Centrum voor Wiskunde en Informatica

Sojourn times in non-homogeneous QBD processes with processor sharing

R. Núñez Queija

Probability, Networks and Algorithms (PNA)

PNA-R9901 February 28, 1999

Report PNA-R9901
ISSN 1386-3711

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

Sojourn Times in Non-Homogeneous QBD Processes with Processor Sharing

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

ABSTRACT

We study sojourn times of customers in a processor sharing model with a service rate that varies over time, depending on the number of customers and on the state of a random environment. An explicit expression is derived for the Laplace-Stieltjes transform of the sojourn time conditional on the state upon arrival and the amount of work brought into the system. Particular attention is given to the conditional mean sojourn time of a customer as a function of his required amount of work, and we establish the existence of an asymptote as the amount of work tends to infinity. The method of random time change is then extended to include the possibility of a varying service rate. By means of this method, we explain the well-established proportionality between the conditional mean sojourn time and required amount of work in processor sharing queues without random environment. Based on numerical experiments, we propose an approximation for the conditional mean sojourn time. Although first presented for exponentially distributed service requirements, the analysis is shown to extend to phase-type services. The service discipline of discriminatory processor sharing is also shown to fall within the framework.

1991 Mathematics Subject Classification: 60K25, 68M20, 90B12, 90B22.
Keywords & Phrases: (discriminatory) processor sharing, varying service rate, sojourn time, random time change, random environment, quasi birth-death process.
Note: This work was carried out in PNA 2.1. as part of the project *Quality in Future Networks* of the Telematics Institute.

1. Introduction

We consider a processor sharing queueing model in which the service speed depends on the state of some underlying Markov chain. To be more precise: Customers arrive at a service station requiring a random amount of service. If at time $t \geq 0$, the number of customers X_t in the system equals $k > 0$, and some (yet to be specified) underlying Markov process $\{Y_t, t \geq 0\}$ is in state $i \in \{1, 2, \ldots, N\}$, then the total service rate at which customers are served, is $r_i^{(k)} \geq 0$. The process Y_t is called the random environment. The offered service speed is divided equally among all customers present. Under the assumption of Poisson arrivals (possibly state-dependent) and exponentially distributed service requirements, the two-dimensional process $\{(X_t, Y_t), t \geq 0\}$ is a non-homogeneous (or level-dependent) Quasi Birth and Death (QBD) process.

The model may be used for the performance analysis of modern telecommunication systems, in which real-time and non real-time (best-effort) traffic share the same resources. Real-time connections (such as telephony and interactive video) have strict delay requirements at the packet- (or cell-) level. Therefore, after accepting such a connection, network resources must be reserved to guarantee a certain transmission rate. Non real-time connections, however, are less sensitive to delays at the packet-level. In file transfers, for example, the transmission time of the *complete* file — i.e. the delay at the connection- (or call-) level — is of interest, and less so the delay of any part of the file. Consequently, the capacity available to non real-time connections may vary over the duration of such a connection. For instance in ATM (Asynchronous Transfer Mode) networks, it was proposed that non real-time data connections should be accommodated in the ABR (Available Bit Rate) service class. In principle, connections using this service are only offered the capacity that is not required by real-time connections using other service classes such as CBR (Constant Bit Rate) or real-time VBR (Variable Bit Rate). Furthermore, data connections should share the available capacity fairly, i.e. each such connection should get an equal share, see for instance The ATM Forum [1]. Also in IP networks (e.g. the internet), it is thought to be useful to make a distinction between real-time and non real-time (best-effort)

connections in order to provide Quality of Service guarantees, see for instance Van der Wal et al. [22] and White [23].
In Núñez Queija et al. [15] the presented model is used for the performance analysis of best-effort and real-time connections in a telecommunication switch, under three different connection acceptance strategies. In that paper the random environment Y_t represents the number of real-time connections on the switch, each of which requires a fixed amount of capacity for its complete duration. The remaining capacity is equally shared among the best-effort connections.

In this paper we study the sojourn times of customers in the processor sharing system with varying total service rate. We are particularly interested in the sojourn times of customers conditioned on their required amount of work. In the context of the above mentioned application in telecommunication systems, these conditional sojourn times of customers correspond to the total transmission time of files of given size. For processor sharing systems with constant service rate it is well-known that the conditional mean sojourn time is proportional to the amount of work. When the server alternates between exponentially distributed activity periods, during which the service rate is constant, and generally distributed unavailability periods, it is shown in Núñez Queija [14] that this proportionality property is lost. However, in that report an asymptotic linearity (for the amount of work tending to infinity) is revealed. In this report we show that this asymptotic result is also valid for the present model, in which the service rate may assume different positive values. Using a time-scale transformation, the problem may be viewed in the context of a Markov reward process. We also discuss the relation of our approach with the branching process approach, which has proven to be valuable in the analysis of traditional processor sharing systems.
The remainder of the paper is organised as follows. We present the basic model in Section 2 and explain how the models of [15] fit into it. In Section 3 the sojourn times of customers are studied. We mainly concentrate on sojourn times conditional on the state upon arrival and on the amount of work brought into the system. An explicit expression is derived for the Laplace-Stieltjes transform of the conditional sojourn time. Particular attention is paid to the conditional mean sojourn time as a function of the amount of work, and we prove the existence of an asymptote, as the amount of work tends to infinity. In Section 4 we extend the method of random time change, originally introduced for the M/G/1 processor sharing queue by Yashkov [24], to our model. This way we translate sojourn times in the queueing system into costs in a Markov reward process. We also discuss the relation between our approach and the branching process often used in the literature to study processor sharing queues. The case where the server may be unavailable for some periods of time is discussed in Section 5. In Section 6 we explain the proportionality property between conditional mean sojourn times and the amount of work brought into the system in processor sharing queues without random environment. We show in Section 7 how the conditional mean sojourn times may be computed. In view of the computational complexity, we propose an approximation based on numerical experiments presented in [15]. In Section 8 it is shown that phase-type services and discriminatory processor sharing essentially fall within the basic model. We also discuss the extension to infinite state spaces. Concluding remarks are made in Section 9.

2. The model

In this section we make precise what we mean by a non-homogeneous (or level-dependent) QBD process with processor sharing.
The generator of a non-homogeneous QBD process is of the form:

$$\mathcal{G} := \begin{bmatrix} D^{(0)} & \Lambda^{(0)} & 0 & \dots & \dots & 0 \\ M^{(1)} & D^{(1)} & \Lambda^{(1)} & 0 & \dots & \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & M^{(L-1)} & D^{(L-1)} & \Lambda^{(L-1)} \\ 0 & \dots & & 0 & M^{(L)} & D^{(L)} \end{bmatrix}. \tag{2.1}$$

It will be assumed throughout this paper that $\mathcal{G}$ is irreducible (all states in the corresponding Markov process communicate). The process is called a homogeneous QBD process if $\forall k$: $M^{(k)} = M$, $D^{(k)} = D$, $\Lambda^{(k)} = \Lambda$.

We consider the case with a finite state space, i.e. L (the maximum *level*) is finite and the submatrices $D^{(k)}$, $\Lambda^{(k)}$, and $M^{(k)}$ are all of finite – but not necessarily the same – dimension. Generalisations to infinite state spaces are possible, but require specific attention regarding ergodicity issues. We address these issues in Section 8.3.
A state in a Markov process with generator (2.1) can be represented by a pair of integers

$$(k,i) \in \mathbf{S} := \left\{(l,j) \;:\; l = 0,1,\ldots,L;\; j \in E^{(l)}\right\},$$

with $E^{(l)} \subseteq \{1,2,\ldots,N\}$, and $N < \infty$. The components k and i are called the level and the phase of the process, respectively.
We denote the steady-state probability vector by $\overline{\pi}$:

$$\overline{\pi}\, \mathcal{G} = \overline{0}, \quad \overline{\pi}\, \overline{1} = 1,$$

with $\overline{0}$ (resp. $\overline{1}$) being the vector with all entries equal to zero (one). Throughout this paper, for any vector $\overline{v}$ its entries $v_i^{(k)}$ are ordered lexicographically, i.e. $v_i^{(k)}$ precedes $v_j^{(l)}$ if $k < l$, or if $k = l$ *and* $i < j$. Another notational convention we adopt, is that any vector multiplying an expression from the left (right) is a row (column) vector. Furthermore we use the symbol $\mathcal{I}$ to denote the identity matrix. Whenever used, the vectors $\overline{1}$ and $\overline{0}$, and the matrix $\mathcal{I}$ are of the right dimension.
In the present paper we do not discuss the computation of the steady-state probability vector $\overline{\pi}$ (or any other probability vector). The interested reader is referred to De Nitto Personè and Grassi [6] for an algorithm to compute the steady-state probabilities in non-homogeneous QBD processes with a finite state space.

We construct a queueing system, which may be modelled as a Markov process having generator (2.1), see Figure 1.

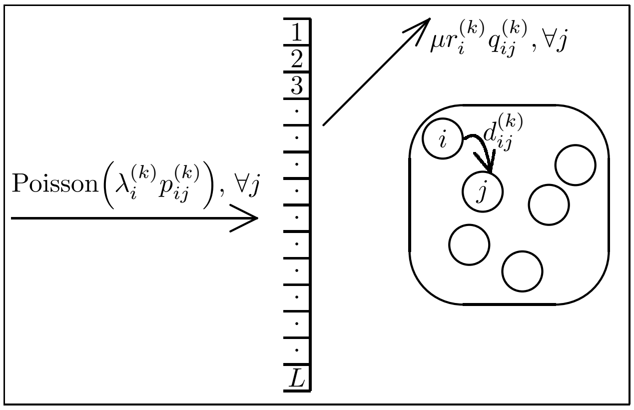

Figure 1: The queueing model

In this queueing model, arrivals and services of customers may be dependent on the queue length, and on the state of a random environment, which itself may be modelled as a Markov process. The random environment may also be dependent on the arrival and departure process of customers. For the time being (we come back to this in Section 8.1), it is assumed that customers have an exponentially distributed service requirement with mean $1/\mu$ (independent of other service requirements, the arrival process, and the random environment). We say that the queueing system of Figure 1 is in state $(k,i) \in \mathbf{S}$ when there are $k \in \{0,1,\ldots,L\}$ customers present and the state of the random environment is $i \in E^{(k)} \subseteq \{1,2,\ldots,N\}$. If the state of the system is (k,i), then new customers arrive according to a Poisson process with rate $\lambda_i^{(k)}$. Upon such an arrival, the number of customers in the system is increased by one, and the random environment changes (immediately) to state $j \in E^{(k+1)}$ with probability $p_{ij}^{(k)}$, where $\sum_{j \in E^{(k+1)}} p_{ij}^{(k)} = 1$ for $0 \leq k \leq L-1$. If $j = i$ then this is the probability that the random environment does not change state. In state (k,i) with $k > 0$, the server works

at rate $r_i^{(k)}$. This service capacity is equally divided among all customers present (processor sharing), such that each customer leaves in an interval of length Δ with probability $\frac{1}{k}\mu r_i^{(k)}\Delta + o(\Delta)$, for $\Delta \downarrow 0$. The total departure rate of customers is therefore $\mu r_i^{(k)}$. Upon such a departure, the random environment changes to state $j \in E^{(k-1)}$ with probability $q_{ij}^{(k)}$, where $\sum_{j\in E^{(k-1)}} q_{ij}^{(k)} = 1$ for $1 \leq k \leq L$. Finally, in state (k,i) the random environment may change to state $j \in E^{(k)}$ – without changing the number of customers – at rate $d_{ij}^{(k)}$, $j \neq i$.
For $(k,i) \in \mathbf{S}$ it is convenient to define $p_{ij}^{(k)} = 0,\ j \notin E^{(k+1)}$, $q_{ij}^{(k)} = 0,\ j \notin E^{(k-1)}$, and $d_{ij}^{(k)} = 0,\ j \notin E^{(k)}$. Note that, $E^{(-1)}$ and $E^{(L+1)}$ are not defined, therefore we further set $\lambda_i^{(L)} := r_i^{(0)} := 0$, for all $i \in \{1,2,\ldots,N\}$.

