Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing

Verloop, I.M.; Ayesta, U.; Nunez Queija, R.

Published in:
Operations Research

DOI:
10.1287/opre.1110.0914

OPERATIONS RESEARCH
Vol. 59, No. 3, May–June 2011, pp. 648–660
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5903 | 0648

DOI 10.1287/opre.1110.0914

Heavy-Traffic Analysis of a Multiple-Phase Network with Discriminatory Processor Sharing

I. M. Verloop
BCAM—Basque Center for Applied Mathematics, 48170 Derio, Spain; CWI, 1090 GB Amsterdam, The Netherlands, verloop@bcamath.org

U. Ayesta
BCAM—Basque Center for Applied Mathematics, 48170 Derio, Spain; IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain, ayesta@bcamath.org

R. Núñez-Queija
CWI, 1090 GB Amsterdam, The Netherlands; University of Amsterdam, 1018 WB Amsterdam, The Netherlands, nunezqueija@uva.nl

We analyze a generalization of the discriminatory processor-sharing (DPS) queue in a heavy-traffic setting. Customers present in the system are served simultaneously at rates controlled by a vector of weights. We assume that customers have phase-type distributed service requirements and allow that customers have different weights in various phases of their service.

In our main result we establish a state-space collapse for the queue-length vector in heavy traffic. The result shows that in the limit, the queue-length vector is the product of an exponentially distributed random variable and a deterministic vector. This generalizes a previous result by Rege and Sengupta [Rege, K. M., B. Sengupta. 1996. Queue length distribution for the discriminatory processor-sharing queue. *Oper. Res.* **44**(4) 653–657], who considered a DPS queue with exponentially distributed service requirements. Their analysis was based on obtaining all moments of the queue-length distributions by solving systems of linear equations. We undertake a more direct approach by showing that the probability-generating function satisfies a partial differential equation that allows a closed-form solution after passing to the heavy-traffic limit.

Making use of the state-space collapse result, we derive interesting properties in heavy traffic: (i) For the DPS queue, we obtain that, conditioned on the number of customers in the system, the residual service requirements are asymptotically independent and distributed according to the forward recurrence times. (ii) We then investigate how the choice for the weights influences the asymptotic performance of the system. In particular, for the DPS queue we show that the scaled holding cost reduces as classes with a higher value for $d_k/E(B_k^{fwd})$ obtain a larger share of the capacity, where d_k is the cost associated to class k, and $E(B_k^{fwd})$ is the forward recurrence time of the class-k service requirement. The applicability of this result for a moderately loaded system is investigated by numerical experiments.

Subject classifications: discriminatory processor sharing; heavy traffic; phase-type service requirements; residual service requirements; scheduling.
Area of review: Stochastic Models.
History: Received March 2009; revisions received December 2009, May 2010; accepted June 2010.

1. Introduction

The discriminatory processor-sharing (DPS) model introduced in Kleinrock (1967) is a versatile generalization of the celebrated (egalitarian) processor-sharing (PS) model. DPS allows class-based differentiation by assigning different weights to customers of different classes. (In this paper we adopt the traditional queueing theoretic terminology; often "customers" are abstract entities such as jobs, flows, packets, etc.) The processing resources are then distributed among all customers, in proportion to their relative weights. As new customers join the system and others leave after having completed their service requirement, the actual resource allocation to each customer fluctuates dynamically over time.

The asymmetric and dynamic fluctuation of the service rates give rise to complex behavior of the stochastic processes describing the numbers of customers in the system and their respective service completion times. The literature devoted to the analysis of DPS has been significantly extended over the past decade as renewed interest in DPS arose due to its relevance in communication networks with distributed control, in particular, the Internet (Altman et al. 2004). An extensive survey of the DPS literature can be found in Altman et al. (2006).

