Sojourn times in a processor sharing queue with service interruptions

R. Núñez-Queija

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
E-mail: sindo@cwi.nl

Received 21 August 1998; revised 4 June 1999

We study the sojourn times of customers in an M/M/1 queue with the processor sharing service discipline and a server that is subject to breakdowns. The lengths of the breakdowns have a general distribution, whereas the "on-periods" are exponentially distributed. A branching process approach leads to a decomposition of the sojourn time, in which the components are independent of each other and can be investigated separately. We derive the Laplace–Stieltjes transform of the sojourn-time distribution in steady state, and show that the expected sojourn time is not proportional to the service requirement. In the heavy-traffic limit, the sojourn time conditioned on the service requirement and scaled by the traffic load is shown to be exponentially distributed. The results can be used for the performance analysis of elastic traffic in communication networks, in particular, the ABR service class in ATM networks, and best-effort services in IP networks.

Keywords: processor sharing, service interruptions, sojourn time, elastic traffic, available bit rate services, best-effort traffic

1. Introduction

We consider a processor sharing queue with a single server which 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. Because of the priority structure,

© J.C. Baltzer AG, Science Publishers

ABR connections will typically 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 (or more) *positive* service speeds. The definition of the ABR service class [31] requires that the available capacity is fairly shared among users. This explains the relevance of processor sharing models for the performance analysis of ABR connections. 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 [23].

Processor sharing queues have been studied extensively in the literature. Already in 1969 Sakata et al. [24] showed that the steady-state queue length distribution of the M/G/1 queue with processor sharing was geometric, and insensitive to the service time distribution except from its first moment. This result was extended by Cohen [3] to a general class of networks, in which 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 (there called *generalised processor sharing*). Cohen [3] 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 distribution 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. [2]. Sengupta and Jagerman [28] found an alternative expression for the LST of the distribution of the sojourn time conditioned only on the number of customers seen upon arrival. In particular they found that the kth 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 [18].

The sojourn time distribution in the M/G/1 processor sharing queue was first analysed by Yashkov [36]. Schassberger [25] considered the M/G/1 processor sharing queue as the limit of the round robin discipline. Ott [21] found the joint LST and generating function of the distribution of the sojourn time and the number of customers left behind. Van den Berg and Boxma [32] 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 [22] gave a decomposition theorem for the sojourn time distribution for the M/G/1 with K classes of customers and *discriminatory* processor sharing. Grishechkin [12,13] 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 overview of the literature on processor sharing queueing models we refer to Yashkov's survey papers [38,39], and references therein.

In the present study we analyse the sojourn times of customers in the M/M/1 processor sharing queue with a server which 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 may be relaxed for some parts of our analysis. For instance, the decomposition result in section 3 may be obtained similarly for generally distributed service requirements. Also, the results of sections 4 and 5 may be generalised for that case by use of the Laplace transform method for solving differential equations. Nevertheless, the results for the exponential services are presented for two reasons. Firstly, the fundamental ideas are the same as for general service requirements, but the presentation is more transparent. Secondly, under the exponentiality assumption for service requirements we get more explicit results, and at some points we are able to carry the analysis further. This may be important for future attempts at analysing models with a more general process for variations in the server availability, for instance a server which alternates between two *positive* service speeds. The latter case does not lend itself (yet) for a similar detailed mathematical analysis, see for instance Núñez-Queija [20].

Queueing models with a *First Come First Served* (FCFS) discipline and servers that are subject to breakdowns have received much attention in the literature. The first ones to consider queueing models with interruptions (and their connection with priority models) were White and Christie [34]. Gaver [11] obtained the steady-state queue length distribution of the $M^X/G/1$ queue with exponentially distributed on-times and general off-times. We further mention the early work of Mitrani and Avi-Itzhak [17] on a queueing model with multiple servers which are subject to breakdowns, and the work of Neuts [19, chapter 6] concerned with queues in a random environment. Bounds and approximations for queue lengths and sojourn times when the on-times have a general distribution as well are studied by Federgruen and Green [6,7] and Sengupta [26]. Recent publications on queues with server breakdowns are, for instance, Takine and Sengupta [29], Li et al. [16], and Lee [15]. For an extensive overview of the literature on queueing models with service interruptions we refer to Federgruen and Green [6,7]. More recent references can be found in Takine and Sengupta [29]. 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 LSTs of their distributions and then solving these in terms of a single integral equation. We derive the first two moments of the two fundamental random variables of section 5, and give the general form of higher moments. In section 6 we use these results to obtain an explicit expression for the first

two moments of the sojourn time of a customer conditioned on his service requirement, the state of the server upon arrival and the number of other customers in the system upon arrival. In particular we extend a result of Sengupta and Jagerman [28] to our model with server breakdowns, proving that the kth moment of the conditional sojourn time is a polynomial of degree k in the number of customers upon arrival. In section 7 we give the LST of the distribution of the sojourn time distribution of a customer conditioned only on his own work requirement, assuming that he arrives to the system in steady state. 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 next two sections are then devoted to an asymptotic analysis of the model. In section 8 we study sojourn times of customers with large service requirements (tending to infinity), and in section 9 we consider the heavy-traffic case. We conclude in section 10 with some final remarks.

2. Model description

We consider a server which 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 $F(t) := \mathbf{P}\{T_{\text{off}} \leqslant t\}$, $t \geqslant 0$. The LST of this distribution will be denoted by

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

and the kth 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 of them 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 $(1/n)\mu\Delta t + \mathrm{o}(\Delta t)$. During off-periods the service of all customers is interrupted until the server becomes active again.

We define the random variable $X(t)$ to be the number of customers at the server at time $t \geqslant 0$. The random variable $Y(t)$ is equal to 1 if at time $t \geqslant 0$ 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 nontrivial 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.

Below we 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 $|z| \leqslant 1$,

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

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

For later use, we give the means of these conditional distributions:

$$\mathbf{E}[X \mid Y = 1] = \frac{\lambda(1 + \nu m_1) + \nu \lambda^2 m_2/2}{\mu - \lambda(1 + \nu m_1)}, \tag{2.4}$$

$$\mathbf{E}[X \mid Y = 0] = \lambda \frac{m_2}{2m_1} + \frac{\lambda(1 + \nu m_1) + \nu \lambda^2 m_2/2}{\mu - \lambda(1 + \nu m_1)}. \tag{2.5}$$

By averaging expressions (2.2) and (2.3) over $\mathbf{P}\{Y = 0\}$ and $\mathbf{P}\{Y = 1\}$ we find the probability generating function (p.g.f.) of the marginal distribution of X:

$$\mathbf{E}\left[z^X\right] = \frac{1}{1 + \nu m_1}\left(1 + \nu \frac{1 - \phi(\lambda(1 - z))}{\lambda(1 - z)}\right) \times \frac{\mu - \lambda(1 + \nu m_1)}{\mu - \lambda z - \nu z(1 - \phi(\lambda(1 - z)))/(1 - z)}. \tag{2.6}$$

In the remainder of this section we give an informal discussion of the derivation of expressions (2.2) and (2.3). In particular, in remark 2.1 we discuss the equivalence of the queue length process in our model with the queue length process of two queueing models with the FCFS queue discipline. Expression (2.2) can be found by considering the queue length process only during on-periods. For this, we "delete" all off-periods and interpret the arrivals during an off-period as a batch arrival. In the resulting transformed model there are three types of events:

(i) departures of customers at rate μ when there is at least one customer present,

(ii) single arrivals according to a Poisson process with rate λ, and

(iii) batch arrivals according to a Poisson process with rate ν and batch sizes having 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 are "empty" with probability $\phi(\lambda)$. To avoid this, we may consider only non-empty batches which arrive with rate $\nu(1-\phi(\lambda))$, having p.g.f. $(\phi(\lambda(1-z))-\phi(\lambda))/(1-\phi(\lambda))$. The balance equations for this transformed model readily lead to equation (2.2).

The factor $(1-\phi(\lambda(1-z)))/(m_1\lambda(1-z))$ in equation (2.3) is the p.g.f. of the number of customers that arrive during the backward recurrence time of an off-period. This can be explained as follows. At an arbitrary time instant at which the server is off, the number of customers in the system is the sum of the customers that were at the server when the server turned off and the number of customers that have arrived since that time. The elapsed time since the server turned off is distributed as the backward recurrence time of an off-period. Moreover, because of the exponentially distributed on-periods, we may use the *Poisson Arrivals See Time Averages* (PASTA) property – see Wolff [35] – to show that the number of customers present when the server turns off has the same distribution as X *given* that $Y=1$.

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 (2.6) can, therefore, be obtained from Gaver [11, 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 a general service time distribution. The analysis is based on *completion times* of customers, see Gaver [11, section 4.2]. In our case the distribution of the completion times has LST

$$\beta(s)=\frac{\mu}{\mu+s+\nu(1-\phi(s))},\quad \mathrm{Re}(s)\geqslant 0. \tag{2.7}$$

These "enlarged" 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. It can be shown that the first customer in a busy period has to wait before his service begins (this corresponds to the server being in the off-state in the original model with breakdowns) with probability $p=(\nu(1-\phi(\lambda)))/(\lambda+\nu(1-\phi(\lambda)))$, in which case the distribution of the residual off-period has LST

$$\delta(s)=\frac{\lambda}{1-\phi(\lambda)}\cdot\frac{\phi(s)-\phi(\lambda)}{\lambda-s},\quad \mathrm{Re}(s)\geqslant 0.$$

Expression (2.6) can also be verified using the LST of the queue-length distribution in an M/G/1 queue with exceptional first service, see Welch [33, theorem 2]. In that queue the distribution of the regular services has LST $\beta(s)$ and that of the exceptional first services has LST $(1-p+p\delta(s))\beta(s)$.