The queueing system under consideration may be modelled as a Markov process with generator (2.1). The level of the process corresponds to the number of customers in the system, and the phase of the process corresponds to the state of the random environment. The submatrices in (2.1) are:

$$\begin{aligned}
\Lambda^{(k)} &= \left[\lambda_i^{(k)} p_{ij}^{(k)}\right]_{i\in E^{(k)}, j\in E^{(k+1)}}, \\
M^{(k)} &= \left[\mu r_i^{(k)} q_{ij}^{(k)}\right]_{i\in E^{(k)}, j\in E^{(k-1)}}, \\
D^{(k)} &= \left[d_{ij}^{(k)}\right]_{i,j\in E^{(k)}},
\end{aligned}$$

where the $d_{ii}^{(k)}$ are such that (2.1) is a true generator (all rows sum to 0).
Usually the service discipline considered for queueing systems which can be modelled as a QBD is FIFO (First In First Out). In the present queueing system the service discipline is processor sharing. Because of the exponentially distributed service requirements, the queue length process obeys the same probabilistic law for all work-conserving service disciplines that do not take into account actual service requirements (including FIFO and processor sharing). The queue-length in non-homogeneous QBD-processes has been studied extensively in the literature, see for instance [6] where an algorithm is provided for the computation of the steady-state queue-length distribution.
For the sojourn times of customers, the service discipline *does* matter. Sojourn times in QBD processes under the FIFO discipline are discussed in Neuts [13, Section 3.9]. For non-homogeneous QBD processes an analogous treatment is possible. The distribution in terms of Laplace-Stieltjes transforms may be found in Li and Sheng [12]. With the processor sharing service discipline, sojourn times for the case of a server alternating between availability periods (during which the service rate is constant) and non-availability periods (during which no service is provided), are studied in [14]. There, the lengths of non-availability periods may be drawn from a general distribution. If these lengths are exponentially distributed, then it coincides with the present model, with $N = 2$ and $L = \infty$, see also Section 8.3.

Example
To conclude this section, we briefly show how the models analysed in [15] fit into the framework of this paper: Our queueing system is used to model a (single) link in a telecommunication system. Two types of traffic are transmitted over the link: Real-time traffic (voice, video) and best-effort traffic (data files). Real-time traffic does not allow for delays, and upon acceptance of such a connection, a fixed amount of service capacity must be reserved for the complete duration of that connection. Best-effort traffic has no delay requirements at the cell level, and the service capacity allocated to a connection of that type, may vary over its duration. Moreover, best-effort traffic sources share equally in the capacity available to them.
The system may be modelled as follows: The number of real-time connections using the link determines the amount of capacity that is left over for best-effort traffic. Therefore, if we identify best-effort connections with the customers in our queueing system, the number of real-time connections determines the random environment. The fair sharing of the available capacity among best-effort sources is modelled by the processor sharing service discipline. In addition, single best-effort connections may have a maximum transmission rate, therefore the total capacity allocated to best-effort traffic may depend on the number of best-effort connections.

In [15] three different strategies to accept new calls are considered. Two of these strategies (called 'integrated' and 'mixed') may be studied using the approach presented in this paper. We remark that in the 'integrated' scenario, whether or not new real-time connections are accepted, depends on the number of best-effort connections, which leads (in a natural way) to dependence of the random environment (c.q. the number of real-time connections) on the number of customers (c.q. the number of best-effort connections).

3. Sojourn times

In this section we study the sojourn time of a customer conditioned on the number of customers and the state of the random environment upon his arrival. Particular attention is paid to the case where we also condition on the amount of work brought into the system.
It will be useful to define the following generator:

$$\mathcal{H} := \left[\begin{array}{c|cccccc} 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ \hline M^{(1)}\overline{1} & D^{(1)} & \Lambda^{(1)} & & & & \\ \frac{1}{2}M^{(2)}\overline{1} & \frac{1}{2}M^{(2)} & D^{(2)} & \Lambda^{(2)} & & & \\ \vdots & & \ddots & \ddots & \ddots & & \\ \frac{1}{k}M^{(k)}\overline{1} & & & \frac{k-1}{k}M^{(k)} & D^{(k)} & \Lambda^{(k)} & \\ \vdots & & & & \ddots & \ddots & \ddots \\ \frac{1}{L}M^{(L)}\overline{1} & & & & & \frac{L-1}{L}M^{(L)} & D^{(L)} \end{array}\right]. \tag{3.1}$$

The state space of a Markov process with generator $\mathcal{H}$ may be denoted by all pairs (k, i), with $k = 1, 2, \dots, L$ and $i \in E^{(k)}$, and an absorbing state 0. Note that there are no states (k, i) with $k = 0$, and that all states (k, i), $k = 1, 2, \dots, L$ are transient. The latter statement follows from the irreducibility of $\mathcal{G}$ defined by (2.1). In the first column of $\mathcal{H}$ we find the transition (absorption) rates from all other states into state 0. From any state (k, i) the absorption rate (into state 0) equals $\frac{1}{k}\mu r_i^{(k)}$.

Theorem 3.1 *The sojourn time of a customer who enters the system with $k-1$ other customers present and the random environment being in state i, is distributed as the absorption time in a Markov process $\mathcal{M}_\mathcal{H}$ with generator $\mathcal{H}$ defined by (3.1), starting from state (k, i).*

Proof
The proof can be given by comparing the evolution of the queueing system of Figure 1, from the moment that the tagged customer arrives (and finds $k-1$ other customers and the random environment in state i), with the evolution of the Markov process $\mathcal{M}_\mathcal{H}$, starting in state (k, i), until absorption in state 0.
In particular, at any moment that the tagged customer is in service with $l-1$ other customers and the random environment in state j, the rate at which he is served is $\frac{1}{l}r_j^{(l)}$, and his 'departure rate' is therefore $\mu \cdot \frac{1}{l}r_j^{(l)}$. The departure of the tagged customer from the queueing system corresponds to absorption in state 0 in the Markov process $\mathcal{M}_\mathcal{H}$ (see the first column of $\mathcal{H}$). ■

Remark 3.1 For the computation of the moments of the absorption time in $\mathcal{M}_\mathcal{H}$ (and hence of the sojourn time in the queueing system) from any initial state we refer to Li and Sheng [12]. □

In this paper we further concentrate on the sojourn time of a customer with an amount of work $\tau > 0$: For $k = 1, 2, \dots, L$ and $i \in E^{(k)}$, let $V_i^{(k)}(\tau)$ be the (remaining) sojourn time of a (tagged) customer, starting with $k-1$ other customers present, the random environment in state i, and and having a (remaining) service requirement of τ. Define the Laplace-Stieltjes transform (LST) of the distribution of $V_i^{(k)}(\tau)$ by

$$v_i^{(k)}(s;\tau) \ := \ \mathbf{E}\left[e^{-sV_i^{(k)}(\tau)}\right], \quad \mathrm{Re}(s) \geq 0,$$

and let $\overline{v}(s;\tau)$ be the vector with the $v_i^{(k)}(s;\tau)$ ordered lexicographically. In the following we derive an explicit expression for $\overline{v}(s;\tau)$.

Remark 3.2 In this section and in Section 4 we concentrate on the case where the $r_i^{(k)}$ are all *strictly positive*. In Section 5 we extend the analysis to the case where some of the $r_i^{(k)}$ may be zero. □

We need to introduce another generator:

$$\mathcal{G}^* := \begin{bmatrix} \tilde{D}^{(1)} & \Lambda^{(1)} & & & & & & \\ \frac{1}{2}M^{(2)} & \tilde{D}^{(2)} & \Lambda^{(2)} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \frac{k-1}{k}M^{(k)} & \tilde{D}^{(k)} & \Lambda^{(k)} & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \frac{L-2}{L-1}M^{(L-1)} & \tilde{D}^{(L-1)} & \Lambda^{(L-1)} \\ & & & & & \frac{L-1}{L}M^{(L)} & \tilde{D}^{(L)} \end{bmatrix}. \tag{3.2}$$

The matrices $\tilde{D}^{(k)}$ differ from the matrices $D^{(k)}$ only in their diagonal elements: These are such that each row of $\mathcal{G}^*$ sums to 0. The state space of a Markov process with generator $\mathcal{G}^*$ will be denoted by

$$\mathbf{S}^* := \{(k,i) \in \mathbf{S} \ : \ k > 0\}.$$

Remark 3.3 The matrix $\mathcal{G}^*$ is the generator of a Markov process, which we call the *process with one permanent customer:* Suppose we consider the queueing system of Figure 1, with the modification that there is one customer who never leaves the system, but shares in the service rate as if he were an ordinary customer. With that modification, the number of customers k ranges from 1 to L, and if we consider the queue length X_t^* and the state of the random environment Y_t^* at time t simultaneously, then the process $\{(X_t^*, Y_t^*),\ t \geq 0\}$ is a non-homogeneous QBD process with generator $\mathcal{G}^*$ defined by (3.2). □

Additionally, we define the diagonal matrix

$$\mathcal{R} := \mathrm{diag}\left[\frac{1}{k} r_i^{(k)}\right]_{(k,i)\in\mathbf{S}^*}.$$

The entries of $\mathcal{R}$ along the diagonal are ordered lexicographically in (k,i).
We may now formulate the following theorem:

Theorem 3.2 *If* $r_i^{(k)} > 0$ *for all* $(k,i) \in \mathbf{S}^*$*, then for* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$*,*

$$\frac{\partial}{\partial \tau}\overline{v}(s;\tau) = \mathcal{R}^{-1}\left[\mathcal{G}^* - s\,\mathcal{I}\right]\overline{v}(s;\tau), \tag{3.3}$$
$$\overline{v}(s;0) = \overline{1}; \tag{3.4}$$

and hence,

$$\overline{v}(s;\tau) = \mathrm{e}^{\tau\mathcal{R}^{-1}\left[\mathcal{G}^* - s\,\mathcal{I}\right]}\,\overline{1}. \tag{3.5}$$

