Centrum voor Wiskunde en Informatica

Sojourn times in a processor sharing queue with service interruptions

R. Núñez Queija

Probability, Networks and Algorithms (PNA)

PNA-R9807 August 31, 1998

Report PNA-R9807
ISSN 1386-3711

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

Sojourn Times in a Processor Sharing Queue with Service Interruptions

R. Núñez-Queija
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
email sindo@cwi.nl

ABSTRACT

We study the sojourn time of customers in an $M/M/1$ queue with processor sharing service discipline and service interruptions. The lengths of the service interruptions have a general distribution, whereas the periods of service availability are assumed to have an exponential distribution. A branching process approach is shown to lead to a decomposition of the sojourn time into independent contributions, that can be investigated separately. The Laplace-Stieltjes Transform of the distribution of the sojourn time is found through an integral equation. We derive the first two moments of the sojourn time conditioned on the amount of work brought into the system and on the number of customers present upon arrival. We show that the expected sojourn time of a customer that arrives at the system in steady state is not linear in the amount of work he brings with him. Finally, we show that the sojourn time conditioned on the amount of work, scaled by the traffic load, converges in heavy traffic to an exponential distribution. This study was motivated by a need for delay analysis of elastic traffic in modern communication networks. Specifically, the results are of interest for the performance analysis of the Available Bit Rate (ABR) service class in ATM networks, as well as for the best-effort services in IP networks.

1991 Mathematics Subject Classification: 60K25, 68M20, 90B12, 90B22.
Keywords & Phrases: processor sharing, service interruptions, sojourn time, elastic traffic, Available Bit Rate, best-effort traffic.
Note: work carried out under project ATM in PNA 2.1.

1. Introduction

In this paper we consider a processor sharing queue with a single server that is subject to breakdowns. For this model we study the sojourn time distribution of customers in the system, that is the time that elapses between the arrival of a customer and his departure from the system. In the (egalitarian) processor sharing service discipline, when there are $n > 0$ customers in the system, all these customers simultaneously get an equal share of the service capacity, i.e. each customer gets a fraction $1/n$ of the capacity. The processor sharing service discipline became of interest as the idealisation of time-sharing queueing models that arose with the introduction of time-sharing computing in the sixties. Today, processor sharing models have many other applications, for instance in the performance analysis of telecommunication networks. The present study was motivated by the Available Bit Rate (ABR) service class in Asynchronous Transfer Mode (ATM) networks. The ABR service class is primarily designed for carrying data-connections with a low priority, in contrast with higher priority Constant Bit Rate (CBR) and Variable Bit Rate (VBR) services. Typically ABR-connections will receive a varying service capacity. In this paper we consider the extreme case where the service capacity available to the ABR traffic alternates between a positive value and zero, as a first step towards analysing the case where, for instance, the server alternates between two *positive* service speeds. The relevance of the processor sharing discipline for the performance analysis of ABR connections is a consequence of the requirement that the available capacity should be shared fairly among ABR users. This requirement was stated in the definition of the ABR service class by the ATM Forum in [1]. The results in this paper are also of interest for best-effort services in Internet Protocol (IP) networks. Similar to the ABR traffic in ATM networks, best-effort traffic streams in IP networks have to share the capacity on the communication links of the network, see for instance Roberts [19].

Processor sharing queues have been studied extensively in the literature. The steady state queue length distribution of the $M/G/1$ queue with processor sharing was readily shown to be geometric, and insensitive of the service time distribution except from its first moment, see Sakata et al. [20]. This result was extended by Cohen [5] for networks with general service time distributions and *generalised processor sharing*, i.e. the rate at which the customers at a certain node are served, is a function of the node and of the number of customers at that node. In that paper, Cohen also gives results for mean sojourn times. However, determining the sojourn time distribution in processor sharing queues turned out to be a very difficult problem.

For the $M/M/1$ queue with processor sharing, a closed-form expression for the Laplace-Stieltjes transform (LST) of the sojourn times – conditional on the amount of service required and the number of customers seen upon arrival – was first derived by Coffman et al. in [3]. Sengupta and Jagerman [23] found an alternative expression for the LST of the sojourn time conditioned only on the number of customers seen upon arrival. In particular they found that the k^{th} moment of the conditional sojourn time is a polynomial of degree k in the number of customers upon arrival. The *distribution* function of the sojourn times, conditioned on the amount of service required, was studied by Morrison in [15].

The sojourn time distribution in the $M/G/1$ processor sharing queue was first analysed by Kitayev and Yashkov in [11] and by Yashkov in [28]. Schassberger [21] considered the $M/G/1$ processor sharing queue as the limit of the round robin schedule. Ott [17] found the joint LST and generating function of the sojourn time and the number of customers left behind. Van den Berg and Boxma [2] exploited the product form structure of an $M/M/1$ queue with general feedback for an alternative derivation of the sojourn time distribution in the $M/G/1$ processor sharing queue. Rege and Sengupta [18] gave a decomposition theorem for the sojourn time distribution for the $M/G/1$ with K classes of customers and *discriminatory* processor sharing. In [10] Grishechkin described the $M/G/1$ queue with batch arrivals and a generalised processor sharing discipline by means of Crump-Mode-Jagers branching processes. For a more extensive treatise on the literature on processor sharing queueing models we refer to Yashkov's survey papers [29] and [30], and the references therein.

In the present study we analyse the sojourn times of customers in the $M/M/1$ processor sharing queue with a server that alternates between an on-state and an off-state (breakdown). When the server is in the off-state there is no service. We assume that the on-periods and the off-periods form an alternating renewal process. We require that the on-periods have an exponentially distributed duration, but make no assumption on the distribution of the duration of the off-periods other than finiteness of moments involved in the analysis, in particular the mean duration of the off-periods.

The assumption of exponentially distributed service requirements is of no essential importance in our analysis. Most of our results can be extended for general service time distributions (at the expense of the explicitness of some expressions). Nevertheless, in this paper the results for the exponential services are presented for two reasons: (i) The fundamental ideas are the same as for general service requirements, but the presentation is more transparent, and (ii) our real interest is in generalising the process that models the server availability (for instance a server that alternates between two *positive* service speeds).

Queueing models with *First Come First Served* (FCFS) discipline and servers that are subject to breakdowns (and repairs) have received much attention in the literature. These models have many practical applications, for instance in production line facilities, traffic (road intersections), computer networks and telecommunication networks. We give a brief overview of the literature on these models. The first ones to consider queueing models with interruptions (and their connection with priority models) were White and Christie [27]. Gaver [9] obtained the steady state queue length distribution of the $\mathrm{M}^{\mathrm{X}}/\mathrm{G}/1$ queue with exponentially distributed on-times and general off-times. We further mention the early work of Mitrani and Avi-Itzhak [14] on a queueing model with multiple servers which are subject to breakdowns, and the work of Neuts concerned with queues in a random environment, see Chapter 6 of [16], and in particular Section 6.3. Federgruen and Green [7, 8] studied bounds and approximations for the case when (also) the on-times have a general distribution. Recent publications on queues with server breakdowns are for instance Takine and Sengupta [24], Li et al. [13], and Lee [12]. For an extensive overview of the literature on queueing models with service interruptions, and further references, we refer to [7] and [8]. More recent references can be found in [24]. To the author's knowledge, there are no previous publications on queues with server breakdowns and processor sharing discipline.

The paper is organised as follows. We define the model in Section 2, and give the joint steady-state distribution of the state of the server and the number of customers in the system. In Section 3, we represent the sojourn time of a customer conditional on his service requirement, by a branching process. In Section 4, we characterise the distributions of two fundamental random variables in the branching process by deriving differential equations for the LST's of their distributions, and then solving these in terms of a single integral equation, which is suitable for numerical evaluation. An asymptotic analysis, when the service requirement of the tagged customer tends to infinity, is performed in Section 5. We derive the first two moments of the two fundamental random variables in Section 6, and give the general form of the higher moments. In Section 7 we use these results to give explicitly the first two moments of the sojourn time of a customer conditioned on his work requirement, the state of the server upon arrival and the number of other customers in the system upon arrival. We show that the polynomiality of the moments of the conditional sojourn times, as observed by Sengupta and Jagerman in [23], also holds in our model with server breakdowns. In Section 8 we give the LST of the sojourn time distribution of a customer conditioned only on his own work requirement, assuming that he arrives to the system when steady state is reached. In particular we see that – unlike the case without server breakdowns – the mean sojourn time of a customer is not a linear function of the amount of work required by that customer. The heavy traffic case is considered in Section 9. We conclude with some final remarks in Section 10.

2. Model description

We consider a server that alternates between an *on*-state and an *off*-state. The on-periods are assumed to be exponentially distributed with mean $1/\nu$, independent of everything else. The off-periods are i.i.d. random variables (generically denoted by T_{off}) having probability distribution $F(t) := \mathbf{P}\{T_{off} \leq t\}$, $t \geq 0$. The LST of this distribution will be denoted by

$$\phi(s) := \int_{t=0}^{\infty} \mathrm{e}^{-st}\mathrm{d}F(t), \qquad \mathrm{Re}(s) \geq 0,$$

and the k^{th} moment of $F(t)$ by

$$m_k := \int_{t=0}^{\infty} t^k \mathrm{d}F(t).$$

Throughout this paper we assume that $m_1 < \infty$.
Customers arrive to the server according to a Poisson process with rate λ, requiring an exponentially distributed amount of service with mean $1/\mu$. There is room for infinitely many customers at the server. When the server is on, all customers present are simultaneously served according to the *(egalitarian) processor sharing* discipline, i.e. when there are $n > 0$ customers present, each receives service at rate $1/n$. Thus, because of the exponentially distributed service requirements, the service of any of the customers is completed within the next Δt time units with probability $\frac{1}{n}\mu\Delta t + \mathrm{o}(\Delta t)$. During off-periods the service of all customers is interrupted until the server again becomes active.

We define the random variable $X(t)$ to be the number of customers at the server at time $t \geq 0$. The random variable $Y(t)$ is equal to 1 if at time $t \geq 0$ the server is on, and $Y(t)$ is equal to 0 otherwise. Under the ergodicity condition,

$$\frac{\lambda}{\mu} < \frac{1}{1+\nu m_1}, \tag{2.1}$$

the pair $(X(t), Y(t))$ has a non-trivial limiting distribution. The left-hand side of Condition (2.1) is the average amount of work that arrives to the system per unit of time. The right-hand side is the average service capacity per unit of time, which is equal to the fraction of time that the server is available.
We now determine the limiting distribution of $(X(t), Y(t))$, under Condition (2.1). Let (X, Y) be distributed according to this distribution, then $\mathbf{P}\{Y=1\} = 1 - \mathbf{P}\{Y=0\} = \frac{1}{1+\nu m_1}$, and for $|z| \leq 1$,

$$\mathbf{E}\left[z^X \,|Y=1\right] = \frac{\mu - \lambda(1+\nu m_1)}{\mu - \lambda z - \nu z \frac{1-\phi(\lambda(1-z))}{1-z}}, \tag{2.2}$$

$$\mathbf{E}\left[z^X \,|Y=0\right] = \frac{1-\phi(\lambda(1-z))}{m_1\lambda(1-z)}\mathbf{E}\left[z^X \,|Y=1\right]. \tag{2.3}$$

In the remainder of this section we discuss the derivation of (2.2) and (2.3). In particular, in Remarks 2.1 and 2.2, we discuss the equivalence of the queue length process in our model with the queue length process of two queueing models with FCFS service discipline.
Expression (2.2) can be found by considering the queue length process only during on-periods, by deleting the off-periods and interpreting the arrivals during an off-period as a batch arrival. Thus in this transformed model there are three events: Departures of customers at rate μ when there is at least one customer present, single arrivals according to a Poisson process with rate λ and batch arrivals according to a Poisson process with rate ν and batch sizes having probability generating function (p.g.f.) $\phi(\lambda(1-z))$, which is the p.g.f. of the number of arrivals during an off-period. Note that batches can be empty (with probability $\phi(\lambda)$). This can be avoided by letting the batches arrive with rate $\nu(1-\phi(\lambda))$ with batch sizes having p.g.f. $\frac{\phi(\lambda(1-z))-\phi(\lambda)}{1-\phi(\lambda)}$. The balance equations for this transformed model readily lead to Equation (2.2).
The factor $\frac{1-\phi(\lambda(1-z))}{m_1\lambda(1-z)}$ in Equation (2.3) is the p.g.f. of the number of customers that arrive during the *overshoot* period of an off-period. When a customer arrives and finds the server off, the number of customers that he finds in the system is the sum of the customers that were at the server when the server went off (using PASTA, because of the exponential on-periods this is distributed as $X\,|\{Y=1\}$) and the number of customers that have arrived since the server went off (again with PASTA, the time that has elapsed is distributed as the overshoot of an off-period).