Remark 2.2. If the breakdowns (off-periods) are exponentially distributed too, i.e., $\phi(s)=1/(1+m_1s)$, the probabilities $\mathbf{P}\{X=i,\ Y=0\}$ and $\mathbf{P}\{X=i,\ Y=1\}$, $i=0,1,\ldots$, are explicitly given by Neuts [19, theorem 6.3.1].

Remark 2.3. Note that our model does not fit into the framework of Fuhrmann and Cooper [10]. Even if we assume a FCFS queue discipline – as in remark 2.1 – their assumption 4 is not satisfied. Nevertheless, it can be seen from expression (2.6) that for the queue-length distribution a similar decomposition as in [10] holds.

3. A branching process representation

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*. The first to use this time-transformation method for the analysis of processor sharing queues apparently was Yashkov [37]. A closely related method, but without transformation of time, was already used in Yashkov [36]. Foley and Klutke [9] use the transformed process to study the case where the total service capacity may depend on the number of customers in the system. Grishechkin [12,13] used the time-transformation method to study queues with a general class of service disciplines – including processor sharing – by means of Crump-Mode–Jagers branching processes. For more references on the time-transformation method and its use in the analysis of processor sharing queues we refer to Yashkov [39, section 2.4].

We present a direct use of the time-transformation technique to analyse sojourn times in the processor sharing queue with service interruptions presented in section 2. However, the same approach is applicable to more general models, for instance those in Grishechkin [12,13]. In remark 3.7 we illustrate how the analysis in this section may be extended to the case with generally distributed service requirements. Restricting ourself to the model of section 2 makes the presentation more transparent, while the fundamental ideas are the same as in the more general cases. Furthermore, we are able to carry the analysis further, and in particular in sections 6 and 7 we obtain more explicit results.

In our presentation we first assume there is a permanent customer which never leaves the system. For this customer we study the accumulation of received service. All other ("nonpermanent") 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 $t \geqslant 0$, *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 $[0,t]$:

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

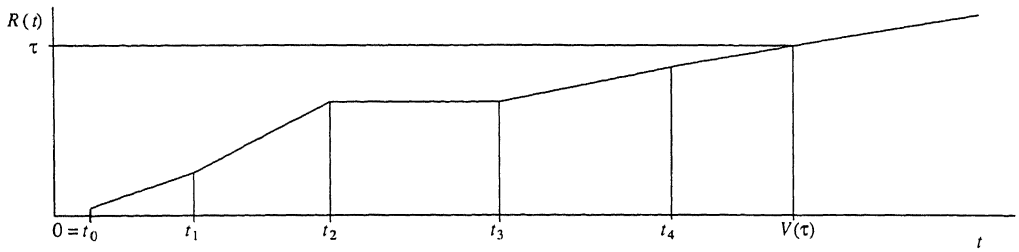

Figure 1. $R(t)$ and $V(\tau)$.

We define, for $\tau \geqslant 0$,

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

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.

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 $\tau > 0$.

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). For $i \in \{0, 1\}$, we denote by $V_i(\tau)$ the process $V(\tau)$ conditional on $Y(0) = i$, or equivalently: $V_i(\tau) := V(\tau) \mid \{Y(0) = i\}$. Similarly we define $Z_i(t) := Z(t) \mid \{Y(0) = i\}$ and $Y_i(t) := Y(t) \mid \{Y(0) = i\}$. We, first, concentrate on $V_1(\tau)$, the conditional sojourn time of a customer which arrives when the server is working, and at the end of this section derive the results for $V_0(\tau)$, the conditional sojourn time of a customer which arrives during an off-period.

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 $x(y)$ for which these limits exist.

Observation 3.1. For arbitrary $n \in \mathbb{N}$, *given* $Z_1(0) = n$ (and $Y(0) = 1$), it holds with probability 1 that $V_1(0+) = 0$. This follows immediately from the fact that, for small τ, $V_1(\tau)$ is equal to $(n+1)\tau$ with probability

$$1 - \left(\lambda + \frac{n}{n+1}\mu + \nu\right)(n+1)\tau + o(\tau).$$

Note that this is *not true* for $V_0(\tau)$, because in that case the server must first become active again.

Denote the number of times that the server turned off during the period $(0, t)$ by the random variable $N(t)$, and the length of the ith off-period started *after* time 0 by D_i, $i \in \{1, 2, \ldots\}$. Note that $\{D_1, D_2, \ldots\}$ is an i.i.d. sequence with distribution $F(t)$. Further, for $\tau > 0$, define

$$N'(\tau) := N\big(V_1(\tau)\big).$$

The random variable $N'(\tau)$ is well defined because $V_1(\tau)$ – also a random variable – is strictly increasing in τ (with probability 1). 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 for $\tau > 0$ the processes $Z_1'(\tau) := Z_1(V_1(\tau))$ and $Y_1'(\tau) := Y_1(V_1(\tau)+)$.

Lemma 3.2. With probability 1, the process $V_1(\tau)$ is related to the processes $Z_1'(\tau)$, $N'(\tau)$, and D_i, through the equation

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

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

Proof. Consider any realisation of the arrival process, the sequence of required services, and the availability of the server. In figure 1 a particular realisation is depicted. We observe 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 $V_1(\tau+) - V_1(\tau) = D_{N'(\tau+)}$. □

With the aid of figure 1, we make the following observation:

Observation 3.3. 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:

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

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 equation (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:

$$\mathbf{E}\big[\mathrm{e}^{-sD_{N'(\tau)}} z^{Z_1'(\tau)-Z_1'(\tau-)} \mid N'(\tau) - N'(\tau-) = 1\big] = \phi\big(s + \lambda(1-z)\big).$$

To study the distribution of $V_1(\tau)$, we construct a branching process that is equivalent to $(Z_1'(\tau), N'(\tau); D_1, \ldots, D_{N'(\tau)})$, and impose a reward structure on this branching process that will turn out to be useful. Consider a population $\mathcal{P}$ of elements which 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 receives a reward at rate 1 (per time unit). An element generates children in two ways, independent of 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), and receives an immediate reward which depends on the number of children in the nest in a stochastic way. The simultaneous distribution of A children in the nest and the immediate reward D is given by

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

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

Observation 3.4. Denote the number of nonpermanent 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 reward of the ith nest by D_i''. By comparing the transition rates of both processes it is seen that the processes $(Z_1'(\tau), N'(\tau); D_1, \ldots, D_{N'(\tau)})$ and $(Z_1''(\tau), N''(\tau); D_1'', \ldots, D_{N''(\tau)}'')$ are equivalent. Also, $V_1(\tau)$ is distributed as the reward of the population from time 0 until time τ.

In the next theorem we formulate the main result of this section. For this, we need to introduce the random variables $C_i(\tau)$, $i \in \{0, 1, 2, \ldots\}$. $C_0(\tau)$ is the reward for the permanent element and his offspring between time instants 0 and τ. Similarly, $C_i(\tau)$, $i = 1, 2, \ldots, Z_1''(0)$, is the reward for the ith nonpermanent individual, who was present at time 0, plus the reward for his offspring between time instants 0 and τ. Note that all $C_i(\tau)$, $i \geqslant 1$, have the same distribution.

The decomposition of sojourn times given in the theorem was established by Yashkov [36, expression (3.4)] for the ordinary M/G/1 processor sharing queue, and

by Rege and Sengupta [22, theorem 6] for the M/G/1 queue with discriminatory processor sharing.

Theorem 3.5. The conditional sojourn time $V_1(\tau)$ of a customer who finds the server working upon arrival can be decomposed as

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

where $\stackrel{\mathrm{d}}{=}$ means equality in distribution. All random variables involved in the right-hand side are mutually independent. In particular, $C_0(\tau)$ is distributed as $V_1(\tau)$ conditional on $Z_1(0) = 0$.

Proof. Using the reward-interpretation of $V_1(\tau)$ given in observation 3.4, we can split $V_1(\tau)$ into the individual rewards of all elements. By construction, the elements of the population $\mathcal{P}$ behave independently of each other. Therefore, the $C_i(\tau)$ – including $C_0(\tau)$ – form an independent sequence. Finally, by definition, $Z_1'(0) = Z_1(0)$, which concludes the proof. □

In section 4, we characterise the LSTs of the distributions of $C_0(\tau)$ and $C_1(\tau)$ by a set of differential equations, which we solve in terms of an integral equation.