Proof
The proof can be given by marginal analysis: When the state of the queueing system is $(k,i) \in \mathbf{S}^*$, the customer with a remaining amount of work τ (as well as all other customers) is served at rate $\frac{1}{k}r_i^{(k)}$.
Consider a small time interval of length $\frac{k\Delta}{r_i^{(k)}}$ and condition on the possible events occurring in this interval:

$$
\begin{aligned}
& v_i^{(k)}(s;\tau+\Delta) \\
= \;& \mathrm{e}^{-s\frac{k\Delta}{r_i^{(k)}}} \times \\
& \left\{ \lambda_i^{(k)} \frac{k\Delta}{r_i^{(k)}} \sum_j p_{ij}^{(k)} v_j^{(k+1)}(s;\tau) + \mu \frac{k-1}{k} r_i^{(k)} \frac{k\Delta}{r_i^{(k)}} \sum_j q_{ij}^{(k)} v_j^{(k-1)}(s;\tau) \right. \\
& \left. + \sum_{j\neq i} d_{ij}^{(k)} \frac{k\Delta}{r_i^{(k)}} v_j^{(k)}(s;\tau) + \left(1 + \tilde{d}_{ii}^{(k)} \frac{k\Delta}{r_i^{(k)}}\right) v_i^{(k)}(s;\tau) \right\} + \mathrm{o}(\Delta),
\end{aligned}
$$

with $\tilde{d}_{ii}^{(k)} = -\mu\frac{k-1}{k} r_i^{(k)} - \lambda_i^{(k)} - \sum_{j\neq i} d_{ij}^{(k)}$.
This leads to the differential equation (3.3). The initial conditions (3.4) follow from the fact that all $r_i^{(k)}$ are positive, and hence $\mathbf{E}\left[V_i^{(k)}(0+)\right] = 0$. It can then be verified that (3.5) is a solution to (3.3) and (3.4). Since the solution must be unique, we are done. ■

In the proof of the following corollary we use standard results for Markov reward processes. These processes fall within the framework of Markov decision theory (here no decisions are to be made). The first to present a systematic treatment of Markov reward processes on a finite state space seems to have been Howard [10]. In particular the results on continuous-time Markov reward processes (pp. 99–104) are of interest to us. The theory of Markov decision processes is presented in Puterman [16, Chapter 11]. Other references are Ross [17, Section 6.7] (only the discrete-time case), and Tijms [21, Section 3.5] (where the close relationship between the continuous-time case and the discrete-time case is exploited).
We use the symbol $\overline{1}_i^{(k)}$ to denote the vector with the entry in position (k,i) equal to 1, and all other entries equal to 0.

Corollary 3.3 *If $r_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \geq 0$,*

$$
\mathbf{E}\left[V_i^{(k)}(\tau)\right] = \frac{\tau}{c^* - \rho^*} + \overline{1}_i^{(k)} \left[\mathcal{I} - \mathrm{e}^{\tau \mathcal{R}^{-1}\mathcal{G}^*}\right] \overline{\gamma}, \tag{3.6}
$$

where

$$
\begin{aligned}
c^* &= \sum_{(k,i)\in\mathbf{S}^*} \pi^{*(k)}_i \cdot r_i^{(k)}, \\
\rho^* &= \sum_{(k,i)\in\mathbf{S}^*} \pi^{*(k)}_i \lambda_i^{(k)} \cdot \frac{1}{\mu},
\end{aligned}
$$

with $\overline{\pi}^* = (\pi^{*(k)}_i)_{(k,i)\in\mathbf{S}^*}$ the steady-state probability vector of the model with one permanent customer:

$$
\overline{\pi}^* \; \mathcal{G}^* = \overline{0}, \quad \overline{\pi}^* \; \overline{1} = 1.
$$

The vector $\overline{\gamma}$ satisfies

$$
-\mathcal{G}^* \; \overline{\gamma} = \overline{1} - \frac{1}{c^* - \rho^*} \; \mathcal{R}\overline{1}, \tag{3.7}
$$

and is unique up to translation by the vector $\overline{1}$. Expression (3.6) is, however, insensitive to such a translation. We may normalise $\overline{\gamma}$ such that $\overline{\pi}^*\mathcal{R}\overline{\gamma} = 0$.

Proof
The result can be obtained by differentiating (3.5) with respect to s, and setting $s = 0$. However, we give a

more direct proof. In the same way as we derived (3.3) and (3.4), we may find the following set of differential equations and initial conditions:

$$\frac{\partial}{\partial\tau}\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{k,i} = \mathcal{R}^{-1}\overline{1} + \mathcal{R}^{-1}\mathcal{G}^*\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{k,i}, \tag{3.8}$$

$$\mathbf{E}\left[V_i^{(k)}(0)\right] = 0, \quad \forall\ (k,i) \in \mathbf{S}^*. \tag{3.9}$$

By $(\cdot)_{k,i}$ we mean the vector with the entries between brackets ordered lexicographically in $(k,i) \in \mathbf{S}^*$. It is left to the reader to verify that there is at most one solution to this set of differential equations and initial conditions.

Suppose for the moment that a vector $\overline{\gamma}$ exists, satisfying (3.7) and normalised as required. By substitution of (3.6) into (3.8) and (3.9), we may verify that these differential equations and initial conditions are satisfied, and hence (3.6) is the unique solution. Note that $\mathcal{G}^*\overline{1} = \overline{0}$, since $\mathcal{G}^*$ is the generator of a Markov process.

From (3.7) we note that if the vector $\overline{\gamma}$ exists, it may be interpreted as the relative-cost vector in a Markov reward process, with generator $\mathcal{G}^*$, and costs at rate $1 - \frac{1}{c^*-\rho^*}\cdot\frac{1}{k}r_i^{(k)}$ when the process is in state $(k,i) \in \mathbf{S}^*$. In order for (3.7) to have a solution, it is necessary that this Markov reward process should have average costs per time unit equal to 0, because the left-hand side is equal to zero if we pre-multiply by $\overline{\pi}^*$. Indeed,

$$\sum_{(k,i)\in\mathbf{S}^*} {\pi^*}_i^{(k)}\frac{1}{k}r_i^{(k)} = c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*}{\pi^*}_i^{(k)}\frac{k-1}{k}r_i^{(k)}\mu = c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*}{\pi^*}_i^{(k)}\lambda_i^{(k)} = c^* - \rho^*,$$

where the one-but-last equality sign is due to the fact that the average number of customers leaving the system per time unit equals the average number of customers entering the system per time unit. Since the state space is finite, the existence of a vector $\overline{\gamma}$, and its uniqueness up to translation along the vector $\overline{1}$, is guaranteed by standard Markov decision theory, see for instance [10, pp. 99–104], [17, Corollary 6.20], [21, Theorem 3.1.1] or [16, Theorem 11.4.3].

Note that translation along $\overline{1}$ of a vector $\overline{\gamma}$ satisfying (3.7) does not alter the solution for $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$, since all rows of the matrix $\mathcal{I} - \mathrm{e}^{\tau\mathcal{R}^{-1}\mathcal{G}^*}$ in (3.6) sum to 0 (because all rows of $\mathcal{G}^*$ do). Therefore, if $\overline{\gamma} = \overline{v}$ satisfies (3.7), then so does

$$\overline{\gamma} := \overline{v} - \frac{\overline{\pi}^*\mathcal{R}\overline{v}}{c^*-\rho^*}\overline{1},$$

which is normalised as required. ■

Remark 3.4 The entities c^* and ρ^* have the following interpretation. In the queueing system with one permanent customer, the service capacity not given to other customers is assigned to the permanent customer. The average capacity per unit of time available for all customers (including the permanent one) is c^*. Per unit of time, on average $\sum_{(k,i)\in\mathbf{S}^*}\lambda_i^{(k)}$ customers enter the system, each requiring an expected amount of work $1/\mu$. Therefore ρ^* is the average amount of work entering the system per unit of time. □

Corollary 3.4 *If $r_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \to \infty$ we have:*

$$\mathbf{E}\left[V_i^{(k)}(\tau)\right] - \frac{\tau}{c^*-\rho^*} \longrightarrow \gamma_i^{(k)}, \quad (k,i) \in \mathbf{S}^*.$$

Proof

Note that $\mathcal{R}^{-1}\mathcal{G}^*$ is the infinitesimal generator of an irreducible Markov process on a finite state space. Its largest eigenvalue is therefore equal to 0 and of multiplicity 1. The left and right eigenvectors corresponding to the eigenvalue 0 are $\frac{1}{c^*-\rho^*}\overline{\pi}^*\mathcal{R}$ and $\overline{1}$ (after proper scaling). As a consequence, the matrix $\lim_{\tau\to\infty}\mathrm{e}^{\tau\mathcal{R}^{-1}\mathcal{G}^*}$ exists and has all rows equal to the probability vector $\frac{1}{c^*-\rho^*}\overline{\pi}^*\mathcal{R}$. The corollary now follows from Expression (3.6). ■

Remark 3.5 For the special case of a constant service rate, $r_i^{(k)} = 1,\ \forall\,(k,i)$, and state independent arrivals (Poisson with rate λ), the model reduces to the M/M/1/L queue with processor sharing. It can be shown that for this case (the subscripts are omitted since there is no random environment):

$$\gamma^{(k+1)} - \gamma^{(k)} = \frac{1}{k\lambda\rho^k}\sum_{j=1}^{k}\left(\frac{1}{1-\rho^*} - j\right)\rho^j > 0, \quad k = 1,2,\dots,L-1.$$

Here, $\rho := \lambda/\mu$, and

$$\rho^* = \frac{\sum_{l=1}^{L-1} l\ \rho^{l-1}}{\sum_{l=1}^{L} l\ \rho^{l-1}} \cdot \rho.$$

Passing $L \to \infty$, we find in case $\rho < 1$:

$$\gamma^{(k+1)} - \gamma^{(k)} \overset{L\to\infty}{\longrightarrow} \frac{1}{\lambda}\cdot\frac{\rho}{1-\rho},$$

which indeed corresponds to the M/M/1 queue with processor sharing. For that model we may even explicitly find:

$$\mathbf{E}\left[V^{(k)}(\tau)\right] = \frac{\tau}{1-\rho} \ + \ \frac{1}{\mu-\lambda}\left(k - \frac{1}{1-\rho}\right)\left(1 - \mathrm{e}^{-\tau\,\{\mu-\lambda\}}\right),$$

cf. Coffman et al. [4, Formula 33] (there the *delay* is studied instead of the sojourn time, which gives a term $\frac{\rho\tau}{1-\rho}$ instead of $\frac{\tau}{1-\rho}$). □

4. Random time change

In the proof of Corollary 3.3 we already mentioned the interpretation of the coefficients $\gamma_i^{(k)}$ as relative costs in a Markov Reward process. In this section we explore such an interpretation further, making a link with the random time change method. In its most essential form, this method was introduced for the analysis of the M/G/1 processor sharing queue by Yashkov [24]. Foley and Klutke [7] studied the queue-length process and the process of accumulated work after applying the random time change. Grishechkin [8, 9] further exploited the method by reformulating it in terms of Crump-Mode-Jagers branching processes. The approach was used in [14] to study a processor sharing queue with an unreliable server.