By averaging over $\mathbf{P}\{Y=0\}$ and $\mathbf{P}\{Y=1\}$ we find the p.g.f. of the marginal distribution of X,

$$\mathbf{E}\left[z^X\right]=\frac{1}{1+\nu m_1}\left(1+\nu\frac{1-\phi(\lambda(1-z))}{\lambda(1-z)}\right)\frac{\mu-\lambda(1+\nu m_1)}{\mu-\lambda z-\nu z\frac{1-\phi(\lambda(1-z))}{1-z}}. \tag{2.4}$$

Remark 2.1 Because of the exponentially distributed services, the queue length process remains unchanged if we replace the processor sharing service discipline by the FCFS discipline. Expression (2.4) can therefore be obtained from Gaver [9, Formula 8.4], where the p.g.f. of the number of customers in the system, at arbitrary points in time, is given for the case of general service time distribution.
If also the breakdowns (off-periods) are exponentially distributed, i.e. $\phi(s)=\frac{1}{1+m_1 s}$, the probabilities $x_{i,off}:=\mathbf{P}\{X=i,Y=0\}$ and $x_{i,on}:=\mathbf{P}\{X=i,Y=1\}$, $i=0,1,\ldots$, are explicitly given by Neuts in Theorem 6.3.1. on p. 277 of [16]:

$$(x_{i,off}\ \ x_{i,on})=(\pi_{off}\ \ \pi_{on})(I-R)R^i,$$

with $\pi_{on}=1-\pi_{off}=\mathbf{P}\{Y=1\}$, I the identity matrix, and

$$R=\frac{\lambda}{\mu}\begin{bmatrix}\frac{m_1(\mu+\nu)}{m_1\lambda+1} & 1\\ \frac{m_1\nu}{m_1\lambda+1} & 1\end{bmatrix}.$$

It can be verified that this corresponds to (2.2) and (2.3) when $\phi(s)=\frac{1}{1+m_1 s}$.

Remark 2.2 The queue length process is also equivalent to that of the $M/G/1$ queue with exceptional first service and LST of the regular service time distribution

$$\beta(s)=\frac{\mu}{\mu+s+\nu\left(1-\phi\left(s\right)\right)},\qquad \mathrm{Re}(s)\geq 0. \tag{2.5}$$

These 'enlarged' regular service times are the sum of the actual time it takes to serve a customer (exponentially distributed with mean $1/\mu$) and all off-periods that occur during such a service. The (exceptional) first service in a busy period is with probability p equal to a regular service and with probability $1-p$ there is an additional waiting period before a regular service can begin. In the original model, p would be the probability that the first customer to arrive after the service station has become empty, finds the server in the on-state. By conditioning on the first event after the system becomes empty (this event is either an arrival or a server breakdown), we find

$$p=\frac{\lambda}{\lambda+\nu}+\frac{\nu}{\lambda+\nu}\phi(\lambda)p,$$

where $\phi(\lambda)$ is the probability that no customers arrive during an off-period, and so $p=\frac{\lambda}{\lambda+\nu(1-\phi(\lambda))}$. The additional waiting time for the fraction $1-p$ of customers then corresponds to the time that a customer in the original model who finds an empty system with the server in the off-state, has to wait before the server becomes active. This waiting time is therefore distributed as $T_{off}-T_{arrival}\left|\{T_{off}>T_{arrival}\}\right.$, where $T_{arrival}$ has an exponential distribution with mean $1/\lambda$, and is independent of T_{off}. It can be seen that

$$\mathbf{E}\left[\mathrm{e}^{-s\left(T_{off}-T_{arrival}\right)}\left|T_{off}>T_{arrival}\right.\right]=\frac{\lambda}{1-\phi(\lambda)}\frac{\phi(s)-\phi(\lambda)}{\lambda-s}.$$

If we denote the LST of the exceptional services by $\overline{\beta}(s)$, we have

$$\overline{\beta}(s) = \left(p + (1-p)\frac{\lambda}{1-\phi(\lambda)}\frac{\phi(s)-\phi(\lambda)}{\lambda - s}\right)\beta(s), \qquad \mathrm{Re}(s) \geq 0.$$

Using this in the queue length distribution of the $M/G/1$ queue with exceptional first service — see Formula 3 in Theorem 2 of Welch [26] — we have for $|z| \leq 1$:

$$\mathbf{E}\left[z^X\right] = \frac{1+\lambda\beta'(0)}{1+\lambda(\beta'(0)-\overline{\beta}'(0))}\frac{\beta(\lambda(1-z)) - z\overline{\beta}(\lambda(1-z))}{\beta(\lambda(1-z))-z},$$

which corresponds to Expression (2.4).

3. A branching process representation

In this section we show how the sojourn time of a customer (that is the total time spent in the system) can be studied by means of a branching process. For this purpose we will observe the process on a *transformed time scale*. This time-scale transformation has already been used by Grishechkin [10] for the M/G/1 queue with batch arrivals and a generalised processor sharing discipline. A method closely related to this branching process approach, without time-scale transformation, was used in Yashkov [28] to find the LST of the sojourn time in the ordinary M/G/1 queue. By some (non-trivial) modifications, the analysis of [10] can be extended to include service interruptions. In this paper we present a more direct use of the time-transformation technique to processor sharing queues with service interruptions. Doing so, we gain more intuitive insights into the dynamics of the model. We choose to show the branching process representation for the case with exponentially distributed service requirements, that is for the model presented in Section 2. One advantage is that, for this case, the complexity of the problem is reduced, because we don't need to condition on the amount of work that the customers present have received. This makes the presentation more transparent, while the fundamental ideas are the same as in the case with general service time distributions. In Remark 3.4 we briefly indicate how the generalisation to the latter case can be obtained. Another advantage is that, for the case of exponential service times, we are able to get more explicit results for the moments of the sojourn times conditioned on the amount of work, the number of competing customers and the state of the server upon arrival (see Section 7) and conditioned only on the amount of work (see Section 8) when the system is in steady state.

In our presentation we first assume there is a permanent customer that never leaves the system. All other customers (that arrive with rate λ) have an exponentially distributed service requirement with mean $1/\mu$. Let $Z(t)$ be the number of customers at the server at time $t \geq 0$, *excluding* the permanent customer. As before, $Y(t)$ is 1 if the server is on at time t and 0 otherwise. Then, at time t, the permanent customer receives service at rate

$$\frac{Y(t)}{1+Z(t)}.$$

Let the random variable $R(t)$ be the amount of service received by the permanent customer during the time interval $[0,t]$:

$$R(t) := \int_{u=0}^{t} \frac{Y(u)}{1+Z(u)}\mathrm{d}u.$$

We now define for $\tau \geq 0$:

$$V(\tau) := \inf \{t \geq 0 \ : \ R(t) \geq \tau\}.$$

Thus, $V(\tau)$ is the moment that the amount of service received by the permanent customer reaches the level τ. In Figure 1 a typical realisation of $R(t)$ and $V(\tau)$ is depicted.

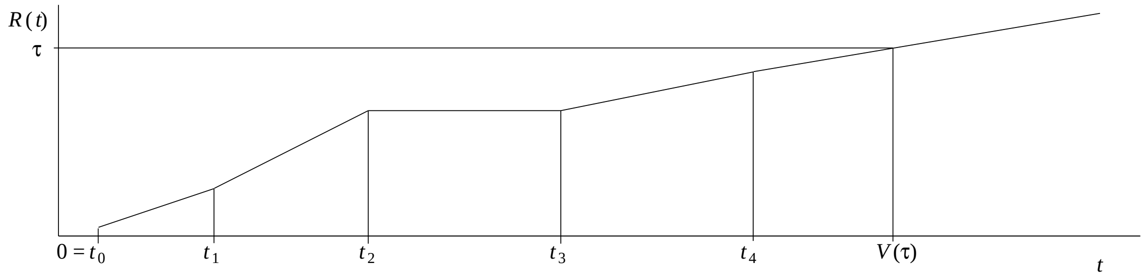

Figure 1: $R(t)$ and $V(\tau)$.

In this example, at time $t_0 = 0$ there are two other customers in the system along with the permanent customer, therefore $R(t)$ increases at rate $1/3$ immediately after time t_0. At time t_1 one of the customers leaves and the rate increases to $1/2$. From t_2 until t_3 the server is off, and during this period 3 customers arrive, leading to a rate $1/5$ immediately after t_3. At time t_4 another customer arrives, etc. $V(\tau)$ is the moment that the service received by the permanent customer reaches the level τ.
Now, if the permanent customer is replaced by a customer requiring an amount of service τ, then $V(\tau)$ is the time at which this customer leaves the system, i.e. $V(\tau)$ is the sojourn time of that customer. Our goal is to determine the distribution of the random variable $V(\tau)$ for an arbitrary $\tau > 0$.

In the sequel we use the notation $x(y+) := \lim_{u \downarrow y} x(u)$ and $x(y-) := \lim_{u \uparrow y} x(u)$ for any function $x(y)$.

We distinguish between the cases where $Y(0) = 1$ (start with a working server) and $Y(0) = 0$ (start with a server in the off-state). With $i \in \{0, 1\}$, we use the notation $V_i(\tau) := V(\tau) \,|\{Y(0) = i\}$, $Z_i(t) := Z(t) \,|\{Y(0) = i\}$, and $Y_i(t) := Y(t) \,|\{Y(0) = i\}$. We first concentrate on $V_1(\tau)$ and at the end of this section derive the results for $V_0(\tau)$.

Observation 3.1 *For arbitrary $n \in \mathbb{N}$, given $Z_1(0) = n$ (and $Y(0) = 1$),*

$$V_1(0+) = 0.$$

This observation follows easily from the fact that, for small τ, $V_1(\tau)$ is equal to $(n+1)\tau$ with probability $1 - \left(\lambda + \frac{n}{n+1}\mu + \nu\right)(n+1)\tau + o(\tau)$ (by conditioning on the events that occur during a time interval of length $(n+1)\tau$). This is *not true* for $V_0(\tau)$ (in that case, the server must first become active again).
Denote the number of times that the server went off during the period $(0, t)$ by the random variable $N(t)$, and the length of the i^{th} off-period started *after* time 0 by D_i, $i = 1, 2, \ldots$. Note that $\{D_1, D_2, \ldots\}$ is an i.i.d. sequence with distribution $F(t)$.
Define for $\tau > 0$:

$$N'(\tau) := N\left(V_1\left(\tau\right)\right).$$

The random variable $N'(\tau)$ is well defined because $V_1(\tau)$ – also a random variable – is strictly increasing in τ for any realisation of the arrival and departure process. Note that $N'(\tau+)-N'(\tau) = 1$ *if and only if* at time $t = V_1(\tau)$ the server turns into the off-state. Otherwise $N'(\tau+) - N'(\tau) = 0$. Similar to $N'(\tau)$, we define the processes:

$$Z_1'(\tau) := Z_1(V_1(\tau)),$$

and

$$Y_1'(\tau) := Y_1(V_1(\tau)+),$$

for $\tau > 0$.

Lemma 3.1 *For any realisation of the arrival and departure process,*

$$V_1(\tau) = \int_{\sigma=0}^{\tau} \left[1 + Z_1'(\sigma)\right] \mathrm{d}\sigma + \sum_{i=1}^{N'(\tau)} D_i, \tag{3.1}$$

with the empty sum being equal to zero (when $N'(\tau) = 0$).

Proof
We observe (with the aid of Figure 1) that if $N'(\tau+) - N'(\tau) = 0$, then

$$\frac{\mathrm{d}V_1(\tau)}{\mathrm{d}\tau} = 1 + Z_1'(\tau),$$

and if $N'(\tau+) - N'(\tau) = 1$, then

$$V_1(\tau+) - V_1(\tau) = D_{N'(\tau+)}.$$

■

Using Figure 1, we make the following observation:

Observation 3.2 *The transformed process $(Z_1'(\tau), N'(\tau))$ is Markovian, with transition rates given in the following table for n, k and $j \in \mathbb{N}_0$,*

from state	*to state*	*transition rate*
(n,k)	$(n+1,k)$	$(n+1)\lambda$
(n,k)	$(n-1,k)$	$n\mu$
(n,k)	$(n+j,k+1)$	$(n+1)\nu p_j$

Here, p_j is the probability that during an off-period, j new customers arrive:

$$\sum_{j=0}^{\infty} z^j p_j = \phi(\lambda(1-z)).$$

In words, the transformation from the process $(Z_1(t), N(t), Y_1(t))$ to the process $(Z_1'(\tau), N'(\tau))$ consists in (i) shrinking the time scale by a factor $n+1$ when $Z_1(t)=n$ and $Y_1(t)=1$, and (ii) replacing off-periods by batch arrivals of customers.
In (3.1), $V_1(\tau)$ also depends on $D_1, \ldots, D_{N'(\tau)}$. We emphasise that, if $N'(\tau)-N'(\tau-)=1$ then $Z_1'(\tau)-Z_1'(\tau-)$ and $D_{N'(\tau)}$ are *not* independent: $D_{N'(\tau)}$ is the length of an off-period in the original process and $Z_1'(\tau)-Z_1'(\tau-)$ is the number of customers that arrived during that period:

$$\mathbf{E}\left[\mathrm{e}^{-sD_{N'(\tau)}} z^{Z_1'(\tau)-Z_1'(\tau-)} \,\middle|\, N'(\tau)-N'(\tau-)=1\right] = \phi(s+\lambda(1-z)).$$

To study the distribution of $V_1(\tau)$, we construct a branching process that is equivalent with $\left(Z_1'(\tau), N'(\tau); D_1, \ldots, D_{N'(\tau)}\right)$, and impose a cost structure on this branching process that will turn out to be beneficial. Consider a population $\mathcal{P}$ of elements that evolves in the following way: The lifetime of an element of the population has an exponential distribution with mean duration $1/\mu$. During its lifetime an element has a cost of 1 per time unit. An element generates children in two ways, independent from all other living elements. According to a Poisson process with rate λ an element gives birth to children, one at a time. In addition, according to another (independent) Poisson process with rate ν, an element generates nests of children (possibly empty nests). A nest of (possibly zero) children has an immediate cost on society that depends on the number of children in the nest in a stochastic way. The simultaneous distribution of A children in the nest and an immediate cost D, is given by

$$\mathbf{E}\left[\mathrm{e}^{-sD} z^{A}\right] = \phi(s+\lambda(1-z)).$$

Finally, there is a permanent element in the population that generates children in the same way as the other elements (but never dies).