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 the distribution of D_0. By conditioning on the length of D_0 and the number of arrivals A_0:

$$V_0(\tau) \mid \{Z_0(0) = n,\ D_0 = d_0,\ A_0 = k\} \stackrel{\mathrm{d}}{=} d_0 + V_1(\tau) \mid \{Z_1(0) = n + k\}. \tag{3.2}$$

Corollary 3.6. $V_0(\tau)$, the conditional sojourn time of a customer who finds the server in the off-state upon arrival, can be written as

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

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

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

Proof. The corollary follows from theorem 3.5 and relation (3.2). □

We define the LSTs of the distributions of $C_0(\tau)$ and $C_i(\tau)$, $i \in \{1, 2, \ldots\}$, by $g_0(\tau; s)$ and $g_1(\tau; s)$: for $\mathrm{Re}(s) \geqslant 0$,

$$g_0(\tau; s) := \mathbf{E}\big[\mathrm{e}^{-sC_0(\tau)}\big], \qquad g_1(\tau; s) := \mathbf{E}\big[\mathrm{e}^{-sC_i(\tau)}\big], \quad i = 1, 2, \ldots.$$

From theorem 3.5 and corollary 3.6 we have, for $\mathrm{Re}(s) \geqslant 0$,

$$\mathbf{E}\big[\mathrm{e}^{-sV(\tau)} \mid Y(0) = 1,\ Z(0) = n\big] = g_0(\tau; s)\big\{g_1(\tau; s)\big\}^n, \tag{3.3}$$

$$\mathbf{E}\big[\mathrm{e}^{-sV(\tau)} \mid Y(0) = 0,\ Z(0) = n\big] = g_0(\tau; s)\big\{g_1(\tau; s)\big\}^n \phi_0\big(s + \lambda\big(1 - g_1(\tau; s)\big)\big). \tag{3.4}$$

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 the distribution 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 3.7. The generalisation of this representation by branching processes for general service time distributions, $B(x)$, $x \geqslant 0$, can be obtained by using the method of supplementary variables. We extend the state space representation with the vector $(x_1, x_2, \ldots, x_n)$ when there are n customers in the system. We again assume that a newly arrived customer with service requirement 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 conditional sojourn time of the new customer $V_1(\tau; n; x_1, \ldots, x_n)$ then

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

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 lifetime distribution $B(x)$. Thus, $C_i(\tau; x_i)$ is the reward for a family until time τ, starting with one individual with a remaining lifetime x_i. We omit the details of this generalisation and refer to Yashkov [36] for a related analysis of the case without service interruptions.

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

We derive a set of differential equations which uniquely determine $g_0(\tau; s)$ and $g_1(\tau; s)$, the LSTs of the distributions of $C_0(\tau)$ and $C_1(\tau)$. We then express $g_0(\tau; s)$ in terms of $g_1(\tau; s)$, and – for real $s > 0$ – derive a useful integral equation for $g_1(\tau; s)$.

Lemma 4.1. For $\mathrm{Re}(s) \geqslant 0$ and $\tau \geqslant 0$, $g_0(\tau; s)$ and $g_1(\tau; s)$ are uniquely determined by the following set of differential equations,

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

$$\begin{aligned}\frac{\partial}{\partial \tau} g_0(\tau; s) = {} & -(s + \lambda + \nu) g_0(\tau; s) + \lambda g_0(\tau; s) g_1(\tau; s) \\ & + \nu g_0(\tau; s)\phi\big(s + \lambda\big(1 - g_1(\tau; s)\big)\big),\end{aligned} \tag{4.2}$$

and initial conditions,

$$g_0(0;s) = g_1(0;s) = 1. \tag{4.3}$$

Proof. See the appendix. □

Theorem 4.2. $g_0(\tau;s)$ can be expressed in terms of $g_1(\tau;s)$ as

$$g_0(\tau;s) = g_1(\tau;s)\exp\left\{\mu\left(\tau - \int_{u=0}^{\tau} g_1(u;s)^{-1}\,\mathrm{d}u\right)\right\}. \tag{4.4}$$

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

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

If we also use (4.1) we may rewrite this as

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

which leads to relation (4.4). □

The remainder of this section is devoted to finding the solution of (4.1) for real $s > 0$. We first define the *clearing period* of the model of section 2 as 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. We generically denote the clearing period by the random variable CP and the LST of its distribution by $r_1(s) = \mathbf{E}[\mathrm{e}^{-sCP}]$.

Lemma 4.3. The clearing period has the same distribution as the busy period of an ordinary M/G/1 queue with arrival rate λ and LST of the service time distribution $\beta(\cdot)$ given by expression (2.7).

As a consequence, for $\mathrm{Re}(s) \geqslant 0$, $x = r_1(s)$ is the unique root – inside (or on) the unit circle in the complex plane – of the equation

$$(s+\lambda+\mu+\nu)x = \lambda x^2 + \mu + \nu x\phi\big(s+\lambda(1-x)\big). \tag{4.6}$$

Proof. Note that for the model with the FCFS queue discipline – described in remark 2.1 – we may define the clearing period as we did above for the model of section 2. Moreover, the clearing periods of both models have the same distribution. It is easily seen that the clearing period of the model in remark 2.1 has the same distribution as the busy period of an ordinary M/G/1 queue with arrival rate λ and LST of the service time distribution $\beta(\cdot)$. This proves the first statement of the lemma.

Furthermore, we immediately have that for $\mathrm{Re}(s) \geqslant 0$, $r_1(s)$ is equal to the (unique) root inside (or on) the unit circle of the equation,

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

see, for instance, Cohen [4, p. 250]. This equation readily leads to (4.6). □

Observation 4.4. Let $s > 0$ be fixed. $C_1(\tau)$ is nondecreasing in τ with probability 1, and so $g_1(\tau; s)$ is non-increasing in τ. Therefore, the right-hand side of (4.1) is negative for $\tau \geqslant 0$. Indeed, for $\tau = 0$ this is easily verified because $g_1(0; s) = 1$. Now it must be that, for $\tau \geqslant 0$,

$$r_1(s) \leqslant g_1(\tau; s).$$

Otherwise, the right-hand side of (4.1) would be positive for some $\tau > 0$, because the zero $r_1(s)$ is of multiplicity 1.

Theorem 4.5. For real $s > 0$, the solution to (4.1) satisfying (4.3) is obtained from

$$\int_{x=1}^{g_1(\tau;s)} \frac{1}{\mu - (s + \lambda + \mu + \nu)x + \lambda x^2 + \nu x \phi(s + \lambda - \lambda x)} \, \mathrm{d}x = \tau. \tag{4.7}$$

Proof. The integral in relation (4.7) is well defined, because the denominator of the integrand has no zeros in $(r_1(s), 1)$ for $s > 0$, see lemma 4.3. 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. □

In section 8 we use relation (4.7) 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 $C_0(\tau)/\tau$ and (more importantly) $V(\tau)/\tau$ for $\tau \to \infty$.

Relation (4.7) is not very practical for determining moments of $C_1(\tau)$ (and $C_0(\tau)$). In section 5 we study these moments directly.

Computational issues

Although our primary focus is on analytical derivations, it is worth pointing out how our results may potentially be used for numerical calculations. We show how the distribution of $C_1(\tau)$ can be computed in the case that the distribution of the off-periods has a rational LST. We also discuss some difficulties regarding the computation of the distribution of $C_0(\tau)$.

Suppose for the moment that – for *real* (and positive) values of s – $g_1(\tau; s)$ can be evaluated from relation (4.7). Then we use the Gaver–Stehfest algorithm, see Abate and Whitt [1, pp. 52–55], to compute the distribution function of $C_1(\tau)$. To evaluate the nth Gaver–Stehfest approximant, one typically needs $2n$-digit precision in the calculations. In general, taking $n = 15$ gives good results – relative errors are

typically less than 3% for tail probabilities of the order 10^{-3} – and comparison with the results using $n = 20$ provides a useful accuracy check.

To illustrate how $g_1(\tau; s)$ can be evaluated, let us first consider the case that the off-periods have a hyperexponential distribution. In that case the LST of the distribution of the off-periods is of the form

$$\phi(s) = \sum_{i=1}^{k} \frac{w^{(i)}}{1 + m_1^{(i)} s},$$

with $w^{(i)} > 0$, $\sum_{i=1}^{k} w^{(i)} = 1$, $m_1^{(1)} > m_1^{(2)} > \cdots > m_1^{(k)} > 0$, and $\mathrm{Re}(s) > -1/m_1^{(1)}$. Note that $m_1 = \sum_{i=1}^{k} w^{(i)} m_1^{(i)}$. After multiplying the numerator and the denominator of the integrand in relation (4.7) by

$$\prod_{i=1}^{k} \{1 + m_1^{(i)}(s + \lambda(1 - x))\},$$

it becomes a rational function in x with the degree of the denominator equal to $k+2$, and that of the numerator equal to k. It can be seen that the denominator is positive for $x = 0$ and for $x = (s + \lambda + 1/m_1^{(i)})/\lambda$ when i is odd, and the denominator is negative for $x = 1$ and for $x = (s + \lambda + 1/m_1^{(i)})/\lambda$ when i is even. Moreover, if $x \to \infty$ then the denominator tends to $+\infty$ when k is even, and to $-\infty$ when k is odd. Therefore, for $s > 0$ and $i = 1, 2, \ldots, k+2$, the roots $r_i(s)$ of the denominator satisfy

$$0 < r_1(s) < 1 < r_2(s) < \frac{s + \lambda + 1/m_1^{(1)}}{\lambda} < r_3(s) < \cdots < \frac{s + \lambda + 1/m_1^{(k)}}{\lambda} < r_{k+2}(s).$$

This relation enables an efficient computation of the roots, for instance using the Newton–Raphson method (combined with the bisection method) on each of the above intervals containing exactly one root. By partial fraction expansion, relation (4.7) can now be written as

$$\tau = \int_{x=1}^{g_1(\tau;s)} \sum_{i=1}^{k+2} \frac{a_i(s)}{r_i(s) - x}\, \mathrm{d}x = -\sum_{i=1}^{k+2} a_i(s) \ln\left(\frac{r_i(s) - g_1(\tau; s)}{r_i(s) - 1}\right). \tag{4.8}$$

The functions $a_i(s)$ are given by

$$a_i(s) = \frac{\prod_{j=1}^{k}(1 + m_1^{(j)}\{s + \lambda(1 - r_i(s))\})}{\lambda^{k+1} \prod_{j=1}^{k} m_1^{(j)} \prod_{j \neq i}(r_j(s) - r_i(s))}.$$

Note that, for $s > 0$, $r_1(s) < g_1(s; \tau) < 1$ and $a_1(s) > 0$, whereas $r_i(s) > 1$ and $a_i(s) < 0$, $i \in \{2, 3, \ldots, k+2\}$. After computing the roots $r_i(s)$ and the coefficients $a_i(s)$, $g_1(\tau; s)$ can be found from expression (4.8), again using the Newton–Raphson method.

