STOCHASTIC MODELS, 17(1), 61–92 (2001)

SOJOURN TIMES IN NON-HOMOGENEOUS QBD PROCESSES WITH PROCESSOR SHARING

Rudesindo Núñez-Queija

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

ABSTRACT

We study sojourn times of customers in a processor sharing queue 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 paid 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.

Key Words: (Discriminatory) processor sharing; Varying service rate; Sojourn time; Random time change; Random environment; Quasi birth–death process.

E-mail: sindo@cwi.nl

1. INTRODUCTION

We study 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 $c_i^{(k)} \geq 0$. The process $Y(t)$ is called the random environment. The offered service capacity is divided equally among all customers present. Under the assumption of Poisson arrivals (possibly state-dependent) and exponentially distributed service requirements, the 2D process $\{(X(t), Y(t)), t \geq 0\}$ is a nonhomogeneous (or level-dependent) Quasi Birth–Death (QBD) process. The QBD structure is not essential to the analysis, but has computationally attractive properties. The QBD structure is preserved throughout the analysis and reflected in the results. In Section 8 we show how the analysis can be extended to the case when service requirements have a phase-type distribution. This destroys the QBD structure, but qualitative properties of sojourn times are preserved.

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 network 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—that is, 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, that is each such connection should get an equal share, see for instance The ATM Forum (1). Also in IP (Internet Protocol) networks it is thought to be useful to make a distinction between real-time and non real-time (best-effort) connections to provide Quality of Service guarantees, see for instance Van der Wal et al. (2) and White (3).

In Núñez Queija et al. (4) 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 (file transfers).

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 (5–9). 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 (10) that this proportionality property is lost. However, in that paper an asymptotic linearity (for the amount of work tending to infinity) is revealed. Here 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 organized as follows. We present the model in Section 2. 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 (11), to our model. This way we translate sojourn times in the queueing system into rewards 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 from (4). In Section 8 it is shown that phase-type services and discriminatory processor sharing essentially fall within the model. We also discuss the extension to infinite state spaces. Concluding remarks are made in Section 9.

2. THE MODEL

Consider a processor-sharing queue in a random environment as depicted in Figure 1. In the queue at most $L \in \mathbb{N}$ customers can be present. We assume that

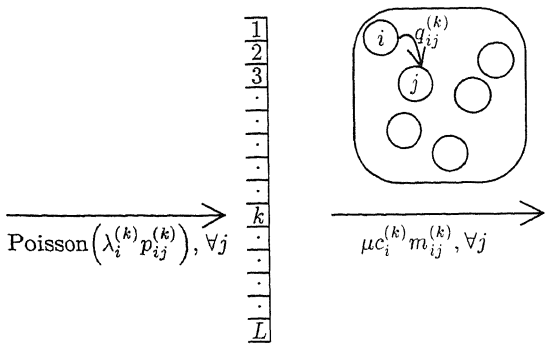

Figure 1. The queueing model.

the random environment may be modeled as a Markov process with state space $\{1, 2, \ldots, N\}$, with $N \in \mathbb{N}$. Changes in the random environment may be dependent on the arrival and departure process of customers. The set of possible states of the random environment when the number of customers is $k \in \{0, 1, 2, \ldots, L\}$ is denoted by the subset $E^{(k)} \subseteq \{1, 2, \ldots, N\}$. We say that the queueing system of Figure 1 is in state (k, i) when there are $k \in \{0, 1, \ldots, L\}$ customers present and the state of the random environment is $i \in E^{(k)}$. The set of all possible system states is denoted by:

$$S := \left\{(k, i) : k = 0, 1, \ldots, L; \quad i \in E^{(k)}\right\}. \tag{1}$$

The arrival rate of new customers and the service rate of customers in the queue are determined by both the queue length and the state of the random environment. 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 customers' service requirements, the arrival process, and the random environment). 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 $p_{ij}^{(k)}$ is the probability that the random environment does not change state. In state (k, i) with $k > 0$, the server works at rate $c_i^{(k)} \geq 0$. This service capacity is equally divided among all customers present (processor sharing). Hence, each customer leaves in an interval of length Δ with probability $\frac{1}{k}\mu c_i^{(k)}\Delta + o(\Delta)$, for $\Delta \downarrow 0$. The total departure rate of customers is therefore $\mu c_i^{(k)}$. Upon such a departure, the random environment changes to state $j \in E^{(k-1)}$ with probability $m_{ij}^{(k)}$, where $\sum_{j \in E^{(k-1)}} m_{ij}^{(k)} = 1$ 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 $q_{ij}^{(k)}$, $j \neq i$. For $(k, i) \in S$ it is convenient to define $p_{ij}^{(k)} = 0$, $j \notin E^{(k+1)}$, $m_{ij}^{(k)} = 0$, $j \notin E^{(k-1)}$, and $q_{ij}^{(k)} = 0$, $j \notin E^{(k)}$. Note that

$E^{(-1)}$ and $E^{(L+1)}$ are not defined, therefore we further set $\lambda_i^{(L)} := c_i^{(0)} := 0$, for all $i \in \{1, 2, \ldots, N\}$.

At time $t \geq 0$, $X(t)$ is the number of customers in the system and $Y(t)$ is the state of the random environment. $\{(X(t), Y(t)), t \geq 0\}$ is a non-homogeneous QBD process. Its infinitesimal generator can be written as:

$$\mathcal{G} := \begin{bmatrix} Q_d^{(0)} & \Lambda^{(0)} & 0 & \ldots & \ldots & 0 \\ M^{(1)} & Q_d^{(1)} & \Lambda^{(1)} & 0 & \ldots & \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & M^{(L-1)} & Q_d^{(L-1)} & \Lambda^{(L-1)} \\ 0 & \ldots & & 0 & M^{(L)} & Q_d^{(L)} \end{bmatrix}. \tag{2}$$

The submatrices in this generator are given by:

$$\Lambda^{(k)} = \left[\lambda_i^{(k)} p_{ij}^{(k)}\right]_{i \in E^{(k)}, j \in E^{(k+1)}},$$
$$M^{(k)} = \left[\mu c_i^{(k)} m_{ij}^{(k)}\right]_{i \in E^{(k)}, j \in E^{(k-1)}},$$
$$Q_d^{(k)} = \left[q_{ij}^{(k)}\right]_{i,j \in E^{(k)}},$$

where the $q_{ii}^{(k)}$ are such that Equation (2) is a true generator (all rows sum to 0).

The state space of the process $(X(t), Y(t))$ is given by $\boldsymbol{S}$ in Equation (1). The components k and i of the state $(k, i) \in \boldsymbol{S}$ are called the level and the phase of the QBD process, respectively. 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.

It will be assumed throughout this chapter that $\mathcal{G}$ is irreducible (all states in the corresponding Markov process communicate). The process is called a homogeneous QBD process if for all $1 \leq k \leq L - 1$: $M^{(k)} = M$, $Q_d^{(k)} = Q_d$ and $\Lambda^{(k)} = \Lambda$. The state space is finite, since we assumed that L, the maximum number of customers in the system, and N, the number of states of the random environment, are both finite. The submatrices $Q_d^{(k)}$, $\Lambda^{(k)}$, and $M^{(k)}$ are all of finite—but not necessarily the same—dimension. Generalizations to infinite state spaces are possible, but require specific attention regarding ergodicity issues. We briefly address these issues in Section 8.3.

We denote the vector of steady-state probabilities of $(X(t), Y(t))$ by π:

$$\pi\mathcal{G} = \mathbf{0}, \qquad \pi\mathbf{1} = 1,$$

with $\mathbf{0}$ being the vector with all entries equal to zero and $\mathbf{1}$ the vector with all entries equal to one. Throughout this paper (article), for any vector $\mathbf{v}$ its entries $v_{k,i}$ are ordered lexicographically, that is, $v_{k,i}$ precedes $v_{l,j}$ if $k < l$, or if $k = l$ and $i < j$. Another notational convention we adopt, is that any vector multiplying a matrix from

the left (right) is a row (column) vector. Furthermore we use the symbol I to denote the identity matrix. Whenever used, the vectors $\mathbf{1}$ and $\mathbf{0}$, and the matrix I are of the appropriate dimension.