The seminal paper of Fayolle et al. (1980) provided the first analysis of the mean sojourn time conditioned on the service requirement, by solving a system of integrodifferential equations. As a by-product, the mean queue lengths of the various classes were shown to depend on

the *entire* service requirement distributions of all customer classes. This is as opposed to the egalitarian PS model, where the marginal queue lengths have a geometric distribution that only depends on the average loads of all classes, thus exhibiting a desirable insensitivity among the various classes. Although not strictly insensitive towards higher moments of service requirement distributions, the DPS model was shown to have finite mean queue lengths irrespective of any higher-order characteristics (Avrachenkov et al. 2005). This is further illustrated by the heavy-traffic bounds on the mean queue lengths reported in Aalto et al. (2007), which depend only on the service weights and the mean traffic loads. Partial insensitivity results have also been demonstrated for other performance criteria such as the class-dependent mean sojourn time conditioned on the service requirement (Avrachenkov et al. 2005), and the tail index of the sojourn time distribution (Borst et al. 2006).

Several papers have analyzed (discriminatory) processor-sharing mechanisms assuming overload conditions with general service requirement distributions. Altman et al. (2004) determine the queue-length growth rates of the standard DPS model by a fixed-point equation, generalizing the analogous result for egalitarian processor sharing (Jean-Marie and Robert 1994). More recently, further extensions to bandwidth-sharing networks (Egorova et al. 2007) and a network setting similar to ours (Ben Tahar and Jean-Marie 2009) have been obtained. In all these references the *transient* behavior of the queue lengths is studied under overload conditions while we investigate the convergence of the (scaled) *steady-state* distribution as the critical load is approached.

In the present paper, we assume that all customer classes have phase-type service requirement distributions and study the heavy-traffic behavior of a generalization of the DPS model, allowing customers to have different weights in various phases of their service. This extension allows, for example, incorporation of sophisticated scheduling techniques that give preferential treatment to customers that are close to service completion, thus reducing the number of customers in the system and their mean response times, (cf. Righter and Shanthikumar 1989). Similar generalizations of DPS were previously considered by Ben Tahar and Jean-Marie (2009), Grishechkin (1992), and Haviv and van der Wal (2008). The analysis in Grishechkin (1992) is particularly relevant for the present study. There, the generalized DPS model was investigated, assuming finite second moments of the service times. Through appropriate choices for quite a general functional of the queue-length process, Grishechkin (1992) determined the heavy-traffic distributions of the marginal queue lengths and response times (after scaling). Our results are complementary to those: on one hand, we restrict the focus to the queue lengths, and on the other hand, we study the *joint* queue-length distribution. Doing so, we establish a *state-space collapse* for the queue-length vector in heavy traffic. The result shows that in the limit, the queue-length vector is the product of an exponentially distributed random variable and a *deterministic* vector. The reduction of dimensionality of a multidimensional stochastic process under heavy-traffic scaling has been demonstrated previously in other queueing models; see, for example, Bell and Williams (2001), Stolyar (2004), and Kang et al. (2009).

Our work is inspired by the heavy-traffic analysis for the traditional DPS model with exponentially distributed service requirements in Rege and Sengupta (1996). After developing a procedure to determine all moments of the queue-length distributions from systems of linear equations, Rege and Sengupta (1996) show that the variability of the queue-length vector is of a lower order than the mean queue lengths, which directly leads to state-space collapse of the multidimensional queue-length vector. In Kessel et al. (2004) it was indicated that a similar approach could be followed for the heavy-traffic analysis of the DPS queue with phase-type distributions. Here we follow a different and more direct approach by investigating the joint probability-generating function of the queue lengths. The probability-generating function is shown to satisfy a partial differential equation that takes a convenient form after passing to the heavy-traffic limit, allowing a closed-form solution in that case. This approach allows an elegant heavy-traffic analysis for the case of phase-type distributions.

Because phase-type distributions lie dense in the class of all probability distributions, in practice the restriction to this class is not seen as being essential. In the present study, an important caveat must be accounted for, however. Because all phase-type distributions (with a finite number of phases) have a finite second moment, this restriction is implicit in our modeling approach. We do believe, however, that our results extend to general service requirements.