We tested the above procedure to compute the distribution function of $C_1(\tau)$ for the case of no service interruptions, and for the case of exponentially distributed off-periods. In the first case an explicit expression for $g_1(\tau; s)$ can be found in Coffman et al. [2, equation (16)]. Using this expression, the Euler algorithm – see Abate and Whitt [1, section 7] – gives a reliable alternative to compare the results. In general, the outcomes of both methods agreed up to a relative difference of at most 3% for tail probabilities of the order 10^{-3}. In the case of exponentially distributed off-periods we compared our results to those generated by simulation, and again found that the relative differences were at most 3%.

We saw above that for hyperexponential off-periods the roots $r_i(s)$ are all real and positive, and we found disjoint intervals on the positive real line, each containing exactly one root. When the distribution of the off-periods has a rational LST, but is not a hyperexponential distribution, the analysis proceeds along the same lines. However, in general some of the roots may be complex. This is, for instance, the case when the off-periods have an Erlang distribution.

Serious complications arise when the distribution of the off-periods does not have a rational LST. In principle, the left-hand side of relation (4.7) can be computed using, for instance, Simpson's rule (or a higher order Newton–Cotes method) for numerical integration. However, like any other inversion method, the Gaver–Stehfest algorithm is highly sensitive to small errors in the computation of the LST that is to be inverted. Therefore, computation of the integral in relation (4.7) requires exceedingly long computation times due to the usual accuracy problems with numerical integration.

The same difficulties are encountered in the computation of $g_0(\tau; s)$ using equation (4.4). Even if $g_1(\tau; s)$ has been computed accurately, for instance using the above procedure for the case that the distribution of the off-periods has a rational LST, evaluating the right-hand side of equation (4.4) requires an additional numerical integration leading to prohibitively long computation times (poor results were obtained even after 2 hours on a Sun Sparc 4 station).

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

In section 4 we saw that $g_0(\tau; s)$ and $g_1(\tau; s)$, the LSTs of the distributions 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 (4.4) and (4.7). However, this solution is not very practical for determining 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 [36] also remarks that, in the M/G/1 processor sharing queue, such an approach leads to a more tractable derivation of moments. First, we state the following theorem which is a consequence of a result of De Meyer and Teugels [5, lemma 3].

Theorem 5.1. If the kth moment of the off-periods, m_k, exists, then also the kth moments of $C_1(\tau)$ and $C_0(\tau)$ exist.

Proof. See the appendix. □

We start by illustrating the derivation of the first and second moments of $C_1(\tau)$ and $C_0(\tau)$. We then formulate and prove theorem 5.2 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

$$\frac{\partial}{\partial\tau}\mathbf{E}\big[C_1(\tau)\big] = 1 + \nu m_1 - \big\{\mu - \lambda(1+\nu m_1)\big\}\mathbf{E}\big[C_1(\tau)\big], \tag{5.1}$$

$$\frac{\partial}{\partial\tau}\mathbf{E}\big[C_0(\tau)\big] = 1 + \nu m_1 + \lambda(1+\nu m_1)\mathbf{E}\big[C_1(\tau)\big]. \tag{5.2}$$

Formally, it should first be verified that interchanging the order of differentiation is allowed. However, in our case, we can also get (5.1) and (5.2) by directly applying the argument of conditioning on the events in a time interval of length Δ to $\mathbf{E}[C_0(\tau)]$ and $\mathbf{E}[C_1(\tau)]$, and then letting $\Delta \downarrow 0$. Using the initial conditions $C_0(0) = C_1(0) = 0$, we find

$$\mathbf{E}\big[C_1(\tau)\big] = \frac{1+\nu m_1}{\mu - \lambda(1+\nu m_1)}\big(1 - \mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\big), \tag{5.3}$$

$$\mathbf{E}\big[C_0(\tau)\big] = \mu\frac{1+\nu m_1}{\mu - \lambda(1+\nu m_1)}\tau - \lambda\left(\frac{1+\nu m_1}{\mu - \lambda(1+\nu m_1)}\right)^2\big(1 - \mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\big). \tag{5.4}$$

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

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\big[C_1(\tau)^2\big] &= -\big\{\mu - \lambda(1+\nu m_1)\big\}\mathbf{E}\big[C_1(\tau)^2\big] + 2(1+\nu m_1)\mathbf{E}\big[C_1(\tau)\big] \\ &\quad + 2\lambda(1+\nu m_1)\mathbf{E}\big[C_1(\tau)\big]^2 + \nu m_2\big\{1 + \lambda\mathbf{E}\big[C_1(\tau)\big]\big\}^2,\end{aligned} \tag{5.5}$$

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\big[C_0(\tau)^2\big] &= 2(1+\nu m_1)\mathbf{E}\big[C_0(\tau)\big] + 2\lambda(1+\nu m_1)\mathbf{E}\big[C_0(\tau)\big]\mathbf{E}\big[C_1(\tau)\big] \\ &\quad + \lambda(1+\nu m_1)\mathbf{E}\big[C_1(\tau)^2\big] + \nu m_2\big\{1 + \lambda\mathbf{E}\big[C_1(\tau)\big]\big\}^2.\end{aligned} \tag{5.6}$$

We can solve this using (5.3) and (5.4):

$$\begin{aligned}\mathbf{E}\big[C_1(\tau)^2\big] &= -(a_1 + 2a_2)\tau\,\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau} \\ &\quad + \frac{a_1 + \nu m_2}{\mu - \lambda(1+\nu m_1)}\big(1 - \mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\big) \\ &\quad + \frac{a_2}{\mu - \lambda(1+\nu m_1)}\big(1 - \mathrm{e}^{-2\{\mu-\lambda(1+\nu m_1)\}\tau}\big),\end{aligned} \tag{5.7}$$

$$\begin{aligned}\mathbf{E}\big[C_0(\tau)^2\big] &= b_1\tau + b_2\tau^2 + b_3\tau\,\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau} \\ &\quad - b_4\big(1 - \mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\big) - b_5\big(1 - \mathrm{e}^{-2\{\mu-\lambda(1+\nu m_1)\}\tau}\big),\end{aligned} \tag{5.8}$$

where

$$a_1 = 2(1 + \nu m_1 + \nu m_2 \lambda) \frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)},$$

$$a_2 = \lambda\big(2(1 + \nu m_1) + \nu m_2 \lambda\big) \left(\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} \right)^2,$$

$$b_1 = \nu m_2 \left(\frac{\mu}{\mu - \lambda(1 + \nu m_1)} \right)^3,$$

$$b_2 = \mu^2 \left(\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} \right)^2,$$

$$b_3 = 2\lambda \left(\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} \right)^3 \left(2\mu + \lambda(1 + \nu m_1) + \lambda \nu m_2 \frac{\mu}{1 + \nu m_1} \right),$$

$$b_4 = 2\lambda \left(\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} \right)^4 \left(\frac{2\mu - \lambda(1 + \nu m_1)}{1 + \nu m_1} + \frac{1}{2} \nu m_2 \frac{3\mu^2 - \lambda^2(1 + \nu m_1)^2}{(1 + \nu m_1)^3} \right),$$

$$b_5 = \lambda^2 \left(\frac{1 + \nu m_1}{\mu - \lambda(1 + \nu m_1)} \right)^4 \left(2 + \frac{1}{2} \nu m_2 \frac{2\lambda(1 + \nu m_1) - \mu}{(1 + \nu m_1)^2} \right).$$

The same approach can be applied to determine higher moments. In theorem 5.2 this is done to reveal the structure of these moments.

Theorem 5.2. For $k \geqslant 1$, provided that $m_k < \infty$, and thus $\mathbf{E}[C_1(\tau)^k] < \infty$ and $\mathbf{E}[C_0(\tau)^k] < \infty$,

$$\mathbf{E}\big[C_1(\tau)^k\big] = \alpha_0^{(k)} + \sum_{m=1}^{k} \mathrm{e}^{-m\{\mu - \lambda(1+\nu m_1)\}\tau} \sum_{n=0}^{k-m} \alpha_{m,n}^{(k)} \tau^n, \tag{5.9}$$

$$\mathbf{E}\big[C_0(\tau)^k\big] = \sum_{m=0}^{k} \mathrm{e}^{-m\{\mu - \lambda(1+\nu m_1)\}\tau} \sum_{n=0}^{k-m} \beta_{m,n}^{(k)} \tau^n, \tag{5.10}$$

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

Proof. See the appendix. □

6. 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 5, we find explicit expressions for the first two moments of the conditional sojourn time.

