Department of Applied Mathematics
Faculty of EEMCS

University of Twente
The Netherlands

P.O. Box 217
7500 AE Enschede
The Netherlands
Phone: +31-53-4893400
Fax: +31-53-4893114
Email: memo@math.utwente.nl
www.math.utwente.nl/publications

Memorandum No. 1753

Decomposing the queue length distribution of processor-sharing models into queue lengths of permanent customer queues

S-K. Cheung, J.L. van den Berg [1]
and R.J. Boucherie

March, 2005

ISSN 0169-2690

[1]TNO Information and communication Technology, P.O. Box 5050, 2600 GB Delft

Decomposing the queue length distribution of processor-sharing models into queue lengths of permanent customer queues

Sing-Kong Cheung*, Hans van den Berg†*, and Richard J. Boucherie*

* Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

† TNO Information and Communication Technology
P.O. Box 5050, 2600 GB Delft, The Netherlands

E-mail: {S.K.Cheung, R.J.Boucherie}@utwente.nl, J.L.vandenBerg@telecom.tno.nl

Abstract: We obtain a decomposition result for the steady state queue length distribution in egalitarian processor-sharing (PS) models. In particular, for an egalitarian PS queue with K customer classes, we show that the marginal queue length distribution for class k factorizes over the number of other customer types. The factorizing coefficients equal the queue length probabilities of a PS queue for type k in isolation, in which the customers of the other types reside *permanently* in the system. Similarly, the (conditional) mean sojourn time for class k can be obtained by conditioning on the number of permanent customers of the other types. The decomposition result implies linear relations between the marginal queue length probabilities, which also hold for other PS models such as the egalitarian processor-sharing models with state-dependent system capacity that only depends on the total number of customers in the system. Based on the exact decomposition result for egalitarian PS queues, we propose a similar decomposition for discriminatory processor-sharing (DPS) models, and numerically show that the approximation is accurate for moderate differences in service weights.

Keywords: Processor-sharing queues, queue length, decomposition, permanent customers, approximation, discriminatory processor-sharing.

AMS Subject Classifications: primary 60K25, 90B22; secondary: 60K37

1 Introduction

The processor-sharing (PS) service discipline is of considerable interest in many application areas in which different users receive a share of a scarce common system resource. In particular, in the field of the performance evaluation of computer and communication systems, the PS discipline has been widely adopted as a convenient paradigm for modelling bandwidth sharing.

Kleinrock [13] introduced the simplest and best known egalitarian PS discipline, in which a single server assigns each customer a fraction $1/n$ of the service capacity when there are n customers in the system. In particular, he showed that the mean sojourn time conditional on the service requirement $x > 0$ is proportional to x. For an extensive body of literature on (egalitarian) PS queues, we refer to Yashkov's survey papers [18, 19] and references therein.

Cohen [8] generalized the standard PS model into a PS model in which each customer receives a service rate according to an arbitrary positive function $\phi(n)$. By appropriate choice of $\phi(n)$ this model describes a very wide class of service disciplines, and this model significantly enhances the

modelling capabilities of the standard PS model. In many practical applications it models the main factors determining the performance, while on the other hand, it is simple enough to be analytically tractable, see e.g. [4, 15].

Another generalization of PS is the discriminatory processor-sharing (DPS) discipline, where a customer of type k receives service rate $w_k/\sum_{j=1}^{K} w_j n_j$, according to the set of weights $\{w_j : j = 1, ..., K\}$, and when n_j customers of type j are present in the system. If all weights w_j are equal, then we have the ordinary PS queue. Under DPS it is possible to give preferential treatment (non-preemptive) to one or more customer classes at the expense of others. The range of applications for DPS is extremely large; see e.g. [1, 6, 12]. Exact analysis of DPS turns out to be difficult compared to ordinary PS. Therefore, results for DPS are scarce in the literature. Most notably, the simple geometric queue length distribution for the ordinary PS queue does not have a counterpart for DPS, and tractable transform results for the sojourn time distribution seem not to exist.

For DPS, Fayolle et. al. [9] showed that the conditional mean sojourn times satisfy a system of integro-differential equations and derived closed form expressions for the case of exponential service requirements. In that case, the unconditional sojourn times can be obtained from a system of linear equations. Rege and Sengupta [17] obtained the moments of the queue length distributions as the solutions to linear equations for the case of exponential service requirements, and they also proved a heavy-traffic limit theorem for the joint queue length distribution. These results were extended to phase type distributions by Van Kessel et. al. [11]. More recently, Avrachenkov et. al. [2] show that the mean queue lengths of all classes are finite under the usual stability condition, regardless of the higher moments of the service requirements. They also showed that the conditional sojourn times of the different customer classes are stochastically ordered according to the DPS weights.

In the present paper, for multi-class egalitarian and discriminatory PS models we investigate a decomposition of the queue length and sojourn time distributions into the marginal queue length distributions for models with permanent customers. In particular, for the egalitarian PS model we obtain an exact and analytically tractable decomposition that is remarkable and interesting on its own and offers additional insights into egalitarian PS queues. We apply this decomposition to discriminatory PS to obtain an efficient and analytically tractable approximation of the queue length distribution and mean sojourn times.

More specific, for a two class egalitarian PS queue with Poisson arrivals λ_1, λ_2, when (N_1, N_2) is the joint steady state queue length distribution, we show that the marginal queue length distribution N_i is in distribution equal to $\widetilde{N}_i^{(N_j+1)*}$, with $i \neq j$, and $\widetilde{N}_i$ is the steady state queue length of a single class $M/G/1$ PS queue with arrival rate λ_i (same as in the two class PS queue), and $\widetilde{N}_i^{(m+1)*}$ denotes the $(m+1)$-fold convolution of $\widetilde{N}_i$. The random variable $\widetilde{N}_i^{(m+1)*}$ can be interpreted as the queue length of a $M/G/1$ PS queue with arrival rate λ_i and with m permanent customers; see e.g. [5]. The decomposition result implies that the marginal queue length distribution for class 1, factorizes over the number of class 2 customers, and where the factorizing coefficients are equal to the queue length probabilities of an isolated PS queue for type 1, given that type 2 customers are permanent in the system. This queue length decomposition result can be generalized for arbitrary number of classes K, and similar results hold for other egalitarian PS models, e.g., PS networks with feedback customers, and PS models with state-dependent but *balanced* class capacities, which are treated in Section 3.

In Section 4, we propose an approximation method for DPS based on the queue length decomposition result. The basic assumption is that an isolated customer class in DPS is considered to behave like an egalitarian PS model with reduced capacity and a *random environment* that is exogenously determined. Similar idea is presented in Lee [14], to assume independence between the various customer classes in light traffic. More specifically, if one type of customers is treated as permanent in a general two class DPS model, then the model is analytically tractable for the non-permanent class type with reduced service capacity that is exogenously given. The approximations are obtained as solutions of linear systems, which can be applied under general DPS frameworks.

2 Model

In this section we introduce a general single server processor-sharing model with K customer classes and we introduce the notation used in this paper. Customers arrive at a single server according to individual and independent Poisson processes with rate $\lambda_k > 0$, for customer class k. The required service times of type k customers are i.i.d. random variables with a general distribution $F_k(x) = \mathbb{P}(X_k \leq x)$ with mean $\mathbb{E}X_k$. Denote the load of class k by $\rho_k = \lambda_k \mathbb{E}X_k$, and the total offered load by $\rho := \sum_{j=1}^{K} \rho_j$. The server shares its capacity among all customers present in the system. Denote $\mathbf{n} = (n_1, ..., n_K)$, with n_j the number of customers of type j present in the system. The server capacity may be dependent on the system state. When the system state is $\mathbf{n}$, the total rate class k receives is $\phi_k(\mathbf{n})$. All customers within a class k type share the capacity $\phi_k(\mathbf{n})$ in an egalitarian manner, i.e., each customer in class k receives rate $\phi_k(\mathbf{n})/n_k$. We assume that $\phi_k(\mathbf{n}) = 0$ if and only if $n_k = 0$. The total server capacity is denoted by $\phi(\mathbf{n}) := \sum_{k=1}^{K} \phi_k(\mathbf{n})$.

This general model describes a very wide class of service disciplines. In particular, it includes the following special cases of processor-sharing models.