Observation 3.3 *Denote the number of non-permanent elements in the population at time τ by $Z_1''(\tau)$, the number of nest-births between time 0 and time τ by $N''(\tau)$ and the cost of the i^{th} nest by D_i''. By comparing the transition rates of both processes it is seen that the processes $\left(Z_1'(\tau), N'(\tau); D_1, \ldots, D_{N'(\tau)}\right)$ and $\left(Z_1''(\tau), N''(\tau); D_1'', \ldots, D_{N''(\tau)}''\right)$ are equivalent.*
Also, $V_1(\tau)$ is distributed as the cost of the population from time 0 until time τ.

In the next theorem we formulate the main result of this section. The decomposition given in the theorem was shown by Yashkov [28] for the ordinary M/G/1 processor sharing queue.

Theorem 3.1 *The conditional sojourn time $V_1(\tau)$ of a customer who finds the server working upon arrival, can be decomposed as,*

$$V_1(\tau) \stackrel{d}{=} C_0(\tau) + \sum_{i=1}^{Z_1(0)} C_i(\tau),$$

where $\stackrel{d}{=}$ means equivalence in distribution. All random variables involved in the right-hand side are mutually independent.
In particular, $C_0(\tau) \stackrel{d}{=} V_1(\tau) \,|\{Z_1(0)=0\}$.

Proof
By construction, the elements of the population $\mathcal{P}$ behave independent from each other. Using the cost-interpretation of $V_1(\tau)$ given in Observation 3.3, we can split $V_1(\tau)$ into the independent

contributions of each element. Let $C_0(\tau)$ be the cost of the permanent element and his offspring between time instants 0 and τ. Also, let $C_i(\tau)$, $i = 1, 2, \dots$, be the cost of a non-permanent individual, who was present at time 0, plus the cost of his offspring between time instants 0 and τ. It is clear that $C_1(\tau) \stackrel{d}{=} C_i(\tau)$, for all $i = 2, 3, \dots$. Moreover, all the $C_i(\tau)$ – including $C_0(\tau)$ – are independent of each other.
By definition, $Z_1'(0) = Z_1(0)$, which concludes the proof. ■

In Section 4, we characterise the LST's of the distributions of $C_0(\tau)$ and $C_1(\tau)$ by a set of differential equations. We also derive an integral equation which is more suitable for the numerical computation of these LST's.

We now turn to $V_0(\tau)$, that is the sojourn time of a customer with τ work and starting with a server in the off-state. Let D_0 be the *residual* off-period at time zero and A_0 be the number of arrivals during D_0. Let $\phi_0(s)$ be the LST of D_0. By conditioning on the length of D_0 and the number of arrivals A_0:

$$V_0(\tau) \,|\{Z_0(0) = n, D_0 = d_0, A_0 = k\} \stackrel{d}{=} d_0 + V_1(\tau) \,|\{Z_1(0) = n + k\} \; .$$

This gives us the following corollary:

Corollary 3.2 *The conditional sojourn time $V_0(\tau)$ of a customer who finds the server in the off-state upon arrival, satisfies,*

$$V_0(\tau) \stackrel{d}{=} D_0 + C_0(\tau) + \sum_{i=1}^{Z_0(0)+A_0} C_i(\tau).$$

All random variables on the right-hand side are mutually independent, except for the pair (D_0, A_0) which has the joint distribution,

$$\mathbf{E}\left[\mathrm{e}^{-sD_0} z^{A_0}\right] = \phi_0(s + \lambda(1 - z)), \quad \mathrm{Re}(s) \geq 0, |z| \leq 1.$$

We define the LST of $C_0(\tau)$ and $C_i(\tau)$ by $g_0(\tau; s)$ and $g_1(\tau; s)$: For $\mathrm{Re}(s) \geq 0$,

$$\begin{aligned} g_0(\tau; s) &:= \mathbf{E}\left[\mathrm{e}^{-sC_0(\tau)}\right], \\ g_1(\tau; s) &:= \mathbf{E}\left[\mathrm{e}^{-sC_i(\tau)}\right], \quad i = 1, 2, \dots. \end{aligned}$$

From Theorem 3.1 and Corollary 3.2 we have, for $\mathrm{Re}(s) \geq 0$,

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)} \,|Y(0) = 1, Z(0) = n\right] = g_0(\tau; s) \left\{g_1(\tau; s)\right\}^n, \tag{3.2}$$

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)} \,|Y(0) = 0, Z(0) = n\right] = g_0(\tau; s) \left\{g_1(\tau; s)\right\}^n \phi_0(s + \lambda(1 - g_1(\tau; s))). \tag{3.3}$$

In Section 4 we characterise $g_0(\tau; s)$ and $g_1(\tau; s)$ by means of a set of differential equations, in order to determine the LST of $V(\tau)$.
We conclude this section with the following remark, which indicates how the representation of the sojourn time by a branching process can be extended to the case of general service time distributions.

Remark 3.4 The generalisation of this representation by branching processes for general service time distributions $B(x)$, $x \geq 0$, can be obtained by extending the state space representation with the vector $\overline{x_n} := (x_1, x_2, \ldots, x_n)$ when there are n customers in the system. We again assume that a newly arrived customer who brings an amount of work τ finds the server available, and we further condition on the number of customers in the system upon arrival (n) and the residual service requirement of each of those customers (x_i, $i = 1, 2, \ldots, n$). If we call the conditioned sojourn time of the new customer $V_1(\tau; n; x_1, \ldots, x_n)$ then

$$V_1(\tau; n; x_1, \ldots, x_n) \stackrel{d}{=} C_0(\tau) + \sum_{i=1}^{n} C_i(\tau; x_i),$$

where the $C_0(\tau)$ and $C_i(\tau; x_i)$, $i = 1, 2, \ldots,$ are the analogues of the earlier $C_0(\tau)$ and $C_i(\tau)$ for the population model with life time distribution $B(x)$. I.e. $C_i(\tau; x_i)$ is the cost of a family until time τ, starting with one individual with a remaining life time x_i. We omit the details of this generalisation in this paper, and refer to Yashkov [28] for a related analysis of the case without service interruptions.

4. Characterisation of $g_0(\tau; s)$ and $g_1(\tau; s)$

In this section we derive a set of differential equations that uniquely determine $g_0(\tau; s)$ and $g_1(\tau; s)$, the LST's of $C_0(\tau)$ and $C_1(\tau)$. We express $g_0(\tau; s)$ in terms of $g_1(\tau; s)$, and then, for real $s > 0$, derive a useful integral equation for $g_1(\tau; s)$.

Lemma 4.1 *The LST's $g_0(\tau; s)$ and $g_1(\tau; s)$ are uniquely determined by the following set of differential equations,*

$$\begin{aligned}
\frac{\partial}{\partial \tau} g_1(\tau; s) &= -(s + \lambda + \mu + \nu)\, g_1(\tau; s) + \lambda \{g_1(\tau; s)\}^2 + \mu \\
&\quad + \nu g_1(\tau; s) \phi(s + \lambda(1 - g_1(\tau; s))), && (4.1) \\
\frac{\partial}{\partial \tau} g_0(\tau; s) &= -(s + \lambda + \nu)\, g_0(\tau; s) + \lambda g_0(\tau; s) g_1(\tau; s) \\
&\quad + \nu g_0(\tau; s) \phi(s + \lambda(1 - g_1(\tau; s))), && (4.2)
\end{aligned}$$

and initial conditions,

$$g_0(0; s) = g_1(0; s) = 1. \qquad (4.3)$$

Proof

By conditioning on the number of 'single' children and the number of nests that a non-permanent element in the population model generates in a time interval of length Δ, as well as on the survival probability of the element itself in that interval, we get,

$$\begin{aligned}
g_1(\tau + \Delta; s) &= \int_{t=0}^{\Delta} \mu \mathrm{e}^{-\mu t} \mathrm{e}^{-st} \sum_{m=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^m}{m!} \left(\frac{1}{t} \int_{u=0}^{t} g_1(\tau + u; s) \mathrm{d}u \right)^m \\
&\quad \times \sum_{n=0}^{\infty} \mathrm{e}^{-\nu t} \frac{(\nu t)^n}{n!} \left(\frac{1}{t} \int_{u=0}^{t} \phi(s + \lambda(1 - g_1(\tau + u; s))) \mathrm{d}u \right)^n \mathrm{d}t \\
&+ \mathrm{e}^{-\mu \Delta} \mathrm{e}^{-s\Delta} g_1(\tau; s) \sum_{m=0}^{\infty} \mathrm{e}^{-\lambda \Delta} \frac{(\lambda \Delta)^m}{m!} \left(\frac{1}{\Delta} \int_{u=0}^{\Delta} g_1(\tau + u; s) \mathrm{d}u \right)^m \\
&\quad \times \sum_{n=0}^{\infty} \mathrm{e}^{-\nu \Delta} \frac{(\nu \Delta)^n}{n!} \left(\frac{1}{\Delta} \int_{u=0}^{\Delta} \phi(s + \lambda(1 - g_1(\tau + u; s))) \mathrm{d}u \right)^n .
\end{aligned}$$

Here we use the fact that 'Poisson arrivals occur homogeneously in time', see for instance Tijms [25, Theorem 1.2.5].
$\phi(s + \lambda(1 - g_1(\tau; s)))$ is the LST of the cost of a nest plus the cost of all children in that nest and their offspring, until time τ.
Equivalently we may write,

$$\begin{aligned} g_1(\tau+\Delta; s) &= \int_{t=0}^{\Delta} \mu \mathrm{e}^{-\mu t - st - \lambda\left(t - \int_{u=0}^{t} g_1(\tau+u;s)\mathrm{d}u\right) - \nu\left(t - \int_{u=0}^{t} \phi(s+\lambda(1-g_1(\tau+u;s)))\mathrm{d}u\right)} \mathrm{d}t \\ &+ g_1(\tau; s)\mathrm{e}^{-\mu\Delta - s\Delta - \lambda\left(\Delta - \int_{u=0}^{\Delta} g_1(\tau+u;s)\mathrm{d}u\right) - \nu\left(\Delta - \int_{u=0}^{\Delta} \phi(s+\lambda(1-g_1(\tau+u;s)))\mathrm{d}u\right)}. \end{aligned} \tag{4.4}$$

By similar arguments we also find,

$$g_0(\tau+\Delta; s) = g_0(\tau; s)\mathrm{e}^{-s\Delta - \lambda\left(\Delta - \int_{u=0}^{\Delta} g_1(\tau+u;s)\mathrm{d}u\right) - \nu\left(\Delta - \int_{u=0}^{\Delta} \phi(s+\lambda(1-g_1(\tau+u;s)))\mathrm{d}u\right)}. \tag{4.5}$$

From (4.4) and (4.5) we can show that, for $\Delta \downarrow 0$,

$$\begin{aligned} g_1(\tau+\Delta; s) &= (1 - (s+\lambda+\mu+\nu)\Delta)\, g_1(\tau; s) + \lambda\Delta \{g_1(\tau; s)\}^2 + \mu\Delta \\ &\quad + \nu\Delta g_1(\tau; s)\phi(s + \lambda(1 - g_1(\tau; s))) + \mathrm{o}(\Delta), \end{aligned} \tag{4.6}$$

$$\begin{aligned} g_0(\tau+\Delta; s) &= (1 - (s+\lambda+\nu)\Delta)\, g_0(\tau; s) + \lambda\Delta g_0(\tau; s) g_1(\tau; s) \\ &\quad + \nu\Delta g_0(\tau; s)\phi(s + \lambda(1 - g_1(\tau; s))) + \mathrm{o}(\Delta). \end{aligned} \tag{4.7}$$

With (4.6) and (4.7) it is immediate that $g_1(\tau; s)$ and $g_0(\tau; s)$ are continuous from the right in τ. If we replace τ in (4.6) and (4.7) by $\tau - \Delta$, the continuity from the left in τ also easily follows. Subsequently it can be shown that,

$$\lim_{\Delta\downarrow 0} \frac{g_1(\tau+\Delta; s) - g_1(\tau; s)}{\Delta} = \lim_{\Delta\downarrow 0} \frac{g_1(\tau; s) - g_1(\tau-\Delta; s)}{\Delta},$$

and

$$\lim_{\Delta\downarrow 0} \frac{g_0(\tau+\Delta; s) - g_0(\tau; s)}{\Delta} = \lim_{\Delta\downarrow 0} \frac{g_0(\tau; s) - g_0(\tau-\Delta; s)}{\Delta},$$

so that $\frac{\partial}{\partial\tau} g_1(\tau; s)$ and $\frac{\partial}{\partial\tau} g_0(\tau; s)$ exist and satisfy (4.1) and (4.2). The initial conditions (4.3) follow from $C_0(0) = C_1(0) = 0$. ■

Theorem 4.1 *$g_0(\tau; s)$ can be expressed in terms of $g_1(\tau; s)$ as,*

$$g_0(\tau; s) = g_1(\tau; s)\mathrm{e}^{\mu\left\{\tau - \int_{u=0}^{\tau} g_1(u;s)^{-1}\mathrm{d}u\right\}}. \tag{4.8}$$