Allowing the relative service weights of customers to change over time as they acquire service opens up a way to implement size-based scheduling by assigning relatively high weights in service phases that are more likely to lead to a quick service completion. A classical result in the size-based literature states that the so-called $c\mu$-rule minimizes the mean holding cost in an (i) $M/G/1$-queue among all nonpreemptive work-conserving disciplines (Gelenbe and Mitrani 1980) and in a (ii) $G/M/1$-queue among all preemptive nonanticipating disciplines (Buyukkoc et al. 1985, Nain and Towsley 1994). We recall that the $c\mu$-rule is the discipline that gives strict priority in descending order of $c_k\mu_k$, where c_k and μ_k refer to a cost and the inverse of the mean service requirement, respectively, of class k. The optimality of the $c\mu$-rule can be understood from the fact that for both systems (i) and (ii), the original mean service requirement $1/\mu_k$ coincides with the expected remaining service requirement of a class-k customer *at a scheduling decision epoch*. Our analysis extends the $c\mu$-rule to DPS-like policies: in heavy traffic we show that the scaled holding cost reduces as more preference is given to customers in

service phases with an expected remaining service requirement that is small compared with its associated cost.

For the case of the standard DPS queue with phase-type service requirement distributions, we show that in the heavy-traffic setting, conditioned on the number of customers present in each class, the remaining service requirements of the various customers are independent, and distributed according to the forward recurrence times, a result that is well known for egalitarian PS (see, for example, Cohen 1979, Kelly 1979). In addition, we derive that the scaled holding cost in a DPS queue reduces as more preference is given to classes according to the cost of a class divided by its mean forward recurrence time. This provides a useful guideline to schedule a multiclass queue close to saturation for the cases not covered by the $c\mu$-rule.

The paper is organized as follows. In §2 we introduce the Markovian framework studied in the paper and state the main result, which establishes a state-space collapse of the joint queue-length vector. As a preparation for the proof of the main result, the functional equation for the generating function of the joint queue-length vector is studied in §3 and, under the heavy-traffic scaling, in §4. The proof of the main result is given in §5. Section 6 discusses size-based scheduling. Section 7 applies the state-space collapse result to the standard DPS queue with phase-type distributed service requirements. In addition, it presents the implications for residual service requirements and monotonicity properties of the holding cost. Concluding remarks can be found in §8.

2. Markovian Framework and Main Result

We consider a Markovian system with J customer types. Customers arrive according to a Poisson arrival process with rate λ, and an arriving customer is of type i with probability p_{0i}, $i = 1, \ldots, J$. Customers of type i have an exponentially distributed service requirement with mean $1/\mu_i$. After service completion, they become of type j with probability p_{ij}, $j = 1, \ldots, J$, and leave the system with probability $p_{i0} := 1 - \sum_{j=1}^{J} p_{ij}$. Let P be a $J \times J$ matrix with $P = (p_{ij})$, $i, j = 1, \ldots, J$. We assume that all customers eventually leave the system. This implies $\lim_{n\to\infty} P^n = 0$, and hence, $(I - P)^{-1}$ is well defined. In addition, we assume that none of the J types are redundant (i.e., eventually all types are observed); this assumption is formalized following Equation (1) below.

The J customer types share a common resource of capacity 1. There are strictly positive weights $g_1, \ldots, g_J$ associated with each of the types. Whenever there are q_i type-i customers, $i = 1, \ldots, J$, present in the system, each type-j customer is served at rate

$$\frac{g_j}{\sum_{i=1}^{J} g_i q_i}, \quad j = 1, \ldots, J.$$

We denote the number of type-j customers in the system by Q_j.

The above-described framework is a generalization of the standard DPS queue with phase-type distributed service requirements: It represents an $M/PH/1$ DPS queue where customers may have different weights in various phases of their service. In §7 we specify how the standard DPS queue fits into our representation.