1. Egalitarian processor-sharing (with fixed capacity): $\phi_k(\mathbf{n}) = \frac{n_k}{\sum_{j=1}^{K} n_j}$.

2. Egalitarian processor-sharing with state-dependent service capacity: $\phi_k(\mathbf{n}) = \frac{n_k \phi(\mathbf{n})}{\sum_{j=1}^{K} n_j}$. Note that in the original *generalized* PS model studied by Cohen [8], $\phi(\mathbf{n})$ only depends on $\mathbf{n}$ through its sum $n_1 + ... + n_K$.

3. Discriminatory processor-sharing (with fixed capacity): $\phi_k(\mathbf{n}) = \frac{w_k n_k}{\sum_{j=1}^{K} w_j n_j}$.

Furthermore, this model framework also covers DPS models with state-dependent service capacity $\phi(\mathbf{n})$, state-dependent weights $w_k(\mathbf{n})$, and state-dependent service rate $\phi_k(\mathbf{n})/n_k$ for each type k customer.[1]

The egalitarian PS models 1 and 2 (when $\phi(\mathbf{n})$ only depends on $\mathbf{n}$ through its sum $n_1 + ... + n_K$) are analytically tractable. In particular, analytical expressions are available for the equilibrium distributions of customers simultaneously present in the system (and marginal distributions), mean number of customers $\mathbb{E}N_k$ of class k, mean sojourn time $\mathbb{E}T_k$ and conditional mean sojourn time $\mathbb{E}T_k(x)$ of a class k customer given its initial service requirement $x > 0$. For DPS models (Model 3), these expressions have not yet been obtained (in general and tractable form).

3 Decomposition of Egalitarian Processor-Sharing models

In this section, we first establish decomposition results for the ordinary egalitarian PS model. Results for more general egalitarian PS models are briefly indicated at the end of this section.

Consider an egalitarian processor-sharing model with two types of customers (indexed by $l = 1, 2$), with $\phi_l(n_1, n_2) = \frac{n_l}{n_1 + n_2}$, where the second class of customers is possibly an aggregate of several other classes. Let (N_1, N_2) denote the joint steady state queue length in this processor-sharing model. The joint steady state queue length distribution has the product form (cf. [8, 10])

$$\mathbb{P}(N_1 = i; N_2 = j) = (1 - \rho) \binom{i + j}{j} \rho_1^i \rho_2^j, \tag{3.1}$$

when the stability condition is satisfied, i.e., $\rho := \rho_1 + \rho_2 < 1$, and is insensitive to the service time distributions apart from their means; see e.g. [7]. From the key identity $\sum_{i=0}^{\infty} \binom{i+j}{i} \rho^i \equiv$

[1] Note that in single class case: state-dependent service capacity is equivalent to state-dependent service rate, while in the case of multiple classes the equivalence does not hold. In the latter case: state-dependent service rate incorporates both state-dependent weights and state-dependent service capacity.

$\sum_{i=0}^{\infty}\binom{i+j}{j}\rho^i \equiv \left(\frac{1}{1-\rho}\right)^{j+1}$, the marginal queue length distributions are easily obtained:

$$\mathbb{P}(N_1 = i) = \frac{1-\rho}{1-\rho_2}\left(\frac{\rho_1}{1-\rho_2}\right)^i, \quad i \in \mathbb{Z}_+, \tag{3.2}$$

$$\mathbb{P}(N_2 = j) = \frac{1-\rho}{1-\rho_1}\left(\frac{\rho_2}{1-\rho_1}\right)^j, \quad j \in \mathbb{Z}_+. \tag{3.3}$$

3.1 Queue length decomposition

Theorem 1 shows how the marginal steady state queue length probabilities of the two class PS queue can be related through the negative binomial probabilities $\alpha(i,j)$ and $\beta(j,i)$, defined as

$$\alpha(i,j) := \mathbb{P}\left(\widetilde{N}_1^{(j+1)*} = i\right) = (1-\rho_1)^{j+1}\binom{i+j}{i}\rho_1^i, \quad \sum_{i=0}^{\infty}\alpha(i,j) = 1, \quad \text{for all } j \in \mathbb{Z}_+, \tag{3.4}$$

$$\beta(j,i) := \mathbb{P}\left(\widetilde{N}_2^{(i+1)*} = j\right) = (1-\rho_2)^{i+1}\binom{i+j}{j}\rho_2^j, \quad \sum_{j=0}^{\infty}\beta(j,i) = 1, \quad \text{for all } i \in \mathbb{Z}_+, \tag{3.5}$$

where $\widetilde{N}_k$ denotes the steady state queue length of an *isolated* $M/G/1$ PS queue with arrival rate λ_k, general service requirement distribution $F_k(x)$, and $\widetilde{N}_k^{m*}$ denotes the m-fold convolution of the random variable $\widetilde{N}_k$. Assume that $\widetilde{N}_i$ is independent of N_j, for $i \neq j$.

Theorem 1 For $i,j = 1,2$ and $i \neq j$, the marginal queue length N_i is in distribution equal to the random variable $\widetilde{N}_i^{(N_j+1)*}$, i.e.,

$$N_i \stackrel{d}{=} \widetilde{N}_i^{(N_j+1)*}, \tag{3.6}$$

where $\widetilde{N}_i^{(N_j+1)*} := \sum_{m=0}^{N_j}\widetilde{N}_{i,m}$, with $\left\{\widetilde{N}_{i,m}\right\}_{m\geq 0}$ i.i.d. and distributed as $\widetilde{N}_i$.

Proof. First observe that the following equality holds by combining (3.1)-(3.5):

$$\mathbb{P}(N_1 = i; N_2 = j) = \alpha(i,j)\mathbb{P}(N_2 = j) = \beta(j,i)\mathbb{P}(N_1 = i). \tag{3.7}$$

Hence, for all $i \in \mathbb{Z}_+$:

$$\begin{aligned}\mathbb{P}\left(\widetilde{N}_1^{(N_2+1)*} = i\right) &= \sum_{j=0}^{\infty}\mathbb{P}\left(\widetilde{N}_1^{(N_2+1)*} = i \mid N_2 = j\right)\mathbb{P}(N_2 = j) \\ &= \sum_{j=0}^{\infty}\alpha(i,j)\mathbb{P}(N_2 = j) = \sum_{j=0}^{\infty}\mathbb{P}(N_1 = i; N_2 = j) = \mathbb{P}(N_1 = i). \end{aligned} \tag{3.8}$$

Analogously, $\mathbb{P}\left(\widetilde{N}_2^{(N_1+1)*} = j\right) = \mathbb{P}(N_2 = j)$ for all $j \in \mathbb{Z}_+$. ■

Corollary 2 From (3.8) we obtain the following set of linear equations:

$$\mathbb{P}(N_1 = i) = \sum_{j=0}^{\infty}\alpha(i,j)\mathbb{P}(N_2 = j), \tag{3.9}$$

$$\mathbb{P}(N_2 = j) = \sum_{i=0}^{\infty}\beta(j,i)\mathbb{P}(N_1 = i). \tag{3.10}$$

The decomposition theorem can be generalized for K classes with joint steady state distribution $\mathbb{P}(N_1 = n_1; ...; N_K = n_K) = (1-\rho)\left(\sum_{i=1}^K n_i\right)! \prod_{i=1}^K \rho_i^{n_i}/n_i!$, and the marginal distributions $\mathbb{P}(N_k = n_k) = \frac{1-\rho}{1-\rho+\rho_k}\left(\frac{\rho_k}{1-\rho+\rho_k}\right)^{n_k}$ can be decomposed into $N_k \stackrel{d}{=} \widetilde{N}_k^{(N_{-k}+1)*}$, for all $k = 1, ..., K$, where we denote $N_{-k} := \sum_{i=1, i\neq k}^K N_i$.

Theorem 1 can be interpreted as follows. When we observe the queue length N_1 in isolation in the two class PS model at an arbitrary moment in time (in *steady state*), then it also seems that with probability $\mathbb{P}(N_2 = j)$ we are observing the queue length in a single class $M/G/1$ PS queue with arrival rate λ_1 and with j permanent customers. To this end, note that the queue length distribution in an ordinary $M/G/1$ PS queue with j permanent customers, is in distribution equal to the $(j+1)$-fold convolution of the queue length distribution of the same model without permanent customers; see [5].