Usually the service discipline considered for queueing systems that can be modeled as a QBD process is FCFS (First Come First Served), see for instance Neuts (12, Sec. 3.9). 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 FCFS and processor sharing). The queue length in non-homogeneous QBD processes has been studied extensively in the literature, see for instance De Nitto Personè and Grassi (13) where an algorithm is described for the computation of the steady-state queue-length distribution in non-homogeneous QBD processes with a finite state space. Here we do not discuss the computation of the steady-state probability vector π.

For the sojourn times of customers, the service discipline *does* matter. Sojourn times in QBD processes under the FCFS discipline are discussed in (12, Sec. 3.9). For non-homogeneous QBD processes an analogous treatment is possible. The distribution in terms of LSTs may be found in (14). Here, our concern is with the sojourn time distribution under the processor-sharing service discipline.

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:

$$\mathcal{H} := \left[\begin{array}{c|cccccc} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \hline M^{(1)}\mathbf{1} & Q_d^{(1)} & \Lambda^{(1)} & & & & \\ \frac{1}{2}M^{(2)}\mathbf{1} & \frac{1}{2}M^{(2)} & Q_d^{(2)} & \Lambda^{(2)} & & & \\ \vdots & & \ddots & \ddots & \ddots & & \\ \frac{1}{k}M^{(k)}\mathbf{1} & & & \frac{k-1}{k}M^{(k)} & Q_d^{(k)} & \Lambda^{(k)} & \\ \vdots & & & & \ddots & \ddots & \ddots \\ \frac{1}{L}M^{(L)}\mathbf{1} & & & & & \frac{L-1}{L}M^{(L)} & Q_d^{(L)} \end{array}\right]. \tag{3}$$

The state space of a Markov process with generator $\mathcal{H}$ may be denoted by all pairs (k, i), with $k = 1, 2, \ldots, L$ 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), $k = 1, 2, \ldots, L$ are transient. The latter statement follows from the irreducibility of $\mathcal{G}$ given by Definition (2). 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 $\frac{1}{k}\mu c_i^{(k)}$.

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 Equation* (3), *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 $\frac{1}{l}c_j^{(l)}$, and his "departure rate" is therefore $\mu\frac{1}{l}c_j^{(l)}$. 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 (14).

We further concentrate on the sojourn time of a customer with service requirement $\tau > 0$. For $k = 1, 2, \ldots, L$ and $i \in E^{(k)}$, let $V_{k,i}(\tau)$ be the (remaining) sojourn time of a (tagged) customer, starting with $k-1$ other customers present, the random environment in state i, and the tagged customer having a (remaining) service requirement of τ. Define the LST (Laplace-Stieltjes Transform) of the distribution of $V_{k,i}(\tau)$ by

$$v_{k,i}(s;\tau) := \boldsymbol{E}\left[\mathrm{e}^{-sV_{k,i}(\tau)}\right], \qquad \mathrm{Re}(s) \geq 0,$$

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

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

We study the sojourn times conditional on the service requirement using a modified model with one *permanent* customer. Suppose we consider the queueing system of Figure 1, with the modification that there is one customer that never leaves the system, but shares in the service rate as an ordinary customer. With that modification, the number of customers k ranges from 1 to L. Having placed a permanent customer in the system at time 0, we denote the total number of customers in the system (including the permanent customer) at time $t \geq 0$ by $X^*(t)$ and the state of the random environment by $Y^*(t)$. The process $\{(X^*(t), Y^*(t)), t \geq 0\}$ is

again a non-homogeneous QBD process with generator $\mathcal{G}^*$ defined by:

$$\mathcal{G}^* := \begin{bmatrix} \widetilde{Q}_d^{(1)} & \Lambda^{(1)} & & & & & \\ \frac{1}{2}M^{(2)} & \widetilde{Q}_d^{(2)} & \Lambda^{(2)} & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \frac{k-1}{k}M^{(k)} & \widetilde{Q}_d^{(k)} & \Lambda^{(k)} & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \frac{L-2}{L-1}M^{(L-1)} & \widetilde{Q}_d^{(L-1)} & \Lambda^{(L-1)} \\ & & & & & \frac{L-1}{L}M^{(L)} & \widetilde{Q}_d^{(L)} \end{bmatrix}. \quad (4)$$

The matrices $\widetilde{Q}_d^{(k)}$ differ from the matrices $Q_d^{(k)}$ only in their diagonal elements: these are such that each row of $\mathcal{G}^*$ sums to 0. The state space of the process $(X^*(t), Y^*(t))$ will be denoted by:

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

In this section we first analytically derive an expression for the vector of LSTs $\mathbf{v}(s; \tau)$. In Section 4 we relate the solution to the permanent-customer model using the method of random time change.

Define the diagonal matrix

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

The entries of $\mathcal{R}$ along the diagonal are ordered lexicographically in (k, i).

The following theorem gives the LSTs of the conditional sojourn times explicitly. Similar expressions were obtained in (15–17).

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

$$\frac{\partial}{\partial \tau}\mathbf{v}(s; \tau) = \mathcal{R}^{-1}[\mathcal{G}^* - sI]\mathbf{v}(s; \tau), \quad (5)$$

$$\mathbf{v}(s; 0) = \mathbf{1}; \quad (6)$$

and hence,

$$\mathbf{v}(s; \tau) = \exp\{\tau\mathcal{R}^{-1}[\mathcal{G}^* - sI]\}\mathbf{1}. \quad (7)$$

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 any other customer) is served at rate $\frac{1}{k}c_i^{(k)}$. Consider a small time interval of

length $\frac{k\Delta}{c_i^{(k)}}$ and condition on the possible events occurring in this interval:

$$v_{k,i}(s;\tau+\Delta) = e^{-s\frac{k\Delta}{c_i^{(k)}}}\left\{\lambda_i^{(k)}\frac{k\Delta}{c_i^{(k)}}\sum_j p_{ij}^{(k)} v_{k+1,j}(s;\tau)\right.$$
$$+\mu\frac{k-1}{k}c_i^{(k)}\frac{k\Delta}{c_i^{(k)}}\sum_j m_{ij}^{(k)} v_{k-1,j}(s;\tau)$$
$$\left.+\sum_{j\neq i} q_{ij}^{(k)}\frac{k\Delta}{c_i^{(k)}} v_{k,j}(s;\tau)+\left(1+\tilde{q}_{ii}^{(k)}\frac{k\Delta}{c_i^{(k)}}\right)v_{k,i}(s;\tau)\right\}+o(\Delta),$$

with $\tilde{q}_{ii}^{(k)} = -\mu\frac{k-1}{k}c_i^{(k)} - \lambda_i^{(k)} - \sum_{j\neq i} q_{ij}^{(k)}$. This leads to the differential Equation (5). The initial conditions (6) follow from the fact that all $c_i^{(k)}$ are positive, and hence $\boldsymbol{E}[V_{k,i}(0+)] = 0$. It can then be verified that Equation (7) is a solution to Equations (5) and (6).

It remains to be shown that the solution is unique. Suppose $\mathbf{w}(s;\tau)$ is a second solution. Then $\mathbf{w}'(s;\tau) := \mathbf{v}(s;\tau) - \mathbf{w}(s;\tau)$ satisfies a set of equations like (5) and (6) with $\mathbf{1}$ replaced by $\mathbf{0}$. From Equation (5) it follows that $\mathbf{v}(s;\tau)$, $\mathbf{w}(s;\tau)$ and $\mathbf{w}'(s;\tau)$ are analytic in $\tau \geq 0$ (it can be shown iteratively that all derivatives exist). Since all derivatives of $\mathbf{w}'(s;\tau)$ vanish at $\tau = 0$ (this again can be shown iteratively) it must be the case that $\mathbf{w}'(s;\tau) \equiv 0$. □

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 (with the difference that 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 (18). In particular the results on continuous-time Markov–Reward processes (pp. 99–104) are of interest to us. The close relationship between the continuous-time case and the discrete-time case is exploited by Tijms (19, Sec. 3.5). In the proof of the following corollary we further rely on Zijm (20).

