution of (14) is $z(\tau) = \frac{1}{1-\rho}$ in the case $\mathbb{E}(R_V) = 0$. To check this let us substitute an arbitrary constant $z(\tau) = Z$ in (14). This gives*

$$Z = 1 + \lambda Z \left(\int_0^\infty \overline{F}(\tau + y)\mathrm{d}y + \int_0^\tau \overline{F}(\tau - y)\mathrm{d}y \right) = 1 + \lambda Z \int_0^\infty \overline{F}(y)\mathrm{d}y = 1 + Z\rho.$$

Hence, $z(\tau) = Z = \frac{1}{1-\rho}$ is the unique solution so that $D(\tau) = \frac{\tau}{1-\rho}$ and from (13) it follows that $\mathbb{E}(W|B = \tau) = \frac{\rho\tau}{1-\rho}$.

4.1 Asymptotic behavior

For general service requirements, we were not able to solve (14) analytically. However, as $\tau \to \infty$, the limiting behavior can be characterized in closed form. We will show that $\mathbb{E}(T|B = \tau)$ has an asymptote with slope $\tau/(1-\rho)$ of which the bias term can be explicitly calculated. The analysis is similar to that in [1].

Before stating the result, we present an auxiliary result for the workload in the system. Sample-path wise, the workload is independent of the work-conserving scheduling discipline being deployed. (We say that a scheduling discipline is work-conserving if the capacity is fully used whenever it is available and there are customers in the system.) We have the following result for the mean workload in the system for any work-conserving discipline.

Lemma 2 *Consider a single server queue (and any work-conserving scheduling discipline) with multiple vacations. The mean workload in the system is*

$$\frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)} + \rho\mathbb{E}(R_V), \tag{24}$$

and for any work-conserving scheduling discipline π the following conservation law holds:

$$\lambda \int_0^\infty \mathbb{E}(T^\pi|B = x)\overline{F}(x)\mathrm{d}x = \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)} + \rho\mathbb{E}(R_V), \tag{25}$$

where T^π represents the sojourn time under discipline π, i.e., $T^\pi = B + W^\pi$.

Proof: The mean sojourn time in a FCFS queue with multiple vacations is, cf. [34, equation (2.2.5)],

$$\mathbb{E}(T^{FCFS}|B = x) = x + \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)} + \mathbb{E}(R_V). \tag{26}$$

In [2] it was shown that for any work-conserving discipline π, the mean workload in the system is equal to

$$\lambda \int_0^\infty \mathbb{E}(T^\pi|B = x)\overline{F}(x)\mathrm{d}x. \tag{27}$$

This expression follows from the generalized Little's law known as $H = \lambda G$ [9] (note that the integral equals the expected contribution of a customer to the workload).

FCFS is a work-conserving discipline, hence substituting (26) in (27), we obtain that the mean workload in the system under any work-conserving discipline is given by (24), and Equation (25) follows directly. □

We note that in the presence of vacations, the mean delay in the FCFS queue (waiting time), i.e., $\mathbb{E}(W^{FCFS}) = \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)} + \mathbb{E}(R_V)$, see (26), does not coincide with the mean workload (Equation (24)). The difference is in the factor ρ in the term corresponding to the vacations.

We now present the asymptotic behavior of the mean conditional sojourn time.

Proposition 4 *The mean conditional sojourn time $\mathbb{E}(T^{PS}|B = \tau)$ has an asymptote of slope $\frac{\tau}{1-\rho}$ and bias term*

$$\lim_{\tau\to\infty} \left(\mathbb{E}(T^{PS}|B = \tau) - \frac{\tau}{1-\rho} \right) = \frac{\mathbb{E}(R_V)}{1-\rho}. \tag{28}$$

Proof: For ease of notation we again use the function $D(\tau) := \mathbb{E}(T^{PS}|B=\tau)$. From Appendix 3 we see that $D(\tau) - \frac{\tau}{1-\rho}$ is increasing with respect to τ and upper bounded, hence, the bias term $\delta D(\tau) := D(\tau) - \frac{\tau}{1-\rho}$ has a proper limit as $\tau \to \infty$. We can write $\lim_{\tau\to\infty}\left(D(\tau) - \frac{\tau}{1-\rho}\right) = \int_0^\infty \delta D'(x)\mathrm{d}x$. Using the relation $D'(\tau) = \delta D'(\tau) + \frac{1}{1-\rho}$, we obtain from (14) that

$$\delta D'(x) = (1-\rho)2\lambda\mathbb{E}(R_V)\overline{F}(x) + \lambda\int_0^\infty \delta D'(y)\overline{F}(x+y)\mathrm{d}y + \lambda\int_0^x \delta D'(y)\overline{F}(x-y)\mathrm{d}y. \tag{29}$$