We have to stress that $a(i,j)$ is not defined as the conditional steady state distribution of the number of type 1 customers conditioned on the number of j type 2 customers in steady state, i.e., $a(i,j)$ is not defined as $\mathbb{P}(N_1 = i \mid N_2 = j) := \mathbb{P}(N_1 = i; N_2 = j)/\mathbb{P}(N_2 = j)$. It is defined as the steady state queue length distribution of an isolated type 1 queue with j *permanent* customers in the system. However, the remarkable fact is that $\mathbb{P}(N_1 = i \mid N_2 = j)$ and $\alpha(i,j) := \mathbb{P}\left(\widetilde{N}_1 = i \mid j \text{ permanent customers}\right)$ are identical.

From the class 1 point-of-view in the original two class PS model, it seems as if class 1 behaves according to an ordinary single class $M_{\lambda_1}/G/1$ PS queue with j permanent customers if j customers of type 2 are present in the system. Furthermore, if there is a customer arrival (resp. departure) for type 2 in the system, then it seems as if class 1 *instantaneously* 'jumps' to a $M_{\lambda_1}/G/1$ model with $j+1$ (resp. $j-1$) permanent customers, and as if the new equilibrium (steady state behavior) is instantaneously attained at the jump epoch.

3.2 Sojourn time decomposition

After establishing the queue length decomposition result, a natural question is whether or not a similar decomposition result holds for the sojourn time distribution. As the sojourn time distribution (conditional and unconditional) in a $M_{\lambda_1}/G/1$ PS queue with j permanent customers, is distributed as the $(j+1)$-fold convolution of the sojourn distribution of the same $M_{\lambda_1}/G/1$ PS queue without permanent customers (see [5]), one could expect that $T_1(x) \stackrel{d}{=} \widetilde{T}_1^{(N_2+1)*}(x)$ and $T_2(x) \stackrel{d}{=} \widetilde{T}_2^{(N_1+1)*}(x)$, where $T_k(x)$ is the conditional sojourn time for customer type k (with initial service requirement $x > 0$) in the original two class PS model, and $\widetilde{T}_k(x)$ is the conditional sojourn time for customer type k in the isolated queue. However, it is easily seen that this is not true in general. For example, take $\lambda_1 = 0$ and $\lambda_2 > 0$, then $\widetilde{T}_1^{(N_2+1)*}(x) \stackrel{d}{=} (N_2+1)x$, and the latter random variable is well-known to be insensitive to the service time distribution $F_2(x)$, apart from its mean. However, since a customer with fixed service requirement is concerned in the original two class PS model, it must hold that $T_1(x) \stackrel{d}{=} T_2(x)$ and this is not insensitive to the distribution $F_2(x)$ (only the mean $\mathbb{E}[T_k(x)]$ is insensitive to the service time distributions), hence $T_1(x) \stackrel{d}{=} \widetilde{T}_1^{(N_2+1)*}(x)$ can not hold in general. However, it can be easily shown that $\mathbb{E}[T_1(x)] = \mathbb{E}\left[\widetilde{T}_1^{(N_2+1)*}(x)\right]$. See Theorem 3.

Theorem 3 The conditional mean sojourn times can be decomposed into

$$\mathbb{E}[T_1(x)] = \mathbb{E}\left[\widetilde{T}_1^{(N_2+1)*}(x)\right] \equiv \sum_{j=0}^{\infty} \frac{(j+1)x}{1-\rho_1}\mathbb{P}(N_2 = j), \tag{3.11}$$

$$\mathbb{E}[T_2(x)] = \mathbb{E}\left[\widetilde{T}_2^{(N_1+1)*}(x)\right] \equiv \sum_{i=0}^{\infty} \frac{(i+1)x}{1-\rho_2}\mathbb{P}(N_1 = i), \tag{3.12}$$

where $(m+1)x/(1-\rho_k)$ is the mean conditional sojourn time of an isolated $M/G/1$ PS queue with arrival rate λ_k, service requirement distribution $F_k(x)$ and with m permanent customers.

Proof. From (3.2), (3.3), (3.11), (3.12) it is readily verified that

$$\mathbb{E}\left[\widetilde{T}_1^{(N_2+1)*}(x)\right] = \mathbb{E}\left[\widetilde{T}_2^{(N_1+1)*}(x)\right] = \frac{x}{1-\rho}, \tag{3.13}$$

which is the same as the well-known result $\mathbb{E}[T_1(x)] = \mathbb{E}[T_2(x)] = x/(1-\rho)$, e.g. see [13, 18]. ■

Obviously, the result for unconditional mean sojourn times is similar; since it also follows directly from the exact decomposition result for queue length distributions and Little's Law.

3.3 Generalization to a feedback network with egalitarian processor-sharing

Consider a processor-sharing network with an egalitarian PS node and a node used by a single feedback customer. Exogenous customer arrivals at the PS node form a Poisson process with rate $\lambda > 0$, and these customers are served at the PS node with i.i.d. service requirements (generally distributed with mean $\mathbb{E}X$). The service requirement for the feedback customer at the PS node is generally distributed and denoted by the random variable Z. After service completion of the feedback customer at the PS node, the feedback customer is routed to the feedback node (with probability 1) where he spends a generally distributed time Y. After this random time Y at the feedback node, the feedback customer joins the PS node for a service requirement Z.

If we denote $\mathbb{P}(N^{PS} = n)$ as the steady state distribution of the number of (non-feedback) customers at the PS node, then it is readily verified that the following decomposition holds:

$$\mathbb{P}(N^{PS} = n) = \xi \cdot \pi_0(n) + (1-\xi) \cdot \pi_1(n), \tag{3.14}$$

where $\pi_i(n)$ is the steady state distribution of the $M_\lambda/G/1$ PS queue with i permanent customers, and ξ is the steady state probability that the feedback customer is at the feedback node in the network, i.e., $\pi_0(n) = (1-\rho)\rho^n$, $\pi_1(n) = (1-\rho)^2(n+1)\rho^n$, $\xi = \frac{\mathbb{E}Y}{\mathbb{E}Y + \mathbb{E}Z/(1-\rho)}$ and $\rho := \lambda\mathbb{E}X$.

The result can be extended to multiple feedback customers where the feedback node is a so-called BCMP [3] node. In fact, the feedback node may be replaced by a BCMP network.

3.4 Generalization to egalitarian processor-sharing queues with state-dependent capacity

Consider the processor-sharing queue with K customer classes served in egalitarian manner, with the total service capacity dependent on the system state $\mathbf{n}$ through its sum $n_1 + .. + n_K$, cf. [8]. More precisely, $\phi(\mathbf{n}) = \varphi(\mathbf{n} \cdot \mathbf{e})$, for all $\mathbf{n} \neq \mathbf{0}$ (null vector), where $\mathbf{e}$ is the vector with 1-entries of appropriate length, $\mathbf{n} \cdot \mathbf{e}$ denotes the inner product, and where $\varphi(x) : \mathbb{N} \to \mathbb{R}_+$ is an arbitrary positive function. Serving the customers in egalitarian manner reads

$$\frac{\phi_i(\mathbf{n})}{n_i} = \frac{\varphi(\mathbf{n} \cdot \mathbf{e})}{\mathbf{n} \cdot \mathbf{e}}, \text{ for all } i = 1, ..., K, \tag{3.15}$$

and the class capacities $\phi_i(\mathbf{n})$ are uniquely characterized and balanced by (see [10, 7])

$$\phi_i(\mathbf{n}) = \frac{\Phi(\mathbf{n} - \mathbf{e}_i)}{\Phi(\mathbf{n})}, \tag{3.16}$$

where $\Phi(\mathbf{n})$ is the so-called *balance function*, and $\mathbf{e}_i$ is the i-th unity vector of appropriate length. It is said that the class capacities $\phi_i(\mathbf{n})$ are balanced if a function $\Phi(\mathbf{n})$ exists such that (3.16) is

satisfied, and equivalently the class capacities $\phi_i(\mathbf{n})$ are balanced if

$$\phi_i(\mathbf{n}-\mathbf{e}_j)/\phi_i(\mathbf{n}) = \phi_j(\mathbf{n}-\mathbf{e}_i)/\phi_j(\mathbf{n}), \text{ for all } i,j, \text{ and } n_i>0, n_j>0. \tag{3.17}$$

From (3.15) and (3.16), we get $\Phi(\mathbf{n}) = \frac{(\mathbf{n}\cdot\mathbf{e})!}{\prod_{i=1}^K n_i!}\left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}}\varphi(j)\right)^{-1}$, and without restriction $\varphi(0)\equiv 1$. The joint steady state queue length distribution $\pi(\mathbf{n}) := \mathbb{P}(N_1=n_1;...;N_K=n_K)$ is given by the product form (see [7])

$$\pi(\mathbf{n}) = (\mathbf{n}\cdot\mathbf{e})!\left(G\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}}\varphi(j)\right)^{-1}\prod_{i=1}^K \rho_i^{n_i}/n_i!, \text{ for } \mathbf{n}\neq\mathbf{0}. \tag{3.18}$$

with $\rho_i = \lambda_i \mathbb{E}X_i$, and a normalizing constant G. It can be shown that the marginal distributions of (3.18) can be decomposed into queue lengths of (isolated) permanent customer queues.