We use the symbol $\mathbf{1}_{k,i}$ 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 $c_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \geq 0$,*

$$\boldsymbol{E}[V_{k,i}(\tau)] = \frac{\tau}{c^* - \rho^*} + \mathbf{1}_{k,i}[I - \exp\{\tau\mathcal{R}^{-1}\mathcal{G}^*\}]\boldsymbol{\gamma}, \tag{8}$$

where

$$c^* = \sum_{(k,i)\in\mathbf{S}^*} \pi_{k,i}^* c_i^{(k)},$$

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

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

$$\pi^*\mathcal{G}^* = \mathbf{0}, \qquad \pi^*\mathbf{1} = 1.$$

The vector γ *satisfies*

$$-\mathcal{G}^*\gamma = \mathbf{1} - \frac{1}{c^* - \rho^*}\mathcal{R}\mathbf{1}, \tag{9}$$

and is unique up to translation by the vector $\mathbf{1}$. *Expression* (8) *is, however, invariant with respect to such a translation. We may normalize* γ *such that* $\pi^*\mathcal{R}\gamma = 0$.

Proof: The result can be obtained by differentiating Equation (7) with respect to s, and setting $s = 0$. However, we give a more direct proof. In the same way as we derived Equations (5) and (6), we may find the following set of differential equations and initial conditions:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\boldsymbol{E}[V_{k,i}(\tau)])_{k,i} = \mathcal{R}^{-1}\mathbf{1} + \mathcal{R}^{-1}\mathcal{G}^*(\boldsymbol{E}[V_{k,i}(\tau)])_{k,i}, \tag{10}$$

$$\boldsymbol{E}[V_{k,i}(0)] = 0, \qquad \forall (k,i) \in \mathbf{S}^*. \tag{11}$$

By $(\cdot)_{k,i}$ we mean the vector with the entries between brackets ordered lexicographically in $(k,i) \in \mathbf{S}^*$. As in the proof of Theorem 3.2 it can be verified that there is at most one solution to this set of differential equations and initial conditions.

Suppose for the moment that a vector γ exists, satisfying Equation (9) and normalized as required. By substitution of Equation (8) into Equations (10) and (11), we may verify that these differential equations and initial conditions are satisfied, and hence Equation (8) is the unique solution. Note that $\mathcal{G}^*\mathbf{1} = \mathbf{0}$, since $\mathcal{G}^*$ is the generator of a Markov process.

From Equation (9) we note that if the vector γ exists, it may be interpreted as the "relative reward" vector in a Markov-Reward process. This vector contains for each state of the process the long-run difference in accumulated rewards when starting in that state relative to those when starting in steady state, see Tijms (19, pp. 187, 188) for a discussion. The generator of this Markov-Reward process is $\mathcal{G}^*$ and rewards are generated at rate $1 - \frac{1}{k}c_i^{(k)}\frac{1}{c^*-\rho^*}$ when the process is in state $(k,i) \in \mathbf{S}^*$.

In order for Equation (9) to have a solution, it is necessary that this Markov-Reward process has average reward per time unit equal to 0, because the left-hand side is equal to zero if we premultiply by π^*. Indeed,

$$\begin{aligned}\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\frac{1}{k}c_i^{(k)} &= c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\frac{k-1}{k}c_i^{(k)}\mu \\ &= c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\lambda_i^{(k)} \\ &= c^* - \rho^*, \end{aligned} \tag{12}$$

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 $\boldsymbol{\gamma}$, and its uniqueness up to translation along the vector $\mathbf{1}$, is guaranteed by Zijm (20, Theorem 4.5).

Note that translation along the vector $\mathbf{1}$ of a vector $\boldsymbol{\gamma}$ satisfying Equation (9) does not alter the solution for $\boldsymbol{E}[V_{k,i}(\tau)]$, since all rows of the matrix $I - \exp\{\tau\mathcal{R}^{-1}\mathcal{G}^*\}$ in Equation (8) sum to 0 (because all rows of $\mathcal{G}^*$ do). Therefore, if $\boldsymbol{\gamma} = \mathbf{v}$ satisfies Equation (9), then so does

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

which is normalized as required. □

Remark 3.3. 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 $\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\lambda_i^{(k)}$ 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 $c_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \to \infty$ we have:*

$$\boldsymbol{E}[V_{k,i}(\tau)] - \frac{\tau}{c^* - \rho^*} \to \gamma_{k,i}, \quad (k,i) \in \mathbf{S}^*.$$

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 scaled left and right eigenvectors corresponding to the eigenvalue 0 are $\frac{1}{c^*-\rho^*}\pi^*\mathcal{R}$ and $\mathbf{1}$. As a consequence, the matrix $\lim_{\tau\to\infty}\exp\{\tau\mathcal{R}^{-1}\mathcal{G}^*\}$ exists and has all rows equal to the probability vector $\frac{1}{c^*-\rho^*}\pi^*\mathcal{R}$. The corollary now follows from Expression (8). □

Remark 3.4. For the special case of a constant service rate, $c_i^{(k)} = 1, \forall(k,i)$, 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 second subscripts are omitted since there is no random environment):

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

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

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

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

$$\gamma_{k+1} - \gamma_k \to \frac{\rho}{\lambda(1-\rho)}, \qquad L \to \infty,$$

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

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

cf. Coffman et al. (21, Formula (33)) (there the *delay* $V_k(\tau) - \tau$ 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 mentioned the interpretation of the coefficients $\gamma_{k,i}$ as relative rewards in a Markov–Reward process. In this section we explore such an interpretation further and link this to the method of random time change, which was introduced for the analysis of processor-sharing systems by Yashkov (22). In essence, but without transformation of time, this method was already used for the analysis of the M/G/1 processor-sharing queue in (11). Foley and Klutke (23) studied the queue-length process and the process of accumulated work after applying the random time change to a processor-sharing model in which the total service capacity may depend on the number of customers in the system. Grishechkin (24,25) further exploited the method by reformulating it in terms of Crump-Mode-Jagers branching processes and applying it to the analysis of queues with a general class of service disciplines, including processor sharing. For more references on the time-transformation method and its use in the analysis of processor-sharing queues we refer to Yashkov (26, Sec. 2.4).

Our starting point is the Markov process $(X^*(t), Y^*(t))$, that is, 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 $k/c_i^{(k)}$. For instance, the new arrival rate of customers in state (k, i) is $\lambda_i^{(k)} \times k/c_i^{(k)}$. 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 nonpermanent 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 $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)), \sigma \geq 0\}$ we denote the process of queue length and state of the random environment in the new time scale. The generator of the Markov process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$, is $\mathcal{R}^{-1}\mathcal{G}^*$. The inverse of the matrix $\mathcal{R}$ exists since we assumed all service rates $c_i^{(k)}$ to be non-zero (see Section 5 for the general case). Note that the processes $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ have the same *jump-chain*. By the

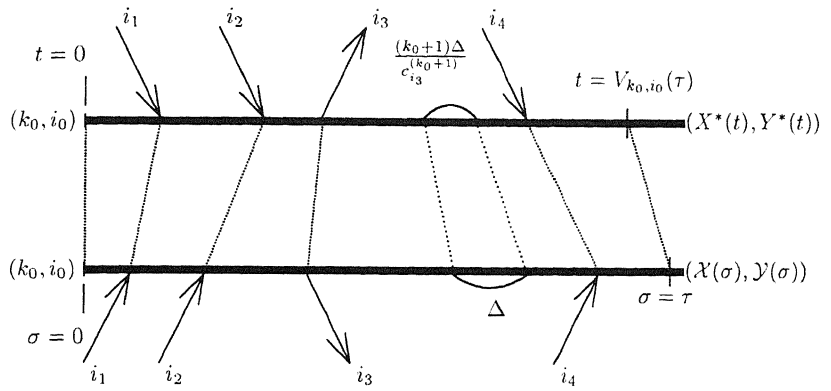

Figure 2. Coupling of the jump-chains of $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$.

jump-chain of a Markov process we mean the Markov chain embedded at transition epochs.