As we mentioned in Remark 3.2, in this section we again restrict ourself to the case where all the service rates $r_i^{(k)}$ are strictly positive. In Section 5 we allow for some of them being equal to zero.
Our starting point is the Markov process (X_t^*, Y_t^*), i.e. the queue length and the state of the random environment in the queueing model of Figure 1, when there is one permanent customer in the system. This permanent customer shares in the service capacity as any other customer, but never leaves the system. We already saw that $\mathcal{G}^*$ is the infinitesimal generator of the process (X_t^*, Y_t^*).
We make a random time change in the following way. When (X_t^*, Y_t^*) is in state (k,i), all transitions out of this state are 'sped up' by a factor $\frac{k}{r_i^{(k)}}$. For instance, the new arrival rate of customers in state (k,i) is $\frac{k}{r_i^{(k)}}\lambda_i^{(k)}$. More importantly, the new departure rate of customers is exactly $(k-1)\mu$ in all states (k,i). Note that $k-1$ is the number of non-permanent customers in state $(k,i) \in \mathbf{S}^*$. Apparently, *in the new time scale*, each customer receives one unit of service per 'time' unit. By $\{(\chi_\sigma, \Psi_\sigma),\ \sigma \geq 0\}$ we denote the process of queue length and state of the random environment in the new time scale. The generator of the Markov process $(\chi_\sigma, \Psi_\sigma)$ is $\mathcal{R}^{-1}\mathcal{G}^*$. Note that the processes (X_t^*, Y_t^*) and $(\chi_\sigma, \Psi_\sigma)$ have *the same jump-chain*. By a jump-chain of a Markov process we mean the embedded Markov chain at transition epochs.

We now explain how the processes $\{(X_t^*, Y_t^*),\ t \geq 0\}$ and $\{(\chi_\sigma, \Psi_\sigma),\ \sigma \geq 0\}$ are related, using a coupling argument on their jump-chains: Suppose that at time $t = 0$, the process (X_t^*, Y_t^*) is in state (k_0, i_0) and observe the process as it evolves over time. For a given path of the process (X_t^*, Y_t^*), we may 'perform' the random time change as indicated above: For any period of time that (X_t^*, Y_t^*) resides in a state (k, i), we 'shrink' the length of this period by a factor $\frac{k}{r_i^{(k)}}$, i.e. we divide the length of the period by this number. We may so construct a path for the process $(\chi_\sigma, \Psi_\sigma)$, starting in (k_0, i_0) for $\sigma = 0$. In Figure 2 such a construction is depicted.

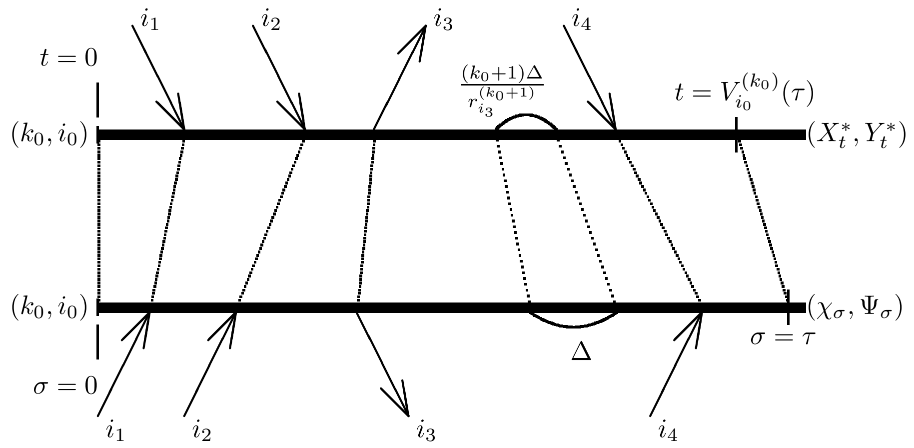

Figure 2: Coupling of the jump-chains of (X_t^*, Y_t^*) and $(\chi_\sigma, \Psi_\sigma)$.

Two horizontal axes are drawn. The upper-axis corresponds to the 'normal' time-axis on which we observe the process (X_t^*, Y_t^*) for $t \geq 0$. The lower-axis corresponds to the new 'time' scale, after the random time change. On this axis we observe the process $(\chi_\sigma, \Psi_\sigma)$.
In the realisation depicted in Figure 2 the following events happen successively: The process starts in (k_0, i_0), then a customer arrives and the random environment changes to state i_1, another customer arrives and the random environment changes to state i_2, a customer departs and the random environment changes to state i_3, and finally another customer arrives and the random environment moves to state i_4. Of course, the random environment may change without changing the number of customers, but for transparency of the picture no such event is drawn. Note that since both processes (X_t^*, Y_t^*) and $(\chi_\sigma, \Psi_\sigma)$ have the same jump-chain, any such realisation (indeed) occurs with the same probability in both processes.
Now concentrate on the indicated time-interval of length $\frac{(k_0+1)\Delta}{r_{i_3}^{(k_0+1)}}$ on the upper axis. This interval lies between the moment of the first departure and the moment of the third arrival. At any point in this interval the number of customers X_t^* (including the permanent one) is $k_0 + 1$, and the random environment Y_t^* is in state i_3. During this interval of time, the amount of service received by the permanent customer (and any other customer in the system), equals Δ. This argument can be used for any time interval during which the state does not change. It is seen that the amount of service received by the permanent customer between time $t = 0$ and the time point (on the upper axis) which corresponds to the point $\sigma = \tau$ (on the lower axis), is exactly τ. Therefore, the point on the upper axis corresponding to $\sigma = \tau$ on the lower axis is exactly $V_{i_0}^{(k_0)}(\tau)$: It is the amount of time that a customer must stay in the system before he has received an amount of service τ, starting at time $t = 0$ with no service received, $k_0 - 1$ other customers, and the random environment in state i_0.

We introduce the following cost structure on the process $(\chi_\sigma, \Psi_\sigma)$: In state (k, i) costs are incurred at rate $\frac{k}{r_i^{(k)}}$.

Theorem 4.1 *The sojourn time $V_i^{(k)}(\tau)$ of a customer in the queueing system of Figure 1, arriving when there are $k-1$ other customers in the system, the random environment being in state i, and bringing an amount of work τ, is distributed as the costs incurred in the process $(\chi_\sigma, \Psi_\sigma)$ over the interval $\sigma \in (0,\tau)$, starting at $\sigma = 0$ in state (k,i).*

Proof
From our construction of the coupled (jump-)processes above, it follows that the accumulated costs in the process $(\chi_\sigma, \Psi_\sigma)$ over the interval $\sigma \in (0,\tau)$ on the lower axis, are equal to $V_{i_0}^{(k_0)}(\tau)$ on the upper axis (Figure 2). As we already remarked, any such realisation has the same probability for both processes (X_t^*, Y_t^*) and $(\chi_\sigma, \Psi_\sigma)$. ■

From Theorem 4.1 we may obtain the result of Corollary 3.4, which is restated in the following corollary:

Corollary 4.2 *With probability 1:*

$$\lim_{\tau\to\infty} \frac{V_i^{(k)}(\tau)}{\tau} = g^* := \mathbf{E}\left[\frac{\chi}{\mu r_\Psi^{(\chi)}}\right], \tag{4.1}$$

where the distribution of (χ, Ψ) is the equilibrium distribution of $(\chi_\sigma, \Psi_\sigma)$.
Furthermore, the limit

$$\lim_{\tau\to\infty} \mathbf{E}\left[V_i^{(k)}(\tau)\right] - g^*\tau, \tag{4.2}$$

exists and is finite.

Proof
The result is standard for irreducible Markov reward processes with a finite state space, see for instance [10, pp. 99–104], [17, Corollary 6.20], [21, Theorem 3.1.1] or [16, Theorem 11.4.3]. ■

Remark 4.1 Equation (4.1) holds under much more general assumptions. In fact, with $L < \infty$, $N < \infty$, and all $r_i^{(k)} > 0$, it can be proved using the Renewal Reward theorem (e.g. [17, Theorem 3.16] or [21, Theorem 1.3.1]) under the sole assumption that the original process (X_t, Y_t) is regenerative with finite expected regeneration time. However, the limit in (4.2) may not exist under such general assumptions. □

Remark 4.2 Corollaries 3.4 and 4.2 imply that $g^* = \frac{1}{c^*-\rho^*}$, or equivalently:

$$\mathbf{E}\left[\frac{\chi}{r_\Psi^{(\chi)}}\right] = \left(\mathbf{E}\left[\frac{r_{Y^*}^{(X^*)}}{X^*}\right]\right)^{-1}.$$

This can be verified by noting that the distribution of (χ, Ψ) is given by the vector $\frac{1}{c^*-\rho^*} \cdot \overline{\pi}^*\mathcal{R}$, and that the cost vector in that process is given by $\mathcal{R}^{-1}\overline{1}$. □

Remark 4.3 The fact that

$$\mathbf{E}\left[\frac{r_{Y^*}^{(X^*)}}{X^*}\right] = c^* - \rho^*,$$

can be argued as follows: As we saw in Remark 3.4, c^* is the average capacity per unit of time available for all customers, and ρ^* is the average amount of work entering the system per unit of time. Since all non-permanent customers eventually leave the system, ρ^* is also the average amount of service capacity assigned to non-permanent customers (in the long run). Hence, $c^* - \rho^*$ is the average capacity per unit of time assigned to the permanent customer. □

Remark 4.4 The process $(\chi_\sigma, \Psi_\sigma)$ is closely related to the branching process which is used in the literature to study sojourn times in traditional processor sharing systems with constant service capacity ($r_i^{(k)} = 1$, there is no random environment), constant arrival rate (λ), and infinite space for customers ($L = \infty$). We refer to [8, 9] and [14] for details regarding the branching process approach. Here, we only briefly address the most important aspects. After the random time change, customers may be seen as individuals in a population, one of them having an infinite life time (corresponding to the permanent customer), and all others with an exponentially distributed life time with mean $1/\mu$ (independent of everything else). Thus, with $k \geq 1$ individuals in the population (including the permanent one) the total 'death' rate is $(k-1)\mu$. Each of the k individuals gives birth to new individuals at rate λ, the total birth rate is thus $k\lambda$. Clearly, the evolution of each individual and all his descendants is independent of all other individuals, which makes this branching process very suitable for analysis. In fact, the approach is applicable to other service disciplines, including *discriminatory processor sharing*, see [9] and Section 8.2.
In our case, the branching process evolves depending on a random environment. The life time of non-permanent individuals is still exponentially distributed with mean $1/\mu$. If the state of the random environment is i then each individual (including the permanent one) gives birth to new individuals with rate $\frac{\lambda_i^{(k)}}{r_i^{(k)}}$. This depends both on the state of the random environment, and on the number of individuals in the population. The random environment also evolves dependent on the number of individuals. With k living individuals, the random environment may change from state i to state j with rate $k\,\frac{d_{ij}^{(k)}}{r_i^{(k)}}$. The mutual dependence of the branching process and the random environment and the dependence among individuals make this approach less suitable for analysis. □

5. Server unavailability

In this section we extend the analysis of Sections 3 and 4 to the case where some of the $r_i^{(k)}$ may be equal to zero, i.e. there are periods during which no service is provided to the customers. The case with $L = \infty$ and the service capacity alternating between a positive value (for exponentially distributed periods of time) and 0 (during generally distributed periods of time) is analysed in [14].
We define the subset of states

$$\mathbf{S}_0^* := \left\{(l,j) \in \mathbf{S}^* \ : \ r_j^{(l)} = 0\right\}.$$

In applications the fact whether $r_i^{(k)} = 0$ will typically only depend on i, but for generality of the presentation we do not assume this. Partition the state space $\mathbf{S}^*$ into $\mathbf{S}_0^*$ and its complement $\mathbf{S}_+^* := \mathbf{S}^* - \mathbf{S}_0^*$, and 're-order' the rows and the columns of the generator $\mathcal{G}^*$ accordingly:

$$\mathcal{G}^* = \begin{bmatrix} \mathcal{G}_+^* & \mathcal{G}_{+0}^* \\ \mathcal{G}_{0+}^* & \mathcal{G}_0^* \end{bmatrix}.$$