Theorem 4 For multi-class egalitarian processor-sharing models, with *balanced* class capacities $\phi_k(\mathbf{n}) = \varphi(\mathbf{n}\cdot\mathbf{e})n_k/(\mathbf{n}\cdot\mathbf{e})$, the marginal steady state queue length distribution can be decomposed into $N_k \overset{d}{=} \widetilde{N}_k^{(N_{-k}+1)*}$, for all $k=1,...,K$, and where we denote $N_{-k} := \sum_{i=1,i\neq k}^K N_i$.

Proof. The decomposition for class k follows from the observation that

$$\left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}}\varphi(j)\right)^{-1} \equiv \left(\prod_{l=1}^{n_k}\varphi(l+(\mathbf{n}\cdot\mathbf{e}-n_k))\right)^{-1}\left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}-n_k}\varphi(j)\right)^{-1}, \text{ for } n_k\geq 1, \tag{3.19}$$

with $\varphi(0)\equiv 1$. Hence, with (3.19) and by appropriate summation of (3.18), the marginal queue length distribution for class k equals

$$\mathbb{P}(N_k=n_k) = \sum_{\substack{n_1,...,n_{k-1}\\ n_{k+1},...,n_K}} \pi(\mathbf{n}) \sim \sum_{\substack{n_1,...,n_{k-1}\\ n_{k+1},...,n_K}} (\mathbf{n}\cdot\mathbf{e})!\left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}}\varphi(j)\right)^{-1}\prod_{i=1}^K \rho_i^{n_i}/n_i! \tag{3.20}$$

$$= \sum \frac{(\mathbf{n}\cdot\mathbf{e})!}{n_k!\prod_{i\neq k}^K n_i!}\left\{\left(\prod_{l=1}^{n_k}\varphi(l+(\mathbf{n}\cdot\mathbf{e}-n_k))\right)^{-1}\rho_k^{n_k}\right\}\cdot\left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}-n_k}\varphi(j)\right)^{-1}\prod_{i\neq k}^K \rho_i^{n_i}, \tag{3.21}$$

where the symbol $\sim$ denotes equality up to a multiplicative constant.

The proof is readily completed, by observing that the expression between parentheses in (3.21) is up to a multiplicative constant (and a combinatorial expression) equivalent to the queue length distribution $\mathbb{P}(\widetilde{N}_k = n_k \mid \mathbf{n}\cdot\mathbf{e}-n_k$ permanent customers), for type k in isolation and with $\mathbf{n}\cdot\mathbf{e}-n_k$ permanent customers of the other types. The expression after the parentheses in (3.21) is equivalent to the marginal steady state probability $\mathbb{P}(N_{-k}=\mathbf{n}\cdot\mathbf{e}-n_k)$, after appropriate summation. ■

4 Approximation for Discriminatory Processor-Sharing models

In this section we propose an approximation method for (unconditional) mean sojourn times in *general* discriminatory processor-sharing models. The basic approximation assumption in the DPS model is, that a class k queue (in isolation) is considered as an egalitarian PS model with (reduced) state-dependent capacity, and where the state-dependent capacity for class k is exogenously determined. In the exact DPS model this is obviously not the case, since the *random environments* for the different isolated queues in DPS are interrelated and not independent. We investigate the 'error' impact if this assumption is made under DPS models. By the exact queue length decomposition results for

egalitarian PS models (with state-dependent and balanced class capacities), our method provides exact results if applied on these egalitarian models.

4.1 General approximation method for mean sojourn times

For sake of notational convenience, first we consider a two class DPS model where $\phi_l(n_1, n_2), l = 1, 2$, can be an arbitrary positive function. This is a *generalized* DPS model with state-dependent service rates (possibly including state-dependent weights $\frac{\phi_1(n_1,n_2)/n_1}{\phi_2(n_1,n_2)/n_2}$). In addition, we assume a finite number of service positions for both customer types separately ($N_1 \leq m$ and $N_2 \leq n$), which is not a crucial assumption.

4.1.1 Approximation method for $K = 2$ customer classes

If one customer type is treated as permanent in the system, then the model is analytically tractable for the non-permanent type. More precisely, the probabilities $\alpha(i, j)$ and $\beta(j, i)$ are easily computed in closed form by (see [8, 15])

$$\begin{aligned} \alpha(i,j) &= \frac{\rho_1^i \varphi_{1,i}(j)}{\sum_{k=0}^m \rho_1^k \varphi_{1,k}(j)}, \qquad (4.1) \\ \text{where } \varphi_{1,i}(j) &= \left(\prod_{k=1}^{i} \phi_1(k,j) \right)^{-1}, \varphi_{1,0}(j) \equiv 1 \text{ for all } j = 0, 1, ..., n, \\ \beta(j,i) &= \frac{\rho_2^j \varphi_{2,j}(i)}{\sum_{k=0}^n \rho_2^k \varphi_{2,k}(i)}, \qquad (4.2) \\ \text{where } \varphi_{2,j}(i) &= \left(\prod_{k=1}^{j} \phi_2(i,k) \right)^{-1}, \varphi_{2,0}(i) \equiv 1 \text{ for all } i = 0, 1, ..., m. \end{aligned}$$

Our basic approximation assumption for DPS models is that the linear system given in Corollary 2 is applicable. Under this assumption, we approximate the marginal distributions $\eta_i = \mathbb{P}(N_1 = i)$ and $\xi_j = \mathbb{P}(N_2 = j)$ by solving the following set of linear equations:

$$\begin{aligned} \eta_i &= \sum_{j=0}^{n} \alpha(i,j)\xi_j, \quad \text{for } i = 0, 1, ..., m, \qquad (4.3) \\ \xi_j &= \sum_{i=0}^{m} \beta(j,i)\eta_i, \quad \text{for } j = 0, 1, ..., n, \qquad (4.4) \end{aligned}$$

cf. Corollary 2, or in matrix form: $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\xi}$, and $\boldsymbol{\xi} = \mathbf{B}\boldsymbol{\eta}$, where $\boldsymbol{\eta} = (\eta_0, \eta_1, ..., \eta_m)^T$, $\boldsymbol{\xi} = (\xi_0, \xi_1, ..., \xi_n)^T$, and the matrices are given by

$$\mathbf{A} = \begin{pmatrix} \alpha(0,0) & \alpha(0,1) & \cdots & \alpha(0,n) \\ \alpha(1,0) & \alpha(1,1) & \cdots & \alpha(1,n) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha(m,0) & \alpha(m,1) & \cdots & \alpha(m,n) \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \beta(0,0) & \beta(0,1) & \cdots & \beta(0,m) \\ \beta(1,0) & \beta(1,1) & \cdots & \beta(1,m) \\ \vdots & \vdots & \ddots & \vdots \\ \beta(n,0) & \beta(n,1) & \cdots & \beta(n,m) \end{pmatrix}. \qquad (4.5)$$

It is not difficult to give conditions such that the (approximated) probability vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are uniquely determined after normalization. The system of equations is also equivalent to $\boldsymbol{\eta} = (\mathbf{AB})\boldsymbol{\eta}$, or $\boldsymbol{\xi} = (\mathbf{BA})\boldsymbol{\xi}$, which can be interpreted as 'solving the equation $\pi = \pi\mathcal{P}$', where $\mathcal{P}$ is a transition matrix of a discrete-time Markov chain. In many practical DPS models, it is easily verified that the product matrices $(\mathbf{AB})^T$ and $(\mathbf{BA})^T$, have row sums equal to one and do not have negative entries

(irreducible, regular stochastic matrices). It is sufficient to have $\phi_j(\mathbf{n}) > 0$ for all j, and for all vectors $\mathbf{n}$ with $n_j > 0$, to guarantee uniqueness of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, up to a multiplicative constant.