We now explain how the process $\{(X^*(t), Y^*(t)), t \geq 0\}$ is related to the process $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)), \sigma \geq 0\}$, 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 $k/c_i^{(k)}$, i.e. we divide the length of the period by this number. We may so construct a path for the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$, starting in (k_0, i_0) for $\sigma = 0$. In Figure 2 such a construction is depicted. 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 $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)), \sigma \geq 0$.

In the realization 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 $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ have the same jump-chain, any such realization (indeed) occurs with the same probability in both processes. Now concentrate on the indicated time-interval of length $(k_0 + 1)\Delta/c_{i_3}^{(k_0+1)}$ 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 $V_{k_0,i_0}(\tau)$: 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 reward structure in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$: In state (k, i) reward is earned at rate $k/c_i^{(k)}$. The accumulation of rewards in this process can now be related to sojourn times in the processor-sharing queue (with exclusively positive service rates).

Theorem 4.1. *The sojourn time $V_{k,i}(\tau)$ 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 cumulative reward in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ 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 reward in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ over the interval $\sigma \in (0, \tau)$ on the lower axis, is equal to $V_{k_0,i_0}(\tau)$ on the upper axis (Fig. 2). As we already remarked, any such realization has the same probability for both processes $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. □

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

Corollary 4.2. *With probability* 1:

$$\lim_{\tau\to\infty} \frac{V_{k,i}(\tau)}{\tau} = g^* := \boldsymbol{E}\left[\frac{\mathcal{X}}{c_{\mathcal{Y}}^{(\mathcal{X})}}\right], \tag{13}$$

where the distribution of $(\mathcal{X}, \mathcal{Y})$ is the equilibrium distribution of $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. Furthermore, the limit

$$\lim_{\tau\to\infty} \boldsymbol{E}[V_{k,i}(\tau)] - g^*\tau, \tag{14}$$

exists and is finite.

Proof: Relation (13) is standard for irreducible Markov–Reward processes with a finite state space, see for instance Ross (27, Corollary 6.20) or Tijms (19, Theorem 3.1.1). We use Zijm (20, Theorem 4.5) to establish the convergence in Relation (14). □

Remark 4.1. Equation (13) holds under much more general assumptions. In fact, with $L < \infty$, $N < \infty$, and all $c_i^{(k)} > 0$, it can be proved using the Renewal Reward theorem (e.g. Ross (27, Theorem 3.16) or Tijms (19, Theorem 1.3.1)) under the sole assumption that the original process $(X(t), Y(t))$ is regenerative with finite expected regeneration time. The interested reader is also referred to Iyer et al. (15) where a Central Limit Theorem is derived for $V(\tau)$.

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

$$\boldsymbol{E}\left[\frac{\mathcal{X}}{c_{\mathcal{Y}}^{(\mathcal{X})}}\right] = \left(\boldsymbol{E}\left[\frac{c_{Y^*}^{(X^*)}}{X^*}\right]\right)^{-1},$$

where we use Equation (12). This can be verified by noting that the distribution of $(\mathcal{X}, \mathcal{Y})$ is given by the vector $\frac{1}{c^* - \rho^*}\pi^*\mathcal{R}$, and that the reward vector in that process is given by $\mathcal{R}^{-1}\mathbf{1}$. The fact that

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

can be argued as follows. As we saw in Remark 3.3, 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.

If the arrival rate and total service rate do not depend on the number of customers in the system and the random environment evolves independently of the history of the queue length, then the average total service capacity and the average amount of work entering the system per unit of time are the same for the original model and the model with one permanent customer, i.e., $c^* = c$ and $\rho^* = \rho$. This is the case in the model of (10).

Remark 4.3. Branching processes that are closely related to the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ have previously been used to study sojourn times in traditional processor-sharing systems with constant service capacity (i.e., $c_i^{(k)} = 1$; there is no random environment), constant arrival rate (λ), and infinite space for customers ($L = \infty$), see Yashkov (22) and Grishechin (24,25). Let us briefly discuss the essentials of that approach. 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 having 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 Section 8.2.

In our case, the branching process is governed by 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 and the population size is k (including the permanent one), then each of the k individuals gives birth to new individuals with rate $\lambda_i^{(k)}/c_i^{(k)}$. This birth rate 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 $k \times q_{ij}^{(k)}/c_i^{(k)}$. The mutual dependence of the branching process and the random environment and the dependence among individuals make this approach less suitable for the analysis of sojourn times in the present model.

5. SERVER UNAVAILABILITY

In this section we extend the analysis of Sections 3 and 4 to the case where some of the $c_i^{(k)}$ may be equal to zero, that is, there are periods during which no service is provided to the customers. In the setting of our model, unavailability periods are exponentially distributed or, more generally, have a phase-type distribution (when two or more states of the random environment for which the service rate is zero, communicate directly with each other).

We define the subset of states

$$\mathbf{S}_0^* := \{(i, j) \in \mathbf{S}^* : c_j^{(l)} = 0\}.$$

In applications, the fact whether $c_i^{(k)} = 0$ 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 "reorder" the rows and the columns of the generator $\mathcal{G}^*$ accordingly:

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

Some 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 reorder the entries of $\pi^* = (\pi_+^*, \pi_0^*)$, with π_+^* and π_0^* vectors with their entries ordered lexicographically. Starting from any $(l, j) \in \mathbf{S}_0^*$, let $U_{l,j}$ be the amount of time the process remains in the set $\mathbf{S}_0^*$, and $W_{l,j} \in \mathbf{S}_+^*$ the first state that is visited after leaving $\mathbf{S}_0^*$. Note that $U_{l,j}$ 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 (14). Furthermore, for $\mathrm{Re}(s) \geq 0$, define the matrix $\mathcal{U}(s)$ of dimension $|\mathbf{S}_0^*| \times |\mathbf{S}_+^*|$ with entries:

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

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

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

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

and hence

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

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

$$\begin{aligned}\mathcal{U}(s) &= \mathrm{e}^{-\Delta s}([I + \Delta\mathcal{G}_0^*]\mathcal{U}(s) + \Delta\mathcal{G}_{0+}^*) + o(\Delta)\\ &= (I + \Delta[\mathcal{G}_0^* - sI])\mathcal{U}(s) + \Delta\mathcal{G}_{0+}^* + o(\Delta),\end{aligned}$$

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

$$-[\mathcal{G}_0^* - sI]\mathcal{U}(s) = \mathcal{G}_{0+}^*, \quad \mathrm{Re}(s) \geq 0. \tag{15}$$

Since $\mathcal{G}_0^*$ is a transient generator, $\mathcal{G}_0^* - sI$ is invertible for all $\mathrm{Re}(s) \geq 0$, and hence the first statement of the lemma follows.