The first integral can be written as

$$\begin{aligned}\lambda\int_0^\infty \delta D'(y)\overline{F}(x+y)\mathrm{d}y &= \lambda\left(\delta D(y)\overline{F}(x+y)\right)|_{y=0}^{y=\infty} + \lambda\int_0^\infty \delta D(y)\mathrm{d}F(x+y)\\ &= \lambda\int_0^\infty \delta D(y)\mathrm{d}F(x+y).\end{aligned} \tag{30}$$

The last step follows from the following two facts: (i) there exists an $L<\infty$ such that $\delta D(x) \le Lx$ for all $x \ge 0$ (see [1, Lemma 4] for details), (ii) since $\int_0^\infty x\mathrm{d}F(x) = \int_0^\infty \overline{F}(x)\mathrm{d}x + \lim_{x\to\infty} x\overline{F}(x)$, and $\mathbb{E}(B) < \infty$, we obtain $\lim_{x\to\infty} x\overline{F}(x) = 0$.

Using (29) and (30), we obtain that the bias term satisfies:

$$\begin{aligned}&\lim_{\tau\to\infty}\left(D(\tau) - \frac{\tau}{1-\rho}\right) = \int_0^\infty \delta D'(x)\mathrm{d}x\\
&= \lambda\int_0^\infty\int_0^\infty \delta D'(y)\overline{F}(x+y)\mathrm{d}y\mathrm{d}x + \lambda\int_0^\infty\int_0^x \delta D'(y)\overline{F}(x-y)\mathrm{d}y\mathrm{d}x + (1-\rho)2\lambda\mathbb{E}(R_V)\int_0^\infty \overline{F}(x)\mathrm{d}x\\
&= \lambda\int_{x=0}^\infty\int_{y=0}^\infty \delta D(y)\mathrm{d}F(x+y)\mathrm{d}x + \lambda\int_0^\infty \delta D'(y)\int_y^\infty \overline{F}(x-y)\mathrm{d}x\mathrm{d}y + (1-\rho)2\lambda\mathbb{E}(R_V)\mathbb{E}(B)\\
&= \lambda\int_{y=0}^\infty \delta D(y)\int_{x=0}^\infty \mathrm{d}F(x+y)\mathrm{d}y + \lambda\int_0^\infty \delta D'(y)\int_0^\infty \overline{F}(h)\mathrm{d}h\mathrm{d}y + (1-\rho)2\lambda\mathbb{E}(R_V)\mathbb{E}(B)\\
&= \lambda\int_0^\infty \delta D(y)\overline{F}(y)\mathrm{d}y + \rho\int_0^\infty \delta D'(y)\mathrm{d}y + (1-\rho)2\lambda\mathbb{E}(R_V)\mathbb{E}(B)\\
&= \mathbb{E}(R_V)\rho^2 + \rho\int_0^\infty \delta D'(y)\mathrm{d}y + (1-\rho)2\lambda\mathbb{E}(R_V)\mathbb{E}(B),\end{aligned} \tag{31}$$

where in the last step we used that $\lambda\int_0^\infty \delta D(x)\overline{F}(x)\mathrm{d}x = \mathbb{E}(R_V)\rho - (1-\rho)\rho\mathbb{E}(R_V) = \mathbb{E}(R_V)\rho^2$, which follows from substituting $\mathbb{E}(T^{PS}|B=\tau) = (1-\rho)\mathbb{E}(R_V) + D(\tau) = (1-\rho)\mathbb{E}(R_V) + \delta D(\tau) + \frac{\tau}{1-\rho}$ (see (13)) into (25).

Solving Equation (31) for $\int_0^\infty \delta D'(x)\mathrm{d}x$ we obtain

$$\lim_{\tau\to\infty}\left(D(\tau) - \frac{\tau}{1-\rho}\right) = \int_0^\infty \delta D'(x)\mathrm{d}x = \mathbb{E}(R_V)\frac{\rho(2-\rho)}{1-\rho},$$

and from (13) we obtain Equation (28). □

5 Processor Sharing in Polling Systems

In this section we consider a polling system consisting of N queues $Q_1, \ldots, Q_N$, cyclically visited by a single server. Customers arrive according to independent Poisson processes with arrival rate λ_i to Q_i. Customers in Q_i have generally distributed service requirements B_i. We define $\rho_i = \lambda_i\mathbb{E}(B_i)$ and we denote by $\rho^* = \sum_{i=1}^N \rho_i$ the total load. The random switch-over time of the server from Q_i to Q_{i+1} is denoted by S_i, and $S = \sum_{i=1}^N S_i$. All inter-arrival times, service requirements and switch-over times are assumed to be independent. Let I_i (R_{I_i}) denote the (residual) length of an intervisit time for Q_i in the polling system. The LST of I_1 is denoted by $\tilde{I}_1(\cdot)$.