We let R_i denote the remaining service requirement until departure for a customer that is now of type i. Note that this includes service in all subsequent stages as the customer changes from one type to another. Because the service time of each type is exponentially distributed, the expected remaining service requirements can be interpreted as absorption times in an appropriate Markov chain and therefore satisfy the following system of linear equations: $\mathbb{E}(R_i) = 1/\mu_i + \sum_{j=1}^{J} p_{ij}\mathbb{E}(R_j)$. Let $\mathbb{E}(\vec{R}) = (\mathbb{E}(R_1), \ldots, \mathbb{E}(R_J))$ and $\vec{m} = (1/\mu_1, \ldots, 1/\mu_J)$, so that we can write

$$\mathbb{E}(\vec{R})^T = (I - P)^{-1}\vec{m}^T.$$

Denote the total traffic load by

$$\rho := \lambda \sum_{j=1}^{J} p_{0j}\mathbb{E}(R_j).$$

Let γ_i represent the expected number of times a customer is of type i during its visit in the network. Hence, $\gamma_1, \ldots, \gamma_J$ satisfy the following equations:

$$\gamma_i = p_{0i} + \sum_{j=1}^{J} \gamma_j p_{ji}, \quad i = 1, \ldots, J, \tag{1}$$

i.e., $\vec{\gamma} = \vec{p}_0(I - P)^{-1}$, with $\vec{\gamma} = (\gamma_1, \ldots, \gamma_J)$ and $\vec{p}_0 = (p_{01}, \ldots, p_{0J})$. Our assumption that none of the J types is redundant entails that $\vec{\gamma}$ is a vector with strictly positive elements. Note that γ_i/μ_i represents the expected cumulative amount of service a customer requires while being of type i during its visit in the network. We denote the load corresponding to customers while they are of type i by

$$\rho_i := \lambda \frac{\gamma_i}{\mu_i}.$$

Hence, for the total traffic load ρ we may equivalently write

$$\begin{aligned}\rho = \lambda \sum_{j=1}^{J} p_{0j}\mathbb{E}(R_j) &= \lambda \vec{p}_0 \mathbb{E}(\vec{R})^T \\ &= \lambda \vec{p}_0 (I - P)^{-1}\vec{m}^T \\ &= \lambda \vec{\gamma}\vec{m}^T = \lambda \sum_{j=1}^{J} \frac{\gamma_j}{\mu_j} = \sum_{j=1}^{J} \rho_j. \end{aligned} \tag{2}$$

Our main result shows that the steady-state distribution of the queue-length vector takes a rather simple form when the system is near saturation, i.e., $\rho \uparrow 1$, which is commonly referred to as the heavy-traffic regime. This regime can be

obtained by fixing the $\vec{p}_0$, P, and $\vec{m}$, and letting

$$\lambda \uparrow \hat{\lambda} := \frac{1}{\vec{p}_0(I-P)^{-1}\vec{m}^T}, \tag{3}$$

because then $\rho = \lambda\vec{p}_0(I-P)^{-1}\vec{m}^T \uparrow 1$. Although approaching heavy traffic in this way is natural, the results remain valid for any other sequence of parameters (belonging to stable systems) that reaches heavy traffic in the limit. In heavy traffic, we denote by

$$\hat{\rho}_i = \hat{\lambda}\frac{\gamma_i}{\mu_i}$$

the load corresponding to customers while they are of type i ($\sum_{j=1}^{J}\hat{\rho}_j = 1$).

We can now state our main result, which establishes a state-space collapse for the queue-length vector in the heavy-traffic regime. We note that throughout the paper we do not explicitly reflect the dependence of the queue length processes on the traffic load ρ, in order to keep notation compact.