Proof
From (4.2) and (4.3) we can immediately express $g_0(\tau; s)$ in terms of $g_1(\tau; s)$:

$$g_0(\tau; s) = \exp\left\{-(s+\lambda+\nu)\tau + \int_{u=0}^{\tau} \left[\lambda g_1(u; s) + \nu\phi(s + \lambda(1 - g_1(u; s)))\right] \mathrm{d}u\right\}. \tag{4.9}$$

If we also use (4.1) we may rewrite this to,

$$g_0(\tau; s) = \exp\left\{\int_{u=0}^{\tau} \frac{\frac{\partial}{\partial u} g_1(u; s) - \mu\,(1 - g_1(u; s))}{g_1(u; s)} \mathrm{d}u\right\},$$

which leads to Relation (4.8). ■

The remainder of this section is dedicated to the solution of (4.1) for real $s > 0$. First we need to study the roots of the following equation:

$$(s + \lambda + \mu + \nu)\, x = \lambda x^2 + \mu + \nu x \phi(s + \lambda(1 - x)). \tag{4.10}$$

Note that if the right-hand side of (4.1) is equal to zero, then $x = g_1(\tau; s)$ is a root of Equation (4.10). We now argue that (4.10) is the relation for the LST of the *clearing period* of the model of Section 2. By the clearing period we mean the time it takes for the system to become empty, starting with one customer and a working server. If there were no off-periods, the clearing period would be equal to the busy period. Note that if the service discipline were FCFS instead of processor sharing, the clearing period would be the same (because of work conservation). Combining this with the well-known branching argument for busy periods, it can be seen that the time it takes to clear the system starting with $n > 0$ customers and a working server, is just the sum of n independent clearing periods. Now, by conditioning on the first event that occurs after a clearing period starts — this can be the arrival of a new customer, the departure of the present customer or the server turning into the off-state — we get Relation (4.10) for $x = \mathbf{E}\left[\mathrm{e}^{-sCP}\right]$, $\mathrm{Re}(s) \geq 0$, where the generic random variable CP stands for the duration of the clearing period. It can be shown that, for $\mathrm{Re}(s) \geq 0$, the root $x = \mathbf{E}\left[\mathrm{e}^{-sCP}\right]$ is the unique root in the unit circle of Equation (4.10). In Remark 4.1 we indicate how this can be done by identifying the clearing period with the busy period of an $M/G/1$ queue.

Remark 4.1 The clearing period equals the busy period of the $M/G/1$ queue with arrival rate λ and LST $\beta(.)$ of the service time distribution given by (2.5). The interpretation of these service times is given in Remark 2.2. The well-known relation for the busy-period of the $M/G/1$ queue,

$$x = \beta(s + \lambda(1 - x)),$$

readily leads to (4.10). We immediately have that for $\mathrm{Re}(s) \geq 0$, $\mathbf{E}\left[\mathrm{e}^{-sCP}\right]$ is equal to the (unique) root of (4.10) inside (or on) the unit circle (see for instance Cohen [4] p. 250).
We emphasise that the non-empty period is not equal to the busy period of the $M/G/1$ queue with exceptional first service as described in Remark 2.2. In fact, in the context of that model, the non-empty period is equal to the busy period *initiated by a customer with a regular service.*

Observation 4.2 *Denote the (unique) root of Equation (4.10) by $r_1(s) = \mathbf{E}\left[\mathrm{e}^{-sCP}\right]$. Since $C_1(\tau)$ is non-decreasing in τ with probability 1, $g_1(\tau; s)$ is non-increasing in τ for fixed $s > 0$. Indeed, the right-hand side of (4.1) is negative for $\tau = 0$ (because $g_1(0; s) = 1$). Now, it must be that,*

$$r_1(s) \leq g_1(\tau; s),$$

because if $g_1(\tau; s)$ would cross $r_1(s)$, the right-hand side of (4.1) would become positive, since the zero $r_1(s)$ is of multiplicity 1.

Theorem 4.2 *For real $s > 0$, the solution to (4.1) satisfying (4.3), is obtained from,*

$$\int_{x=1}^{g_1(\tau;s)} \frac{1}{\mu - (s + \lambda + \mu + \nu)x + \lambda x^2 + \nu x \phi(s + \lambda - \lambda x)} \mathrm{d}x = \tau. \tag{4.11}$$

Proof
The integral in Relation (4.11) is well defined, because the denominator of the integrand has no zeros in $(r_1(s), 1)$ for $s > 0$, see Remark 4.1. The integral is taken for x from 1 to $g_1(\tau; s)$ so that the initial Condition (4.3) is satisfied. By differentiating with respect to τ, it is readily seen that (4.1) is also satisfied. ■

Relation (4.11) can be used for numerical evaluation of $g_1(\tau; s)$. In Section 5 we use Relation (4.11) to study the asymptotics of $g_1(\tau; s)$ as $\tau \to \infty$. This in turn enables us to prove the convergence in probability of $\frac{C_0(\tau)}{\tau}$ and (more importantly) $\frac{V(\tau)}{\tau}$ for $\tau \to \infty$.
For the determination of moments, Relation (4.11) is not very practical. In Section 6 we study the moments of $C_1(\tau)$ (and $C_0(\tau)$) directly.

Remark 4.3 When the off-periods have an exponential distribution, i.e.

$$\phi(s) = \frac{1}{1 + m_1 s}, \qquad \mathrm{Re}(s) > -\frac{1}{m_1},$$

the integrand of (4.11) becomes

$$\frac{1 + m_1 \{s + \lambda(1-x)\}}{\{1 + m_1 \{s + \lambda(1-x)\}\} \{\lambda x^2 - (s + \lambda + \mu + \nu) x + \mu\} + \nu x}.$$

The denominator of this rational function is a polynomial of degree 3 in x. For $s > 0$, it can be shown that this polynomial has three real roots: As before, the root in the interval (0,1) will be $r_1(s)$, the ones in $(1, \infty)$ are denoted by $r_2(s)$ and $r_3(s)$ and we order them — for real $s > 0$ — as $r_2(s) < \frac{1+m_1(s+\lambda)}{m_1\lambda} < r_3(s)$.
The left-hand side of (4.11) becomes

$$\begin{aligned} &\int_{x=1}^{g_1(\tau;s)} \left(\frac{a_1(s)}{r_1(s) - x} + \frac{a_2(s)}{r_2(s) - x} + \frac{a_3(s)}{r_3(s) - x} \right) \mathrm{d}x \\ &= -a_1(s) \ln\left(\frac{g_1(\tau; s) - r_1(s)}{1 - r_1(s)} \right) - a_2(s) \ln\left(\frac{r_2(s) - g_1(\tau; s)}{r_2(s) - 1} \right) - a_3(s) \ln\left(\frac{r_3(s) - g_1(\tau; s)}{r_3(s) - 1} \right), \end{aligned}$$

where

$$\begin{aligned} a_1(s) &= \frac{1 + m_1 \{s + \lambda(1 - r_1(s))\}}{m_1 \lambda^2 (r_2(s) - r_1(s))(r_3(s) - r_1(s))} > 0, \\ a_2(s) &= \frac{1 + m_1 \{s + \lambda(1 - r_2(s))\}}{m_1 \lambda^2 (r_1(s) - r_2(s))(r_3(s) - r_2(s))} < 0, \\ a_3(s) &= \frac{1 + m_1 \{s + \lambda(1 - r_3(s))\}}{m_1 \lambda^2 (r_1(s) - r_3(s))(r_2(s) - r_3(s))} < 0. \end{aligned}$$

We finally remark that, for off-periods having a phase-type distribution, the solution given in Theorem 4.2 can be treated along the same lines. The number of zeros in the denominator of the integrand depends on the number of phases of the off-periods.

5. Asymptotic analysis for $\tau \to \infty$.

In this section we study the behaviour of $g_1(\tau; s)$ as $\tau \to \infty$. We will then use these asymptotics to show the convergence of $\frac{V(\tau)}{\tau}$ for $\tau \to \infty$.

Our starting point is Relation (4.11). By partial fraction expansion,

$$\frac{1}{\mu-(s+\lambda+\mu+\nu)x+\lambda x^2+\nu x\phi(s+\lambda-\lambda x)}=\frac{k_1(s)}{x-r_1(s)}+k_2(x;s), \tag{5.1}$$

where

$$k_1(s):=\lim_{x\to r_1(s)}\frac{x-r_1(s)}{\mu-(s+\lambda+\mu+\nu)x+\lambda x^2+\nu x\phi(s+\lambda-\lambda x)}, \tag{5.2}$$

exists and the function $k_2(x;s)$ is analytic in x, for $|x|\le 1$ and $\mathrm{Re}(s)\ge 0$.
Using (5.1) in (4.11) we get, for $s>0$,

$$k_1(s)\cdot\ln(g_1(\tau;s)-r_1(s))+\int_{x=1}^{g_1(\tau;s)}k_2(x;s)\mathrm{d}x=k_1(s)\cdot\ln(1-r_1(s))+\tau. \tag{5.3}$$

If we let $\tau\to\infty$ in (5.3), we may conclude that

$$\lim_{\tau\to\infty}g_1(\tau;s)=r_1(s),\qquad s>0. \tag{5.4}$$

This is an immediate consequence of the analyticity in x of $k_2(x;s)$ and the boundedness of $g_1(\tau;s)$, which imply that the second term on the left-hand side of (5.3) is bounded. In Remark 5.1 we discuss how this limiting property can be seen probabilistically in our model.

Remark 5.1 If we concentrate on a non-permanent element of the population model of Section 3 *and* his offspring (we call this a *family*), then under the ergodicity Condition (2.1), this family dies out with probability 1. Consider the cost that this family generates until its extinction. This cost is equal to the sum of the lifetimes of all the members of this family *plus* the cost of all nests in this family. By assigning the cost of a nest to the individual that generated it, and concatenating the lifetimes of all family members, it can be seen that the total cost of this family is distributed as a clearing period of the model of Section 2:

$$\lim_{\tau\to\infty}C_1(\tau)\stackrel{d}{=}CP.$$

This corresponds to (5.4).
Further exploiting (5.3), we can carry our asymptotic analysis one step further:

$$\lim_{\tau\to\infty}\left\{k_1(s)\cdot\ln\left(\frac{g_1(\tau;s)-r_1(s)}{1-r_1(s)}\right)-\tau\right\}=-\int_{x=1}^{r_1(s)}k_2(x;s)\mathrm{d}x. \tag{5.5}$$

Using (5.5) we can prove the following Lemma:

Lemma 5.1 *For $s>0$,*

$$\lim_{\tau\to\infty}\int_{u=0}^{\tau}\left(g_1(u;\frac{s}{\tau})-r_1(\frac{s}{\tau})\right)\mathrm{d}u=0,$$

and consequently,

$$\lim_{\tau\to\infty}\int_{u=0}^{\tau}\left(\phi(\frac{s}{\tau}+\lambda-\lambda g_1(u;\frac{s}{\tau}))-\phi(\frac{s}{\tau}+\lambda-\lambda r_1(\frac{s}{\tau}))\right)\mathrm{d}u=0.$$

Proof
Using Relation (5.3) we may write:

$$\int_{u=0}^{\tau} \left(g_1(u;\frac{s}{\tau}) - r_1(\frac{s}{\tau})\right) \mathrm{d}u = \left(1 - r_1(\frac{s}{\tau})\right) \int_{u=0}^{\tau} \exp\left\{\frac{1}{k_1(\frac{s}{\tau})}\left(u - \int_{x=1}^{g_1(u;\frac{s}{\tau})} k_2(x;\frac{s}{\tau})\mathrm{d}x\right)\right\} \mathrm{d}u.$$

It is clear from (5.2) that $k_1(s) < 0$, for $s > 0$: for $x = 0$ the numerator on the right-hand side of (5.2) is negative and the denominator is positive, and as $x \uparrow r_1(s)$ neither the numerator, nor the denominator changes sign.
For $s > 0$, let $M(s) \in [r_1(s), 1]$ be such that $\int_{x=1}^{M(s)} k_2(x;s)\mathrm{d}x$ is maximal. Then we may write,

$$\begin{aligned} 0 &\le \int_{u=0}^{\tau} \left(g_1(u;\frac{s}{\tau}) - r_1(\frac{s}{\tau})\right) \mathrm{d}u \\ &\le \left(1 - r_1(\frac{s}{\tau})\right) \mathrm{e}^{\frac{-1}{k_1(\frac{s}{\tau})}\int_{x=1}^{M(\frac{s}{\tau})} k_2(x;\frac{s}{\tau})\mathrm{d}x} \cdot k_1(\frac{s}{\tau})\left(\mathrm{e}^{\frac{\tau}{k_1(\frac{s}{\tau})}} - 1\right). \end{aligned} \tag{5.6}$$

Now, if we take $\tau \to \infty$ then $r_1(\frac{s}{\tau})$ and $M(\frac{s}{\tau})$ go to 1, $k_2(x;\frac{s}{\tau})$ remains bounded for $r_1(\frac{s}{\tau}) \le x \le 1$ and,

$$\lim_{s \downarrow 0} k_1(s) = \frac{-1}{\mu - \lambda(1 + \nu m_1)}.$$

Thus, if we let $\tau \to \infty$ in (5.6) then the upperbound goes to 0.
The second part of the lemma follows from the first part by noting that $\phi(s)$ is a decreasing and convex function for $s \ge 0$, and $\frac{\mathrm{d}}{\mathrm{d}s}\phi(s)|_{s=0} = -m_1$. Therefore it holds that $\phi(s_1) - \phi(s_2) \le m_1(s_2 - s_1)$, whenever $0 \le s_1 \le s_2$. ■

Lemma 5.1 allows us to prove the following theorem:

Theorem 5.1 *For $s \ge 0$,*

$$\lim_{\tau \to \infty} g_0(\tau;\frac{s}{\tau}) = \mathrm{e}^{-s\mu\frac{1+\nu m_1}{\mu - \lambda(1+\nu m_1)}},$$

and hence,

$$\frac{C_0(\tau)}{\tau} \xrightarrow{\mathrm{P}} \mu\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} = \frac{1}{\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}},$$

as $\tau \to \infty$. Here $\xrightarrow{\mathrm{P}}$ means convergence in probability.