The approximated (unconditional) mean sojourn time for each class follows from Little's law, and in our case with finite capacity (blocking) we have the approximation:

$$\mathbb{E}\widehat{T}_1 = \frac{1}{\lambda_1(1-\eta_m)} \sum_{i=0}^{m} i \cdot \eta_i, \tag{4.6}$$

$$\mathbb{E}\widehat{T}_2 = \frac{1}{\lambda_2(1-\xi_n)} \sum_{j=0}^{n} j \cdot \xi_j, \tag{4.7}$$

Remark 5 The proposed approximation is exact for egalitarian PS models with *balanced* class capacities. The steady state queue length distribution is insensitive to the service time distributions if and only if the class capacities are *balanced* (see [7]), hence the approximation (4.1)-(4.7) can not be exact for models with *unbalanced* capacities, since the approximation is insensitive to the service time distributions.

4.1.2 Outline of the approximation method for $K > 2$ customer classes

In principle, our approximation can be applied for general number of customer classes K. The method seems very efficient, since only linear systems have to be solved. However, with increasing K significantly more computational effort needs to be done. To illustrate the complexity, let us consider the case of $K = 3$ classes. Suppose that the class capacities $\phi_l(n_1, n_2, n_3)$, $l = 1, 2, 3$, are given in a three class PS model with system states $(N_1, N_2, N_3) = (i, j, k)$. The approximated marginal steady state probabilities $\eta_i = \mathbb{P}(N_1 = i), \xi_j = \mathbb{P}(N_2 = j), \zeta_k = \mathbb{P}(N_3 = k)$ are uniquely obtained (up to a multiplicative constant) from the linear equations (4.8), (4.9), (4.10):

$$\left\{ \begin{array}{rcl} \eta_i & = & \sum_j (\sum_k \alpha(i \mid j,k)\pi_{3,2}(k \mid j))\xi_j \\ \eta_i & = & \sum_k (\sum_j \alpha(i \mid j,k)\pi_{2,3}(j \mid k))\zeta_k \end{array} \right. \begin{array}{rcl} =: & \sum_j a_{i,j}\xi_j \\ =: & \sum_k b_{i,k}\zeta_k \end{array}, \tag{4.8}$$

$$\left\{ \begin{array}{rcl} \xi_j & = & \sum_i (\sum_k \beta(j \mid i,k)\pi_{3,1}(k \mid i))\eta_i \\ \xi_j & = & \sum_k (\sum_i \beta(j \mid i,k)\pi_{1,3}(i \mid k))\zeta_k \end{array} \right. \begin{array}{rcl} =: & \sum_i c_{j,i}\eta_i \\ =: & \sum_k d_{j,k}\zeta_k \end{array}, \tag{4.9}$$

$$\left\{ \begin{array}{rcl} \zeta_k & = & \sum_i (\sum_j \gamma(k \mid i,j)\pi_{2,1}(j \mid i))\eta_i \\ \zeta_k & = & \sum_j (\sum_i \gamma(k \mid i,j)\pi_{1,2}(i \mid j))\xi_j \end{array} \right. \begin{array}{rcl} =: & \sum_i e_{k,i}\eta_i \\ =: & \sum_j f_{k,j}\xi_j \end{array}. \tag{4.10}$$

The coefficients $\alpha(i \mid j,k)$ are easily computed, similar to (4.1), since $\alpha(i \mid j,k)$ is the steady state queue length distribution for the isolated type 1 queue given that type 2 and 3 customers reside permanently in the system. Analogously, the coefficients $\beta(j \mid i,k)$ and $\gamma(k \mid i,j)$ are also easily computed. The pairs of coefficients $\{\pi_{2,1}(j \mid i), \pi_{3,1}(k \mid i)\}$, $\{\pi_{1,2}(i \mid j), \pi_{3,2}(k \mid j)\}$, and $\{\pi_{2,3}(j \mid k), \pi_{1,3}(i \mid k)\}$ are obtained as unique solutions (up to multiplicative constant) from the linear systems (4.11), (4.12), (4.13):

$$\left\{ \begin{array}{rcl} \pi_{2,1}(j \mid i) & = & \sum_k \beta(j \mid i,k)\pi_{3,1}(k \mid i) \\ \pi_{3,1}(k \mid i) & = & \sum_j \gamma(k \mid i,j)\pi_{2,1}(j \mid i) \end{array} \right., \text{ for all } i, \tag{4.11}$$

$$\left\{ \begin{array}{rcl} \pi_{1,2}(i \mid j) & = & \sum_k \alpha(i \mid j,k)\pi_{3,2}(k \mid j) \\ \pi_{3,2}(k \mid j) & = & \sum_i \gamma(k \mid i,j)\pi_{1,2}(i \mid j) \end{array} \right., \text{ for all } j, \tag{4.12}$$

$$\begin{cases} \pi_{1,3}(i \mid k) &= \sum_j \alpha(i \mid j,k)\pi_{2,3}(j \mid k) \\ \pi_{2,3}(j \mid k) &= \sum_i \beta(j \mid i,k)\pi_{1,3}(i \mid k) \end{cases}, \text{ for all } k. \tag{4.13}$$

The systems (4.8), (4.9), (4.10) written in matrix form: $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\xi}$, $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\zeta}$, $\boldsymbol{\xi} = \mathbf{C}\boldsymbol{\eta}$, $\boldsymbol{\xi} = \mathbf{D}\boldsymbol{\zeta}$, $\boldsymbol{\zeta} = \mathbf{E}\boldsymbol{\eta}$, $\boldsymbol{\zeta} = \mathbf{F}\boldsymbol{\xi}$, are efficiently solved by e.g. the following two systems:

$$\begin{aligned} \boldsymbol{\eta} &= (\mathbf{ACBFDE})\,\boldsymbol{\eta}, \\ \boldsymbol{\xi} &= (\mathbf{CADEBF})\,\boldsymbol{\xi}, \end{aligned}$$

with normalization $\boldsymbol{\eta} \cdot \mathbf{e} = 1$ and $\boldsymbol{\xi} \cdot \mathbf{e} = 1$, and where the system for determining $\boldsymbol{\zeta}$ is automatically satisfied and normalized. For increasing K, it seems that convenient notation may overcome the increase in complexity.

4.2 Conservation law

In this section, we obtain a conservation law for unconditional mean sojourn times in a DPS queue, which turns out to be useful in improving the approximations for the lowest priority class.

Theorem 6 For a K class DPS queue with fixed capacity, Poisson arrivals λ_i, and exponential service requirements with mean $\mathbb{E}X_i$, $i = 1, ..., K$, the following conservation law for unconditional mean sojourn times holds:

$$\sum_{j=1}^{K} \rho_j \mathbb{E}T_j = \sum_{i=1}^{K} \frac{\rho_i}{1-\rho} \mathbb{E}X_i, \tag{4.14}$$

and independently of $(w_1, ..., w_K)$.

Proof. Provided that $\mathbb{E}X_j^2 < \infty$ we can use the conservation law for DPS models from [2]:

$$\sum_{j=1}^{K} \lambda_j \int_0^\infty \mathbb{E}T_j(x)\,(1 - F_j(x))\,dx = \overline{U},$$

where $\overline{U}$ is the time average unfinished work in the system. In case of Poisson arrivals we have

$$\overline{U} = \frac{\sum_{i=1}^{K} \lambda_i \mathbb{E}X_i^2}{2(1-\rho)},$$

see [2]. The proof is readily completed, since in case of exponential service requirements, we have $(1 - F_j(x))\,dx = \frac{1}{\mu_j} dF_j(x)$, and $\mathbb{E}X_j^2 = 2\,(\mathbb{E}X_j)^2 = 2/\mu_j^2$. ■

Remark 7 The practical use of a conservation law is that if we are able to obtain accurate approximations of $\mathbb{E}T_k$ for customer classes $k = 1, ..., K-1$, then an accurate approximation for class K follows automatically.

4.3 Numerical Results

In this section, we numerically investigate our approximation method with exact results in case of exponential service requirements, for the two and three class DPS models and with fixed capacity.