Differentiating Equation (15) with respect to s, setting $s = 0$, and using the fact that $\mathcal{U}(0)$ is a probability matrix (so that $\mathcal{U}(0)\mathbf{1} = \mathbf{1}$), we may prove the second statement of the lemma. □

As before, denote by $v_{k,i}(s;\tau)$ the LST of $V_{k,i}(\tau)$, 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

$$\mathbf{v}_0(s;\tau) = (v_{l,j}(s;\tau))_{(l,j)\in\mathbf{S}_0^*} \quad \text{and} \quad \mathbf{v}_+(s;\tau) = (v_{k,i}(s;\tau))_{(k,i)\in\mathbf{S}_+^*},$$

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 $\mathrm{Re}(s) \geq 0$:*

$$\mathbf{v}_0(s;\tau) = \mathcal{U}(s)\mathbf{v}_+(s;\tau), \tag{16}$$

and in particular,

$$(\boldsymbol{E}[\mathrm{V}_{l,j}(\tau)])_{(l,j)\in\mathbf{S}_0^*} = (\boldsymbol{E}[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\,(\boldsymbol{E}[V_{k,i}(\tau)])_{(k,i)\in\mathbf{S}_+^*}. \tag{17}$$

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 generalizes Theorem 3.2 to the case $\mathbf{S}_0^* \neq \emptyset$. Before proceeding, we define the matrix

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

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

Theorem 5.3. *For* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$,

$$\frac{\partial}{\partial\tau}\mathbf{v}_+(s;\tau) = \mathcal{R}_+^{-1}[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(s) - sI]\mathbf{v}_+(s;\tau),$$
$$\mathbf{v}_+(s;0) = \mathbf{1};$$

and hence,

$$\mathbf{v}_+(s;\tau) = \exp\{\tau\mathcal{R}_+^{-1}[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(s) - sI]\}\mathbf{1}.$$

Proof: The proof may proceed as that for Theorem 3.2: For any $(k, i) \in \mathbf{S}_+^*$ derive the differential equation for $v_{k,i}(s;\tau)$ by conditioning on the possible events in a time interval $\frac{k}{c_i^{(k)}}\Delta$, and then take $\Delta \downarrow 0$. Substituting Equation (16) for $\mathbf{v}_0(s;\tau)$ readily leads to the desired result. □

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

Corollary 5.4. *For* $\tau \geq 0$,

$$(\boldsymbol{E}[V_{k,i}(\tau)])_{(k,i)\in\mathbf{S}_+^*} = \frac{\tau}{c^* - \rho^*}\mathbf{1} + [I - \exp\{\tau\mathcal{R}_+^{-1}[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)]\}]\boldsymbol{\gamma},$$

where c^* *and* ρ^* *are as in Corollary* 3.3. *The vector* $\boldsymbol{\gamma}$ *satisfies*

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

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

$$\boldsymbol{E}[V_{k,i}(\tau)] - \frac{\tau}{c^* - \rho^*} \to \gamma_{k,i}.$$

Proof: Similar to Equations (10) and (11) we may derive differential equations for $\boldsymbol{E}[V_{k,i}(\tau)]$, $(k,i) \in \mathbf{S}_+^*$. Using Equation (17) we get:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\boldsymbol{E}[V_{k,i}(\tau)])_{(k,i)\in\mathbf{S}_+^*} = \mathcal{R}_+^{-1}\left[\mathbf{1} + \mathcal{G}_{+0}^*(\boldsymbol{E}[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*}\right] + \mathcal{R}_+^{-1}[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)]\,(\boldsymbol{E}[V_{k,i}(\tau)])_{(k,i)\in\mathbf{S}_+^*}.$$

Of course, $\boldsymbol{E}[V_{k,i}(0)] = 0$, for $(k,i) \in \mathbf{S}_+^*$. To see that the solution given in the corollary satisfies this set of differential equations and initial conditions, note that the vector $\boldsymbol{\gamma}$ may be interpreted as the vector of relative rewards in a Markov–Reward process with generator $\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)$, rewards being earned according to the vector $\mathbf{1} + \mathcal{G}_{+0}^*(\boldsymbol{E}[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*} - \frac{1}{c^*-\rho^*}\mathcal{R}_+\mathbf{1}$. It remains to be shown that the average rewards in this Markov–Reward process equals 0. The steady-state probability vector of this process is given by $\frac{1}{\pi_+^*\mathbf{1}}\pi_+^*$, and using (cf. Lemma 5.1),

$$\pi_+^*\mathcal{G}_{+0}^*(\boldsymbol{E}[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*} = \pi_+^*\mathcal{G}_{+0}^*[-\mathcal{G}_0^*]^{-1}\mathbf{1} = \pi_0^*\mathbf{1},$$

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

Corollary 5.5. For $\tau \to \infty$,

$$(\boldsymbol{E}[V_{l,j}(\tau)])_{(l,j)\in\mathbf{S}_0^*} - \frac{\tau}{c^*-\rho^*}\mathbf{1} \to (\boldsymbol{E}[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\boldsymbol{\gamma}.$$

Proof: By Lemma 5.2, Corollary 5.4, and using the fact that $\mathcal{U}(0)$ is a probability matrix. □

The remainder of this section is devoted to the method of random time change in the case that $\mathbf{S}_0^* \neq \emptyset$, unifying the branching process approach with that of random time change. We use the same arguments as in Section 4, with the following modifications: All transitions that 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 $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. 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 $k/c_i^{(k)}$. When the state is $(l,j) \in \mathbf{S}_0^*$ this time transformation can not be done since $c_j^{(l)} = 0$. 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 $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ makes a (direct) transition at the point σ from $(k,i) \in \mathbf{S}_+^*$ to $(l,j) \in \mathbf{S}_+^*$. Thus $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ *is not observed* on the states in $\mathbf{S}_0^*$. When $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ makes such a transition corresponding to a visit of $(X^*(t), Y^*(t))$ to the set $\mathbf{S}_0^*$, an immediate reward is earned that is equal to the time that $(X^*(t), Y^*(t))$ spends within the set $\mathbf{S}_0^*$.

In the process $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)), \sigma \geq 0\}$ with state space $\mathbf{S}_+^*$ and generator $\mathcal{R}_+^{-1}[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)]$ there are two types of transitions and two types of rewards.

"Ordinary" transitions occur according to the transient generator $\mathcal{G}_+^*$, and "ordinary" rewards are earned at rate $\frac{k}{c_i^{(k)}}$ in state $(k, i) \in \mathbf{S}_+^*$. The other transitions and rewards 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) an instantaneous reward is earned. The instantaneous reward, and the state after an (l, j)-event are jointly distributed as the pair $(U_{l,j}, W_{l,j})$, 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 emphasize that the jump-chains of the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ and the process $(X^*(t), Y^*(t))$ *restricted to the set* $\mathbf{S}_+^*$ are identical.

Theorem 4.1 remains true with the above modifications, and so does Corollary 4.2 if we redefine the constant g^* as

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

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 constant g^*. As in Remark 4.2, in the queueing system with one permanent customer, $c^* - \rho^*$ is the average capacity per unit of time assigned to the permanent customer. Remark 4.3 only needs to be modified to account for the transitions with instantaneous rewards. This can be done, as in (10), by "attaching" the instantaneous rewards to the birth of a nest of children. This way, the analysis can proceed even if the periods of service unavailability are generally distributed.

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. (7, Formula (10)), Wolff (28), or Kleinrock (8, Formula (4.17)). Cohen (9, Formula (7.27)) found the proportionality property for the M/G/1/L queue with processor sharing and queue-dependent total service capacity (there called generalized 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 $S_0^* = \emptyset$: If $c^{(k)} = 0$ 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 (23). 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 (29, Remark 5.10, p. 115), and Van den Berg and Boxma (30, 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 3.4 (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, that is, the $c^{(k)}$ 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:

$$\pi_k^* \frac{1}{k} c^{(k)} \sim \pi_{k-1}, \quad k = 1, 2, \ldots, L,$$

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 $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ is given by the vector $\frac{1}{c^* - \rho^*} \pi^* \mathcal{R}$. For the present case, in the absence of a random environment, we thus have for $k = 1, 2, \ldots, L$, that $\mathbf{P}\{\mathcal{X} = k\} = \frac{1}{c^* - \rho^*} \pi_k^* \frac{c^{(k)}}{k}$, and hence, $\mathbf{P}\{\mathcal{X} = k\} = \frac{\pi_{k-1}}{1 - \pi_L}$. This property has an interesting consequence: Suppose the queueing system under consideration is in steady state, and let the random variable X have 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

$$\int_{\sigma=0}^{\tau} \frac{\mathcal{X}(\sigma)}{c^{(\mathcal{X}(\sigma))}} \, \mathrm{d}\sigma,$$

with $\mathcal{X}_0$ distributed as $X + 1$ given that $X < L$. However, this distribution is the steady-state distribution of the process $\mathcal{X}(\sigma)$, and so $\mathbf{P}\{\mathcal{X}(\sigma) = k\} = \frac{1}{1 - \pi_L} \pi_{k-1}$, $k = 1, 2, \ldots, L$, 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 service requirement τ):