Proof
Using the first part of Lemma 5.1 we can write, for $s \ge 0$,

$$\lim_{\tau \to \infty} \int_{u=0}^{\tau} \left(1 - g_1(u;\frac{s}{\tau})\right) \mathrm{d}u = \lim_{\tau \to \infty} \tau \left(1 - r_1(\frac{s}{\tau})\right) = s\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)},$$

where we use that $\lim_{s\downarrow 0} \frac{1-r_1(s)}{s} = \mathbf{E}[CP]$. We can find $\mathbf{E}[CP] = \frac{1+\nu m_1}{\mu - \lambda(1+\nu m_1)}$ from Relation (4.10).
Similarly, using the second part of Lemma 5.1 we have, again for $s \ge 0$,

$$\begin{aligned} \lim_{\tau \to \infty} \int_{u=0}^{\tau} \left(1 - \phi(\frac{s}{\tau} + \lambda - \lambda g_1(u;\frac{s}{\tau}))\right) \mathrm{d}u &= \lim_{\tau \to \infty} \tau \left(1 - \phi(\frac{s}{\tau} + \lambda - \lambda r_1(\frac{s}{\tau}))\right) \\ &= s m_1 \frac{\mu}{\mu - \lambda(1 + \nu m_1)}. \end{aligned}$$

Using this in Relation (4.9) gives the convergence in distribution. The convergence in probability then follows immediately, because the limit is a constant.

■

Theorem 5.1 has the following corollary:

Corollary 5.2 *The sojourn time $V(\tau)$ of a customer with an amount of work τ, satisfies*

$$\frac{V(\tau)}{\tau} \xrightarrow{\text{P}} \frac{1}{\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}},$$

as $\tau \to \infty$.

Proof
The proof is immediate from Theorem 5.1 and Formulas (3.2) and (3.3). ■

Remark 5.2 Using the Renewal Reward Theorem, see [25, Theorem 1.3.1], it can be shown that the convergence of $\frac{V(\tau)}{\tau}$, and $\frac{C_0(\tau)}{\tau}$, is in fact convergence with probability 1. To see this, note that $N''(\tau)$, the process counting the number of elements in the population $\mathcal{P}$ at time τ, is regenerative. The regeneration points can be taken to be the times at which the permanent element *becomes* the only element of the population. It then can be shown that the lengths of the regeneration cycles have a finite expectation.

Remark 5.3 The term $\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}$ can be seen as the average speed at which the permanent customer receives service, when the system *with the permanent customer* is in steady state: The average service capacity is $\frac{1}{1+\nu m_1}$ per time unit, and on average an amount of capacity $\frac{\lambda}{\mu}$ per time unit is required to serve other customers (since the system with a permanent customer is ergodic, all non-permanent customers eventually leave the system).

Remark 5.4 In addition to Theorem 5.1 and Corollary 5.2, it can be shown that

$$\frac{V(\tau) - C_0(\tau)}{\tau} \xrightarrow{\text{P}} 0, \quad \tau \to \infty.$$

This is a consequence of Theorem 3.1, Corollary 3.2, and Remark 5.1.

6. Moments of $C_0(\tau)$ and $C_1(\tau)$

In Section 4 we saw that the LST's of $C_0(\tau)$ and $C_1(\tau)$ are determined by a set of differential equations. The solution for these differential equations is given by (4.9) and (4.11). However, this solution is not very practical to determine moments of $C_0(\tau)$ and $C_1(\tau)$. In this section we show how the moments of C_0 and C_1 can be found by directly solving an alternative system of differential equations. Yashkov [28] also remarks that, in the M/G/1 processor sharing queue, such an approach leads to a more tractable derivation of moments.
The next theorem is a consequence of a result of De Meyer and Teugels [6].

Theorem 6.1 *If the k^{th} moment of the off-periods, m_k, exists, then also the k^{th} moments of $C_1(\tau)$ and $C_0(\tau)$ exist.*

Proof
It is known for the $M/G/1$ queue that the k^{th} moment of the busy period exists if and only if the k^{th} moment of the service time exists, see De Meyer and Teugels [6, Lemma 3]. In view of Remark 4.1, this implies that the k^{th} moment of the clearing period exists, if and only if $m_k < \infty$. Since $C_1(\tau)$ is non-decreasing in τ with probability 1, and $C_1(\tau)$ converges to the clearing period CP, as $\tau \to \infty$, it must be that

$$\mathbf{E}\left[C_1(\tau)^k\right] \le \mathbf{E}\left[CP^k\right],$$

($C_1(\tau)$ is stochastically smaller than CP), and hence the k^{th} moment of $C_1(\tau)$ exists when $m_k < \infty$. To prove the result for $C_0(\tau)$, we first write the following identity:

$$C_0(\tau) = \sum_{i=1}^{N^{(\lambda)}(\tau)} C_i(\tau - T_i^{(\lambda)}) + \sum_{j=1}^{N^{(\nu)}(\tau)} D_j + \sum_{n=1}^{N^{(\lambda)}(D_j)} C_{j,n}(\tau - T_j^{(\nu)}).$$

Here, $N^{(\lambda)}(\tau)$ is the number of 'regular' children that the permanent element, in the population $\mathcal{P}$, generates (at rate λ) over a time span of length τ. $T_i^{(\lambda)}$ is the time at which the i^{th} regular child is born, and $C_i(\tau - T_i^{(\lambda)})$ is the cost of this child and his offspring until time τ. Similarly, $N^{(\nu)}(\tau)$ is the number of batches of children of the permanent element (generated at rate ν) until time τ. D_j is the direct cost of the j^{th} batch, $N^{(\lambda)}(D_j)$ is the number of children in the j^{th} batch, $T_j^{(\nu)}$ is the time at which the j^{th} batch is generated, and $C_{j,n}(\tau - T_j^{(\nu)})$ is the cost associated with the n^{th} child in the j^{th} batch and his offspring, until time τ. The above identity was given in terms of LST's in Relation (4.9).
If we replace each of the costs until time τ associated with a child of the permanent customer and his offspring, by the cost of the family of that child over a total time-span of length τ, we clearly have an upperbound for $C_0(\tau)$:

$$C_0(\tau) \le \overline{C_0}(\tau) := \sum_{i=1}^{N^{(\lambda)}(\tau)} C_i(\tau) + \sum_{j=1}^{N^{(\nu)}(\tau)} D_j + \sum_{n=1}^{N^{(\lambda)}(D_j)} C_{j,n}(\tau).$$

For $\mathrm{Re}(s) > 0$, the LST of $\overline{C_0}(\tau)$ is given by,

$$\mathbf{E}\left[\mathrm{e}^{-s\overline{C_0}(\tau)}\right] = \mathrm{e}^{-\tau\{s+\lambda(1-g_1(\tau;s))+\nu(1-\phi(s+\lambda-\lambda g_1(\tau;s)))\}}. \tag{6.1}$$

If $m_k < \infty$, and hence by the first part of the theorem $\mathbf{E}\left[C_1(\tau)^k\right] < \infty$, we can write, for $s \downarrow 0$,

$$\begin{aligned}
\phi(s) &= 1 + \sum_{i=1}^{k} m_i \frac{(-s)^i}{i!} + \mathrm{o}(s^k), \\
g_1(\tau; s) &= 1 + \sum_{j=1}^{k} \mathbf{E}\left[C_1(\tau)^j\right] \frac{(-s)^j}{j!} + \mathrm{o}(s^k),
\end{aligned} \tag{6.2}$$

see De Meyer and Teugels [6, Lemma 1]. Combining these, we get,

$$\phi(s + \lambda - \lambda g_1(\tau; s)) = 1 + \sum_{i=1}^{k} m_i \frac{\left(-s + \lambda \sum_{j=1}^{k} \mathbf{E}\left[C_1(\tau)^j\right] \frac{(-s)^j}{j!}\right)^i}{i!} + \mathrm{o}(s^k). \tag{6.3}$$

From Equation (6.1), it is now straightforward to see that the LST of $\overline{C_0}(\tau)$ has a finite k^{th} derivative in $s = 0$. Therefore, the k^{th} moment of $\overline{C_0}(\tau)$, and hence the k^{th} moment of $C_0(\tau)$, exists.

■

We start by illustrating the derivation for the first and second moments of $C_1(\tau)$ and $C_0(\tau)$. We then formulate and prove Theorem 6.2 which reveals the structure of the higher moments, as a function of τ.
By differentiating (4.1) and (4.2) with respect to s and then setting $s = 0$ we get,

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)\right] &= 1+\nu m_1 - \{\mu-\lambda(1+\nu m_1)\}\,\mathbf{E}\left[C_1(\tau)\right], && (6.4)\\ \frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)\right] &= 1+\nu m_1 + \lambda(1+\nu m_1)\mathbf{E}\left[C_1(\tau)\right]. && (6.5)\end{aligned}$$

Strictly speaking it first should be verified that this differentiation and interchanging the order of differentiation is allowed. However, in our case, we can get (6.4) and (6.5) also by directly applying the argument of conditioning on the events in a time interval of length Δ to $\mathbf{E}\left[C_0(\tau)\right]$ and $\mathbf{E}\left[C_1(\tau)\right]$, and then letting $\Delta \downarrow 0$.
Using the initial conditions

$$C_0(0) = C_1(0) = 0,$$

we find

$$\begin{aligned}\mathbf{E}\left[C_1(\tau)\right] &= \frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right), && (6.6)\\ \mathbf{E}\left[C_0(\tau)\right] &= \mu\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\tau - \lambda\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^2\left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right). && (6.7)\end{aligned}$$

If $m_2 < \infty$, we can repeat this procedure to find $\mathbf{E}\left[C_0(\tau)^2\right]$ and $\mathbf{E}\left[C_1(\tau)^2\right]$. Differentiating (4.1) and (4.2) twice w.r.t. s and then setting $s = 0$ (or by a direct conditioning argument) we find

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)^2\right] &= -\{\mu-\lambda(1+\nu m_1)\}\,\mathbf{E}\left[C_1(\tau)^2\right] + 2(1+\nu m_1)\mathbf{E}\left[C_1(\tau)\right]\\ &\quad +2\lambda(1+\nu m_1)\mathbf{E}\left[C_1(\tau)\right]^2 + \nu m_2\left\{1+\lambda\mathbf{E}\left[C_1(\tau)\right]\right\}^2, && (6.8)\\ \frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)^2\right] &= 2(1+\nu m_1)\mathbf{E}\left[C_0(\tau)\right] + 2\lambda(1+\nu m_1)\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right]\\ &\quad +\lambda(1+\nu m_1)\mathbf{E}\left[C_1(\tau)^2\right] + \nu m_2\left\{1+\lambda\mathbf{E}\left[C_1(\tau)\right]\right\}^2. && (6.9)\end{aligned}$$

We can solve this using (6.6) and (6.7):

$$\begin{aligned}\mathbf{E}\left[C_1(\tau)^2\right] &= -(a_1+2a_2)\tau\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\\ &\quad +\frac{a_1+\nu m_2}{\mu-\lambda(1+\nu m_1)}\left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right) && (6.10)\\ &\quad +\frac{a_2}{\mu-\lambda(1+\nu m_1)}\left(1-\mathrm{e}^{-2\{\mu-\lambda(1+\nu m_1)\}\tau}\right),\\ \mathbf{E}\left[C_0(\tau)^2\right] &= b_1\tau + b_2\tau^2 + b_3\tau\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\\ &\quad -b_4\left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right) - b_5\left(1-\mathrm{e}^{-2\{\mu-\lambda(1+\nu m_1)\}\tau}\right), && (6.11)\end{aligned}$$

where

$$\begin{aligned}
a_1 &= 2(1+\nu m_1+\nu m_2\lambda)\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)},\\
a_2 &= \lambda(2(1+\nu m_1)+\nu m_2\lambda)\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^2,\\
b_1 &= \nu m_2\left(\frac{\mu}{\mu-\lambda(1+\nu m_1)}\right)^3,\\
b_2 &= \mu^2\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^2,\\
b_3 &= 2\lambda\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^3(2\mu+\lambda(1+\nu m_1))+2\nu m_2\left(\frac{\lambda(1+\nu m_1)}{\mu-\lambda(1+\nu m_1)}\right)^2\frac{\mu}{\mu-\lambda(1+\nu m_1)},\\
b_4 &= 2\lambda\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^3\frac{2\mu-\lambda(1+\nu m_1)}{\mu-\lambda(1+\nu m_1)}+\nu m_2\lambda\frac{1+\nu m_1}{(\mu-\lambda(1+\nu m_1))^4}\left(3\mu^2-\lambda^2(1+\nu m_1)^2\right),\\
b_5 &= 2\lambda^2\left(\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\right)^4+\frac{1}{2}\nu m_2\lambda^2\frac{(1+\nu m_1)^2}{(\mu-\lambda(1+\nu m_1))^3}\left(\lambda\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}-1\right).
\end{aligned}$$

The same approach can be applied to determine higher moments. In Theorem 6.2 this is done to reveal the structure of these moments.