For compactness, we use the following notation:

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

Observation 6.1. From theorem 3.5, we have, for $k, n \in \mathbb{N}$,

$$\begin{aligned}\mathbf{E}_n\left[V_1(\tau)^k\right] &= \mathbf{E}\left[\left(C_0(\tau)+\cdots+C_n(\tau)\right)^k\right]\\ &= \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[C_n(\tau)^{k-j}\right]\mathbf{E}\left[\left(C_0(\tau)+\cdots+C_{n-1}(\tau)\right)^j\right]\\ &= \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[C_1(\tau)^{k-j}\right]\mathbf{E}_{n-1}\left[V_1(\tau)^j\right], \qquad (6.1)\\ \mathbf{E}_0\left[V_1(\tau)^k\right] &= \mathbf{E}\left[C_0(\tau)^k\right]. \qquad (6.2)\end{aligned}$$

Moreover, combining theorem 3.5 and corollary 3.6, we find

$$\mathbf{E}_n\left[V_0(\tau)^k\right] = \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[\left(D_0+\sum_{i=1}^{A_0}C_i(\tau)\right)^j\right]\mathbf{E}_n\left[V_1(\tau)^{k-j}\right]. \qquad (6.3)$$

We remind the reader that D_0 is the residual off-period at time zero, with LST $\phi_0(\cdot)$, and that A_0 is the number of arrivals during D_0. We find

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

These derivatives can be found by using [5, lemma 1] to expand $\phi_0(s+\lambda-\lambda g_1(\tau;s))$ in a Taylor series, analogous to equation (A.7) below.

From (6.1) and (6.2) we can compute the conditional moments $\mathbf{E}_n[V_1(\tau)^k]$ recursively, once we have the moments of $C_0(\tau)$ and $C_1(\tau)$. The moments of $V_0(\tau)$ are then found from (6.3). In particular, we have for $k=1$, see also equations (3.3) and (3.4),

$$\mathbf{E}_n\left[V_1(\tau)\right] = \mathbf{E}\left[C_0(\tau)\right] + n\mathbf{E}\left[C_1(\tau)\right], \qquad (6.4)$$
$$\mathbf{E}_n\left[V_0(\tau)\right] = \mathbf{E}\left[D_0\right] + \mathbf{E}\left[C_0(\tau)\right] + \left(n+\lambda\mathbf{E}[D_0]\right)\mathbf{E}\left[C_1(\tau)\right], \qquad (6.5)$$

and for $k=2$,

$$\begin{aligned}\mathbf{E}_n\left[V_1(\tau)^2\right] &= \mathbf{E}\left[C_0(\tau)^2\right] + n\mathbf{E}\left[C_1(\tau)^2\right] + 2n\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right]\\ &\quad + n(n-1)\mathbf{E}\left[C_1(\tau)\right]^2, \qquad (6.6)\\ \mathbf{E}_n\left[V_0(\tau)^2\right] &= \mathbf{E}\left[D_0^2\right] + 2\mathbf{E}\left[D_0\right]\left(\mathbf{E}\left[C_0(\tau)\right] + n\mathbf{E}\left[C_1(\tau)\right]\right)\\ &\quad + 2\lambda\mathbf{E}\left[D_0^2\right]\mathbf{E}\left[C_1(\tau)\right] + \mathbf{E}\left[C_0(\tau)^2\right]\\ &\quad + 2\left(n+\lambda\mathbf{E}[D_0]\right)\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right]\end{aligned}$$

$$+ \big(n + \lambda \mathbf{E}[D_0]\big)\mathbf{E}\big[C_1(\tau)^2\big]$$
$$+ \big(n(n-1) + (2n-1)\lambda\mathbf{E}[D_0] + \lambda^2\mathbf{E}\big[D_0^2\big]\big)\mathbf{E}\big[C_1(\tau)\big]^2. \quad (6.7)$$

Using (5.3), (5.4), (5.7) and (5.8) we have explicit formulas for these first and second moments.

Theorem 6.2. For fixed $k \in \mathbb{N}$, if $m_k < \infty$ and $\mathbf{E}[D_0^k] < \infty$, then $\mathbf{E}_n[V_1(\tau)^k]$ and $\mathbf{E}_n[V_0(\tau)^k]$ are polynomials in n of degree k:

$$\mathbf{E}_n\big[V_i(\tau)^k\big] = \sum_{l=0}^{k} c_{k,l}^{(i)}(\tau) n^l, \quad i \in \{0,1\}. \quad (6.8)$$

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

$$c_{k,l+1}^{(1)}(\tau) = \frac{1}{l+1}\left\{\sum_{i=l+2}^{k}(-1)^{i-l}\binom{i}{l}c_{k,i}^{(1)}(\tau) + \sum_{j=l}^{k-1}\sum_{i=l}^{j}(-1)^{i-l}\binom{i}{l}\binom{k}{j}\mathbf{E}\big[C_1(\tau)^{k-j}\big]c_{j,i}^{(1)}(\tau)\right\}, \quad (6.9)$$
$$c_{k,0}^{(1)}(\tau) = \mathbf{E}\big[C_0(\tau)^k\big],$$

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

$$c_{k,l}^{(0)}(\tau) = \sum_{j=l}^{k}\binom{k}{j}c_{j,l}^{(1)}(\tau)\mathbf{E}\left[\left(D_0 + \sum_{i=1}^{A_0} C_i(\tau)\right)^{k-j}\right]. \quad (6.10)$$

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 $\mathbf{E}[C_0(\tau)^k]$ in theorem 5.2.

Proof. Expression (6.8), for $i = 1$, can be proved by arguing that (6.1) and (6.2) uniquely determine the $\mathbf{E}_n[V_1(\tau)^k]$ for $k, n \in \mathbb{N}$, and that (6.8), for $i = 1$, with the $c_{k,l}^{(1)}(\tau)$ defined by (6.9), satisfies (6.1) and (6.2). Then, (6.8), for $i = 0$, and (6.10), follow from relation (6.3).

The last statement follows from the fact that a product of two functions of the class defined by relation (5.10), 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 [28, theorem 1] proved that, in the M/M/1 processor sharing queue without server breakdowns, the kth moment of the sojourn time conditional on n competing customers is a polynomial in n of degree k. As a corollary

of theorem 6.2 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 6.3. If $m_k < \infty$ and $\mathbf{E}[D_0^k] < \infty$, 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 6.2, $\int_{\tau=0}^{\infty} c_{k,l}^{(i)}(\tau)\mu\,\mathrm{e}^{-\mu\tau}\,\mathrm{d}\tau < \infty$, for $i \in \{0,1\}$. The corollary then follows from expression (6.8). □

7. Sojourn times in steady state

We study the sojourn time distribution 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

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

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

Theorem 7.1. For $\mathrm{Re}(s) \geqslant 0$, the LST of the distribution of $V(\tau)$ is given by

$$\begin{aligned}
&\mathbf{E}\big[\mathrm{e}^{-sV(\tau)} \mid Y(0) = 1\big] \\
&\quad = g_0(\tau; s)\,\frac{\mu - \lambda(1+\nu m_1)}{\mu - \lambda g_1(\tau; s) - \nu g_1(\tau; s)(1 - \phi(\lambda(1 - g_1(\tau; s))))/(1 - g_1(\tau; s))}, \qquad (7.1)\\
&\mathbf{E}\big[\mathrm{e}^{-sV(\tau)} \mid Y(0) = 0\big] \\
&\quad = \mathbf{E}\big[\mathrm{e}^{-sV(\tau)} \mid Y(0) = 1\big]\,\frac{1 - \phi(s + \lambda - \lambda g_1(\tau; s))}{m_1(s + \lambda - \lambda g_1(\tau; s))}\,\frac{1 - \phi(\lambda - \lambda g_1(\tau; s))}{m_1\lambda(1 - g_1(\tau; s))}. \qquad (7.2)
\end{aligned}$$

Proof. Equation (7.1) is found from (2.2) and (3.3). To find the sojourn times starting with an off-period, we remark that the residual length of that off-period is distributed as the forward recurrence time of the off-periods, i.e., $\phi_0(s) = (1 - \phi(s))/(m_1 s)$. Then using (2.3) and (3.4) we get (7.2). □

Corollary 7.2. The mean sojourn time is given by

$$\begin{aligned}
\mathbf{E}\big[V(\tau)\big] = {}& \frac{\tau}{1/(1+\nu m_1) - \lambda/\mu} + \frac{\nu m_2/2}{1+\nu m_1} \\
&+ \frac{1}{2}\,\nu\lambda m_2\,\frac{2\mu - \lambda(1+\nu m_1)}{\{\mu - \lambda(1+\nu m_1)\}^2}\big(1 - \mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\big). \qquad (7.3)
\end{aligned}$$

Proof. From theorem 7.1, by differentiating with respect to s and putting $s = 0$, we find

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

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

Alternatively, we may find $\mathbf{E}[V(\tau) \mid Y(0) = 1]$ more directly by substituting expression (2.4) for n in (6.4), and using expressions (5.3) and (5.4). Similarly, we can find $\mathbf{E}[V(\tau) \mid Y(0) = 0]$ by substituting expression (2.5) for n in (6.5), and using $\mathbf{E}[D_0] = m_2/(2m_1)$.