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

So, for the model with exponentially distributed service requirements we have explained why this proportionality occurs, namely because the stationary distribution of $\mathcal{X}(\sigma)$ is the same as that of X given that $X < L$. We can generalize our arguments to the $M/G/1/L$ queue with processor sharing and queue-dependent service rates, cf. Cohen (9, Formula (7.27)) (for $L = \infty$ this was done in (23)). We give a brief outline of the proof: If the service requirements are distributed according to the distribution $B(x)$, $x \geq 0$, then

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

is the density function of there being k customers in the system with respective remaining service requirements $x_1, \ldots x_k$, see Cohen (9, Formula (5.9)). For this

model we may apply the random time change to the system with one permanent customer, as described above: we "shrink" the time-scale by a factor $k/c^{(k)}$ 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 normalization) 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. If we allow the arrival rate to depend on the queue length then the proportionality property is lost. 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 (9, 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 (9) is fundamentally 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. In this section we show how $\boldsymbol{E}[V(\tau)]$ can be computed. Computation of transient rewards in Markov–Reward models has received quite some attention in the area of *performability* of computing systems, see for instance Smith et al. (16) for mean transient rewards (which correspond to $\boldsymbol{E}[V(\tau)]$. The issue of computation of the entire distribution was addressed in Iyer et al. (15), Reibman et al. (17), Donatiello and Grassi (31), Nabli and Sericola (32), and De Souza e Silva and Gail (33).

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 $\lambda^{(k)} = \lambda$, $k = 0, 1, \ldots, L-1$.

7.1. Computation

For $(k, i) \in \mathbf{S}^*$, denote by $a_{k,i}$ the steady-state probability that the system is in state (k, i) immediately after the arrival of a customer. The $a_{k,i}$ 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, for $k = 1, \ldots, L-1$,

$$T^{(k,0)} = \left[-Q_d^{(k)}\right]^{-1} \Lambda^{(k)},$$

and for $k = 1, \ldots, L$; $n = 1, \ldots, k$,

$$T^{(k,n)} = \prod_{m=0}^{n-1} \left(\left[-Q_d^{(k-m)}\right]^{-1} M^{(k-m)}\right)\left[-Q_d^{(k-n)}\right]^{-1} \Lambda^{(k-n)}.$$

We now show how $\boldsymbol{E}[V(\tau)]$ can be computed after having determined the steady-state probabilities immediately after the arrival of a customer. We focus again on the case $\mathbf{S}_0^* = \emptyset$, see Remark 7.3 below for the case $\mathbf{S}_0^* \neq \emptyset$. Our starting point is the set of differential equations and initial conditions given in Equations (10) and (11). Obviously, for $n \geq 1$,

$$\frac{\mathrm{d}^n}{\mathrm{d}\tau^n}(\boldsymbol{E}[V_{k,i}(\tau)])_{k,i} \mid_{\tau=0} = (\mathcal{R}^{-1}\mathcal{G}^*)^{n-1}\mathcal{R}^{-1}\mathbf{1} \tag{18}$$

We use Jensen's method, see also Reibman et al. (17) and Donatiello and Grassi (31), to uniformize the generator $\mathcal{R}^{-1}\mathcal{G}^*$, and define the probability matrix

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

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 $\boldsymbol{E}[V_{k,i}(\tau)]$ around $\tau = 0$ exists (at the end we verify the result), and using Equation (18) we may find:

$$(\boldsymbol{E}[V_{k,i}(\tau)])_{k,i} = \frac{1}{\eta}\sum_{l=0}^{\infty}\left(1 - \mathrm{e}^{-\eta\tau}\sum_{k=0}^{l}\frac{(\eta\tau)^k}{k!}\right)(\mathcal{P}^*)^l\mathcal{R}^{-1}\mathbf{1}. \tag{19}$$

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

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

and hence the infinite sum in Equation (19) exists for every $\tau \geq 0$. Moreover, by substitution it may be seen that it satisfies the differential equations and initial conditions given in Equations (10) and (11).

Expression (19) for the $\boldsymbol{E}[V_{k,i}(\tau)]$ provides a numerically stable algorithm, since it only involves multiplication and addition of positive terms. Within the summation one needs to evaluate the "coefficients" $e^{-\eta\tau} \sum_{k=l+1}^{\infty} \frac{(\eta\tau)^k}{k!}$, 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 (10), we may start from the final Expression (8) in Corollary 3.3. Again we may use Jensen's uniformization method to derive:

$$(\boldsymbol{E}[V_{k,i}(\tau)])_{k,i} = \frac{\tau}{c^* - \rho^*}\mathbf{1} + \boldsymbol{\gamma} - e^{-\eta\tau} \exp\{\eta\tau\mathcal{P}^*\}\boldsymbol{\gamma}. \tag{20}$$

However, for this approach one first needs to compute the vector $\boldsymbol{\gamma}$. Moreover, the vector $\boldsymbol{\gamma}$ contains negative elements that may cause the evaluation of $e^{-\eta\tau} \exp\{\eta\tau\mathcal{P}^*\}\boldsymbol{\gamma}$ to be numerically unstable. No problems were encountered, though, in the numerical experiments of Núñez Queija et al. (4), where both methods were used to compute the exact value of $\boldsymbol{E}[V_{k,i}(\tau)]$. 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 5.3. For $(k,i) \in \mathbf{S}_+^*$, the $\boldsymbol{E}[V_{k,i}(\tau)]$ are found as before. However, first the $\boldsymbol{E}[u_{l,j}]$, for $(l,j) \in \mathbf{S}_0^*$, and the probability matrix $\mathcal{U}(0)$, need to be computed. Using Lemma 5.2, from the $\boldsymbol{E}[V_{k,i}(\tau)]$, $(k,i) \in \mathbf{S}_+^*$, also the $\boldsymbol{E}[V_{l,j}(\tau)]$, for $(l,j) \in \mathbf{S}_0^*$ can be computed. Note that $\boldsymbol{E}[V_{l,j}(0+)] = \boldsymbol{E}[u_{l,j}] > 0$, for $(l,j) \in \mathbf{S}_0^*$. As a consequence, $\boldsymbol{E}[V(0+)] > 0$, unless $p_{i,j}^{(l-1)} = 0, \forall (l,j) \in \mathbf{S}_0^*, i \in E^{(l-1)}$.

7.2. Approximation

Although Expression (19) provides a numerically stable algorithm to compute the $\boldsymbol{E}[V_{k,i}(\tau)]$, in general this task requires considerable computation time and memory space. Therefore it would be convenient to have a good approximation which is less computationally demanding. From Corollary 3.4 we have for the mean of $V(\tau)$:

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

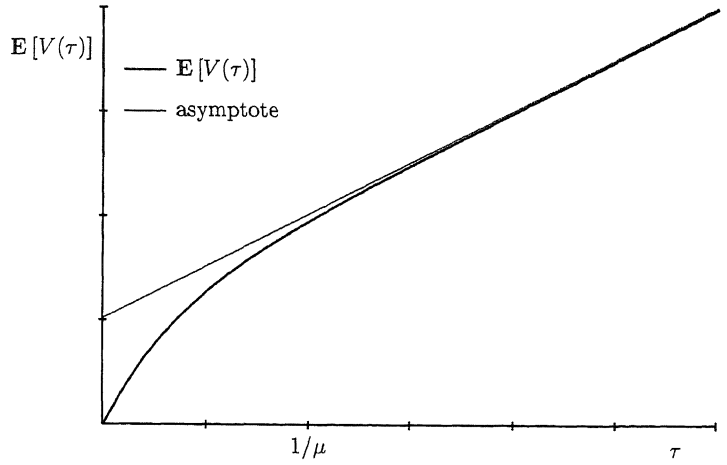

Figure 3. Example with $N = 5$.