Theorem 6.2 *For $k \geq 1$, provided that $m_k < \infty$, and thus $\mathbf{E}\left[C_1(\tau)^k\right] < \infty$ and $\mathbf{E}\left[C_0(\tau)^k\right] < \infty$,*

$$\mathbf{E}\left[C_1(\tau)^k\right] = \alpha_0^{(k)} + \sum_{m=1}^{k} \mathrm{e}^{-m\{\mu-\lambda(1+\nu m_1)\}\tau} \sum_{n=0}^{k-m} \alpha_{m,n}^{(k)}\tau^n, \tag{6.12}$$

$$\mathbf{E}\left[C_0(\tau)^k\right] = \sum_{m=0}^{k} \mathrm{e}^{-m\{\mu-\lambda(1+\nu m_1)\}\tau} \sum_{n=0}^{k-m} \beta_{m,n}^{(k)}\tau^n, \tag{6.13}$$

where the $\alpha_0^{(k)}$, $\alpha_{m,n}^{(k)}$ and $\beta_{m,n}^{(k)}$ are coefficients that are independent of τ.

Note that because of the initial Condition (4.3),

$$\sum_{m=1}^{k} \alpha_{m,0}^{(k)} = -\alpha_0^{(k)} \text{ and } \sum_{m=0}^{k} \beta_{m,0}^{(k)} = 0.$$

Proof of Theorem 6.2

Let T_{off} be as before and $N(T_{off})$ be the number of Poisson arrivals (with rate λ) during the period T_{off}. If $C_1(\tau), C_2(\tau), \ldots$ is an i.i.d. sequence with LST $g_1(\tau; s)$, then using Equations (6.2) and (6.3),

$$\begin{aligned}
\mathbf{E}\left[\mathrm{e}^{-s\left(T_{off}+C_1(\tau)+\ldots C_{1+N(T_{off})}(\tau)\right)}\right] &= g_1(\tau;s)\cdot\phi(s+\lambda-\lambda g_1(\tau;s))\\
&= 1+\frac{s}{\lambda}+\frac{1}{\lambda}\sum_{i=1}^{k}\left((s+\lambda)\frac{m_i}{i!}+\frac{m_{i-1}}{(i-1)!}\right)\left(-s+\lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i+\mathrm{o}(s^k)
\end{aligned} \tag{6.14}$$

We write out the terms in the summation as,

$$
\begin{aligned}
&\left(-s+\lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i \\
&= \sum_{i_0+i_1+\cdots+i_k=i}\binom{i}{i_0,\ldots,i_k}(-s)^{i_0}\prod_{j=1}^{k}\left(\lambda\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^{i_j} \\
&= \sum_{n=0}^{k}(-s)^n\sum_{\substack{i_0+i_1+i_2+\cdots+i_k=i\\ i_0+i_1+2i_2+\cdots+ki_k=n}}\binom{i}{i_0,\ldots,i_k}\prod_{j=1}^{k}\left(\lambda\mathbf{E}\left[C_1(\tau)^j\right]\frac{1}{j!}\right)^{i_j}+\mathrm{o}(s^k). \quad (6.15)
\end{aligned}
$$

Note that there are combinations of $k, i, n \in \mathbb{N}$, for which

$$\left\{(i_0,i_1,\ldots,i_k)\in\mathbb{N}_0^{k+1}:\sum_{j=0}^{k}i_j=i, i_0+\sum_{j=1}^{k}ji_j=n\right\}=\emptyset.$$

We now prove the theorem by induction on k. From (6.14) and (6.15) we can show that if $\mathbf{E}\left[C_1(\tau)^j\right]$ has the form of (6.12) for $j=1,2,\ldots,k-1$, then:

$$
\begin{aligned}
&\mathbf{E}\left[\left(T_{off}+C_1(\tau)+\cdots+C_{1+N(T_{off})}(\tau)\right)^k\right] \\
&= (\lambda m_1+1)\mathbf{E}\left[C_1(\tau)^k\right]+\gamma_0^{(k)}+\sum_{m=1}^{k}\mathrm{e}^{-m\{\mu-\lambda(1+\nu m_1)\}\tau}\sum_{n=0}^{k-m}\gamma_{m,n}^{(k)}\tau^n.
\end{aligned}
$$

This can be verified by noting that the only contribution of $\mathbf{E}\left[C_1(\tau)^k\right]$ to the coefficient of $(-s)^k$ in (6.14), is through the term with $i=1$. All other contributions to the coefficient of $(-s)^k$ are either zero, or come from products of the $\mathbf{E}\left[C_1(\tau)^j\right]$, for $j=1,2,\ldots,k-1$. Apart from a constant in τ, they all consist of terms of the form $\mathrm{e}^{-m\{\mu-\lambda(1+\nu m_1)\}\tau}\tau^n$, with $m\geq 1$, $n\geq 0$ and $m+n\leq k$. Writing out the terms, it is seen that $\gamma_{1,k-1}^{(k)}=0$. This is a consequence of the fact that for $l_1,l_2=1,2,\ldots$, the product $\mathbf{E}\left[C_1(\tau)^{l_1}\right]\times\mathbf{E}\left[C_1(\tau)^{l_2}\right]$ is of the same form as $\mathbf{E}\left[C_1(\tau)^{l_1+l_2}\right]$ in (6.12), except for the terms containing $\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\tau^n$, with $n\geq\max(l_1,l_2)$, which do not appear. The other coefficients $\gamma_{m,n}^{(k)}$ can be found from the $\alpha_{m,n}^{(j)}$ for $j<k$, by use of (6.14) and (6.15).
As before, we can derive a differential equation for $\mathbf{E}\left[C_1(\tau)^k\right]$:

$$
\begin{aligned}
\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)^k\right] &= -(\lambda+\mu+\nu)\mathbf{E}\left[C_1(\tau)^k\right]+k\mathbf{E}\left[C_1(\tau)^{k-1}\right]+\lambda\mathbf{E}\left[(C_1(\tau)+C_2(\tau))^k\right] \\
&\quad+\nu\mathbf{E}\left[\left(T_{off}+C_1(\tau)+\cdots+C_{1+N(T_{off})}(\tau)\right)^k\right] \\
&= -\{\mu-\lambda(1+\nu m_1)\}\mathbf{E}\left[C_1(\tau)^k\right]+k\mathbf{E}\left[C_1(\tau)^{k-1}\right] \\
&\quad+\lambda\sum_{l=1}^{k-1}\binom{k}{l}\mathbf{E}\left[C_1(\tau)^l\right]\mathbf{E}\left[C_1(\tau)^{k-l}\right] \\
&\quad+\nu\gamma_0^{(k)}+\nu\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\sum_{n=0}^{k-2}\gamma_{1,n}^{(k)}\tau^n+\nu\sum_{m=2}^{k}\mathrm{e}^{-m\{\mu-\lambda(1+\nu m_1)\}\tau}\sum_{n=0}^{k-m}\gamma_{m,n}^{(k)}\tau^n.
\end{aligned}
$$

Note that in the right-hand side of this differential equation, no term with $e^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\tau^{k-1}$ appears. Solving for $\mathbf{E}\left[C_1(\tau)^k\right]$, indeed leads to the form of Relation (6.12). The coefficients $\alpha_0^{(k)}$ and $\alpha_{m,n}^{(k)}$ are recursively determined by the $\alpha_0^{(j)}$ and $\alpha_{m,n}^{(j)}$ for $j < k$.
To prove the second part of the theorem we use the differential equation for $\mathbf{E}\left[C_0(\tau)^k\right]$:

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)^k\right] &= -(\lambda+\nu)\mathbf{E}\left[C_0(\tau)^k\right] + k\mathbf{E}\left[C_0(\tau)^{k-1}\right] + \lambda\mathbf{E}\left[(C_0(\tau)+C_1(\tau))^k\right] \\ &\quad +\nu\mathbf{E}\left[\left(T_{off}+C_0(\tau)+C_1(\tau)+\cdots+C_{N(T_{off})}(\tau)\right)^k\right] \\ &= k\mathbf{E}\left[C_0(\tau)^{k-1}\right] + \lambda\sum_{l=0}^{k-1}\binom{k}{l}\mathbf{E}\left[C_0(\tau)^l\right]\mathbf{E}\left[C_1(\tau)^{k-l}\right] \\ &\quad +\nu\sum_{l=0}^{k-1}\binom{k}{l}\mathbf{E}\left[C_0(\tau)^l\right]\mathbf{E}\left[\left(T_{off}+C_1(\tau)+\cdots+C_{N(T_{off})}(\tau)\right)^{k-l}\right].\end{aligned}$$

By similar arguments as before, we find Relation (6.13). ■

7. Moments of the conditional sojourn time

In this section we study the moments of the sojourn time of a customer conditioned on the service requirement, the state of the server upon arrival, and the number of other customers in the system. We give these moments in terms of the moments of $C_1(\tau)$ and $C_0(\tau)$. In particular, using the expressions for the first two moments of $C_1(\tau)$ and $C_0(\tau)$ found in Section 6, we find explicit expressions for the first two moments of the conditional sojourn time.
For compactness, we will use the notation

$$\begin{aligned}\mathbf{E}_n\left[V_1(\tau)^k\right] &:= \mathbf{E}\left[V(\tau)^k|\{Y(0)=1,Z(0)=n\}\right], \\ \mathbf{E}_n\left[V_0(\tau)^k\right] &:= \mathbf{E}\left[V(\tau)^k|\{Y(0)=0,Z(0)=n\}\right].\end{aligned}$$

Observation 7.1 *From Theorem 3.1, we have for $k, n \in \mathbb{N}$,*

$$\begin{aligned}\mathbf{E}_n\left[V_1(\tau)^k\right] &= \mathbf{E}\left[(C_0(\tau)+\ldots C_n(\tau))^k\right] \\ &= \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[C_n(\tau)^{k-j}\right]\ \mathbf{E}\left[(C_0(\tau)+\ldots C_{n-1}(\tau))^j\right] \\ &= \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[C_1(\tau)^{k-j}\right]\ \mathbf{E}_{n-1}\left[V_1(\tau)^j\right], \qquad (7.1) \\ \mathbf{E}_0\left[V_1(\tau)^k\right] &= \mathbf{E}\left[C_0(\tau)^k\right]. \qquad (7.2)\end{aligned}$$

Moreover, combining Theorem 3.1 and Corollary 3.2, we find,

$$\mathbf{E}_n\left[V_0(\tau)^k\right] = \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[\left(D_0+\sum_{i=1}^{A_0}C_i(\tau)\right)^j\right]\mathbf{E}_n\left[V_1(\tau)^{k-j}\right]. \qquad (7.3)$$

We remind the reader that D_0 is the residual off-period at time zero, with LST $\phi_0(.)$, and that A_0 is the number of arrivals during D_0. We can find $\mathbf{E}\left[\left(D_0+\sum_{i=1}^{A_0} C_i(\tau)\right)^j\right]$ as

$$\mathbf{E}\left[\left(D_0+\sum_{i=1}^{A_0} C_i(\tau)\right)^j\right] = (-1)^j \frac{\partial^j}{\partial s^j}\phi_0(s+\lambda-\lambda g_1(\tau;s))\bigg|_{s=0}.$$

These derivatives can be found by using Lemma 1 of [6] to expand $\phi_0(s+\lambda-\lambda g_1(\tau;s))$ in a Taylor series, analogous to Equation (6.3).

From (7.1) and (7.2) we can compute the conditional moments $\mathbf{E}_n\left[V_1(\tau)^k\right]$ recursively, once we have the moments of $C_0(\tau)$ and $C_1(\tau)$. The moments of $V_0(\tau)$ are then found from (7.3). In particular we have for $k=1$, also cf. Equations (3.2) and (3.3),

$$\begin{aligned}
\mathbf{E}_n\left[V_1(\tau)\right] &= \mathbf{E}\left[C_0(\tau)\right]+n\mathbf{E}\left[C_1(\tau)\right], &(7.4)\\
\mathbf{E}_n\left[V_0(\tau)\right] &= \mathbf{E}\left[D_0\right]+\mathbf{E}\left[C_0(\tau)\right]+(n+\lambda\mathbf{E}\left[D_0\right])\,\mathbf{E}\left[C_1(\tau)\right], &(7.5)\\
& &(7.6)
\end{aligned}$$

and for $k=2$,

$$\begin{aligned}
\mathbf{E}_n\left[V_1(\tau)^2\right] &= \mathbf{E}\left[C_0(\tau)^2\right]+n\mathbf{E}\left[C_1(\tau)^2\right]+2n\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right]\\
&\quad +n(n-1)\mathbf{E}\left[C_1(\tau)\right]^2, &(7.7)\\
\mathbf{E}_n\left[V_0(\tau)^2\right] &= \mathbf{E}\left[D_0^2\right]+2\mathbf{E}\left[D_0\right]\left(\mathbf{E}\left[C_0(\tau)\right]+n\mathbf{E}\left[C_1(\tau)\right]\right)\\
&\quad +2\lambda\mathbf{E}\left[D_0^2\right]\mathbf{E}\left[C_1(\tau)\right]+\mathbf{E}\left[C_0(\tau)^2\right]\\
&\quad +2\left(n+\lambda\mathbf{E}\left[D_0\right]\right)\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right] &(7.8)\\
&\quad +\left(n+\lambda\mathbf{E}\left[D_0\right]\right)\mathbf{E}\left[C_1(\tau)^2\right]\\
&\quad +\left(n(n-1)+(2n-1)\lambda\mathbf{E}\left[D_0\right]+\lambda^2\mathbf{E}\left[D_0^2\right]\right)\mathbf{E}\left[C_1(\tau)\right]^2.
\end{aligned}$$

Using (6.6), (6.7), (6.10) and (6.11) we have explicit formulas for these first and second moments.