When the server arrives at Q_i it serves a number of customers according to a certain visit discipline. We assume that Q_1 uses the exhaustive visit discipline (the server serves the queue until it has become empty), and any other queue Q_i uses any visit discipline that has the branching property as defined

in [13, 28] (this includes the exhaustive and gated disciplines). We assume $\sum_{i=1}^{N} \rho^* < 1$ throughout the section in order to guarantee stability of the system, see [28]. The queue Q_1 uses PS as scheduling policy, and Q_i, $i \neq 1$, employs a work-conserving scheduling policy.

Let W_1 be the delay (sojourn time minus service requirement) of a tagged customer in Q_1 with size B_1. The conditional sojourn time in Q_1 can be studied using the theory developed in Section 3. This can be seen as follows. From the point of view of customers arriving at Q_1, the server is a PS queue where, once Q_1 empties, the server is unavailable during an intervisit time. When the server returns from vacation but finds no customers in Q_1, the server is again unavailable during an intervisit time, etc. Hence, Q_1 can be modeled as an $M/G/1$-PS queue with traffic load ρ_1 and multiple vacations, where an arbitrary vacation length is distributed as I_1.

Remark 6 *Notice that lengths of successive intervisit times will be dependent, and that the length of an intervisit time will depend on the length of the preceding visit time. However, as observed in Remark 3, it was not required in Section 3 that successive vacations are independent of each other and of previous visit periods). Hence, in the next subsection, we shall be able to use the same reasoning as was used in Section 3 for an $M/M/1$ PS system with exhaustive service and vacations to obtain the sojourn time LST in Q_1 of a polling system – provided Q_1 receives exhaustive service and has exponential service requirements.*

5.1 Exponential service requirements in Q_1

In this section we assume that a customer in Q_1 has an exponentially distributed service requirement denoted by B_1 with $\mathbb{E}(B_1) = 1/\mu_1$. The proof of the following Proposition proceeds just like the proof of Proposition 1 and yields the LST of the conditional delay for a customer in Q_1.

Proposition 5 *Assume customers in Q_1 have exponentially distributed service requirements. Then,*

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sW_1}|B_1=\tau) &= \rho_1 G_1(\tau,s)\frac{\beta_1(\tau,s)(1-\rho_1)}{1-\rho_1\beta_1(\tau,s)}\frac{1-\tilde{I}_1(\lambda_1(1-\beta_1(\tau,s)))}{\lambda_1(1-\beta_1(\tau,s))\mathbb{E}(I_1)}\\ &\quad +(1-\rho_1)G_1(\tau,s)\frac{\tilde{I}_1(\lambda(1-\beta(\tau,s)))-\tilde{I}_1(\lambda(1-\beta(\tau,s))+s)}{s\mathbb{E}(I_1)},\end{aligned}$$

with $G_1(\tau,s)$ and $\beta_1(\tau,s)$ replacing $G(\tau,s)$ and $\beta(\tau,s)$ as defined in Section 2 when replacing λ,μ and ρ by λ_1,μ_1 and ρ_1. In particular, the mean conditional delay is given by

$$\mathbb{E}(W_1|B_1=\tau) = \frac{\rho_1\tau}{1-\rho_1} + \frac{\rho_1(2-\rho_1)\mathbb{E}(R_{I_1})}{1-\rho_1}(1-\mathrm{e}^{-\mu_1\tau(1-\rho_1)}) + (1-\rho_1)\mathbb{E}(R_{I_1}).$$

The LST of the sojourn time depends on the LST of the intervisit times I_1. The latter is given by

$$\tilde{I}_1(s) = \tilde{L}(1-\frac{s}{\lambda_1},1,\ldots,1),$$

where $\tilde{L}(z_1,\ldots,z_N)$ denotes the probability generating function (PGF) of the joint queue length distribution at the beginning of a visit to Q_1, see [28]. We denote by C_i the cycle length of queue i and $\mathbb{E}(C_i) = \frac{\mathbb{E}(S)}{1-\rho^*}$, $i = 1,\ldots,N$. The expected length of a visit to Q_i is $\mathbb{E}(C_i)\rho_i$, hence $\mathbb{E}(I_i) = (1-\rho_i)\mathbb{E}(C_i)$. The residual intervisit time is given in [35].

Closed-form expressions for the distribution of the intervisit time I_1 have been obtained for asymptotic regimes, which allows to further simplify Proposition 5. This will be done in Subsection 5.1.1 and Subsection 5.1.2 for the polling systems with large switch-over times and in heavy traffic, respectively.