A little reflection shows that if the states within $\mathbf{S}_+^*$ and those within $\mathbf{S}_0^*$ are ordered lexicographically, then $\mathcal{G}_+^*$ and $\mathcal{G}_0^*$ are the generators of (possibly reducible) transient QBD processes. We also re-order the entries of $\overline{\pi}^* = \left(\overline{\pi}_+^*, \overline{\pi}_0^*\right)$, with $\overline{\pi}_+^*$ and $\overline{\pi}_0^*$ vectors with their entries ordered lexicographically. Starting from any $(l,j) \in \mathbf{S}_0^*$, let $U_j^{(l)}$ be the amount of time the process remains in the set $\mathbf{S}_0^*$, and $W_j^{(l)} \in \mathbf{S}_+^*$ the first state that is visited after leaving $\mathbf{S}_0^*$. Note that $U_j^{(l)}$ is the sojourn time (or time until exit) in a transient QBD process, for which an efficient routine to compute moments of the distribution can be found in [12]. Furthermore, for $\mathrm{Re}(s) \geq 0$, define the matrix $\mathcal{U}(s)$ of dimension $|\mathbf{S}_0^*| \times |\mathbf{S}_+^*|$ with entries:

$$\mathcal{U}_{(l,j),(k,i)}(s) := \mathbf{E}\left[\mathrm{e}^{-sU_j^{(l)}} \mathbf{1}_{\left\{W_j^{(l)}=(k,i)\right\}}\right], \quad (l,j) \in \mathbf{S}_0^*,\ (k,i) \in \mathbf{S}_+^*.$$

Here $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Note that, in particular, $\mathcal{U}(0)$ is a probability matrix, and that $-\frac{\partial}{\partial s}\mathcal{U}(s)|_{s=0}\overline{1} = \left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*}$.

Lemma 5.1 *The matrix* $\mathcal{U}(s)$ *is given by*

$$\mathcal{U}(s) = -\left[\mathcal{G}_0^* - s\mathcal{I}\right]^{-1} \mathcal{G}_{0+}^*, \quad \mathrm{Re}(s) \geq 0,$$

and hence

$$\left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} = [-\mathcal{G}_0^*]^{-1}\,\overline{1}.$$

Proof
By conditioning on the possible transitions in an interval Δ when we start from any state in $\mathbf{S}_0^*$ we find for $\Delta \downarrow 0$:

$$\mathcal{U}(s) = \mathrm{e}^{-\Delta s}\left([\mathcal{I} + \Delta \cdot \mathcal{G}_0^*]\mathcal{U}(s) + \Delta \cdot \mathcal{G}_{0+}^*\right) + \mathrm{o}(\Delta) = \left(\mathcal{I} + \Delta \cdot [\mathcal{G}_0^* - s\mathcal{I}]\right)\, \mathcal{U}(s) + \Delta \cdot \mathcal{G}_{0+}^* + \mathrm{o}(\Delta),$$

where $\mathrm{o}(\Delta)$ applies to each entry in the matrix equations. Canceling terms, dividing by Δ, and taking $\Delta \downarrow 0$ we have:

$$-\left[\mathcal{G}_0^* - s\mathcal{I}\right]\, \mathcal{U}(s) = \mathcal{G}_{0+}^*, \quad \mathrm{Re}(s) \geq 0. \tag{5.1}$$

Since $\mathcal{G}_0^*$ is a transient generator, $\mathcal{G}_0^* - s\mathcal{I}$ is invertible for all $\mathrm{Re}(s) \geq 0$, and hence the first statement of the lemma follows.
Differentiating (5.1) with respect to s, setting $s = 0$, and using the fact that $\mathcal{U}(0)$ is a probability matrix (so that $\mathcal{U}(0)\overline{1} = \overline{1}$), we may prove the second statement of the lemma. ∎

As before, denote by $v_i^{(k)}(s;\tau)$ the LST of $V_i^{(k)}(\tau)$, the (remaining) sojourn time of a customer with a (remaining) amount of work τ, starting in state $(k,i) \in \mathbf{S}^*$. Construct the vectors $\overline{v}_0(s;\tau) = \left(v_j^{(l)}(s;\tau)\right)_{(l,j)\in\mathbf{S}_0^*}$ and $\overline{v}_+(s;\tau) = \left(v_i^{(k)}(s;\tau)\right)_{(k,i)\in\mathbf{S}_+^*}$ according to the partitioning $\mathbf{S}^* = \mathbf{S}_0^* \cup \mathbf{S}_+^*$. The following lemma gives the relation between the two vectors.

Lemma 5.2 *For* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$*:*

$$\overline{v}_0(s;\tau) = \mathcal{U}(s) \cdot \overline{v}_+(s;\tau), \tag{5.2}$$

and in particular,

$$\left(\mathbf{E}\left[V_j^{(l)}(\tau)\right]\right)_{(l,j)\in\mathbf{S}_0^*} = \left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*}. \tag{5.3}$$

Proof
The proof of the first part is immediate by noting that (i) as long as the system is in $\mathbf{S}_0^*$ no service is received, and (ii) the LST of the joint distribution of the first state visited in $\mathbf{S}_+^*$ and the time until that moment is given by the matrix $\mathcal{U}(s)$. The second part follows by differentiating with respect to s and putting $s = 0$. ∎

With the aid of the two preceding lemmas we are able to prove the following theorem, which generalises Theorem 3.2 to the case $\mathbf{S}_0^* \neq \emptyset$. Before proceeding, we define the matrix

$$\mathcal{R}_+ := \mathrm{diag}\left[\frac{1}{k} r_i^{(k)}\right]_{(k,i)\in\mathbf{S}_+^*},$$

with the entries along the diagonal ordered lexicographically in $(k,i) \in \mathbf{S}_+^*$. Note that $\mathcal{R}_+^{-1}$ is well-defined.

Theorem 5.1 *For* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$*,*

$$\begin{aligned} \frac{\partial}{\partial\tau}\overline{v}_+(s;\tau) &= \mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\, \mathcal{U}(s) - s\mathcal{I}\right]\overline{v}_+(s;\tau), \\ \overline{v}_+(s;0) &= \overline{1}; \end{aligned}$$

and hence,

$$\overline{v}_+(s;\tau) = \mathrm{e}^{\tau\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\, \mathcal{U}(s) - s\mathcal{I}\right]}\,\overline{1}.$$

Proof
The proof may proceed as that for Theorem 3.2: For any $(k,i) \in \mathbf{S}_+^*$ derive the differential equation for $v_i^{(k)}(s;\tau)$ by conditioning on the possible events in a time interval $\frac{k}{r_i^{(k)}}\Delta$, and then take $\Delta \downarrow 0$. Substituting (5.2) for $\overline{v}_0(s;\tau)$ readily leads to the desired result. ■

Consequently we have the following corollary, which generalises Corollaries 3.3 and 3.4:

Corollary 5.2 *For $\tau \geq 0$,*

$$\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*} = \frac{\tau}{c^*-\rho^*}\overline{1} + \left[\mathcal{I} - \mathrm{e}^{\tau\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]}\right]\overline{\gamma},$$

where

$$c^* = \sum_{(k,i)\in\mathbf{S}^*} \pi^{*(k)}_i \cdot r_i^{(k)},$$

$$\rho^* = \sum_{(k,i)\in\mathbf{S}^*} \pi^{*(k)}_i \lambda_i^{(k)} \cdot \frac{1}{\mu}.$$

The vector $\overline{\gamma}$ satisfies

$$-\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]\ \overline{\gamma} = \overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{1}{c^*-\rho^*}\ \mathcal{R}_+\overline{1},$$

and is uniquely determined by normalising such that $\overline{\pi}^*\mathcal{R}_+\overline{\gamma} = 0$.
Consequently we have for $(k,i) \in \mathbf{S}_+^*$:

$$\mathbf{E}\left[V_i^{(k)}(\tau)\right] - \frac{\tau}{c^*-\rho^*} \longrightarrow \gamma_i^{(k)}.$$

Proof
Similar to (3.8) and (3.9) we may derive differential equations for $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$, $(k,i) \in \mathbf{S}_+^*$. Using (5.3) we get:

$$\begin{aligned}\frac{\partial}{\partial\tau}\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*} &= \mathcal{R}_+^{-1}\left[\overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*}\right] \\ &\quad + \mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*}.\end{aligned}$$

Of course, $\mathbf{E}\left[V_i^{(k)}(0)\right] = 0$, for $(k,i) \in \mathbf{S}_+^*$. To see that the solution given in the corollary satisfies this set of differential equations and initial solutions note that the vector $\overline{\gamma}$ may be interpreted as the vector of relative costs in a Markov reward process with generator $\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)$, costs being incurred according to the vector $\overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{1}{c^*-\rho^*}\ \mathcal{R}_+\overline{1}$. It remains to be shown that the average costs in this Markov reward process equal zero. The steady-state probability vector of this process is given by $\frac{1}{\overline{\pi}_+^*\overline{1}} \cdot \overline{\pi}_+^*$, and using (cf. Lemma 5.1),

$$\overline{\pi}_+^*\mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} = \overline{\pi}_+^*\mathcal{G}_{+0}^*\left[-\mathcal{G}_0^*\right]^{-1}\overline{1} = \overline{\pi}_0^*\overline{1},$$

we indeed find that the costs per unit time in steady state equal 0.
The limit as $\tau \to \infty$ can be obtained as in the proof of Corollary 3.4. ■

Corollary 5.3 *For $\tau \to \infty$,*

$$\left(\mathbf{E}\left[V_j^{(l)}(\tau)\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{\tau}{c^*-\rho^*}\overline{1} \longrightarrow \left(\mathbf{E}\left[U_j^{(l)}\right]\right)_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\overline{\gamma}.$$

Proof
By Lemma 5.2, Corollary 5.2, and using the fact that $\mathcal{U}(0)$ is a probability matrix. ■

The remainder of this section is dedicated to the method of random time change in the case that $\mathbf{S}_0^* \neq \emptyset$. A similar approach is used in [14] for the M/M/1 queue with processor sharing and server breakdowns. We mimic the arguments of Section 4, with the following modifications: All transitions which occur when the process (X_t^*, Y_t^*) is in the set $\mathbf{S}_0^*$ are 'collapsed' into one single event when constructing the process $(\chi_\sigma, \Psi_\sigma)$. More precisely: When the process (X_t^*, Y_t^*) is in some state in $\mathbf{S}_+^*$, we change the time scale as before, speeding up all transitions out of state $(k,i) \in \mathbf{S}_+^*$ by a factor $\frac{k}{r_i^{(k)}}$. The difference with the case $\mathbf{S}_0^* = \emptyset$ is that when the state is $(l,j) \in \mathbf{S}_0^*$ this time transformation can not be done since $r_j^{(l)} = 0$. Suppose that at some time $t \geq 0$ the process (X_t^*, Y_t^*) changes from state $(k,i) \in \mathbf{S}_+^*$ to some state in $\mathbf{S}_0^*$. Suppose further that the first state within $\mathbf{S}_+^*$ visited thereafter is $(l,j) \in \mathbf{S}_+^*$. If $\sigma \geq 0$ is the point on the transformed time-scale corresponding to time t, then the process $(\chi_\sigma, \Psi_\sigma)$ makes a (direct) transition at the point σ from $(k,i) \in \mathbf{S}_+^*$ to $(l,j) \in \mathbf{S}_+^*$. Thus $(\chi_\sigma, \Psi_\sigma)$ *is not observed* on the states in $\mathbf{S}_0^*$. When $(\chi_\sigma, \Psi_\sigma)$ makes such a transition corresponding to a visit of (X_t^*, Y_t^*) to the set $\mathbf{S}_0^*$, an immediate cost is incurred which is equal to the time that (X_t^*, Y_t^*) spends within the set $\mathbf{S}_0^*$.