Theorem 7.1 *For fixed $k\in\mathbb{N}$, if $m_k<\infty$ and $\mathbf{E}\left[D_0^k\right]<\infty$, then $\mathbf{E}_n\left[V_1(\tau)^k\right]$ and $\mathbf{E}_n\left[V_0(\tau)^k\right]$ are polynomials in n of degree k:*

$$\mathbf{E}_n\left[V_i(\tau)^k\right]=\sum_{l=0}^{k} c_{k,l}^{(i)}(\tau)n^l,\quad i\in\{0,1\}. \tag{7.9}$$

The coefficients $c_{k,l}^{(1)}(\tau)$ are recursively defined by

$$\begin{aligned}
c_{k,l+1}^{(1)}(\tau) &= \frac{1}{l+1}\left\{\sum_{i=l+2}^{k}(-1)^{i-l}\binom{i}{l}c_{k,i}^{(1)}(\tau)\right.\\
&\quad\left.+\sum_{j=l}^{k-1}\sum_{i=l}^{j}(-1)^{i-l}\binom{i}{l}\binom{k}{j}\mathbf{E}\left[C_1(\tau)^{k-j}\right]c_{j,i}^{(1)}(\tau)\right\}, &(7.10)\\
c_{k,0}^{(1)}(\tau) &= \mathbf{E}\left[C_0(\tau)^k\right],
\end{aligned}$$

with $k \in \mathbb{N}$, and $l = 0, 1, \ldots, k-1$. The empty sum (when $l+2 = k+1$) is equal to zero.
For $k \in \mathbb{N}$, and $l = 0, 1, \ldots, k$, the $c_{k,l}^{(0)}(\tau)$ are given by

$$c_{k,l}^{(0)}(\tau) = \sum_{j=l}^{k} \binom{k}{j} c_{j,l}^{(1)}(\tau) \mathbf{E}\left[\left(D_0 + \sum_{i=1}^{A_0} C_i(\tau)\right)^{k-j}\right]. \tag{7.11}$$

Hence, for $i \in \{0,1\}$, $k \in \mathbb{N}$, and $l \in \{0,1,\ldots,k\}$, the functions $c_{k,l}^{(i)}(\tau)$ are of the same form as $\mathbf{E}\left[C_0(\tau)^k\right]$ in Theorem 6.2.

Proof
Expression (7.9), for $i = 1$, can be proved by arguing that (7.1) and (7.2) uniquely determine the $\mathbf{E}_n\left[V_1(\tau)^k\right]$ for $k, n \in \mathbb{N}$, and that (7.9), for $i = 1$, with the $c_{k,l}^{(1)}(\tau)$ defined by (7.10), satisfies (7.1) and (7.2). Then, (7.9), for $i = 0$, and (7.11), follow from Relation (7.3).
The last statement follows from the fact that a product of two functions of the class defined by Relation (6.13), one with $k = l_1$, and the other with $k = l_2$, gives a function of the same class, with $k = l_1 + l_2$. ■

Sengupta and Jagerman [23, Theorem 1] proved that, in the $M/M/1$ processor sharing queue without server breakdowns, the k^{th} moment of the sojourn time conditional on n competing customers, is a polynomial in n of degree k. As a corollary of Theorem 7.1 we have that the result of Sengupta and Jagerman is also true for the $M/M/1$ processor sharing queue *with generally distributed server breakdowns.*

Corollary 7.2 *If $m_k < \infty$ and $\mathbf{E}\left[D_0^k\right] < \infty$, then*

$$\mathbf{E}_n\left[(V_i)^k\right] := \int_{\tau=0}^{\infty} \mathbf{E}_n\left[V_i(\tau)^k\right] \mu \mathrm{e}^{-\mu\tau} \mathrm{d}\tau = \sum_{l=0}^{k} n^l \int_{\tau=0}^{\infty} c_{k,l}^{(i)}(\tau) \mu \mathrm{e}^{-\mu\tau} \mathrm{d}\tau, \quad i \in \{0,1\}.$$

Proof
Obviously, from the last statement of Theorem 7.1, $\int_{\tau=0}^{\infty} c_{k,l}^{(i)}(\tau)\mu \mathrm{e}^{-\mu\tau}\mathrm{d}\tau < \infty$, for $i \in \{0,1\}$. The corollary then follows from Expression (7.9). ■

8. Sojourn time in steady state

In this section we derive results for the sojourn time of a customer with an amount of work τ, arriving to the system in steady state. If the number of competing customers in the system at the beginning of the sojourn time is (as before) denoted by $Z(0)$ and the state of the server by $Y(0)$, then

$$(Z(0), Y(0)) \stackrel{d}{=} (X, Y),$$

and the distribution of (X, Y) is given by (2.2) and (2.3).

Theorem 8.1 *For $\mathrm{Re}(s) \geq 0$, the LST of $V(\tau)$ is given by,*

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)}\,|Y(0)=1\right] = g_0(\tau;s)\frac{\mu-\lambda(1+\nu m_1)}{\mu-\lambda g_1(\tau;s)-\nu g_1(\tau;s)\frac{1-\phi(\lambda(1-g_1(\tau;s)))}{1-g_1(\tau;s)}}, \tag{8.1}$$

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)}\,|Y(0)=0\right] = \mathbf{E}\left[\mathrm{e}^{-sV(\tau)}\,|Y(0)=1\right] \times\frac{1-\phi(s+\lambda-\lambda g_1(\tau;s))}{m_1\,(s+\lambda-\lambda g_1(\tau;s))}\times\frac{1-\phi(\lambda-\lambda g_1(\tau;s))}{m_1\lambda(1-g_1(\tau;s))}. \tag{8.2}$$

Proof
Equation (8.1) is found from (2.2) and (3.2). To find the sojourn times starting with an off-period, we remark that the residual length of that off-period is distributed as the overshoot of the off-periods, i.e. $\phi_0(s)=\frac{1-\phi(s)}{m_1 s}$. Then using (2.3) and (3.3) we get (8.2). ■

Corollary 8.2 *The mean sojourn time is given by*

$$\mathbf{E}[V(\tau)] = \frac{\tau}{\frac{1}{1+\nu m_1}-\frac{\lambda}{\mu}}+\frac{\frac{1}{2}\nu m_2}{1+\nu m_1} + \left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right)\frac{\frac{1}{2}\nu\lambda m_2}{\mu-\lambda(1+\nu m_1)}\left\{1+\frac{\mu}{\mu-\lambda(1+\nu m_1)}\right\}. \tag{8.3}$$

Proof
From Theorem 8.1, by differentiating w.r.t. s and putting $s=0$, we find

$$\mathbf{E}[V(\tau)\,|Y(0)=1] = \frac{\tau}{\frac{1}{1+\nu m_1}-\frac{\lambda}{\mu}}+\left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right)\frac{\frac{1}{2}\nu\lambda^2 m_2(1+\nu m_1)}{\{\mu-\lambda(1+\nu m_1)\}^2},$$

$$\mathbf{E}[V(\tau)\,|Y(0)=0] = \frac{m_2}{2m_1}+\frac{\tau}{\frac{1}{1+\nu m_1}-\frac{\lambda}{\mu}} + \left(1-\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\right)\frac{1+\nu m_1}{\mu-\lambda(1+\nu m_1)}\left\{\lambda\frac{m_2}{m_1}+\frac{\frac{1}{2}\nu\lambda^2 m_2}{\mu-\lambda(1+\nu m_1)}\right\}.$$

Alternatively, we may find $\mathbf{E}[V(\tau)\,|Y(0)=1]$ more directly by substituting

$$\mathbf{E}[X\,|Y=1]=\frac{\lambda(1+\nu m_1)+\frac{1}{2}\nu\lambda^2 m_2}{\mu-\lambda(1+\nu m_1)},$$

for n in (7.4), and using (6.6) and (6.7). Similarly, we can find $\mathbf{E}[V(\tau)\,|Y(0)=0]$ by substituting

$$\mathbf{E}[X\,|Y=0]=\lambda\frac{m_2}{2m_1}+\frac{\lambda(1+\nu m_1)+\frac{1}{2}\nu\lambda^2 m_2}{\mu-\lambda(1+\nu m_1)},$$

for n in (7.5), and with $\mathbf{E}[D_0]=\frac{m_2}{2m_1}$.
By averaging over $\mathbf{P}\{Y=1\}=\frac{1}{1+\nu m_1}$ and $\mathbf{P}\{Y=0\}=\frac{\nu m_1}{1+\nu m_1}$ we get $\mathbf{E}[V(\tau)]$. ■

Note that $\mathbf{E}[V(\tau)]$ is not linear in τ unless $\nu m_2=0$, i.e. when there are no off-periods. The linearity in that case was already observed for the $M/M/1$ processor sharing queue by Coffman et al. [3], and for the $M/G/1$ processor sharing queue by Kitayev and Yashkov [11].

We conclude this section with two remarks, discussing two cases in which the conditional mean sojourn time is approximately linear in τ.

Remark 8.1 $\mathbf{E}[V(\tau)]$ is 'almost linear' in τ when the on- and off-periods alternate rapidly. Formally, construct a new sequence of on- and off-periods by multiplying each on- and off-period by a factor $\alpha \in (0,\infty)$. So in the new sequence, the on-periods are exponentially distributed with mean α/ν, and the new off-periods, generically denoted by $T_{off}^{(\alpha)}$, have LST

$$\mathbf{E}\left[\mathrm{e}^{-sT_{off}^{(\alpha)}}\right] = \phi(\alpha s).$$

In particular, the first two moments of $T_{off}^{(\alpha)}$ are $m_1^{(\alpha)} = \alpha m_1$ and $m_2^{(\alpha)} = \alpha^2 m_2$. Obviously, $\frac{\nu}{\alpha} m_1^{(\alpha)} = \nu m_1$ is independent of α, and so is the probability that the server is on (with the new sequence of on- and off-periods). Therefore the ergodicity condition remains unchanged. If $V^{(\alpha)}(\tau)$ is the sojourn time of a customer with τ work under the new on- and off-periods, then

$$\lim_{\alpha\downarrow 0} \mathbf{E}\left[V^{(\alpha)}(\tau)\right] = \frac{\tau}{\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}}.$$

This limiting case ($\alpha \downarrow 0$) corresponds to the case where the server is always available and works at the *constant* speed $\frac{1}{1+\nu m_1}$ instead of 1. Note that $\frac{1}{1+\nu m_1}$ is the *average* service speed in the model with the alternating server.

On the other hand, when the server alternates very slowly, the expected sojourn time can become arbitrarily large (irrespective of the amount of work the customer carries with him): $\lim_{\alpha\to\infty} \mathbf{E}\left[V^{(\alpha)}(\tau)\right] = \infty$.

Remark 8.2 From (8.3) we also conclude that $\mathbf{E}[V(\tau)]$ is approximately linear for large τ. This is in agreement with the result of Corollary 5.2. This can intuitively be explained by noting that if τ is large, then also the sojourn time will be large. Over a long period of time, the fluctuations in the server availability average out, and for large τ an additional amount of work $\Delta\tau$ requires $\frac{1}{\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}}\Delta\tau$ time units. For an interpretation of $\frac{1}{1+\nu m_1} - \frac{\lambda}{\mu}$, see Remark 5.3.

9. Heavy traffic

In this section we prove the following:

Theorem 9.1 *Provided that the second moment of the off-periods, m_2, is finite,*

$$\lim_{\rho\uparrow 1} \mathbf{E}\left[\mathrm{e}^{-(1-\rho)sV(\tau)}\right] = \frac{1}{1 + (1+\nu m_1 + \frac{1}{2}\nu\lambda m_2)s\tau}, \qquad \mathrm{Re}(s) \geq 0,$$

where the traffic load ρ is defined by

$$\rho := \frac{\lambda(1+\nu m_1)}{\mu}.$$

This result is also known for the ordinary M/G/1 queue (without service interruptions), see Sengupta [22] and Yashkov [31].

Thus, in heavy traffic, the distribution of $(1-\rho)V(\tau)$ converges to the exponential distribution with mean $\left(1+\nu m_1 + \frac{1}{2}\nu\lambda m_2\right)\tau$. Note that the limiting mean is linear in τ. To prove the theorem we need the following lemma:

Lemma 9.1 *For* $\mathrm{Re}(s) \geq 0$,

(i) $\lim_{\rho\uparrow 1} g_1(\tau;(1-\rho)s) = 1$, *and* $\lim_{\rho\uparrow 1} g_0(\tau;(1-\rho)s) = 1$,

(ii) $\lim_{\rho\uparrow 1} \frac{1-g_1(\tau;(1-\rho)s)}{1-\rho} = (1+\nu m_1)s\tau$.