5.1.1 Large switch-over times

In this subsection we assume the switch-over times are deterministic and we consider the polling system as they grow large, i.e., we let $\mathbb{E}(S) \to \infty$. Under the assumption that the exhaustive visit discipline is applied in all queues, it was shown in [36, Section 4] that $\frac{I_1}{\mathbb{E}(S)}$ converges in probability to $\hat{I}_1 := \frac{1-\rho_1}{1-\rho^*}$. Hence, from Proposition 2 we obtain that the scaled delay $\frac{W_1}{\mathbb{E}(S)}$ of a customer with service requirement τ satisfies the following:

Corollary 1 *Assume customers in Q_1 have exponentially distributed service requirements and the switch-over times are deterministic. As $\mathbb{E}(S) \to \infty$, the LST of the scaled conditional delay for a customer in Q_1 is given by*

$$\lim_{\mathbb{E}(S)\to\infty} \mathbb{E}(\mathrm{e}^{-sW_1/\mathbb{E}(S)}|B_1 = \tau) = \rho_1 \tilde{U}_{[0,\omega(\tau)\hat{I}_1]}(s) + (1-\rho_1)\tilde{U}_{[\omega(\tau)\hat{I}_1,(\omega(\tau)+1)\hat{I}_1]}(s),$$

with $\omega(\tau) := \frac{\rho_1}{1-\rho_1}(1-\mathrm{e}^{-\mu_1\tau(1-\rho_1)})$ and $\tilde{U}_{[a,b]}(s)$ the LST of a uniform random variable on $[a,b]$.

We note that the scaled conditional delay can be described as follows: With probability $1-\rho_1$, the tagged customer arrives in a visit to Q_1 and its scaled delay is distributed as a uniform random variable on $[0,\omega(\tau)\hat{I}_1]$. With probability $1-\rho_1$ the tagged customer arrives in an intervisit period and needs to wait a uniform distributed amount of time on $[0,\hat{I}_1]$, i.e., the scaled residual intervisit time, *plus* $\omega(\tau)\hat{I}_1$.

5.1.2 Heavy-traffic regime

In this subsection we consider the polling system in heavy traffic, i.e., we let $\rho_i \uparrow \hat{\rho}_i$ such that $\rho^* \uparrow 1$. Under the assumption that the exhaustive visit discipline is applied in Q_1 and only gated and exhaustive visit disciplines are allowed in all the other queues, it was shown in [25, Theorem 5] that $(1-\rho^*)I_1$ converges in distribution to a Gamma distributed random variable with parameters $\kappa = \frac{\mathbb{E}(S)}{\mathbb{E}(B)}\delta$ and $\theta = \frac{\delta}{\hat{\pi}_1(1-\hat{\rho}_1)}\frac{1}{\mathbb{E}(B)}$, where δ is as defined in [25, Lemma 1]. Hence, from Proposition 2 we obtain that the scaled delay $(1-\rho^*)W_1$ of a customer with service requirement τ satisfies the following:

Corollary 2 *Assume customers in Q_1 have exponentially distributed service requirements. The LST of the scaled conditional delay for a customer in Q_1 in a heavy-traffic setting is given by*

$$\lim_{\rho^*\uparrow 1} \mathbb{E}(\mathrm{e}^{-s(1-\rho^*)W_1}|B_1 = \tau) = \hat{\rho}_1 \frac{1-\tilde{G}^{\kappa,\theta}(s\omega(\tau))}{s\omega(\tau)\kappa/\theta} + (1-\hat{\rho}_1)\frac{\tilde{G}^{\kappa,\theta}(s\omega(\tau)) - \tilde{G}^{\kappa,\theta}(s(\omega(\tau)+1))}{s\kappa/\theta},$$

with $\omega(\tau) := \frac{\hat{\rho}_1}{1-\hat{\rho}_1}(1-\mathrm{e}^{-\mu_1\tau(1-\hat{\rho}_1)})$ and $\tilde{G}^{\kappa,\theta}(s) := \left(\frac{\theta}{\theta+s}\right)^{\kappa}$ the LST of the Gamma distribution.

5.2 General service requirements in Q_1

In this section we allow customers in Q_1 to have generally distributed service requirements with distribution function $F_1(\cdot)$. The expected conditional sojourn time for customers in Q_1 satisfies the integro-differential equation in the following Corollary which is a direct consequence of Proposition 3 and Proposition 4.