4.3.1 Two class DPS queue

In the two class DPS model we refer to type 1 customers as the *high* priority customers and to type 2 customers as the *low* priority customers ($w_1 > w_2$). In case of exponential service requirements,

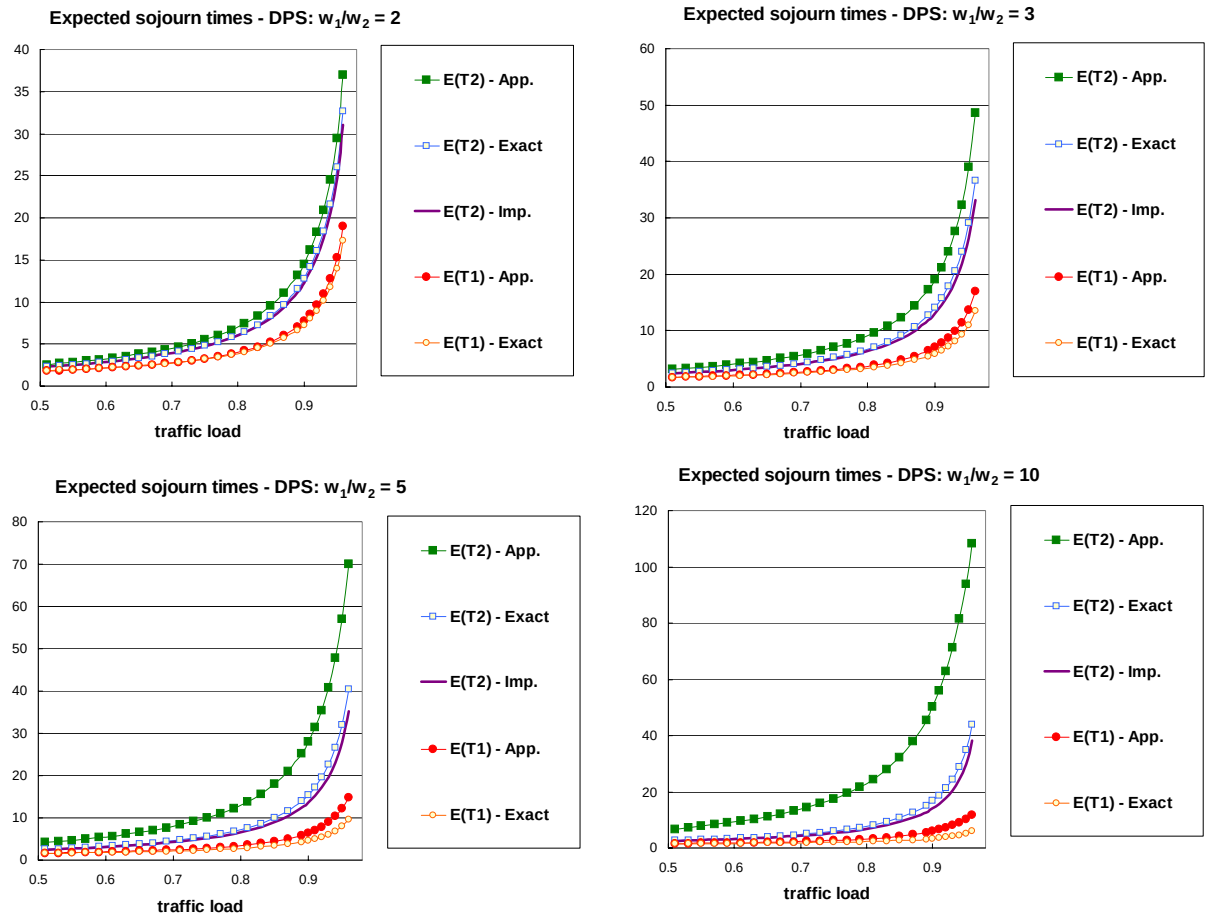

Figure 1: Exact and approximated mean sojourn times $\mathbb{E}T_j$ for 2-class DPS with exponential service requirements with $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$, for weight ratios $\frac{w_1}{w_2} \in \{2, 3, 5, 10\}$.

exact closed form expressions are given by (see [9])

$$\mathbb{E}T_1 = \frac{1/\mu_1}{1-\rho_1-\rho_2}\left(1 + \frac{\mu_1\rho_2\,(w_2 - w_1)}{\mu_1 w_1\,(1-\rho_1) + \mu_2 w_2\,(1-\rho_2)}\right), \tag{4.15}$$

$$\mathbb{E}T_2 = \frac{1/\mu_2}{1-\rho_1-\rho_2}\left(1 + \frac{\mu_2\rho_1\,(w_1 - w_2)}{\mu_1 w_1\,(1-\rho_1) + \mu_2 w_2\,(1-\rho_2)}\right). \tag{4.16}$$

The approximated mean sojourn times $\mathbb{E}\widehat{T}_1$ and $\mathbb{E}\widehat{T}_2$ are calculated from (4.1)-(4.7) with $\phi_1(i,j) = i \cdot w_1/(i \cdot w_1 + j \cdot w_2)$, $\phi_2(i,j) = j \cdot w_2/(i \cdot w_1 + j \cdot w_2)$ and with infinite buffer capacity ($m = n = \infty$). The direct approximation $\mathbb{E}\widehat{T}_2$ (based on decomposition) can be improved by using the conservation law and the direct approximation $\mathbb{E}\widehat{T}_1$ for the high priority class. The improvement $\mathbb{E}\overline{T}_2$ for the low priority class is given by

$$\mathbb{E}\overline{T}_2 = \frac{\rho_1}{\rho_2}\left(\frac{1/\mu_1}{1-\rho} - \mathbb{E}\widehat{T}_1\right) + \frac{1/\mu_2}{1-\rho}. \tag{4.17}$$

Figure 1 provides graphs for the exact and approximated mean sojourn times for both classes with $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$, and for different values of the weight ratio w_1/w_2. For class 2, in addition, the improved approximation (4.17) is included. Figure 1 gives results as function of $\rho = \rho_1 + \rho_2$, with $\rho_1 = \rho_2 = \rho/2$. As can be seen from these graphs, the approximation for $\mathbb{E}T_1$ is good up to a traffic load $\rho = 0.9$ for all weight ratios. The approximation for $\mathbb{E}T_2$ breaks down with increasing