In the process $(\chi_\sigma, \Psi_\sigma)$ with state space $\mathbf{S}_+^*$ and generator $\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]$ there are two types of transitions and two types of costs. 'Ordinary' transitions occur according to the transient generator $\mathcal{G}_+^*$, and 'ordinary' costs are incurred at rate $\frac{k}{r_i^{(k)}}$ in state $(k,i) \in \mathbf{S}_+^*$. The other transitions and costs are related as follows. The entry in row $(k,i) \in \mathbf{S}_+^*$ and column $(l,j) \in \mathbf{S}_0^*$ of the matrix $\mathcal{G}_{+0}^*$ gives the rate with which an (l,j)-event occurs. An (l,j)-event has two consequences: (i) a transition is made, and (ii) instantaneous costs are incurred. The instantaneous costs incurred, and the state after an (l,j)-event are jointly distributed as the pair $(U_j^{(l)}, W_j^{(l)})$, and (the LST of) their joint distribution is given by the matrix $\mathcal{U}(s)$ for $\mathrm{Re}(s) \geq 0$. Note that the state after the (l,j) event may be the same as the state before the event. We emphasise that the jump-chains of the process $(\chi_\sigma, \Psi_\sigma)$ and the process (X_t^*, Y_t^*) *restricted to the set* $\mathbf{S}_+^*$ (often called a *censored process*) are identical.
Theorem 4.1 remains true with the above modifications, and so does Corollary 4.2 if we redefine the constant g^* as

$$g^* := \frac{1}{\overline{\pi}^* \mathcal{R}_+ \overline{1}}.$$

Remark 5.1 The discussion in Remark 4.1 also applies to this section. The relation $g^* = \frac{1}{c^* - \rho^*}$ discussed in Remark 4.2 is true for the redefined constants g^*, c^*, and ρ^*. As in Remark 4.3, in the queueing system with one permanent customer, $c^* - \rho^*$ is the average capacity per unit of time assigned to the permanent customer. Remark 4.4 only needs to be modified to account for the transitions with instantaneous costs, see for instance [14]. □

6. The proportionality result

We now discuss the proportionality between the conditional mean sojourn time and the amount of work brought into the system, in processor sharing systems without random environment. This result is well-known for the M/G/1 queue with processor sharing, see for instance Sakata et al. [18, Formula 10], Sakata et al. [19, Formula 49], or Kleinrock [11, Formula 4.17]. Cohen [5, Formula 7.27] found the proportionality property for the M/G/1/L with processor sharing and queue-dependent total service capacity (there called generalised processor sharing).
In this section we explain *why* this proportionality property holds, using the results from the random time change method of Section 4. Note that, since there is no random environment, this discussion only applies to the case with $\mathbf{S}_0^* = \emptyset$: If $r^{(k)} = 0$ for some $k \geq 1$, then the states with less than k customers are transient. For the M/G/1 queue with queue-dependent service rates the same arguments were used by Foley and Klutke [7]. We show that the arguments also apply to the M/G/1/L processor sharing system with queue-dependent

total service rates. A related discussion for the M/G/1 queue is given in Van den Berg [2, Remark 5.10, p. 115], and Van den Berg and Boxma [3, Remark 8.2].

In the absence of a random environment and with queue-independent arrivals (at rate λ), the queue length process $\{X_t,\ t \geq 0\}$ is an ordinary birth-death process. The queueing models of Remark 3.5 (M/M/1/L and M/M/1) possess these properties. Note however, that (unlike the M/M/1/L and M/M/1 models) the service rates may be queue-dependent, i.e. the $r^{(k)}$ may be different for different $k = 1, 2, \ldots, L$. The steady-state probabilities $\pi^{(k)},\ k = 0, 1, \ldots, L$ of the process X_t, and the steady-state probabilities $\pi^{*(k)}_{,}$ $k = 1, 2, \ldots, L$ – *not including* $k = 0$ – of the process X_t^*, satisfy:

$$\pi^{*(k)} \frac{1}{k} r^{(k)} \sim \pi^{(k-1)}, \quad k = 1, 2, \ldots, L,$$

where the symbol $\sim$ means equality up to multiplication by a constant (independent of k). We already saw in Remark 4.2 that the steady-state distribution of the process $(\chi_\sigma, \Psi_\sigma)$ is given by the vector $\frac{1}{c^*-\rho^*} \cdot \overline{\pi}^* \mathcal{R}$. For the present case, in the absence of a random environment, we thus have for $k = 1, 2, \ldots, L$, that $\mathbf{P}\{\chi = k\} = \frac{1}{c^*-\rho^*} \pi^{*(k)} \cdot \frac{r^{(k)}}{k}$, and hence, $\mathbf{P}\{\chi = k\} = \frac{\pi^{(k-1)}}{1-\pi^{(L)}}$.
This property has an interesting consequence: Suppose the queueing system under consideration is in steady state, and let the random variable X be distributed according to this distribution: $\mathbf{P}\{X = k\} = \pi^{(k)}$. Since we assumed Poisson arrivals, from the PASTA (Poisson Arrivals See Time Averages) property, the number of customers seen by a newly arrived customer is distributed as X. Condition on the fact that the new customer is accepted, which occurs with probability $\mathbf{P}\{X < L\} = 1 - \pi^{(L)}$. Let the amount of work of the new (tagged) customer be $\tau > 0$, and denote his sojourn time by the random variable $V(\tau)$. Theorem 4.1 tells us that $V(\tau)$ is distributed as

$$\int_{\sigma=0}^{\tau} \frac{\chi_\sigma}{r^{(\chi_\sigma)}} \mathrm{d}\sigma,$$

with χ_0 distributed as $X+1$ *given that* $X < L$. However, this distribution is the steady-state distribution of the process χ_σ, and so $\mathbf{P}\{\chi_\sigma = k\} = \frac{1}{1-\pi^{(L)}} \pi^{(k-1)}$, $k = 1, 2, \ldots, L$, for any $\sigma \in [0, \tau]$. Therefore, in steady state we find for the mean of the sojourn time $V(\tau)$ (of an *accepted* customer with an amount of work τ):

$$\mathbf{E}[V(\tau)] = g^*\ \tau.$$

So, for the model with exponentially distributed service requirements we have explained why this proportionality occurs, namely because the stationary distribution of χ_σ is the same as that of X *given that* $X < L$. We can generalise our arguments to the M/G/1/L queue with processor sharing and queue-dependent service rates, cf. [5, Formula 7.27] (for $L = \infty$ this was done in [7]). We give a brief outline of the proof: If the service requirements are distributed according to the distribution $B(x)$, $x \geq 0$, then

$$p_k(x_1, \ldots, x_k) = p_0 \frac{\lambda^k}{\prod_{j=1}^{k} r^{(j)}} \cdot \prod_{j=1}^{k} \left(1 - B(x_j)\right), \quad k = 1, 2, \ldots, L,$$

is the density function of there being k customers in the system with respective remaining service requirements $x_1, \ldots, x_k$, see [5, Formula (5.9)]. For this model we may apply the random time change to the system with one permanent customer, as described above: I.e., we 'shrink' the time-scale by a factor $\frac{k}{r^{(k)}}$ when there are k customers in the system. Viewing the resulting process as a branching process, then $p_k(x_1, \ldots, x_k)$, for $k < L$, is also the density function (up to normalisation) of there being $k+1$ living individuals, the k non-permanent ones having respective remaining life times $x_1, \ldots, x_k$. It is beyond our purposes to work out the details at this point.

Remark 6.1 Queue-dependent arrival rates destroy the proportionality property. The steady-state distribution seen upon arrival, or at arbitrary time points, no longer equals the steady-state distribution of the time-changed process. Under exponentiality assumptions this is easily checked by comparing the balance equations. □

Remark 6.2 A related result regarding the proportionality property was obtained in [5, Theorem 5.3]. The model studied there is a closed queueing model with L customers, who are served according to the processor sharing discipline with queue-dependent service rates. After having completed his service, a customer waits for a generally distributed time, and then enters the system again with a new (independently drawn) service requirement. It is shown that if an exogenous customer with an amount of work τ is brought into the system in steady state, his mean sojourn time is proportional to τ. In this model, the arrival process is obviously queue-dependent, and hence the proportionality result seems to contradict Remark 6.1. However, the considered model in [5] is essentially different from the above models: The exogenous customer may cause the number of customers in service to become $L+1$. Moreover, the arrival process is still determined by the ordinary customers, so that the queue-dependent arrivals in the original process and the time-changed process 'cancel out'. Again, under exponentiality assumptions this is easily seen from the balance equations. □

7. Computation and approximation

We return to the general queueing model of Figure 1. Let $V(\tau)$ be the sojourn time of a customer with an amount of work τ, arriving to the system in steady state. For $(k,i) \in \mathbf{S}^*$, denote by $a_i^{(k)}$ the steady-state probability that the system is in state (k,i) immediately after the arrival of a customer.