Corollary 3 *Assume customers in Q_1 have generally distributed service requirements. The mean conditional delay for a customer in Q_1 is given by*

$$\mathbb{E}(W_1|B_1 = \tau) = (1-\rho_1)\mathbb{E}(R_{I_1}) + D_1(\tau) - \tau,$$

where $\frac{\mathrm{d}}{\mathrm{d}\tau}D_1(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}(D_1|B_1 = \tau)$ is the unique solution $z(\tau)$ of

$$z(\tau) = 1 + (1-\rho_1)2\lambda_1\mathbb{E}(R_{I_1})\overline{F}_1(\tau) + \lambda_1\int_0^\infty z(y)\overline{F}_1(\tau+y)\mathrm{d}y + \lambda_1\int_0^\tau z(y)\overline{F}_1(\tau-y)\mathrm{d}y. \qquad (32)$$

In addition, the mean conditional sojourn time $\mathbb{E}(T_1|B_1 = \tau)$ has an asymptote of slope $\frac{\tau}{1-\rho_1}$ and bias term

$$\lim_{\tau\to\infty}\left(\mathbb{E}(T_1|B_1 = \tau) - \frac{\tau}{1-\rho_1}\right) = \frac{\mathbb{E}(R_{I_1})}{1-\rho_1}.$$

Remark 7 (Gated visit discipline) *In the case that Q_1 employs the gated visit discipline instead of the exhaustive visit discipline, the expected sojourn time under various scheduling disciplines (including PS) is derived in [35]. In this remark we show that, in the case of PS, the same result can be obtained using the integro-differential analysis.*

For a customer in Q_1 (with a gated visit discipline) the conditional mean sojourn time is given by

$$\mathbb{E}(T_1|B=\tau) = \mathbb{E}(R_{C_1}) + \mathbb{E}(D_1|B_1=\tau) = \mathbb{E}(R_{C_1}) + \int_0^\tau D_1'(y)\mathrm{d}y, \tag{33}$$

where $D_1'(y) := \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{E}(D_1|B_1 = y)$ and $\mathbb{E}(R_{C_1})$ is the mean residual cycle length and $\mathbb{E}(D_1|B_1 = \tau)$ denotes the expected sojourn time to get τ units of service starting from the moment the server visits Q_1. In particular we have that $\mathbb{E}(D_1|B_1 = 0) = 0$. The value of $D_1'(y)$ can be derived as follows. We have that $\mathbb{E}(D_1|B_1 = \tau + \Delta) - \mathbb{E}(D_1|B_1 = \tau) \approx \Delta + \Delta b_1 \overline{F}_1(\tau)$, with b_1 the expected number of customers present in Q_1 in addition to the tagged customer when the server starts serving customers in Q_1. Hence, $D_1'(y) = 1 + b_1\overline{F}_1(y)$. By Little's law it follows that $b_1 = 2\lambda_1\mathbb{E}(R_{C_1})$, thus we obtain from (33) that

$$\mathbb{E}(T_1|B=\tau) = \tau + \mathbb{E}(R_{C_1})\left(1 + 2\lambda_1\int_0^\tau \overline{F}_1(s)\mathrm{d}s\right),$$

which is equivalent to [35, Equation (7)]. We note that in [35, Section 3.2] the authors present a method, based on Mean Value Analysis, to derive the value of $\mathbb{E}(R_{C_1})$.

References

[1] K. Avrachenkov, U. Ayesta, and P. Brown. Batch arrival processor sharing with application to multilevel processor sharing scheduling. *Queueing Systems*, 50(4):459–480, 2005.

[2] U. Ayesta. A unifying conservation law for single server queues. *Journal of Applied Probability*, 44(4):1078–1087, 2007.

[3] F. Baccelli and P. Brémaud. *Elements of Queuing Theory: Palm Martingale Calculus and Stochastic Recurrences*. Springer, 2003.

[4] N. Bansal. Analysis of the $M/G/1$ processor sharing queue with bulk arrivals. *Operations Research Letters*, 31(5):401–405, 2003.

[5] 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*, 9:365–401, 1991.

[6] U.N. Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhäuser, 2008.

[7] S.C. Borst. *Polling Systems*. PhD thesis, Tilburg University, 1994.

[8] O.J. Boxma, J. Bruin, and B.H. Fralix. Sojourn times in polling systems with various service disciplines. *Performance Evaluation*, 66:621–639, 2009.

[9] S.L. Brumelle. On the relation between customer and time average in queues. *Journal of Applied Probability*, 2:508–520, 1971.

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

[11] J.W. Cohen. *The Single Server Queue*. North-Holland, 1982.

[12] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27(3):519–532, 1980.

[13] S.W. Fuhrmann. Performance analysis of a class of cyclic schedules. *Bell Laboratories Technical Memorandum 81-59531-1*, 1981.

[14] S.W. Fuhrmann and R.B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[15] J. Kim and B. Kim. Concavity of the conditional mean sojourn time in the M/G/1 processor-sharing queue with batch arrivals. *Queueing Systems*, 58(1):57–64, 2008.