Proposition 1. *Consider the general Markovian framework. When scaled by $1-\rho$, the queue-length vector has a proper limiting distribution as $(\rho_1,\dots,\rho_J)\to(\hat{\rho}_1,\dots,\hat{\rho}_J)$, such that $\rho\uparrow 1$,*

$$(1-\rho)(Q_1,Q_2,\dots,Q_J) \xrightarrow{d} (\hat{Q}_1,\hat{Q}_2,\dots,\hat{Q}_J) \stackrel{d}{=} X\cdot\left(\frac{\hat{\rho}_1}{g_1},\frac{\hat{\rho}_2}{g_2},\dots,\frac{\hat{\rho}_J}{g_J}\right), \tag{4}$$

where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean

$$\mathbb{E}(X) = \frac{\sum_{j=1}^{J}\hat{\rho}_j\mathbb{E}(R_j)}{\sum_{j=1}^{J}(\hat{\rho}_j/g_j)\mathbb{E}(R_j)}. \tag{5}$$

The proof will be given in §5. Here we give some intuition for the result. Proposition 1 shows that in heavy traffic, the multidimensional queue-length process essentially reduces to a one-dimensional random process: it can be expressed as a random variable X times a deterministic vector. Given this reduced variability of the process, the value of the deterministic vector can be understood as follows. When the queue is stable, the rate conservation law (see, for example, Sigman 1991, Theorem 2.1) implies that

$$\rho_j = \mathbb{E}\left(\frac{g_jQ_j}{\sum_{i=1}^{J}g_iQ_i}\cdot\mathbf{1}_{(\sum_{i=1}^{J}Q_i>0)}\right), \tag{6}$$

because the expression within the expectation operator reflects the capacity allocated to type j. Here the function $\mathbf{1}_A$ denotes the indicator function, i.e., $\mathbf{1}_A = 1$ if A is true, and 0 otherwise. Using that the process reduces to one dimension in heavy traffic, in the limit we may replace Q_j/Q_i by a ratio of constants a_j/a_i. Together with (6) and the fact that the scaled queue length will be strictly positive in heavy traffic, this implies

$$a_j = \left(\sum_{i=1}^{J}g_ia_i\right)\frac{\hat{\rho}_j}{g_j}.$$

The prefactor $\sum_i g_ia_i$ is common to all a_j, which explains the appearance of the vector $(\hat{\rho}_1/g_1,\hat{\rho}_2/g_2,\dots,\hat{\rho}_J/g_J)$ in Proposition 1.

Numerical illustration of Proposition 1: We consider two types of customers and choose $g_1=2$, $g_2=1$, $\mu_1=2$, $\mu_2=5$, $p_{01}=0.6$, $p_{02}=0.4$, $p_{12}=0.3$, and $p_{21}=0.1$. In Figure 1 we plot the joint queue-length probabilities (obtained by simulation) for loads $\rho=0.8$ ($\rho_1\approx0.59$, $\rho_2\approx0.21$), $\rho=0.90$ ($\rho_1\approx0.66$, $\rho_2\approx0.24$), and $\rho=0.99$ ($\rho_1\approx0.73$, $\rho_2\approx0.26$), respectively. The horizontal and vertical axes correspond to Q_1 and Q_2, respectively. As a consequence of the state-space collapse stated in Proposition 1, in heavy traffic the probabilities will lie on a straight line with slope $(g_1/\hat{\rho}_1)(\hat{\rho}_2/g_2)\approx0.72$, starting from the origin. In Figure 1 we see that as the load increases, the likely states indeed tend to concentrate more around this line. For load $\rho=0.99$, this effect is clearly visible; the likely queue-length states are strongly concentrated around the line with slope 0.72.