By averaging over $\mathbf{P}\{Y = 1\} = 1/(1+\nu m_1)$ and $\mathbf{P}\{Y = 0\} = \nu m_1/(1+\nu m_1)$ we get $\mathbf{E}[V(\tau)]$. □

It is well known that in "standard" processor sharing queues the conditional mean sojourn time, $\mathbf{E}[V(\tau)]$, is proportional to the amount of work τ. For the M/M/1 processor sharing queue this was first observed by Sakata et al. [24], and for the M/G/1 processor sharing queue by Kitayev and Yashkov [14]. From expression (7.3) we conclude that this is not the case with an unreliable server. If we replace the unreliable server by one that works with *constant* capacity $c = 1/(1+\nu m_1)$, i.e., the average service capacity of the unreliable server, $\mathbf{E}[V(\tau)]$ will be equal to $\tau/(c - \lambda/\mu)$. This corresponds to the linear term in expression (7.3). Note that for fixed τ, λ, μ, and average capacity c, expression (7.3) is fully determined by $m_2/(2m_1)$, the mean backward recurrence time of the off-periods. $\mathbf{E}[V(\tau)]$ is the smallest for deterministic off-periods, i.e., when $m_2 = (m_1)^2$, and can become arbitrarily large for increasing $m_2/(2m_1)$.

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

Remark 7.3. $\mathbf{E}[V(\tau)]$ 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 distribution of the new off-periods, which are generically denoted by $T_{\text{off}}^{(\alpha)}$, has LST

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

In particular, the first two moments of $T_{\text{off}}^{(\alpha)}$ are $m_1^{(\alpha)} = \alpha m_1$ and $m_2^{(\alpha)} = \alpha^2 m_2$. Obviously, $(\nu/\alpha)m_1^{(\alpha)} = \nu m_1$ 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

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

This limiting case ($\alpha \downarrow 0$) corresponds to the case where the server is always available and works at the constant speed $1/(1+\nu m_1)$ instead of 1.

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): $\lim_{\alpha\to\infty} \mathbf{E}[V^{(\alpha)}(\tau)] = \infty$.

Remark 7.4. From expression (7.3) we also conclude that $\mathbf{E}[V(\tau)]$ is approximately linear for large τ. 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 $\Delta\tau/(1/(1+\nu m_1) - \lambda/\mu)$ time units. The term $1/(1+\nu m_1) - \lambda/\mu$ 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 $1/(1+\nu m_1)$ per time unit, and on average an amount of capacity λ/μ per time unit is required to serve other customers (since the system with a permanent customer is ergodic, all nonpermanent customers eventually leave the system). In the next section we study the case with $\tau \to \infty$ in greater detail.

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

We study the behaviour of $g_1(\tau; s)$ as $\tau \to \infty$. Then we use these asymptotics to show the convergence of $V(\tau)/\tau$ for $\tau \to \infty$.

Our starting point is relation (4.7). By partial fraction expansion,

$$\frac{1}{\mu - (s+\lambda+\mu+\nu)x + \lambda x^2 + \nu x \phi(s+\lambda-\lambda x)} = \frac{k_1(s)}{x - r_1(s)} + k_2(x; s), \tag{8.1}$$

where

$$k_1(s) := \lim_{x \to r_1(s)} \frac{x - r_1(s)}{\mu - (s+\lambda+\mu+\nu)x + \lambda x^2 + \nu x \phi(s+\lambda-\lambda x)}, \tag{8.2}$$

exists and the function $k_2(x; s)$ is analytic in x, for $|x| \leqslant 1$ and $\mathrm{Re}(s) \geqslant 0$. Using (8.1) in (4.7) we get, for $s > 0$,

$$k_1(s) \cdot \ln\big(g_1(\tau; s) - r_1(s)\big) + \int_{x=1}^{g_1(\tau;s)} k_2(x; s)\,\mathrm{d}x = k_1(s) \cdot \ln\big(1 - r_1(s)\big) + \tau. \tag{8.3}$$

If we let $\tau \to \infty$ in (8.3), we may conclude that

$$\lim_{\tau\to\infty} g_1(\tau; s) = r_1(s), \quad s > 0. \tag{8.4}$$

This is an immediate consequence of the analyticity of $k_2(x; s)$ in x and the boundedness of $g_1(\tau; s)$, which imply that the second term on the left-hand side of (8.3) is bounded. In remark 8.1 we discuss how this limiting property can be obtained probabilistically in our model.

Remark 8.1. 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 reward that this family generates until its extinction. This reward is equal to the sum of the lifetimes of all the members of this family *plus* the reward of all nests in this family. By assigning the reward of a nest to the individual that generated it, and concatenating the lifetimes of all family members, it can be seen that the total reward of this family is distributed as a clearing period of the model of section 2:

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

This corresponds to (8.4).

Further exploiting (8.3), we can carry our asymptotic analysis one step further: For $s > 0$,

$$\lim_{\tau\to\infty} \left\{ k_1(s) \cdot \ln\left(\frac{g_1(\tau; s) - r_1(s)}{1 - r_1(s)} \right) - \tau \right\} = -\int_{x=1}^{r_1(s)} k_2(x; s)\,\mathrm{d}x. \tag{8.5}$$

Using (8.5) we can prove the following lemma:

Lemma 8.2. For $s > 0$,

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

and, consequently,

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

Proof. See the appendix. □

Theorem 8.3. For $s \geqslant 0$,

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

and, hence,

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

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

Proof. Using the first part of lemma 8.2 we can write, for $s \geqslant 0$,

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

where we use that $\lim_{s\downarrow 0}(1 - r_1(s))/s = \mathbf{E}[CP]$. We can find $\mathbf{E}[CP] = (1+\nu m_1)/(\mu - \lambda(1+\nu m_1))$ from relation (4.6). Similarly, using the second part of lemma 8.2 we have, again for $s \geqslant 0$,

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

Using this in relation (4.5) gives the convergence in distribution by the continuity theorem for LSTs of probability distributions, see Feller [8, theorem 2, p. 408]. The convergence in probability then follows immediately, because the limit is a constant. □

Using formulas (3.3) and (3.4), theorem 8.3 immediately gives the following corollary. The result is in agreement with remark 7.4.

Corollary 8.4. The sojourn time $V(\tau)$ of a customer with an amount of work τ satisfies

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

as $\tau \to \infty$.

Remark 8.5. Using the Renewal Reward Theorem, see, for instance, Tijms [30, theorem 1.3.1], it can be shown that the convergence of $V(\tau)/\tau$, and $C_0(\tau)/\tau$, 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 can then be shown that the lengths of the regeneration cycles have a finite expectation.

Remark 8.6. In addition to theorem 8.3 and corollary 8.4, it can be shown that

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

This is a consequence of theorem 3.5, corollary 3.6, and remark 8.1.

9. Heavy traffic

We now analyse the behaviour of the conditional sojourn time in heavy traffic. The main result of this section is stated in the next theorem.

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

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

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 [27] and Yashkov [40]. Thus, in heavy traffic, the distribution of $(1-\rho)V(\tau)$ converges to the exponential distribution with mean $(1+\nu m_1+\nu\lambda m_2/2)\tau$. Note that the limiting mean is linear in τ. To prove the theorem we use the following lemma:

Lemma 9.2. For $\mathrm{Re}(s) \geqslant 0$,

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

Proof. See the appendix. □

Note that (ii) of lemma 9.2 can be rewritten in terms of the LST of the distribution of the backward (or forward) recurrence time of $C_1(\tau)$ as follows:

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

see formula (5.3).

Using lemma 9.2, theorem 9.1 can be proved by substituting $(1-\rho)s$ for s in (7.1) and (7.2), and letting $\rho \uparrow 1$.

10. Final remarks

We 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 reward structure on it. We indicated how the same

transformation can be applied for 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 explicit form of the results may be valuable in future research when studying M/M/1 processor sharing models with a more general varying service process, see, for instance, Núñez-Queija [20].

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 standard M/G/1 queue with processor sharing. The LSTs of the distributions of these "fundamental" random variables composing $V(\tau)$ were characterised through an integral equation. We computed the first two moments of the fundamental 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 [28, theorem 1]. We gave an explicit expression for the LST of the sojourn time distribution in steady state, in terms of the LSTs of the distributions of the fundamental random variables. The mean of the steady-state sojourn times was found in terms of the input parameters.

We further studied asymptotics of the queueing model. First, we analysed the case for $\tau \to \infty$, proving that $V(\tau)/\tau$ converges (with probability 1) to a constant. Then we proved under heavy-traffic conditions, that is for the traffic load $\rho \uparrow 1$, 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 (asymptotically) linear in three cases: (i) when the on- and off-periods alternate rapidly, (ii) when τ is large, and (iii) in heavy traffic. An intuitive explanation for this linearity in all three cases is that the sojourn times are large compared to the lengths of the on- and off-periods, so that fluctuations in the service availability average out.

Acknowledgements

The author would like to thank S.C. Borst, O.J. Boxma, J.W. Cohen, J.A.C. Resing, B. Sengupta, S.F. Yashkov, A.P. Zwart, and the referees, for many helpful comments that led to improvement of the paper. In particular, the author is indebted to Professor J.W. Cohen for the discussions that led to the asymptotic analysis in section 8.