[16] L. Kleinrock. *Queueing Systems, vol. 2.* John Wiley and Sons, 1976.

[17] L. Kleinrock and E.G. Coffman. Distribution of attained service in time-shared computer systems. *Journal of Computer System Science*, (1):287–298, 1967.

[18] L. Kleinrock, R.R. Muntz, and E. Rodemich. The processor sharing queueing model for time-shared systems with bulk arrivals. *Networks Journal*, 1(1):1–13, 1971.

[19] R.Y.W. Lam, V.C.M. Leung, and H.C.B. Chan. Polling-based protocols for packet voice transport over IEEE 802.11 wireless local area networks. *IEEE Wireless Communications*, 13:22–29, 2006.

[20] D. Miorandi, A. Zanella, and G. Pierobon. Performance evaluation of Bluetooth polling schemes: an analytical approach. *Mobile Networks and Applications*, 9:63–72, 2004.

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

[22] R. Núñez-Queija. Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems*, 34:351–386, 2000.

[23] T.M. O'Donovan. Direct solutions of $M/G/1$ processor sharing models. *Operations Research*, 22:1232–1235, 1974.

[24] T.M. O'Donovan. Distribution of attained service and residual service in general queueing systems. *Operations Research*, 22:570–575, 1974.

[25] T.L. Olsen and R.D. van der Mei. Periodic polling systems in heavy-traffic: distribution of the delay. *Journal of Applied Probability*, 40:305–326, 2003.

[26] N. Osipova. Batch processor sharing with hyper-exponential service time. *Operations Research Letters*, 36(3):372–376, 2008.

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

[28] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409–426, 1993.

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

[30] B. Sengupta and D.L. Jagerman. A conditional response time of the $M/M/1$ processor sharing queue. *AT&T Techn. J.*, 64:409–421, 1985.

[31] R. Serfozo. *Introduction to Stochastic Networks.* Springer, 1999.

[32] H. Takagi. *Queueing Analysis, Vol. 1.* North-Holland, Amsterdam, 1991.

[33] H. Thörisson. *Coupling, Stationarity and Regeneration.* Springer, 2000.

[34] N. Tian and Z.G. Zhang. *Vacation Queueing Models: Theory and Applications.* Springer, 2006.

[35] A. Wierman, E.M.M. Winands, and O.J. Boxma. Scheduling in polling systems. *Performance Evaluation*, 64:1009–1028, 2007.

[36] E.M.M. Winands. On polling systems with large setups. *Oper. Res. Letters*, 35:584–590, 2007.

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

Appendix 1: Proof of Lemma 1

Note that $r(s)$ represents the LST of the length of a busy period in a standard $M/M/1$ queue [6]. Hence, $r(0) = 1$ and $-\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}$ equals the mean length of the busy period, i.e., $1/(\mu(1-\rho))$.

Recall $G(\tau, s) := \frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}$, so $G(\tau, 0) = 1$ and the derivative is

$$\begin{aligned}\frac{\partial G(\tau,s)}{\partial s} &= \frac{-2\rho r\frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\lambda\tau(1-r)} + \lambda\tau\frac{\mathrm{d}r(s)}{\mathrm{d}s}(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}} \\ &\quad -\frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}(-\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s} + \rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r} - \rho r\frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})}{(1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})^2} \\ &\quad -\frac{(1-\rho r^2)\mathrm{e}^{-\lambda\tau(1-r)}\lambda\tau r(1-r)(\rho+1/r^2)\frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}{(1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})^2}.\end{aligned}$$

Hence, setting $s = 0$ we get

$$\begin{aligned}&\frac{\partial G(\tau,s)}{\partial s}|_{s=0} \\ &= \frac{-2\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0} + \lambda\tau\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}(1-\rho)}{1-\rho} + \frac{(1-\rho)\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0} + (1-\rho)\rho\mathrm{e}^{-\mu\tau(1-\rho)}\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}}{(1-\rho)^2} \\ &= \frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}\left(\lambda\tau - \frac{\rho}{1-\rho}(1-\mathrm{e}^{-\mu\tau(1-\rho)}\right) = -\frac{1}{\mu(1-\rho)}\left(\lambda\tau - \frac{\rho}{1-\rho}(1-\mathrm{e}^{-\mu\tau(1-\rho)}\right).\end{aligned}$$