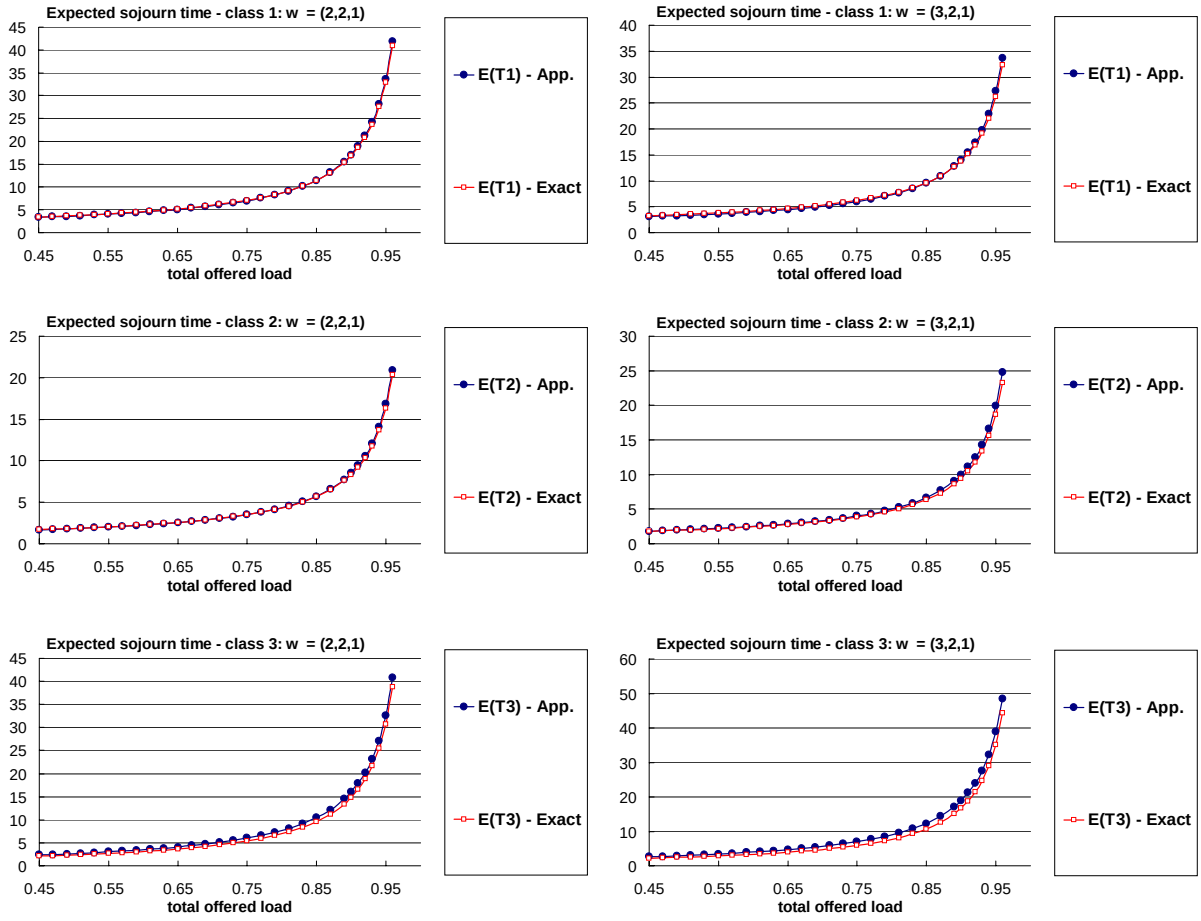

Figure 2: Exact and approximated mean sojourn times $\mathbb{E}T_j$ for 3-class DPS with exponential service requirements ($\mathbb{E}X_1 = 2, \mathbb{E}X_2 = \mathbb{E}X_3 = 1$), for weights $w = (2, 2, 1)$ and $w = (3, 2, 1)$.

weight ratio. However, the approximation (4.17) that uses $\mathbb{E}\widehat{T}_1$ to approximate $\mathbb{E}T_2$ is accurate for all weight ratios. For a discussion of the quality of the approximation, we refer to Section 4.4.

4.3.2 Three class DPS queue

For the three class DPS model with fixed capacity, infinite buffer, and with exponential service requirements, we consider the following numerical examples with mean service requirements $\mathbb{E}X_1 = 2$, and $\mathbb{E}X_2 = \mathbb{E}X_3 = 1$. The exact value for $\mathbb{E}T_j, j = 1, 2, 3$, can be obtained from [9] as solution of a linear system.

Figure 2 provides graphs for the exact and approximated mean sojourn times for the three classes and for two sets of weights $w = (w_1, w_2, w_3)$, respectively for $w = (2, 2, 1)$ and $w = (3, 2, 1)$. Figure 3 provides approximated and exact mean sojourn times for $w = (5, 3, 1)$ and $w = (10, 3, 1)$. The approximated mean sojourn times $\mathbb{E}\widehat{T}_j, j = 1, 2, 3$, are calculated according to the equations (4.8)-(4.13) in Section 4.1.2 and by applying Little's law. The figures are provided as function of the total load $\rho := \rho_1 + \rho_2 + \rho_3$, with $\rho_1 = \rho_2 = \rho_3 = \rho/3$. In addition, in Figure 3, an improved approximation $\mathbb{E}\overline{T}_3$ is included, based on the conservation law (4.14) and based on the direct approximations $\mathbb{E}\widehat{T}_1$ and $\mathbb{E}\widehat{T}_2$ of the other types.

As can be seen from the graphs (Figure 2 and 3), the approximations for $\mathbb{E}\widehat{T}_j$, are accurate as long as the set of weights is 'more or less balanced', see Remark 5. It seems that our approximation improves for $K = 3$ customer classes. This can be explained by the fact that adding an additional customer class can increase the *balance* between the classes.

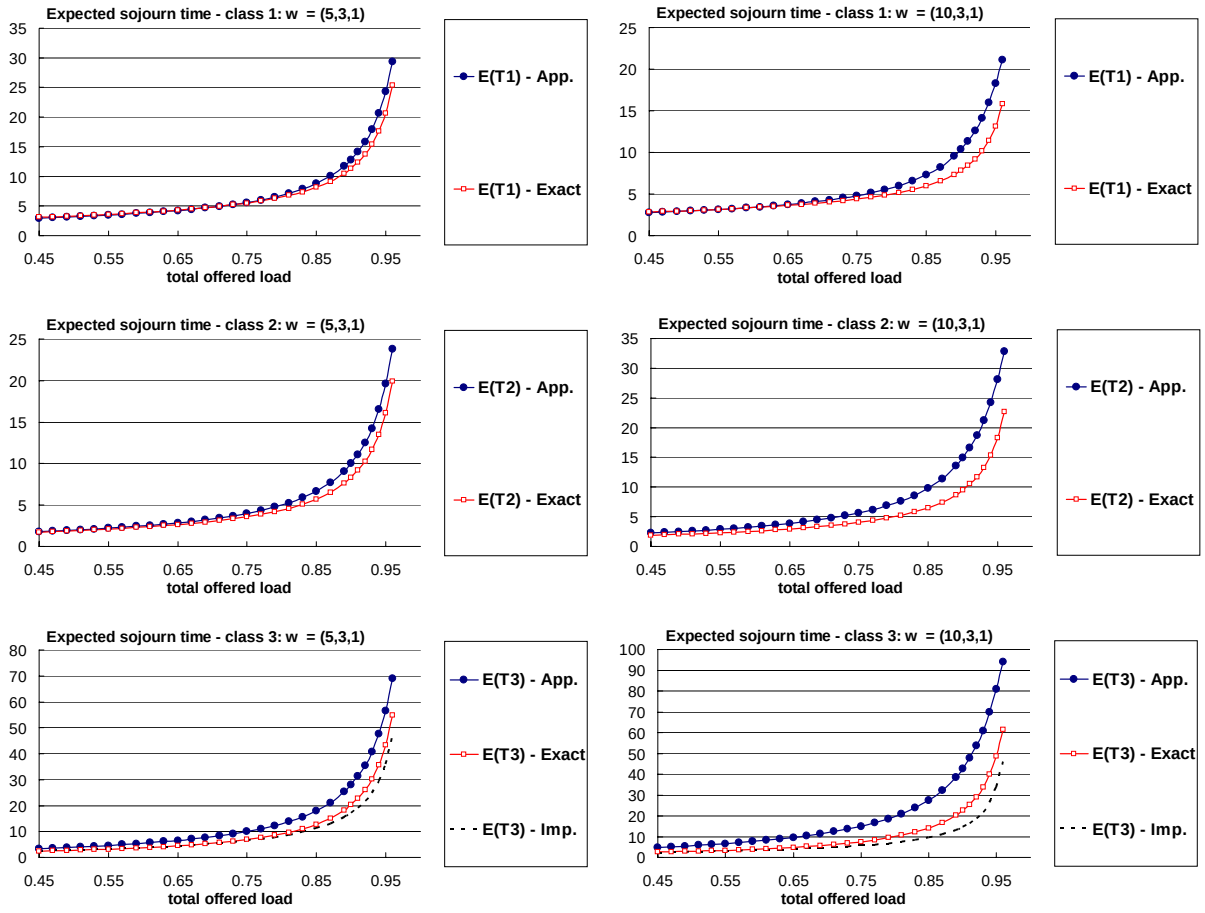

Figure 3: Exact and approximated mean sojourn times $\mathbb{E}T_j$ for 3-class DPS with exponential service requirements $(\mathbb{E}X_1 = 2, \mathbb{E}X_2 = \mathbb{E}X_3 = 1)$, for weights $w = (5, 3, 1)$ and $w = (10, 3, 1)$.

4.4 Discussion

In this section we discuss the quality of our approximation $\mathbb{E}\widehat{T}_j$ for $\mathbb{E}T_j$. In particular, in the case of $K = 2$ customer classes, numerical examples indicate that the approximation for the lower priority class $\mathbb{E}\widehat{T}_2$ is poor when the ratio of weight w_1/w_2 is extremely large (unbalanced), whereas the improved approximation $\mathbb{E}\overline{T}_2$ is very accurate.

Our basic approximation assumption is that the various customer classes in DPS models are treated as isolated customer classes that behave independently of the other classes and according to single class egalitarian PS queues with state-dependent (and reduced) service capacity. Supported by the queue length decomposition result for egalitarian PS models, the isolated single class PS queues in a multi-class egalitarian PS queue are exactly related to the other isolated customer queues.