Appendix: Technical proofs

A.1. Proof of lemma 4.1

By conditioning on the number of "single" children and the number of nests that a nonpermanent 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

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

Here we use the fact that "Poisson arrivals occur homogeneously in time", see, for instance, Tijms [30, theorem 1.2.5]. Note that $\phi(s+\lambda(1-g_1(\tau;s)))$ is the LST of the distribution of the reward of a nest plus the rewards of all children in that nest and their offsprings, until time τ.

Equivalently we may write

$$\begin{aligned} g_1(\tau+\Delta;s) = &\int_{t=0}^{\Delta} \mu \exp\Bigg\{-\mu t - st - \lambda\Bigg(t - \int_{u=0}^{t} g_1(\tau+u;s)\,\mathrm{d}u\Bigg) \\ &\qquad - \nu\Bigg(t - \int_{u=0}^{t} \phi\big(s+\lambda\big(1-g_1(\tau+u;s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}\mathrm{d}t \\ &+ g_1(\tau;s)\exp\Bigg\{-\mu\Delta - s\Delta - \lambda\Bigg(\Delta - \int_{u=0}^{\Delta} g_1(\tau+u;s)\,\mathrm{d}u\Bigg) \\ &\qquad - \nu\Bigg(\Delta - \int_{u=0}^{\Delta} \phi\big(s+\lambda\big(1-g_1(\tau+u;s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}. \end{aligned} \tag{A.1}$$

By similar arguments we also find

$$\begin{aligned} g_0(\tau+\Delta;s) = g_0(\tau;s)\exp\Bigg\{&-s\Delta - \lambda\Bigg(\Delta - \int_{u=0}^{\Delta} g_1(\tau+u;s)\,\mathrm{d}u\Bigg) \\ &- \nu\Bigg(\Delta - \int_{u=0}^{\Delta} \phi\big(s+\lambda\big(1-g_1(\tau+u;s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}. \end{aligned} \tag{A.2}$$

From (A.1) and (A.2) we can show that, for $\Delta \downarrow 0$,

$$\begin{aligned} g_1(\tau+\Delta;s) = &\big(1-(s+\lambda+\mu+\nu)\Delta\big)g_1(\tau;s) + \lambda\Delta\big\{g_1(\tau;s)\big\}^2 + \mu\Delta \\ &+ \nu\Delta g_1(\tau;s)\phi\big(s+\lambda\big(1-g_1(\tau;s)\big)\big) + \mathrm{o}(\Delta), \end{aligned} \tag{A.3}$$

$$\begin{aligned} g_0(\tau+\Delta;s) = &\big(1-(s+\lambda+\nu)\Delta\big)g_0(\tau;s) + \lambda\Delta g_0(\tau;s)g_1(\tau;s) \\ &+ \nu\Delta g_0(\tau;s)\phi\big(s+\lambda\big(1-g_1(\tau;s)\big)\big) + \mathrm{o}(\Delta). \end{aligned} \tag{A.4}$$

With (A.3) and (A.4) 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 (A.3) and (A.4) by $\tau - \Delta$, the continuity from the left in τ also easily follows. Subsequently it can be shown that, for $i \in \{0, 1\}$,

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

so that $(\partial/\partial\tau)g_1(\tau; s)$ and $(\partial/\partial\tau)g_0(\tau; s)$ exist and satisfy equations (4.1) and (4.2). Condition (4.3) follows from $C_0(0) = C_1(0) = 0$.

A.2. *Proof of theorem 5.1*

It is known for the M/G/1 queue that the kth moment of the busy period exists if and only if the kth moment of the service time exists, see De Meyer and Teugels [5, lemma 3]. With lemma 4.3, this implies that the kth moment of the clearing period exists, if and only if $m_k < \infty$. Since $C_1(\tau)$ is nondecreasing in τ with probability 1, and $C_1(\tau)$ converges to the clearing period CP, as $\tau \to \infty$, it must be that

$$\mathbf{E}\big[C_1(\tau)^k\big] \leqslant \mathbf{E}\big[CP^k\big],$$

($C_1(\tau)$ is stochastically smaller than CP), and hence the kth moment of $C_1(\tau)$ exists when $m_k < \infty$.

To prove the result for $C_0(\tau)$, we first write the following identity:

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

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 ith regular child is born, and $C_i(\tau - T_i^{(\lambda)})$ is the reward 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 reward of the jth batch, $N^{(\lambda)}(D_j)$ is the number of children in the jth batch, $T_j^{(\nu)}$ is the time at which the jth batch is generated, and $C_{j,n}(\tau - T_j^{(\nu)})$ is the reward of the nth child in the jth batch and his offspring, until time τ. The above identity was given in terms of LSTs in relation (4.5).

If we replace each of the rewards until time τ associated with a child of the permanent customer and his offspring by the reward of the family of that child over a total time-span of length τ, we clearly have an upper bound for $C_0(\tau)$:

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

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

$$\mathbf{E}\left[\mathrm{e}^{-s\overline{C_0}(\tau)}\right] = \exp\left\{-\tau\left(s + \lambda\left(1 - g_1(\tau; s)\right) + \nu\left[1 - \phi\left(s + \lambda - \lambda g_1(\tau; s)\right)\right]\right)\right\}. \quad \text{(A.5)}$$

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

$$\begin{aligned} \phi(s) &= 1 + \sum_{i=1}^{k} m_i \frac{(-s)^i}{i!} + \mathrm{o}\left(s^k\right), \\ g_1(\tau; s) &= 1 + \sum_{j=1}^{k} \mathbf{E}\left[C_1(\tau)^j\right] \frac{(-s)^j}{j!} + \mathrm{o}\left(s^k\right), \end{aligned} \quad \text{(A.6)}$$

see De Meyer and Teugels [5, lemma 1]. Combining these, we get

$$\begin{aligned} &\phi\left(s + \lambda - \lambda g_1(\tau; s)\right) \\ &\quad = 1 + \sum_{i=1}^{k} m_i \frac{\left(-s + \lambda \sum_{j=1}^{k} \mathbf{E}[C_1(\tau)^j](-s)^j/j!\right)^i}{i!} + \mathrm{o}\left(s^k\right). \end{aligned} \quad \text{(A.7)}$$

From equation (A.5) it is now straightforward to see that the LST of the distribution of $\overline{C_0}(\tau)$ has a finite kth derivative in $s = 0$. Therefore, the kth moment of $\overline{C_0}(\tau)$ and, hence, the kth moment of $C_0(\tau)$, exists.

A.3. *Proof of theorem 5.2*

Let T_{off} be as before and $N(T_{\mathrm{off}})$ 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 of its distribution $g_1(\tau; s)$, then using equations (A.6) and (A.7),

$$\begin{aligned} &\mathbf{E}\left[\mathrm{e}^{-s(T_{\mathrm{off}} + C_1(\tau) + \cdots + C_{1+N(T_{\mathrm{off}})}(\tau))}\right] \\ &\quad = g_1(\tau; s)\phi\left(s + \lambda - \lambda g_1(\tau; s)\right) \\ &\quad = 1 + \frac{s}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{k}\left((s+\lambda)\frac{m_i}{i!} + \frac{m_{i-1}}{(i-1)!}\right)\left(-s + \lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i \\ &\qquad + \mathrm{o}\left(s^k\right). \end{aligned} \quad \text{(A.8)}$$

We write out the terms in the summation as

$$\begin{aligned} &\left(-s + \lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i \\ &\quad = \sum_{i_0 + i_1 + \ldots + i_k = i} \binom{i}{i_0, \ldots, i_k} (-s)^{i_0} \prod_{j=1}^{k}\left(\lambda\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^{i_j} \end{aligned}$$

$$= \sum_{n=0}^{k} (-s)^n \sum_{\substack{i_0+i_1+i_2+\cdots+i_k=i \\ i_0+i_1+2i_2+\cdots+ki_k=n}} \binom{i}{i_0,\ldots,i_k} \prod_{j=1}^{k} \left(\lambda \mathbf{E}\big[C_1(\tau)^j\big] \frac{1}{j!}\right)^{i_j} + o\big(s^k\big). \quad \text{(A.9)}$$

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

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

We now prove the theorem by induction on k. From (A.8) and (A.9) we can show that if $\mathbf{E}[C_1(\tau)^j]$ has the form of (5.9) for $j = 1, 2, \ldots, k-1$, then

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

This can be verified by noting that the only contribution of $\mathbf{E}[C_1(\tau)^k]$ to the coefficient of $(-s)^k$ in (A.8) 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 $\mathbf{E}[C_1(\tau)^j]$, 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 $m \geqslant 1$, $n \geqslant 0$ and $m + n \leqslant k$. 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 $\mathbf{E}[C_1(\tau)^{l_1}] \times \mathbf{E}[C_1(\tau)^{l_2}]$ is of the same form as $\mathbf{E}[C_1(\tau)^{l_1+l_2}]$ in (5.9), except for the terms containing $\mathrm{e}^{-\{\mu-\lambda(1+\nu m_1)\}\tau}\tau^n$, with $n \geqslant \max(l_1, l_2)$, 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 (A.8) and (A.9).

As before, we can derive a differential equation for $\mathbf{E}[C_1(\tau)^k]$:

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

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 $\mathbf{E}[C_1(\tau)^k]$ indeed leads to the form of relation (5.9). 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 $\mathbf{E}[C_0(\tau)^k]$:

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

By similar arguments as before, we find relation (5.10).