3. Functional Equation

Before focusing on the heavy-traffic regime, we derive a functional equation for the generating function of the joint queue-length process. Denote by $\vec{Q}$ and $\vec{q}$ the vectors $(Q_1,Q_2,\dots,Q_J)\geqslant\vec{0}$ and $(q_1,q_2,\dots,q_J)\geqslant\vec{0}$, respectively. The equilibrium distribution $\pi(\vec{q}) := \mathbb{P}(\vec{Q}=\vec{q})$ satisfies

$$\lambda\pi(\vec{0}) = \sum_{i=1}^{J}\mu_ip_{i0}\pi(\vec{e}_i), \tag{7}$$

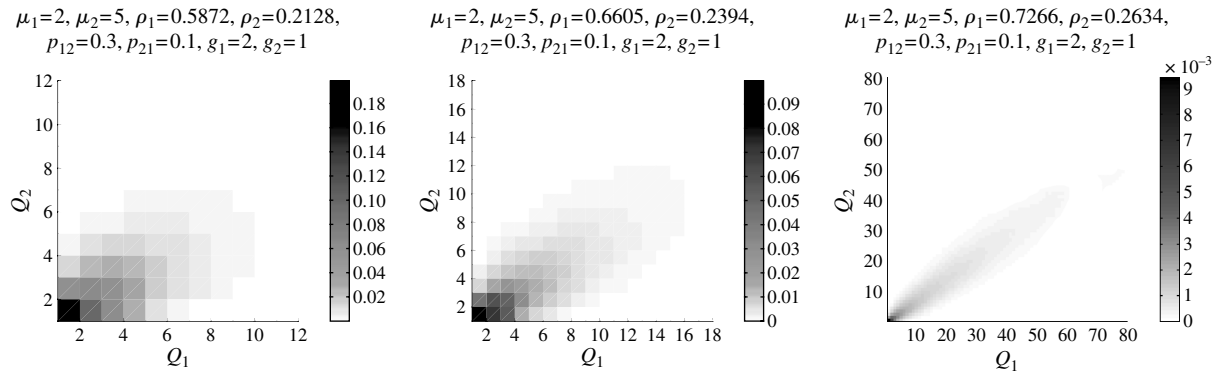
and for $\vec{q}\neq\vec{0}$,

$$\begin{aligned}\left(\lambda+\frac{\sum_{i=1}^{J}g_iq_i\mu_i}{\sum_{i=1}^{J}g_iq_i}\right)\pi(\vec{q}) &= \sum_{i=1}^{J}\lambda p_{0i}\delta_{q_i}\pi(\vec{q}-\vec{e}_i) + \sum_{i=1}^{J}\frac{g_i(q_i+1)}{\sum_{j=1}^{J}g_jq_j+g_i}\cdot\mu_ip_{i0}\pi(\vec{q}+\vec{e}_i)\\ &\quad+\sum_{i=1}^{J}\sum_{j=1}^{J}\delta_{q_j}\cdot\frac{g_i(q_i+1)}{\sum_{m=1}^{J}g_mq_m+g_i-g_j}\cdot\mu_ip_{ij}\pi(\vec{q}+\vec{e}_i-\vec{e}_j),\end{aligned} \tag{8}$$

where $\delta_q=1$ if $q>0$, and $\delta_q=0$ otherwise, and with $\vec{e}_i$ the ith unit vector. It will be notationally convenient to use the following transformation:

$$R(\vec{0})=0 \quad\text{and}\quad R(\vec{q}) = \frac{\pi(\vec{q})}{\sum_{j=1}^{J}g_jq_j}, \quad \text{for } \vec{q}\neq\vec{0}.$$

Figure 1. Joint queue-length probabilities for load $\rho = 0.8$ (left), $\rho = 0.90$ (center), and $\rho = 0.99$ (right), respectively.