This asymptotic relation can be used for a first approximation, that is, $\boldsymbol{E}[V(\tau)] \approx \frac{\tau}{c^*-\rho^*} + \gamma$. Indeed, when the number of states of the random environment is small ($N \leq 5$), the asymptotic result may serve as a useful approximation for $\boldsymbol{E}[V(\tau)]$. For this case, the exact value and the asymptote typically look as shown in Figure 3. However, we shall see in the example below that for larger values of N (≥ 30) the asymptote may give a poor approximation, whereas the tangent in the origin is an excellent approximation of $\boldsymbol{E}[V(\tau)]$, even for relatively large values of τ (Fig. 4).

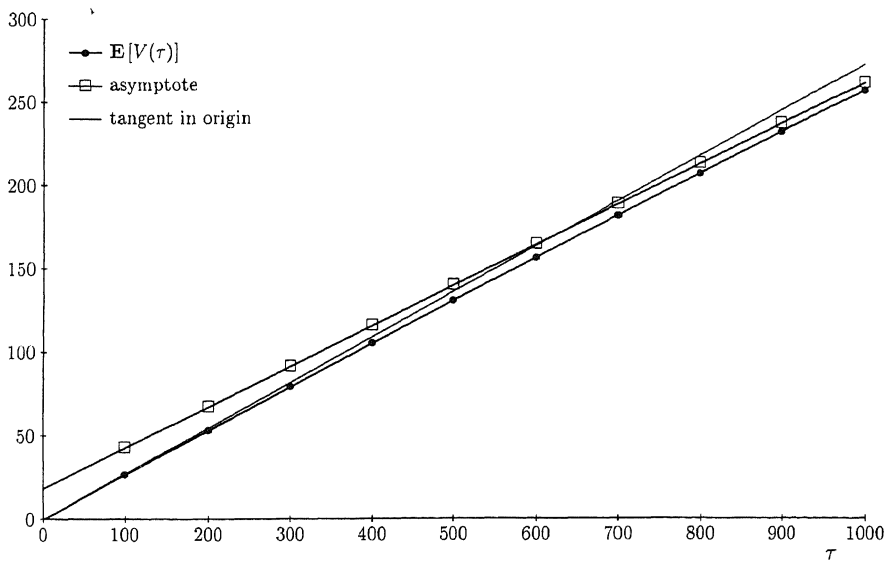

Figure 4. Asymptote and tangent line of $\boldsymbol{E}[V(\tau)]$.

Example. In Núñez Queija et al. (4) the results are applied to a telecommunications model. There, the capacity available to the so-called *elastic* customers in a processor-sharing queue depends on the number of another type of customers (*stream* customers) in the system. Stream customers have preemptive priority over elastic customers. The state of the random environment $Y(t)$ is given by the number of stream customers and behaves as an independent M/M/N/N queue with arrival rate $\lambda^{(s)} = 0.81$ customers per second, mean service time $h^{(s)} = 10$ s and $N = 31$. Elastic customers also arrive according to a Poisson process with rate $\lambda^{(e)} = 2.17$ customers per second and have mean service requirement $1/\mu = 50$ Mbit (service requirements correspond to data file sizes). Stream customers require a bandwidth 5 Mbit/s, while individual elastic customers can use at most a *peak rate* $r_+^{(e)} = 10$ Mbit/s. The total bandwidth is 155 Mbit/s, hence, the maximum number of stream customers $N = 155/5 = 31$. $\boldsymbol{E}[V(\tau)]$ (the mean sojourn time of elastic customers) is displayed in Figure 4. For values of τ up to an order of magnitude larger than the mean service requirement, the tangent in the origin is closer to the actual curve than the asymptote.

The slope δ of the tangent line is equal to the initial expected "delay per unit of service" upon arrival of a customer in steady state:

$$\delta := \sum_{(k,i)\in\mathbf{S}^*} a_{k,i} \frac{k}{c_i^{(k)}}. \tag{21}$$

Note that when the tangent line in the origin is close to the exact value, the mean of $V(\tau)$ is "almost" proportional to τ, the proportionality constant being given by δ.

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 (4) it is observed that both approximations are an *upper bound* for $\boldsymbol{E}[V(\tau)]$, and that for practical purposes the minimum of the two gives a useful approximation. Therefore we propose to use the following refined approximation, by combining the two previously mentioned ones (for the case that $\mathbf{S}_0^* = \emptyset$).

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

The experiments in (4) support this approximation, which only depends on steady-state characteristics and can therefore be efficiently computed. The approximation can be improved by computing more than the first coefficient (δ) of the Taylor-series. This may be done iteratively by using Expression (18), 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.

Remark 7.4. The models evaluated in (4) form a special subclass of the general framework depicted in Figure 1. In particular, the capacity allocated to a single

customer, $c_i^{(k)}/k$, is a nonincreasing function of k (the total number of customers). For practical situations this seems to be a reasonable assumption.

8. GENERALIZATIONS

In Section 2 we made some assumptions that 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 into the framework, or how they can be included in an analogous but generalized analysis. A general drawback in these generalizations is that the QBD structure, which was convenient for computation of various entities, is lost. This may render computations impractical for realistic system sizes.* However, this section shows that qualitative properties derived within the QBD setting are retained by several generalizations. In particular we find in all cases that the conditional mean sojourn time as a function of the amount of work τ, has an asymptote for $\tau \to \infty$.

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. The class of phase-type distributions was described by Neuts (12, 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:

$$\mathbf{S}^* := \left\{ (k_1, k_2, \ldots, k_P; i) \;\middle|\; \begin{array}{l} 0 \le k_1 + k_2 + \ldots + k_P \le L - 1, \\ k_j \in \{0, 1, 2, \ldots, L-1\}, \\ i \in \{1, 2, 3, \ldots, N\} \end{array} \right\}.$$

Here the QBD structure is lost. Note that when studying the process with one permanent customer, the role of the random environment and the service phases of nonpermanent customers is not fundamentally different. We may redefine the random environment such that it also contains the service phases of nonpermanent customers, and then view the resulting model as a special case of the earlier model with $L = 1$.

*These effects have not been studied with rigor in this paper. Within the QBD setting, systems with $L \approx 1500$ and $N \approx 30$ could be evaluated within minutes on a regular personal computer (Intel Pentium processor).

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's service. 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. This discipline was introduced by Kleinrock (6) and already studied via Crump-Mode-Jagers branching processes by Grishechkin (25). 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 $w_j/(k_1 w_1 + \cdots + k_J w_J)$ 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 FCFS and LCFS (Last Come First Served)—may be incorporated by a proper definition of the random environment.

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 (article) may be generalized to infinite state spaces, under recurrence conditions that are stronger than requiring ergodicity. For instance, the existence of the vector γ in Corollaries 3.3 and 5.4 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 (34).

In applications, it is usually the case that the $c_i^{(k)}$ are uniformly bounded from above, so that the rewards in the Markov–Reward processes of the proofs

of Corollaries 3.3 and 5.4 are uniformly bounded. In that case the vector $\boldsymbol{\gamma}$ exists under the assumption of ergodicity. To see this we may proceed as in (19, p. 188) to construct a relative reward vector that satisfies the conditions given for $\boldsymbol{\gamma}$ in Corollaries 3.3 and 5.4. In the same way, we may show that the mean of these constructed relative rewards exists and is finite, so that we may normalize as required in Corollaries 3.3 and 5.4.

Moreover, in applications when $L = \infty$ and $N < \infty$, it is often the case that the QBD process with generator $\mathcal{G}$ given by Definition (2) is homogeneous beyond some level, that is, there is a positive integer K such that $M^{(k)} = M$, $Q_d^{(k)} = Q_d$, and $\Lambda^{(k)} = \Lambda$, for all $k \geq K$ (see for instance Núñez Queija et al. (4). The ergodicity condition is then $\mathbf{p}\Lambda\mathbf{1} < \mathbf{p}M\mathbf{1}$, with $\mathbf{p}[M + Q_d + \Lambda] = \mathbf{0}$, where $\mathbf{0}$ is a vector of zeroes, see Neuts (12, Theorem 3.1.1).