Consider $\beta(\tau, s) = \frac{r(1-\rho r)+(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}}$. Its derivative is

$$\begin{aligned}\frac{\partial\beta(\tau,s)}{\partial s} &= \frac{\frac{\mathrm{d}r(s)}{\mathrm{d}s}(1-\rho r) - r\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s} - \frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\mu\tau(1-\rho r^2)/r} + (1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}\mu\tau(\rho+1/r^2)\frac{\mathrm{d}r(s)}{\mathrm{d}s}}{1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r}} \\ &\quad -\frac{(r(1-\rho r)+(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})(-\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s} + \rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})}{(1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})^2} \\ &\quad -\frac{(r(1-\rho r)+(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})(-\rho r\frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\mu\tau(1-\rho r^2)/r} + (1-r)r\lambda\tau(\rho+1/r^2)\frac{\mathrm{d}r(s)}{\mathrm{d}s}\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})}{(1-\rho r+\rho r(1-r)\mathrm{e}^{-\mu\tau(1-\rho r^2)/r})^2}.\end{aligned}$$

Setting $s = 0$ we get

$$\begin{aligned}&\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0} \\ &= \frac{\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}(1-\rho) - \rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0} - \frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}\mathrm{e}^{-\mu\tau(1-\rho)}}{1-\rho} + \frac{\rho\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0} + \rho\mathrm{e}^{-\mu\tau(1-\rho)}\frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}}{1-\rho} \\ &= \frac{\mathrm{d}r(s)}{\mathrm{d}s}|_{s=0}(1-\mathrm{e}^{-\mu\tau(1-\rho)}) = -\frac{1}{\mu(1-\rho)}(1-\mathrm{e}^{-\mu\tau(1-\rho)}).\end{aligned}$$

This concludes the proof. □

Appendix 2: Proof of Equation (4)

Recall that

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sW}|B=\tau) &= \rho G(\tau,s)\frac{\beta(\tau,s)(1-\rho)}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)} \\ &\quad +(1-\rho)G(\tau,s)\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s\mathbb{E}(V)}.\end{aligned}$$

We note that

$$\tilde{V}(y) = \tilde{V}(0) + y\tilde{V}'(0) + \frac{y^2}{2}\tilde{V}''(0) + O(y^3) = 1 - y\mathbb{E}(V) + \frac{y^2}{2}\mathbb{E}(V^2) + O(y^3), \text{ as } y \to 0. \tag{34}$$

We define $d_1(s) = G(\tau,s)\frac{\beta(\tau,s)(1-\rho)}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}$. Then

$$\begin{aligned}\frac{\mathrm{d}d_1(s)}{\mathrm{d}s} = \quad & \frac{\partial G(\tau,s)}{\partial s}\frac{\beta(\tau,s)(1-\rho)}{1-\rho\beta(\tau,s)}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)} \\ & +G(\tau,s)\frac{(1-\rho\beta(\tau,s))((1-\rho)\frac{\partial\beta(\tau,s)}{\partial s}) - \beta(\tau,s)(1-\rho)(-\rho\frac{\partial\beta(\tau,s)}{\partial s})}{(1-\rho\beta(\tau,s))^2}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)} \\ & +G(\tau,s)\frac{\beta(\tau,s)(1-\rho)}{1-\rho\beta(\tau,s)}\frac{\mathrm{d}}{\mathrm{d}s}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}.\end{aligned}$$

Since $\beta(\tau,0)=1$, from (34) we get directly that $\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}|_{s=0}=1$ and $\frac{\mathrm{d}}{\mathrm{d}s}\frac{1-\tilde{V}(\lambda(1-\beta(\tau,s)))}{\lambda(1-\beta(\tau,s))\mathbb{E}(V)}|_{s=0} = \lambda\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0}\mathbb{E}(R_V)$, hence

$$\frac{\mathrm{d}d_1(s)}{\mathrm{d}s}|_{s=0} = \frac{\partial G(\tau,s)}{\partial s}|_{s=0} + \frac{\partial\beta(\tau,s)}{\partial s}|_{s=0}\frac{1}{1-\rho} + \lambda\mathbb{E}(R_V)\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0},$$

where we used that $\beta(\tau,s)|_{s=0}=1$ and $G(\tau,0)=1$.

Now we define

$$d_2(s) = G(\tau,s)\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s}.$$

Hence, the derivative is equal to

$$\begin{aligned}\frac{\mathrm{d}d_2(s)}{\mathrm{d}s} = & \frac{\partial G(\tau,s)}{\partial s}\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s} \\ & +G(\tau,s)\frac{\mathrm{d}}{\mathrm{d}s}\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s}.\end{aligned}$$

Using (34), we obtain that

$$\frac{\tilde{V}(\lambda(1-\beta(\tau,s))) - \tilde{V}(\lambda(1-\beta(\tau,s))+s)}{s} = \quad \mathbb{E}(V) - \frac{s\mathbb{E}(V^2)}{2} - \lambda(1-\beta(\tau,s))\mathbb{E}(V^2) + O(s^2). \quad (35)$$