Also, let $p(\vec{z})$ and $r(\vec{z})$ denote the generating functions of $\pi(\vec{q})$ and $R(\vec{q})$, respectively, where $\vec{z} = (z_1, \ldots, z_J)$ and $|z_i| < 1$ for $i = 1, \ldots, J$:

$$p(\vec{z}) = \mathbb{E}(z_1^{Q_1} \cdot \cdots \cdot z_J^{Q_J}) = \sum_{q_1=0}^{\infty} \cdots \sum_{q_J=0}^{\infty} z_1^{q_1} \cdot \cdots \cdot z_J^{q_J} \pi(\vec{q}),$$

$$r(\vec{z}) = \mathbb{E}\left(\frac{z_1^{Q_1} \cdot \cdots \cdot z_J^{Q_J}}{\sum_{i=1}^{J} Q_i g_i \cdot \mathbf{1}_{(\sum_{j=1}^{J} Q_j > 0)}} \right)$$

$$= \sum_{q_1=0}^{\infty} \cdots \sum_{q_J=0}^{\infty} z_1^{q_1} \cdot \cdots \cdot z_J^{q_J} R(\vec{q}).$$

Note that

$$g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} = \sum_{\vec{q}:\, \sum_{j=1}^{J} q_j > 0} \frac{g_i q_i}{\sum_{j=1}^{J} g_j q_j} z_1^{q_1} \cdot \cdots \cdot z_J^{q_J} \pi(\vec{q}). \tag{9}$$

Multiplying (8) by $z_1^{q_1} \cdots z_J^{q_J}$, summing both sides over $q_1, q_2, \ldots, q_J$ and adding Equation (7), we obtain from (9) that

$$\lambda p(\vec{z}) + \sum_{i=1}^{J} \mu_i g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} = \sum_{i=1}^{J} \lambda p_{0i} z_i p(\vec{z}) + \sum_{i=1}^{J} \mu_i g_i p_{i0} \frac{\partial r(\vec{z})}{\partial z_i} + \sum_{i=1}^{J} \sum_{j=1}^{J} \mu_i g_i p_{ij} z_j \frac{\partial r(\vec{z})}{\partial z_i}. \tag{10}$$

Because $\pi(\vec{0}) = 1 - \rho$, it follows from (9) that

$$\sum_{i=1}^{J} g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} + 1 - \rho = p(\vec{z}). \tag{11}$$

Together with (10) this gives the following partial differential equation for $r(\vec{z})$:

$$\lambda(1-\rho)\left(1 - \sum_{i=1}^{J} p_{0i} z_i\right) = \sum_{i=1}^{J} \left(\mu_i g_i \left(p_{i0} + \sum_{j=1}^{J} p_{ij} z_j - z_i \right) - \lambda g_i z_i \left(1 - \sum_{j=1}^{J} p_{0j} z_j \right) \right) \frac{\partial r(\vec{z})}{\partial z_i}. \tag{12}$$

This equation turns out to be very useful to analyze the joint queue-length distribution in heavy traffic because it allows for an explicit solution in that asymptotic regime. That is the topic of the next two sections. Note that Equation (12) was derived in Rege and Sengupta (1996) for the case of exponentially distributed service requirements.

4. Heavy-Traffic Scaling

It will be convenient to use the change of variables $z_i = \mathrm{e}^{-s_i}$ with $s_i > 0$, $i = 1, \ldots, J$. Denote $\vec{s} = (s_1, \ldots, s_J)$ and $\mathrm{e}^{-(1-\rho)\vec{s}} = (\mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_J})$. If

$$\lim_{\rho \uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \lim_{\rho \uparrow 1} \mathbb{E}(\mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \cdots \cdot \mathrm{e}^{-(1-\rho)s_J Q_J}) \tag{13}$$

exists, then there is a (possibly defective) random vector $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ such that $(1-\rho)(Q_1, Q_2, \ldots, Q_J)$ converges in distribution to $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$, and the distribution of $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ is uniquely determined by the limit in (13) (cf. the continuity theorem, see Feller 1971). For now, we assume that the limit exists; we come back to this assumption in §5. In this section we give two lemmas that describe properties of $\lim_{\rho \uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$. In particular, in Lemma 2 we obtain a partial differential equation that will be the key element in the proof of the main result stated in Proposition 1.