A.4. *Proof of lemma 8.2*

Using relation (8.3) we may write

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

It is clear from (8.2) that $k_1(s) < 0$, for $s > 0$: for $x = 0$ the numerator on the right-hand side of (8.2) 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

$$\begin{aligned}0 &\leqslant \int_{u=0}^{\tau}\left(g_1\left(u;\frac{s}{\tau}\right) - r_1\left(\frac{s}{\tau}\right)\right)\mathrm{d}u \\ &\leqslant \left(1 - r_1\left(\frac{s}{\tau}\right)\right)\exp\left\{-\frac{1}{k_1(s/\tau)}\int_{x=1}^{M(s/\tau)} k_2\left(x;\frac{s}{\tau}\right)\mathrm{d}x\right\} \\ &\quad \times k_1\left(\frac{s}{\tau}\right)\left(\exp\left\{\frac{\tau}{k_1(s/\tau)}\right\} - 1\right).\end{aligned} \tag{A.10}$$

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

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

Thus, if we let $\tau \to \infty$ in (A.10) then the upper bound 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 $s \geqslant 0$, and $(\mathrm{d}/\mathrm{d}s)\phi(s)|_{s=0} = -m_1$. Therefore, it holds that $\phi(s_1) - \phi(s_2) \leqslant m_1(s_2 - s_1)$, whenever $0 \leqslant s_1 \leqslant s_2$.

A.5. *Proof of lemma 9.2*

Part (i). Substitute $(1-\rho)s$ for s in equations (A.1) and (A.2), 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)

$$\begin{aligned}
h_1(\tau + \Delta; s) = \int_{t=0}^{\Delta} \mu \exp\Bigg\{ -\mu t - \lambda\Bigg(t - \int_{u=0}^{t} h_1(\tau + u; s)\,\mathrm{d}u\Bigg) \\
- \nu\Bigg(t - \int_{u=0}^{t} \phi\big(\lambda\big(1 - h_1(\tau + u; s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}\mathrm{d}t \\
+ h_1(\tau; s) \exp\Bigg\{ -\mu\Delta - \lambda\Bigg(\Delta - \int_{u=0}^{\Delta} h_1(\tau + u; s)\,\mathrm{d}u\Bigg) \\
- \nu\Bigg(\Delta - \int_{u=0}^{\Delta} \phi\big(\lambda\big(1 - h_1(\tau + u; s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}, \\
h_0(\tau + \Delta; s) = h_0(\tau; s) \exp\Bigg\{ -\lambda\Bigg(\Delta - \int_{u=0}^{\Delta} h_1(\tau + u; s)\,\mathrm{d}u\Bigg) \\
- \nu\Bigg(\Delta - \int_{u=0}^{\Delta} \phi\big(\lambda\big(1 - h_1(\tau + u; s)\big)\big)\mathrm{d}u\Bigg)\Bigg\}.
\end{aligned}$$

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

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

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 $\mathrm{Re}(s) \geqslant 0$, $|g_1(\tau; s)| \leqslant 1$, we can find a sequence $(\rho_k)_{k \in \mathbb{N}}$ 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 $(1-g_1(\tau;(1-\rho_k)s))/(1-\rho_k)$. Such a sequence exists because $(1-g_1(\tau;\omega))/\omega \leqslant \mathbf{E}[C_1(\tau)]$ for any $\mathrm{Re}(\omega) \geqslant 0$, and $\mathbf{E}[C_1(\tau)]$ is bounded in $\rho \in [0,1]$, see formula (5.3).

Substitute $(1-\rho)s$ for s in (A.1), subtract both sides of this equation from 1, and use

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

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

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

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 $l_1(\tau;s) = (1+\nu m_1)s\tau$.

References

[1] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, Queueing Systems 10 (1992) 5–87.

[2] E.G. Coffman, R.R. Muntz and H. Trotter, Waiting time distributions for processor sharing systems, J. Assoc. Comput. Mach. 17 (1970) 123–130.

[3] J.W. Cohen, The multiple phase service network with generalised processor sharing, Acta Informatica 12 (1979) 245–284.

[4] J.W. Cohen, *The Single Server Queue*, 2nd ed. (North-Holland, Amsterdam, 1982).
[5] A. de Meyer and J.L. Teugels, On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1, J. Appl. Probab. 17 (1980) 802–813.
[6] A. Federgruen and L. Green, Queueing systems with service interruptions, Oper. Res. 34 (1986) 752–768.
[7] A. Federgruen and L. Green, Queueing systems with service interruptions II, Naval Res. Logist. 35 (1988) 345–358.
[8] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1966).
[9] R.D. Foley and G.-A. Klutke, Stationary increments in the accumulated work process in processor-sharing queues, J. Appl. Probab. 26 (1989) 671–677.
[10] S.W. Fuhrmann and R.B. Cooper, Stochastic decompositions in the M/G/1 queue with generalised vacations, Oper. Res. 33 (1985) 1117–1129.
[11] D.P. Gaver, Jr., A waiting line with interrupted service, including priorities, J. Roy. Statist. Soc. 24 (1962) 73–90.
[12] S.A. Grishechkin, Crump-Mode–Jagers branching processes as a method of investigating M/G/1 systems with processor sharing, Theory Probab. Appl. 36 (1991) 19–35 (translated from Teor. Veroyatnost. i Primenen. 36 (1991) 16–33 (in Russian)).
[13] S.A. Grishechkin, On a relationship between processor sharing queues and Crump-Mode–Jagers branching processes, Adv. in Appl. Probab. 24 (1992) 653–698.
[14] M.Yu. Kitayev and S.F. Yashkov, Analysis of a single-channel queueing system with the discipline of uniform sharing of a device, Engrg. Cybernet. 17 (1979) 42–49.
[15] D.-S. Lee, Analysis of a single server queue with semi-Markovian service interruption, Queueing Systems 27 (1997) 153–178.
[16] W. Li, D. Shi and X. Chao, Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations, J. Appl. Probab. 34 (1997) 546–555.
[17] I. Mitrani and B. Avi-Itzhak, A many server queue with service interruptions, Oper. Res. 16 (1968) 628–638.
[18] J.A. Morrison, Response-time distribution for a processor sharing system, SIAM J. Appl. Math. 45 (1985) 152–167.
[19] M.F. Neuts, *Matrix-geometric Solutions in Stochastic Models – An Algorithmic Approach* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).
[20] R. Núñez-Queija, Sojourn times in non-homogeneous QBD processes with processor sharing, CWI Report PNA-R9901 (1999), submitted for publication.
[21] T.J. Ott, The sojourn time distributions in the M/G/1 queue with processor sharing, J. Appl. Probab. 21 (1984) 360–378.
[22] K.M. Rege and B. Sengupta, A decomposition theorem and related results for the discriminatory processor sharing queue, Queueing Systems 18 (1994) 333–351.
[23] J.W. Roberts, Realising quality of service guarantees in multiservice networks, in: *Proc. IFIP Seminar PMCCN'97* (1997).
[24] M. Sakata, S. Noguchi and J. Oizumi, Analysis of a processor shared queueing model for time sharing systems, in: *Proc. of the 2nd Hawaii Internat. Conf. on System Sciences* (1969) pp. 625–628.
[25] R. Schassberger, A new approach to the M/G/1 processor sharing queue, Adv. in Appl. Probab. 16 (1984) 202–213.
[26] B. Sengupta, A queue with service interruptions in an alternating random environment, Oper. Res. 38 (1990) 308–318.
[27] 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.
[28] B. Sengupta and D.L. Jagerman, A conditional response time of the M/M/1 processor sharing queue, AT&T Techn. J. 64 (1985) 409–421.

[29] T. Takine and B. Sengupta, A single server queue with server interruptions, Queueing Systems 26 (1997) 285–300.
[30] H.C. Tijms, *Stochastic Models – An Algorithmic Approach* (Wiley, Chichester, UK, 1994).
[31] Traffic management specification, Version 4.0, The ATM Forum Technical Committee (April 1996).
[32] J.L. van den Berg and O.J. Boxma, The M/G/1 queue with processor sharing and its relation to a feedback queue, Queueing Systems 9 (1991) 365–401.
[33] P.D. Welch, On a generalised M/G/1 queueing process in which the first customer of each busy period receives exceptional service, Oper. Res. 12 (1964) 736–752.
[34] H. White and L.S. Christie, Queuing with preemptive priorities or with breakdown, Oper. Res. 6 (1958) 79–95.
[35] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 223–231.
[36] S.F. Yashkov, A derivation of response time distribution for a M/G/1 processor-sharing queue, Probl. Control. Inform. Theory 12 (1983) 133–148.
[37] S.F. Yashkov, New applications of random time change to the analysis of processor sharing queues, in: *Proc. of the 4th Internat. Vilnius Conf. on Probability Theory and Mathematical Statistics* (1985) pp. 343–345.
[38] S.F. Yashkov, Processor-sharing queues: Some progress in analysis, Queueing Systems 2 (1987) 1–17.
[39] S.F. Yashkov, Mathematical problems in the theory of processor-sharing queueing systems, J. Soviet Math. 58 (1992) 101–147.
[40] S.F. Yashkov, On a heavy-traffic limit theorem for the M/G/1 processor-sharing queue, Comm. Statist. Stochastic Models 9(3) (1993) 467–471.