Together with (35) and since $G(\tau,0)=1$, we obtain that

$$\frac{\mathrm{d}d_2(s)}{\mathrm{d}s}|_{s=0} = \frac{\partial G(\tau,s)}{\partial s}|_{s=0}\mathbb{E}(V) - \frac{\mathbb{E}(V^2)}{2} + \lambda\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0}\mathbb{E}(V^2).$$

The mean delay is given by

$$\begin{aligned}\mathbb{E}(W|B=\tau) &= -\frac{\partial}{\partial s}\mathbb{E}(\mathrm{e}^{-sW}|B=\tau)|_{s=0} = -\rho\frac{\mathrm{d}d_1(s)}{\mathrm{d}s}|_{s=0} - \frac{(1-\rho)}{\mathbb{E}(V)}\frac{\mathrm{d}d_2(s)}{\mathrm{d}s}|_{s=0} \\ &= -\frac{\partial G(\tau,s)}{\partial s}|_{s=0} - \frac{\rho}{1-\rho}\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0} - (2-\rho)\lambda\mathbb{E}(R_V)\frac{\partial\beta(\tau,s)}{\partial s}|_{s=0} + (1-\rho)\mathbb{E}(R_V).\end{aligned}$$

Hence,

$$\mathbb{E}(W|B=\tau) = \frac{\rho\tau}{1-\rho} + \frac{\rho(2-\rho)\mathbb{E}(R_V)}{1-\rho}(1-\mathrm{e}^{-\mu\tau(1-\rho)}) + (1-\rho)\mathbb{E}(R_V), \quad (36)$$

where in the last step we used Lemma 1. □

Appendix 3: Properties of the solution of (14).

In [1] a PS queue with batch arrivals was studied. In particular, an integro-differential equation (see [1, Equation (1)]) was obtained that models the sojourn time. Comparing the integro-differential equation of [1] with the integro-differential equation (14), we observe that (14) coincides with the integro-differential equation of a batch processor sharing queue where the batch arrival rate is λ, and the first and second moment of the batch size distribution are given by 1 and $(1-\rho)2\lambda\mathbb{E}(R_V)+1$,

respectively. This observation allows us to directly obtain several interesting properties for the solution of Equation (14). We obtain that (i) if $\rho < 1$, then the solution of (14) exists ([1, Theorem 1]) and is unique, ([1, Lemma 3]), (ii) $D(x) - \frac{x}{1-\rho}$ is increasing with respect to x, and (iii) $D(x) - \frac{x}{1-\rho}$ is upper bounded ([1, Lemma 4]).

We do not reproduce the proofs of [1], but it is interesting to highlight the main idea used in [1] to show uniqueness, which consists in showing that the operator on the right hand side of (14) is a contraction mapping. In order to do so consider the fixed point iterations

$$D'_{k+1}(x) = 1 + (1-\rho)2\lambda\mathbb{E}(R_V)\overline{F}(x) + \lambda\int_0^\infty D'_k(y)\overline{F}(x+y)dy + \lambda\int_0^x D'_k(y)\overline{F}(x-y)dy \qquad (37)$$

on the complete functional space of continuous bounded non-negative functions $\mathcal{C}[0,\infty)$ with the supremum metric. Let $\left\|D'\right\| = \sup_x\{D'(x)\} < \infty$. Define the linear integral operator $\mathcal{A}[\beta(x)]$ as follows:

$$\mathcal{A}(\beta(x)) \quad = \quad 1 + (1-\rho)2\lambda\mathbb{E}(R_V)\overline{F}(x) + \lambda\int_0^\infty \beta(y)\overline{F}(x+y)dy + \lambda\int_0^x \beta(y)\overline{F}(x-y)dy. \quad (38)$$

Clearly the operator $\mathcal{A}(\beta(x))$ maps the space $\mathcal{C}[0,\infty)$ into itself.

If we show that the linear integral operator $\mathcal{A}(\beta(x))$ is a contraction, then the integral equation (14) has a unique solution in $\mathcal{C}[0,\infty)$. Let d denote the distance in the metric space $\mathcal{C}[0,\infty)$, that is, $d(\beta_1,\beta_2) = \sup_x|\beta_1(x) - \beta_2(x)|$. In [1] the authors show that $d(\mathcal{A}(\beta_1),\mathcal{A}(\beta_2)) \leq \rho d(\beta_1,\beta_2)$ which proves that the operator is a contraction mapping since $\rho < 1$. The key to show this result consists in noting that, after taking the supremum in both integrals, the term $\sup_x|\beta_1 - \beta_2|$ comes out of the integral and $\lambda\left(\int_0^\infty \overline{F}(x+y)dy + \int_0^x \overline{F}(x-y)dy\right) = \lambda\int_0^\infty \overline{F}(y)dy = \rho$.