Remark 7.1 In our presentation we required $\lambda^{(L)} = 0$ so that no customers are lost. In many applications the arrival process is a Poisson process, and customers arriving when there are L other customers present are lost. Then it must be explicitly stated that the sojourn time of a customer is conditional on this customer not being rejected. As said before, this conditioning is inherent to our formulation. Poisson arrivals are thus incorporated by defining $\lambda^{(L)} = 0$ and $\lambda^{(k)} = \lambda$, $k = 0, 1, \ldots, L-1$. □

The $a_i^{(k)}$ are the steady-state probabilities of a discrete-time Markov chain with transition probability matrix:

$$\mathcal{A} := \begin{bmatrix} T^{(1,1)} & T^{(1,0)} & 0 & \cdots & \cdots & 0 \\ T^{(2,2)} & T^{(2,1)} & T^{(2,0)} & 0 & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & T^{(L-1,2)} & T^{(L-1,1)} & T^{(L-1,0)} \\ T^{(L,L)} & \cdots & & \cdots & T^{(L,2)} & T^{(L,1)} \end{bmatrix}.$$

Here,

$$\begin{aligned} T^{(k,0)} &= \left[-D^{(k)}\right]^{-1} \Lambda^{(k)}, \quad k = 1, \ldots, L-1, \\ T^{(k,n)} &= \prod_{m=0}^{n-1} \left(\left[-D^{(k-m)}\right]^{-1} M^{(k-m)} \right) \left[-D^{(k-n)}\right]^{-1} \Lambda^{(k-n)}, \quad k = 1, \ldots, L, \; n = 1, \ldots, k. \end{aligned}$$

From Corollary 3.4 we have for the mean of $V(\tau)$:

$$\lim_{\tau \to \infty} \mathbf{E}\left[V(\tau)\right] - \frac{\tau}{c^* - \rho^*} = \gamma := \sum_{(k,i) \in \mathbf{S}^*} a_i^{(k)} \gamma_i^{(k)}.$$

We now discuss the possibility of using this asymptotic result to approximate $\mathbf{E}\left[V(\tau)\right]$ for $\tau \geq 0$. We propose an approximation based on numerical experiments for the special models of [15]. Before doing so we show how the exact value of $\mathbf{E}\left[V(\tau)\right]$ can be computed. We do this for the case $\mathbf{S}_0^* = \emptyset$, see Remark 7.3 for the case $\mathbf{S}_0^* \neq \emptyset$. Our starting point is the set of differential equations and initial conditions given in (3.8) and (3.9). Obviously, for $n \geq 1$,

$$\frac{\partial^n}{\partial \tau^n} \left(\mathbf{E}\left[V_i^{(k)}(\tau) \right] \right)_{k,i} |_{\tau=0} = \left(\mathcal{R}^{-1} \mathcal{G}^* \right)^{n-1} \mathcal{R}^{-1} \overline{1}. \tag{7.1}$$

We use Jensen's method to uniformise the generator $\mathcal{R}^{-1}\mathcal{G}^*$, and define the probability matrix

$$\mathcal{P}^* := \mathcal{I} + \frac{1}{\eta}\mathcal{R}^{-1}\mathcal{G}^*,$$

with the scalar $\eta > 0$ being equal to minus the entry with largest absolute value (along the diagonal) of $\mathcal{R}^{-1}\mathcal{G}^*$. Assuming the Taylor-series of $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$ around $\tau = 0$ exists (at the end we verify the result), and using (7.1) we may find:

$$\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{k,i} = \frac{1}{\eta}\sum_{l=0}^{\infty}\left(1 - \mathrm{e}^{-\eta\tau}\sum_{k=0}^{l}\frac{(\eta\tau)^k}{k!}\right)(\mathcal{P}^*)^l \cdot \mathcal{R}^{-1}\overline{1}. \tag{7.2}$$

Noting that $k! \geq (l+1)! \cdot (k-l-1)!$, when $0 < l+1 \leq k$, we have:

$$0 \leq 1 - \mathrm{e}^{-\eta\tau}\sum_{k=0}^{l}\frac{(\eta\tau)^k}{k!} = \frac{\sum_{k=l+1}^{\infty}\frac{(\eta\tau)^k}{k!}}{\mathrm{e}^{\eta\tau}} \leq \frac{(\eta\tau)^{l+1}}{(l+1)!},$$

and hence the infinite sum in (7.2) exists for every $\tau \geq 0$. Moreover, by substitution it may be seen that it satisfies the differential equations and initial conditions (3.8) and (3.9).

Expression (7.2) for the $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$ provides a numerically stable algorithm, since it only involves multiplication and addition of positive terms. Within the summation one needs to evaluate the 'coefficients' $\mathrm{e}^{-\eta\tau}\sum_{k=l+1}^{\infty}\frac{(\eta\tau)^k}{k!}$, which can be done accurately by proper scaling of the terms (to avoid problems when $\eta\tau$ is large).

Remark 7.2 Instead of starting from the differential equations (3.8), we may start from the final Expression (3.6) in Corollary 3.3. Again we may use Jensen's uniformisation method to derive:

$$\left(\mathbf{E}\left[V_i^{(k)}(\tau)\right]\right)_{k,i} = \frac{\tau}{c^* - \rho^*}\overline{1} + \overline{\gamma} - \mathrm{e}^{-\eta\tau} \cdot \mathrm{e}^{\eta\tau\mathcal{P}^*} \cdot \overline{\gamma}. \tag{7.3}$$

However, for this approach one first needs to compute the vector $\overline{\gamma}$. Moreover, the vector $\overline{\gamma}$ contains negative elements which may cause the evaluation of $\mathrm{e}^{-\eta\tau} \cdot \mathrm{e}^{\eta\tau\mathcal{P}^*} \cdot \overline{\gamma}$ to be numerically unstable. However, no problems were encountered in the numerical experiments of [15], where both methods were used to compute the exact value of $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$. In all cases the relative difference between the outcomes was of the order 10^{-8} or smaller (with values of τ up to 10 times the mean $1/\mu$). □

Remark 7.3 When $\mathbf{S}_0^* \neq \emptyset$ we may proceed in a similar way. The starting point is then the set of differential equations mentioned in the proof of Theorem 5.1. For $(k,i) \in \mathbf{S}_+^*$, the $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$ are found as before. However, first the $\mathbf{E}\left[U_j^{(l)}\right]$, for $(l,j) \in \mathbf{S}_0^*$, and the probability matrix $\mathcal{U}(0)$, need to be computed. Using Lemma 5.2, from the $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$, $(k,i) \in \mathbf{S}_+^*$, also the $\mathbf{E}\left[V_j^{(l)}(\tau)\right]$, for $(l,j) \in \mathbf{S}_0^*$ can be computed. Note that $\mathbf{E}\left[V_j^{(l)}(0+)\right] = \mathbf{E}\left[U_j^{(l)}\right] > 0$, for $(l,j) \in \mathbf{S}_0^*$. As a consequence $\mathbf{E}\left[V(0+)\right] > 0$, unless $p_{ij}^{(l-1)} = 0$, $\forall (l,j) \in \mathbf{S}_0^*$, $i \in E^{(l-1)}$. □

Although Expression (7.2) provides a numerically stable algorithm to compute the $\mathbf{E}\left[V_i^{(k)}(\tau)\right]$, in general this task requires considerable computation time and memory space. Therefore we proceed to find a good approximation which is less computationally demanding. We do this for the models of [15], for which we have done extensive numerical experiments (only for $\mathbf{S}_0^* = \emptyset$).

Remark 7.4 Note that the models of [15] form a very restricted subclass of the general framework depicted in Figure 1. In particular, there the capacity allocated to a single customer, $\frac{r_i^{(k)}}{k}$, is a non-increasing function of k (the total number of customers). However, for practical situations this seems to be a reasonable assumption. □

When the number of states of the random environment is small ($N \leq 5$), the asymptotic result may serve as a useful approximation for $\mathbf{E}[V(\tau)]$. For this case, the exact value and the asymptote typically look as shown in Figure 3. For larger values of N the asymptote may give a very poor approximation. As an illustration, Figure 4 is taken from [15], with $N = 35$.

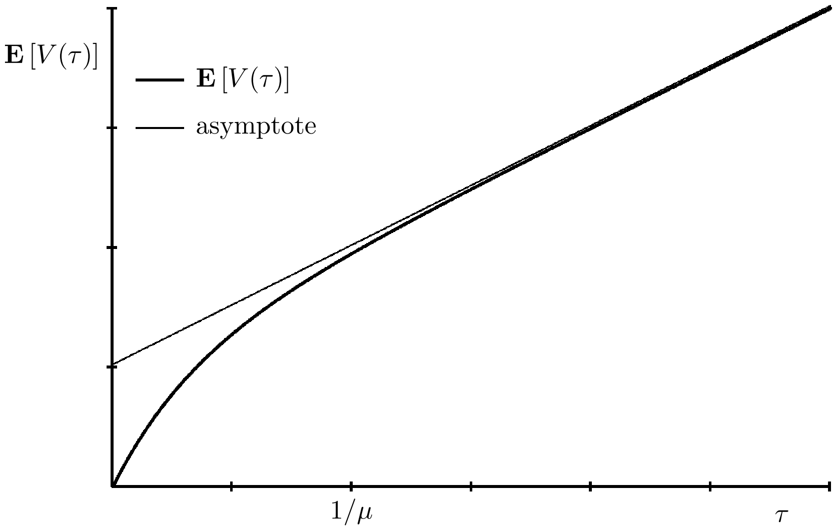

Figure 3: Example with $N = 5$

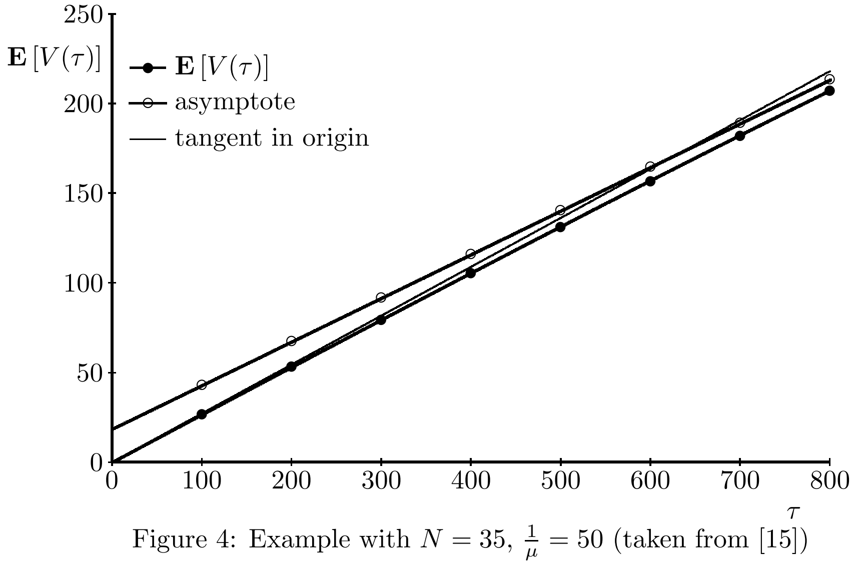

Figure 4: Example with $N = 35$, $\frac{1}{\mu} = 50$ (taken from [15])

From the numerical experiments of [15], it turns out that for larger values of N (≥ 30), the tangent in the origin is often an excellent approximation for $\mathbf{E}[V(\tau)]$: For $\tau \leq 5/\mu$ (with exponential services this is the case for 99.9% of the customers), the relative error is typically less than 2.5%. In Figure 4, the tangent in the origin is also drawn.

Remark 7.5 The fact that the tangent in the origin is close to the exact value implies that in that case the mean of $V(\tau)$ is 'almost' proportional to τ. The proportionality constant is given by the slope (δ) of the tangent, which is equal to the initial expected 'delay per unit service' upon arrival of a customer in steady state:

$$\delta := \sum_{(k,i)\in\mathbf{S}^*} a_i^{(k)} \frac{k}{r_i^{(k)}}.$$

□

In practice it is not clear beforehand which of the two approximations (the asymptote or the tangent in the origin) is best, the more since the quality of both approximations also depends on the transition rates of the random environment. However, in all experiments of [15], it is observed that both approximations are an *upper-bound* for $\mathbf{E}[V(\tau)]$, and that for practical situations the minimum of the two gives a useful approximation. Therefore we propose to use

$$\mathbf{E}[V(\tau)] \approx \min\left(\frac{\tau}{c^*-\rho^*}+\gamma, \delta\tau\right),$$

as an approximation for $\mathbf{E}[V(\tau)]$. This approximation can be improved by computing more than the first coefficient (δ) of the Taylor-series. This may be done iteratively by using (7.1), until two subsequent approximations are considered to be close enough. Note however that this procedure is not guaranteed to be numerically stable, since positive and negative numbers are added in each step. Therefore the roundoff errors may accumulate significantly in the iterative procedure.