We finally remark that for infinite generators, the exponential function as in Equation (7) may be defined by its Taylor-series representation.

9. CONCLUDING REMARKS

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 service requirement, 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 rewards in a particular Markov–Reward process. The latter was shown to be closely related to the traditional branching process approach to study processor-sharing systems with constant total 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 model with fluctuating service capacity) was explained by comparing the steady-state distributions of the original queueing model and the model obtained by the random time change.

We discussed how the conditional mean of the sojourn times as a function of the service requirement may be computed. A numerically stable algorithm was developed, but the computational complexity calls for reliable and efficient approximations. Numerical results motivated an approximation that only depends on steady-state characteristics.

The analysis was shown to include the case of service requirements with a phase-type distribution. We also saw that the more general discriminatory

processor-sharing service discipline fits into our framework. We discussed extensions to infinite state spaces, and showed that for bounded service rates the analysis still applies.

ACKNOWLEDGMENT

The author thanks Onno Boxma, Sem Borst, and Hans van den Berg for helpful comments.

REFERENCES

1. ATM Forum Technical Committee. Traffic Management Specification; Version 4.0; April, 1996.
2. Van der Wal, K.; Mandjes, M.R.H.; Bastiaansen, H. Delay Performance Analysis of the new Internet Services with Guaranteed QoS. Proc. of the IEEE **1997**, *85*, 1947–1957.
3. White, P.; Crowcroft, J. The Integrated Services in the Internet: State of the Art. Proc. of the IEEE **1997**, *85*, 1934–1946.
4. Núñez Queija, R.; Van den Berg, J.L.; Mandjes, M.R.H. Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic. In *Teletraffic Engineering in a Competitive World*, Proceedings of the 16th International Teletraffic Congress, Edinburgh; Key, P., Smith, D., Eds.; Elsevier: Amsterdam, 1999.
5. Kleinrock, L. Analysis of a Time-Shared Processor. Naval Res. Logistics Quarterly **1964**, *11*, 59–73.
6. Kleinrock, L. Time-Shared Systems: A Theoretical Treatment. J. Assoc. Comput. Mach. **1967**, *14*, 242–261.
7. Sakata, M.; Noguchi, S.; Oizumi, J. Analysis of a Processor Shared Queueing Model for Time Sharing Systems. In Proc. 2nd Int. Conf. on *System Sciences*, Hawaii, 1969, 625–628.
8. Kleinrock, L. *Queueing Systems, Vol. II: Computer Applications*; Wiley: New York, 1976.
9. Cohen, J.W. The Multiple Phase Service Network with Generalized Processor Sharing. Acta Informatica **1979**, *12*, 245–284.
10. Núñez Queija, R. Sojourn Times in a Processor Sharing Queue with Service Interruptions. Queueing Syst. **2000**, *34*, 351–386.
11. Yashkov, S.F. A Derivation of Response Time Distribution for a M/G/1 Processor Sharing Queue. Probl. Control and Inf. Theory **1983**, *12*, 133–148.
12. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Johns Hopkins: Baltimore, U.S., 1981.
13. De Nitto Personè, V.; Grassi, V. Solution of Finite QBD Processes. J. Appl. Probab. **1996**, *33*, 1003–1010.
14. Li, S.-Q.; Sheng, H.-D. Generalized Folding-Algorithm for Sojourn Time Analysis of Finite QBD Processes and its Queueing Applications. Comm. Statist. Stochastic Models **1996**, *12*, 507–522.
15. Iyer, B.R.; Donatiello, L.; Heidelberger, P. Analysis of Performability for Stochastic Models of Fault-Tolerant Systems. IEEE Trans. on Comp. **1986**, *C-35*, 902–907.

16. Smith, R.M.; Trivedi, K.S.; Ramesh, A.V. Performability Analysis: Measures, an Algorithm, and a Case Study. IEEE Trans. on Comp. **1988**, *37*, 406–417.
17. Reibman, A.; Smith, R.M.; Trivedi, K.S. Markov and Markov Reward Model Transient Analysis: An Overview of Numerical Approaches. Eur. J. Oper. Res. **1989**, *40*, 257–267.
18. Howard, R.A. *Dynamic Programming and Markov Processes*; M.I.T. Press: New York, 1960.
19. Tijms, H.C. *Stochastic Models: An Algorithmic Approach*; Wiley: Chichester, England, 1994.
20. Zijm, W.H.M. Exponential Convergence in Undiscounted Continuous-Time Markov Decision Chains. Math. of Operations Res. **1987**, *12*, 700–717.
21. Coffman, E.G.; Muntz, R.R.; Trotter, H. Waiting Time Distributions for Processor-Sharing Systems. J. Assoc. Comput. Mach. **1970**, *17*, 123–130.
22. Yashkov, S.F. New Applications of Random Time Change to the Analysis of Processor-Sharing Queues. In Proceedings of the 4th International Conference on *Probability Theory and Mathematical Statistics*, Vilnius, 1985, 343–345.
23. Foley, R.D.; Klutke, G.-A. Stationary Increments in the Accumulated Work Process in Processor-Sharing Queues. J. Appl. Probab. **1989**, *26*, 671–677.
24. Grishechkin, S.A. Crump-Mode-Jagers Branching Processes as a Method of Investigating M/G/1 Systems with Processor Sharing. Theory Probab. Appl. **1991**, *36*, 19–35; translated from Teor. Veroyatnost. i Primenen. **1991**, *36*, 16–33 (in Russian).
25. Grishechkin, S.A. On a Relationship between Processor-Sharing Queues and Crump-Mode-Jagers Branching Processes. Adv. in Appl. Probab. **1992**, *24*, 653–698.
26. Yashkov, S.F. Mathematical Problems in the Theory of Processor-Sharing Queueing Systems. J. Soviet Math. **1992**, *58*, 101–147.
27. Ross, S.M. *Applied Probability Models with Optimization Applications*; Holden-Day: San Francisco, 1970.
28. Wolff, R.W. Time Sharing with Priorities. SIAM J. Appl. Math. **1970**, *19*, 566–574.
29. Van den Berg, J.L. *Sojourn Times in Feedback and Processor Sharing Queues*, Ph.D. Thesis; Rijksuniversiteit Utrecht, 1990.
30. Van den Berg, J.L.; Boxma, O.J. The M/G/1 Queue with Processor Sharing and its Relation to a Feedback Queue. Queueing Syst. **1991**, *9*, 365–401.
31. Donatiello, L.; Grassi, V. On Evaluating the Cumulative Performance Distribution of Fault-Tolerant Computer Systems. IEEE. Trans. on Comp. **1991**, *40*, 1301–1307.
32. Nabli, H.; Sericola, B. Performability Analysis: A New Algorithm. IEEE Trans. on Comp. **1996**, *45*, 491–494.
33. De Souza e Silva, E.; Gail, H.R. An Algorithm to Calculate Transient Distributions of Cumulative Rate and Impulse Based Reward. Comm. Statist. Stochastic Models **1998**, *14*, 509–536.
34. Sennott, L.I. Value Iteration in Countable State Average Cost Markov Decision Processes with Unbounded Costs. Ann. Oper. Res. **1991**, *28*, 261–271.
35. Krieger, U.R.; Naoumov, V.; Wagner, D. Analysis of a Versatile Multi-Class Delay-Loss System with a Superimposed Markovian Arrival Process. Eur. J. Oper. Res. **1998**, *108*, 425–437.
36. Núñez Queija, R. *Processor-Sharing Models for Integrated-Services Networks*, Ph.D. Thesis; Eindhoven University of Technology, 2000; ISBN 90-646-4667-8, http://www.cwi.nl/~sindo.

37. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; Wiley: New York, 1994.
38. Yashkov, S.F. Processor-Sharing Queues: Some Progress in Analysis. Queueing Syst. **1987**, *2*, 1–17.

Received April 29, 1999
Returned for revisions August 8, 2000
Accepted September 11, 2000