When the ratio of weights w_1/w_2 is large, then from class 2 point-of-view the queue behaves as an ON-OFF processor-sharing queue [16]. As an illustration, Figure 4 shows the typical behavior of the queue length processes $N_i(t)$ for a two class DPS under heavy load and large ratio w_1/w_2. From a class 2 point-of-view, it seems as if a *burst of permanent customers* (of size w_1/w_2) arrives instantaneously when a single customer of type 1 arrives in the original two class DPS model. Therefore, when the number of class 1 customers gets large enough, then the service process for class 2 may seem frozen (OFF period), and the queue length process for class 2 increases rapidly. However, since also the high priority customers (class 1) reside in the system for a relatively short time period (class 1 gets a large share of the capacity), the queue length for the high priority class will decrease rapidly. In this case, when there is no high priority customer in the system, the low priority class receives all

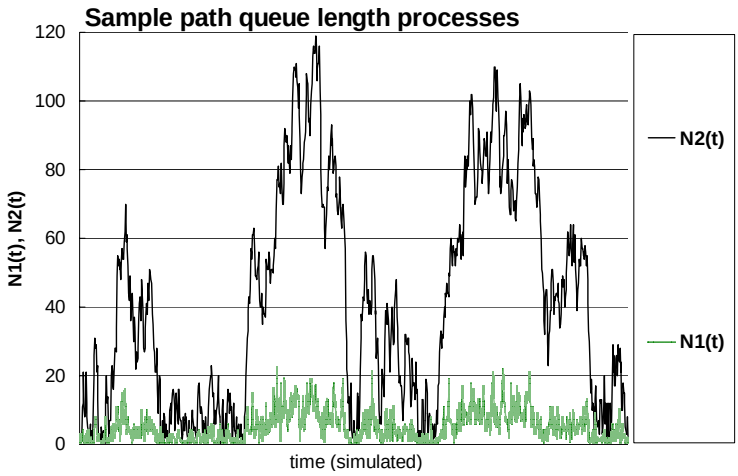

Figure 4: A sample path of the queue length process $N_1(t)$ and $N_2(t)$, for a 2 class DPS model with $w_1/w_2 = 10$, $\lambda_1 = \lambda_2 = 0.49$, and $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$ (exponential service requirements).

the available service capacity despite the large ratio of weights w_1/w_2 (ON period), and the queue length for the low priority class decreases significantly.

If the traffic load of the system is near its saturation (i.e., $\rho < 1, \rho \approx 1$), then the queue length (and sojourn times) for class 2 will become very large, whereas the queue length (and sojourn times) for class 1 stays relatively small (compared to class 2). In the original two class DPS model, the isolated customer class 2 has a *random environment* that is severely influenced by the '*burstiness*' of class 1 (seen from class 2 point-of-view), while the random environment for class 1 is much less dependent on the queue length process of class 2. From an isolated class 1 point-of-view, it seems as if class 1 behaves according to its own single class and isolated (egalitarian) PS queue, with a random environment that is less fluctuating over time (compared to the isolated class 2 point-of-view).

For the case of $K \geq 3$ customer classes, similar behavior is present in the DPS model. The queue length process of the highest priority class has a significant influence on the queue length process of the lowest priority class, and not in the other way round. However, in the case that more customer classes are present in the system, with service weights that are in between the highest and lowest priority class, the marginal influence of the highest priority class on the random environment of lowest priority class may be less than in the case of $K = 2$.

5 Conclusion

In this paper, we obtained a decomposition result for the queue length distributions in the egalitarian processor-sharing models. In particular, for a K class egalitarian processor-sharing model, the marginal steady state distribution N_k for class k, satisfies $N_k \stackrel{d}{=} \widetilde{N}_k^{(N_{-k}+1)*}$. The latter random variable can be interpreted as a random variable denoting the queue length of an isolated class k processor-sharing queue, where the other customer types are permanent customers in the system and N_{-k} represents the total number of permanent customers. This result remains valid for egalitarian processor-sharing models with state-dependent system capacity that only depends on the sum $n_1 + ... + n_K$.

Motivated by these results, we have proposed an approximation for mean sojourn times in *general* DPS models. The numerically efficient method is also applicable for DPS models with state-dependent service rates and state-dependent weights. Numerical results have indicated that our approximation is accurate for a wide range of the weight ratios and for moderate loads. The approximation error is small for all loads if the DPS queue has 'nearly balanced' class capacities, which is

in agreement with the exact queue length decomposition results. In heavy traffic and for extreme weight ratios w_1/w_2 (in case $K = 2$) or extreme *unbalanced set of weights* $\{w_k : k = 1, ..., K\}$, insights provided in this paper suggests other approximations, e.g., exploit processor-sharing models with ON-OFF periods. This remains a topic for further research.

Acknowledgment

This research has been carried out partly within the project Beyond 3G supported by the Dutch Ministry of Economic Affairs under the program Technologische Samenwerking ICT-Doorbraakprojecten, project number TSIT1025.

References

[1] E. Altman, T. Jiménez, and D. Kofman (2004). DPS queues with stationary ergodic service times and the performance of TCP in overload. In Proceedings of IEEE INFOCOM 2004, Hong Kong.

[2] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija (2005). Discriminatory Processor Sharing Revisited. In Proceedings of IEEE INFOCOM 2005, Miami, USA.

[3] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios (1975). "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. Journal of the Association for Computing Machinery, vol 22, no. 2, April 1975, pp 248-260.

[4] J.V.L. Beckers, I. Hendrawan, R.E. Kooij, and R.D. van der Mei (2001). Generalized Processor Sharing models for Internet access lines. In Proceedings 9th IFIP conference on Performance Modeling and Evaluation of ATM & IP Networks, Budapest, June, 101-112.

[5] J.L. van den Berg (1990). Sojourn times in Feedback and Processor Sharing Queues. Ph.D. thesis, Utrecht University, The Netherlands.

[6] T. Bonald and L. Massoulié (2001). Impact of fairness on internet performance. In ACM Sigmetrics, Cambridge, Massachusetts, USA.

[7] T. Bonald, and A. Proutière (2002). Insensitivity in processor-sharing networks. Performance Evaluation 49: 193-209.

[8] J.W. Cohen (1979). The multiple phase service network with generalized processor sharing. Acta Informatica 12: 245-284.

[9] G. Fayolle, I. Mitrani, and R. Iasnogorodski (1980). Sharing a processor among many job classes. Journal of the Association for Computing Machinery 27: 519-532.

[10] F.P. Kelly (1979). Reversibility and Stochastic Networks, Wiley, Chichester.

[11] G. van Kessel, R. Núñez-Queija, and S.C. Borst (2004). Asymptotic Regimes and Approximations for Discriminatory Processor Sharing. Performance Evaluation Review 32, 44-46. Special issue on MAMA 2004, workshop on Mathematical Performance Modeling and Analysis.

[12] G. van Kessel, R. Núñez-Queija, and S.C. Borst (2005). Differentiated bandwidth sharing with disparate flow sizes. In Proceedings of IEEE INFOCOM 2005, Miami, USA.

[13] L. Kleinrock (1967). Time-shared systems: A theoretical treatment. Journal of the Association for Computing Machinery 14: 242-261.

[14] D.-S. Lee (1997). Generalized longest queue first: An adaptive scheduling discipline for ATM networks. In Proceedings of IEEE INFOCOM 1997, Kobe, Japan.

[15] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie, and M. Fleuren (2004). Analysis of flow transfer times in IEEE 802.11 wireless LANs. Annals of telecommunications, vol. 59, no. 11-12, nov.-dec. 2004, Traffic engineering and routing.

[16] R. Núñez-Queija (2000). Sojourn times in a processor sharing queue with service interruptions. Queueing Systems: Theory and Applications, vol. 34, no. 1-4, p. 351-386.

[17] K.M. Rege, and B. Sengupta (1996). Queue length distribution for the discriminatory processor-sharing queue. Operations Research 44: 653-657.

[18] S.F. Yashkov (1987). Processor-sharing queues: Some progress in analysis. Queueing Systems 2: 1-17.

[19] S.F. Yashkov (1992). Mathematical problems in the theory of processor-sharing queueing systems. Journal of Soviet Mathematics 58: 101-147.