8. Generalisations

In Section 2 we made some assumptions which are not essential for the analysis, but facilitated the presentation and discussion. In this section we relax some of the assumptions and show how the resulting models either fit the framework, or how they can be included in an analogous but generalised analysis.

8.1 Service requirements of phase-type

We may allow the service requirements of customers in the queueing system of Figure 1 to be of phase-type. By phase-type distributions we mean the class of distributions presented in Neuts [13, Chapter 2]. In order to preserve the Markovian description of our model, some additional state-descriptors must be added. For the analysis of the sojourn time conditioned on the amount of work, in the queueing model with one permanent customer, a state is determined by the number of customers (excluding the permanent one) in each service phase together with the state of the random environment. Thus if the service requirement distribution consists of P phases, then the state space is given by:

$$\mathbf{S}^* := \{(k_1,k_2,\ldots,k_P;i) \mid 0 \le k_1+k_2+\cdots+k_P \le L-1,\ k_j \in \{0,1,2,\ldots,L-1\},\ i\in\{1,2,3,\ldots,N\}\}.$$

Note that we lose the QBD structure, which was convenient for computation of various entities (the new process could be called a multi-dimensional QBD process). Note also that, when studying the process with one permanent customer, the role of the random environment and the service phases of non-permanent customers is not essentially different. We may redefine the random environment such that it also contains the service phases of non-permanent customers, and then view the resulting model as a special case of the earlier model with $L=1$. In particular we find that, also for phase-type services, the conditional mean sojourn time as a function of the amount of work τ, has an asymptote for $\tau\to\infty$.

For the representation of sojourn times (not conditioned on the amount of work) as absorption times in an appropriate Markov process, we need to add yet another descriptor to the state space, namely the phase of the tagged customer. Then Theorem 3.1 again applies.

8.2 Other service disciplines

Our model of Section 2 also includes other service disciplines. For instance discriminatory processor sharing (sometimes called weighted processor sharing) which contains (ordinary) processor sharing as a special case. Discriminatory processor sharing is of great interest for applications. For this service discipline several classes of customers are identified, numbered as $1, 2, \ldots, J$. With customer class j a weight $w_j > 0$ is associated. If there are k_j customers of class j, $j = 1, 2, \ldots, J$, then each of these gets a fraction $\frac{w_j}{k_1 w_1 + \cdots + k_J w_J}$ of the total (available) capacity. In our model this capacity may be a function of the state of a random environment and the numbers k_j, $j = 1, 2, \ldots, J$. If we are interested in the (conditional) sojourn time of customers of class 1, then we may view the model in the framework of Section 2 by extending the random environment with the tuples $(k_2, \ldots, k_J)$ containing the number of customers of all other classes.

As in Section 8.1, we may allow for phase-type distributions for each of the customer classes. In our state description we need to record the number of customers of any class in each particular service-phase. The number of (other) customers of the class under consideration in each possible service-phase also needs to be incorporated in the random environment.

Similarly, other service disciplines (including First Come First Served and Last Come First Served) may be incorporated by a proper definition of the random environment. Again we emphasise that with these generalisations, the computational complexity is increased tremendously. These generalisations, however, retain the qualitative properties such as the existence of an asymptote for the conditional mean sojourn time.

8.3 Infinite state space

In Section 2 we assumed $L < \infty$ and $N < \infty$. Here we discuss the case where either of these, or both, are infinite. The results obtained in this paper may be generalised to infinite state spaces, under recurrence conditions which are stronger than requiring ergodicity. For instance, the existence of the vector $\overline{\gamma}$ in Corollaries 3.3 and 5.2 is not ensured if we only assume ergodicity. This issue is related to convergence of the value iteration algorithm for Markov reward (decision) processes on countable state spaces, see for instance Sennott [20].

In applications, it is usually the case that the $r_i^{(k)}$ are uniformly bounded from above, so that the costs in the Markov reward processes of the proofs of Corollaries 3.3 and 5.2 are uniformly bounded in absolute value. In that case the vector $\overline{\gamma}$ exists under the assumption of ergodicity. To see this we may proceed as in [21, p. 188] to construct a relative-cost vector which satisfies the conditions given for $\overline{\gamma}$ in Corollaries 3.3 and 5.2. In the same way, we may show that the mean of these constructed relative costs exists and is finite, so that we may normalise as required in Corollaries 3.3 and 5.2.

Moreover, in applications when $L = \infty$ and $N < \infty$, it is often the case that the QBD process with generator (2.1) is homogeneous beyond some level, i.e. there is a positive integer K such that $M^{(k)} = M$, $D^{(k)} = D$, and $\Lambda^{(k)} = \Lambda$, for all $k \geq K$ (see for instance [15]). The ergodicity condition is then $\overline{p}\Lambda\overline{1} < \overline{p}M\overline{1}$, with $\overline{p}\,[M + D + \Lambda] = \overline{0}$, where $\overline{0}$ is a vector of zeroes, see [13, Theorem 3.1.1].

We finally remark that for infinite generators, the exponential function as in (3.5) may be defined by its Taylor-series representation.

9. Concluding remarks

In this paper we studied sojourn times of customers in a Markovian queueing system with processor sharing, in which arrival and service rates may depend on the number of customers already in the system *and* on the state of a random environment. The random environment itself may be dependent on the number of customers in the system. For this model we first represented the sojourn time as the absorption time in an appropriate Markov process.

Particular attention was paid to sojourn times conditioned on the amount of work. For these, we found a closed-form solution for the LST, and in particular for its mean. We showed that as a function of the amount of work, the conditional mean sojourn time has a linear asymptote. By means of the method of random time change, the conditional sojourn times were represented by transient costs in a particular Markov reward process. The latter was shown to be closely related to the branching process which is used in the literature to study processor sharing systems with constant (available) service capacity. For those systems it is known that the conditional mean sojourn time is proportional to the amount of work. This property (which does not hold for our general model) was explained by comparing the steady-state distributions of the original queueing model and the model obtained by the random time change.

We discussed the possibility of computing the conditional mean of the sojourn times as a function of the amount of work. A numerically stable algorithm was developed, but the computational complexity calls for

reliable and efficient approximations. For a special class of models an approximation was proposed, which in all tested cases proved to be both conservative and useful for practical purposes.
The analysis was shown to include the case of service requirements with a phase-type distribution (leading to a 'multi-dimensional' QBD process). We also saw that the more general discriminatory processor sharing service discipline fits into our framework. We discussed extensions to infinite state spaces, and in particular showed that for uniformly bounded service rates the analysis still applies. In particular we find that, for these generalisations, the conditional mean sojourn time as a function of the amount of work τ, has an asymptote for $\tau \to \infty$. In view of the current interest in service requirement distributions with heavy tails, we might wonder whether this is also the case when the service requirements do not have a finite second moment. It is our conjecture that the conditional mean sojourn time still has a linear asymptote. For the model with service interruptions in [14] this statement can be verified to be true. However, in that same paper it is seen that if the lengths of the service interruptions have an infinite second moment, then (in steady state) even $\mathbf{E}[V(\tau)] = \infty$ *for all* $\tau > 0$.

Acknowledgement

The author thanks Onno Boxma, Sem Borst and Hans van den Berg for reading previous versions of the paper, and providing many comments that led to improvements.

References

1. The ATM Forum Technical Committee. *Traffic Management Specification.* Version 4.0, April 1996.
2. J.L. van den Berg. *Sojourn times in feedback and processor sharing queues.* Ph.D. thesis, Rijksuniversiteit Utrecht, 1990.
3. J.L. van den Berg, O.J. Boxma. *The $M/G/1$ queue with processor sharing and its relation to a feedback queue.* Queueing Systems 9 (1991), 365–401.
4. E.G. Coffman, R.R. Muntz, H. Trotter. *Waiting time distributions for processor-sharing systems.* J. Assoc. Comput. Mach. 17 (1970), 123–130.
5. J.W. Cohen. *The multiple phase service network with generalized processor sharing.* Acta Informatica 12 (1979), 245–284.
6. V. De Nitto Personè, V. Grassi. *Solution of finite QBD processes.* J. Appl. Probab. 33 (1996), 1003–1010.
7. R.D. Foley, G.-A. Klutke. *Stationary increments in the accumulated work process in processor-sharing queues.* J. Appl. Probab. 26 (1989), 671–677.
8. 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).
9. S. Grishechkin. *On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes.* Adv. in Appl. Probab. 24 (1992), 653–698.
10. R.A. Howard. *Dynamic programming and Markov processes.* M.I.T. Press, New York, 1960.
11. L. Kleinrock. *Queueing systems, Vol. II: Computer applications.* Wiley, New York, 1976.
12. S.-Q. Li, H.-D. Sheng. *Generalized folding-algorithm for sojourn time analysis of finite QBD processes and its queueing applications.* Comm. Statist. Stochastic Models 12 (1996), 507–522.
13. M.F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins U.P., Baltimore, 1981.
14. R. Núñez Queija. *Sojourn time in a processor sharing queue with service interruptions.* CWI Report PNA-R9807 (1998).
15. R. Núñez Queija, J.L. van den Berg, M.R.H. Mandjes. *Performance evaluation of strategies for integration of elastic and stream traffic.* Teletraffic Engineering in a Competitive World: Proceedings of the 16th International Teletraffic Congress, Edinburgh. Eds. D. Smith and P. Key, Elsevier Science B.V., Amsterdam. Pre-print will be available as CWI Report.

16. M.L. PUTERMAN. *Markov decision processes: discrete stochastic dynamic programming.* Wiley, New York, 1994.
17. S.M. ROSS. *Applied probability models with optimization applications.* Holden-Day, San Francisco [etc.], 1970.
18. M. SAKATA, S. NOGUCHI, J. OIZUMI. *Analysis of a processor shared queueing model for time sharing systems.* Proc. 2nd Hawaii Int. Conf. on System Sciences (1969), 625–628.
19. M. SAKATA, S. NOGUCHI, J. OIZUMI. *An analysis of the M/G/1 queue under round-robin scheduling.* Oper. Res. 19 (1971), 371–385.
20. L.I. SENNOTT. *Value iteration in countable state average cost Markov decision processes with unbounded costs.* Ann. Oper. Res. 28 (1991), 261–271.
21. H.C. TIJMS. *Stochastic models: an algorithmic approach.* Wiley, Chichester, 1994.
22. K. VAN DER WAL, M.R.H. MANDJES, H. BASTIAANSEN. *Delay performance analysis of the new Internet services with guaranteed QoS.* Proceedings of the IEEE 85 (1997), 1947–1957.
23. P. WHITE, J. CROWCROFT. *The integrated services in the Internet: State of the art.* Proceedings of the IEEE 85 (1997), 1934–1946.
24. S.F. YASHKOV. *A derivation of response time distribution for a M/G/1 processor sharing queue.* Problems in Control and Information Theory 12 (1983), 133–148.