Note that (ii) of Lemma 9.1 can be rewritten in terms of the LST of the overshoot distribution of $C_1(\tau)$ as follows:

$$\lim_{\rho\uparrow 1} \frac{1-g_1(\tau;(1-\rho)s)}{\mathbf{E}\left[C_1(\tau)\right](1-\rho)s} = 1,$$

see Formula (6.6).

Proof of Lemma 9.1
Part (i): Substitute $(1-\rho)s$ for s in Equations (4.4) and (4.5), and let $\rho \uparrow 1$. Assuming that $h_1(\tau;s) := \lim_{\rho\uparrow 1} g_1(\tau;(1-\rho)s)$ and $h_0(\tau;s) := \lim_{\rho\uparrow 1} g_0(\tau;(1-\rho)s)$ exist we find (using the Dominated Convergence Theorem for the interchange of limit and integrals),

$$\begin{aligned}
h_1(\tau+\Delta;s) &= \int_{t=0}^{\Delta} \mu \mathrm{e}^{-\mu t-\lambda\left(t-\int_{u=0}^{t} h_1(\tau+u;s)du\right)-\nu\left(t-\int_{u=0}^{t}\phi(\lambda(1-h_1(\tau+u;s)))du\right)}\mathrm{d}t \\
&\quad + h_1(\tau;s)\mathrm{e}^{-\mu\Delta-\lambda\left(\Delta-\int_{u=0}^{\Delta} h_1(\tau+u;s)du\right)-\nu\left(\Delta-\int_{u=0}^{\Delta}\phi(\lambda(1-h_1(\tau+u;s)))du\right)}, \\
h_0(\tau+\Delta;s) &= h_0(\tau;s)\mathrm{e}^{-\lambda\left(\Delta-\int_{u=0}^{\Delta} h_1(\tau+u;s)du\right)-\nu\left(\Delta-\int_{u=0}^{\Delta}\phi(\lambda(1-h_1(\tau+u;s)))du\right)}.
\end{aligned} \tag{9.1}$$

From this we can (as in Section 4) derive the following differential equations:

$$\begin{aligned}
\frac{\partial}{\partial\tau}h_1(\tau;s) &= -(\lambda+\nu+\mu)\,h_1(\tau;s) + \lambda\left\{h_1(\tau;s)\right\}^2 + \mu + \nu h_1(\tau;s)\phi(\lambda(1-h_1(\tau;s))), \\
\frac{\partial}{\partial\tau}h_0(\tau;s) &= -(\lambda+\nu)\,h_0(\tau;s) + \lambda h_0(\tau;s)h_1(\tau;s) + \nu h_0(\tau;s)\phi(\lambda(1-h_1(\tau;s))).
\end{aligned}$$

Together with the boundary conditions $h_1(0;s) = h_0(0;s) = 1$, these differential equations uniquely determine $h_1(\tau;s)$ and $h_0(\tau;s)$. Part (i) is now proved by noting that $h_1(\tau;s) \equiv 1$ and $h_0(\tau;s) \equiv 1$ satisfy these equations. A comment should however be made about the assumption on the existence of $h_1(\tau;s)$ and $h_0(\tau;s)$: Since, for any $\mathrm{Re}(s) \geq 0$, $|g_1(\tau;s)| \leq 1$, we can find a sequence $(\rho_k)_{k\in\mathbf{N}}$ in the interval [0,1] such that $\lim_{k\to\infty}\rho_k = 1$ and $\overline{h}_1(\tau;s) := \lim_{k\to\infty} g_1(\tau;(1-\rho_k)s)$ exists. For $\overline{h}_1(\tau;s)$ we can formulate the differential equations, leading to $\overline{h}_1(\tau;s) \equiv 1$. Since the limit is the same for all convergent sequences, $h_1(\tau;s)$ exists. In the same way it can be argued that $h_0(\tau;s)$ exists.
Part (ii): The proof proceeds along the same lines as for Part (i). We assume the existence of

$$l_1(\tau;s) := \lim_{\rho\uparrow 1} \frac{1-g_1(\tau;(1-\rho)s)}{1-\rho}.$$

Again this existence can be shown by following the subsequent steps for the limit of a convergent sequence $\frac{1-g_1(\tau;(1-\rho_k)s)}{1-\rho_k}$. Such a sequence exists because $\frac{1-g_1(\tau;\omega)}{\omega} \leq \mathbf{E}\left[C_1(\tau)\right]$ for any $\mathrm{Re}(\omega) \geq 0$, and $\mathbf{E}\left[C_1(\tau)\right]$ is bounded in $\rho \in [0,1]$, see Formula (6.6).
Substitute $(1-\rho)s$ for s in (4.4), subtract both sides of this equation from 1 and use,

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \frac{1-\mathrm{e}^{-(1-\rho)sx-\lambda\int_{u=0}^{x}(1-g_1(\tau+u;(1-\rho)s))du-\nu\int_{u=0}^{x}(1-\phi((1-\rho)s+\lambda(1-g_1(\tau+u;(1-\rho)s))))du}}{1-\rho} \\
&= sx + \lambda\int_{u=0}^{x} l_1(\tau+u;s)\mathrm{d}u + \nu\int_{u=0}^{x} m_1\left(s+\lambda l_1(\tau+u;s)\right)\mathrm{d}u,
\end{aligned}$$

(again with the Dominated Convergence Theorem to interchange limit and integrals), to find,

$$\begin{aligned} l_1(\tau+\Delta;s) &= \int_{t=0}^{\Delta} \mu \mathrm{e}^{-\mu t}\left((1+\nu m_1)st + \lambda(1+\nu m_1)\int_{u=0}^{t} l_1(\tau+u;s)\mathrm{d}u\right)\mathrm{d}t \\ &\quad + \mathrm{e}^{-\mu\Delta}\left(l_1(\tau;s) + (1+\nu m_1)s\Delta + \lambda(1+\nu m_1)\int_{u=0}^{\Delta} l_1(\tau+u;s)\mathrm{d}u\right). \end{aligned}$$

For $\Delta \downarrow 0$ we may now write:

$$\begin{aligned} l_1(\tau+\Delta;s) &= l_1(\tau;s) - \Delta\mu l_1(\tau;s) + \Delta(1+\nu m_1)s + \Delta\lambda(1+\nu m_1)l_1(\tau;s) + \mathrm{o}(\Delta) \\ &= l_1(\tau;s) + \Delta(1+\nu m_1)s + \mathrm{o}(\Delta), \end{aligned}$$

where for the last equality we have used that $\mu = \lambda(1+\nu m_1)$ when $\rho = 1$.
Using the boundary condition $l_1(0;s) \equiv 0$ we readily find $l_1(\tau;s) = (1+\nu m_1)s\tau$. ■

Using Lemma 9.1, Theorem 9.1 can now be proved by substituting $(1-\rho)s$ for s in (8.1) and (8.2), and letting $\rho \uparrow 1$.

10. Final Remarks

In this paper we have studied the sojourn time of a customer in the M/M/1 queue with processor sharing service discipline, and the server alternating between exponentially distributed on-periods and generally distributed off-periods. This model is of interest for the performance analysis of the ABR service in ATM networks, and best-effort services in IP networks. By using a time scale transformation, we formulated the problem in terms of a branching process with a cost structure on it. We indicated how the same transformation can be applied to general service times, but for the sake of simplicity and notational convenience, we restricted the analysis to the case of exponentially distributed service requirements. The explicitness of the results may be valuable in future research when studying $M/M/1$ processor sharing models with a more general varying service process.

The sojourn time $V(\tau)$ of a customer, conditional on his service requirement τ, was decomposed into a sum of independent random variables, thus generalising the result known for the regular M/G/1 queue with processor sharing. The LST's of the random variables composing $V(\tau)$, were characterised through an integral equation, which is suitable for numerical evaluation. We computed the first two moments of these random variables, and identified the structure of higher moments. We used these to find the moments of $V(\tau)$, conditional on the number of competing customers, and generalised a result of Sengupta and Jagerman [23, Theorem 1]. We further performed an asymptotic analysis for $\tau \to \infty$, proving that $V(\tau)/\tau$ converges (with probability 1) to a constant. In heavy traffic, that is for the traffic load $\rho \uparrow 1$, it was shown that $(1-\rho)V(\tau)$ converges to an exponential distribution, of which the mean is linear in τ.

In particular, we found that $\mathbf{E}[V(\tau)]$ is not linear in τ, unlike in processor sharing queues without service interruptions. We saw that $\mathbf{E}[V(\tau)]$ is approximately linear in three cases: (i) when the on- and off-periods alternate rapidly, (ii) when τ is large, and (iii) in heavy traffic.

Acknowledgement

The author would like to thank J.W. Cohen, O.J. Boxma, J.A.C. Resing and A.P. Zwart for interesting discussions and many helpful comments on previous versions that led to improvement of the paper. In particular the author is indebted to Professor J.W. Cohen for the comments that led to the asymptotic analysis in Section 5. Also, J.L. van den Berg and M.R.H. Mandjes are acknowledged for pointing the author to useful references in the area of application.

References

1. The ATM Forum Technical Committee. *Traffic Management Specification.* Version 4.0, April 1996.
2. J.L. van den Berg, O.J. Boxma. *The $M/G/1$ queue with processor sharing and its relation to a feedback queue.* Queueing Systems 9 (1991), 365–401.
3. E.G. Coffman, R.R. Muntz, H. Trotter. *Waiting time distributions for processor sharing systems.* Journal of the Association for Computing Machinery 17 (1970), 123–130.
4. J.W. Cohen. *The Single Server Queue.* North-Holland Publishing Company, Amsterdam, 2nd edition, 1982.
5. J.W. Cohen. *The multiple phase service network with generalized processor sharing.* Acta Informatica 12 (1979), 245–284.
6. A. De Meyer, J.L. Teugels. *On the asymptotic behaviour of the distributions of the busy period and service time in* M/G/1. J. Applied Probability 17 (1980), 802–813.
7. A. Federgruen, L. Green. *Queueing systems with service interruptions.* Operations Research 34 (1986), 752–768.
8. A. Federgruen, L. Green. *Queueing systems with service interruptions II.* Naval Res. Logist. 35 (1988), 345–358.
9. D.P. Gaver, Jr. *A waiting line with interrupted service, including priorities.* J. Roy. Statist. Soc. 24 (1962), 73–90.
10. S. Grishechkin. *On a relationship between processor sharing queues and Crump-Mode-Jagers branching processes.* Adv. Applied Probability 24 (1992), 653–698.
11. M.Yu. Kitayev, S.F. Yashkov. *Analysis of a single-channel queueing system with the discipline of uniform sharing of a device.* Engineering Cybernetics 17 (1979), 42–49.
12. D-S. Lee. *Analysis of a single server queue with semi-Markovian service interruption.* Queueing Systems 27 (1997), 153–178.
13. W. Li, D. Shi, X. Chao. *Reliability analysis of $M/G/1$ queueing systems with server breakdowns and vacations.* J. Applied Probability 34 (1997), 546–555.

14. I. Mitrani, B. Avi-Itzhak. *A many server queue with service interruptions.* Operations Research 16 (1968), 628–638.
15. J.A. Morrison. *Response-time distribution for a processor sharing system.* SIAM Journal of Applied Mathematics 45 (1985), 152–167.
16. M.F. Neuts. *Matrix-geometric Solutions in Stochastic Models - An Algorithmic Approach.* The Johns Hopkins University Press, Baltimore, 1981.
17. T.J. Ott. *The sojourn time distributions in the $M/G/1$ queue with processor sharing.* J. Applied Probability 21 (1984), 360–378.
18. K.M. Rege, B. Sengupta. *A decomposition theorem and related results for the discriminatory processor sharing queue.* Queueing Systems 18 (1994), 333–351.
19. J. Roberts. *Realizing quality of service guarantees in multiservice networks.* Proceedings of IFIP Seminar PMCCN'97 (1998), to appear.
20. M. Sakata, S. Noguchi, J. Oizumi. *Analysis of a processor shared queueing model for time sharing systems.* Proc. 2nd Hawaii International Conference on System Sciences (1969), 625–628.
21. R. Schassberger. *A new approach to the $M/G/1$ processor sharing queue.* Adv. Applied Probability 16 (1984), 202–213.
22. B. Sengupta. *An approximation for the sojourn-time distribution for the $GI/G/1$ processor-sharing queue.* Comm. Statist. – Stochastic Models 8 (1992), 35–57.
23. B. Sengupta, D.L. Jagerman. *A conditional response time of the $M/M/1$ processor sharing queue.* AT&T Technical Journal 64 (1985), 409–421.
24. T. Takine, B. Sengupta. *A single server queue with server interruptions.* Queueing Systems 26 (1997), 285–300.
25. H.C. Tijms. *Stochastic Models – An Algorithmic Approach.* John Wiley & Sons Ltd, Chichester (England), 1994.
26. P.D. Welch. *On a generalized $M/G/1$ queueing process in which the first customer of each busy period receives exceptional service.* Operations Research 12 (1964), 736–752.
27. H. White, L.S. Christie. *Queuing with preemptive priorities or with breakdown.* Operations Research 6 (1958), 79–95.
28. 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.
29. S.F. Yashkov. *Processor-sharing queues: Some progress in analysis.* Queueing Systems 2 (1987), 1–17.
30. S.F. Yashkov. *Mathematical problems in the theory of processor-sharing queueing systems.* J. Soviet Mathematics 58 (1992), 101–147.
31. S.F. Yashkov. *On a heavy traffic limit theorem for the $M/G/1$ processor-sharing queue.* Comm. Statist. – Stochastic Models 9 (1993), no. 3, 467–471.