In order to describe the behavior of the generating function, we define

$$\hat{r}(\vec{s}) = \mathbb{E}\left(\frac{1 - \mathrm{e}^{-s_1 \hat{Q}_1} \cdot \cdots \cdot \mathrm{e}^{-s_J \hat{Q}_J}}{\sum_{j=1}^{J} \hat{Q}_j g_j} \cdot \mathbf{1}_{(\sum_{j=1}^{J} \hat{Q}_j > 0)} \right).$$

The "1" in the numerator is to ensure that the expression between brackets remains bounded when the $\hat{Q}_j$s are all near zero. We can now state the following lemma. The proof of this lemma can be found in the electronic companion (e-companion). The e-companion is available as part of the online version that can be found at http://or.journal.informs.org/.

LEMMA 1. *If* $\lim_{\rho \uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ *exists, then it satisfies* $\lim_{\rho \uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{i=1}^{J} g_i (\partial \hat{r}(\vec{s}) / \partial s_i)$.

In the following lemma we show that the partial differential equation as given in (12) simplifies considerably in the heavy-traffic regime. The proof may be found in the e-companion.

LEMMA 2. *If* $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ *exists, then the function* $\hat{r}(\vec{s})$ *satisfies the following partial differential equation*:

$$0=\sum_{i=1}^{J} F_i(\vec{s})\frac{\partial \hat{r}(\vec{s})}{\partial s_i}=\vec{F}(\vec{s})\cdot\nabla\hat{r}(\vec{s}),\quad \forall\,\vec{s}\geqslant\vec{0},$$

where $\vec{F}(\vec{s})=(F_1(\vec{s}),\ldots,F_J(\vec{s}))$, *and*

$$F_i(\vec{s})=g_i\left(\mu_i\left(-s_i+\sum_{j=1}^{J}p_{ij}s_j\right)+\hat{\lambda}\sum_{j=1}^{J}p_{0j}s_j\right),\quad i=1,\ldots,J,$$

with $\hat{\lambda}$ *as defined in* (3).

5. Proof of the Main Result

This section contains the proof of the main result stated in Proposition 1. It consists of two steps, which will be treated separately. First, we show in §5.1 that

$$(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)\overset{d}{=}\left(\frac{\hat{\rho}_1}{g_1},\frac{\hat{\rho}_2}{g_2},\ldots,\frac{\hat{\rho}_J}{g_J}\right)\cdot X \tag{14}$$

for some random variable X. Second, we demonstrate in §5.2 that X is exponentially distributed with mean as given in (5).

With these two partial results, the proof can be completed as follows: In §4 we assumed that $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists, thereby showing in §§5.1 and 5.2 that there is a unique limit. From tightness of the scaled queue-lengths (which follows from tightness of the scaled workload, see §5.2) we obtain that there exists a subsequence of ρ such that $(1-\rho)Q_i$ converges in distribution; cf. Prohorov's theorem (Billingsley 1999), and hence for this subsequence $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists. Because for any converging subsequence we obtain the same limit, this implies that the limit itself exists (see corollary in Billingsley 1999, p. 59). This establishes the state-space collapse $(1-\rho)(Q_1,Q_2,\ldots,Q_J)\overset{d}{\to}(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)$ with $(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)$ taking only values on the line described in (14).

5.1. State-Space Collapse

In this section we give the proof of (14). The proof is based on the fact that the probability-generating function satisfies the partial differential equation as described in Lemma 2. From this partial differential equation it can be derived that the function $\hat{r}(\vec{s})$ is constant on the $(J-1)$-dimensional hyperplane

$$H_c:=\left\{\vec{s}\geqslant\vec{0}\colon\ \sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}s_j=c\right\},\quad c>0.$$

LEMMA 3. *For any* $c>0$, *the function* $\hat{r}(\vec{s})$ *is constant on* H_c.

Hence, the function $\hat{r}(\vec{s})$ depends on $\vec{s}$ only through $\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j$, so there exists a function $\hat{r}^*\colon\mathbb{R}\to\mathbb{R}$ such that $\hat{r}(\vec{s})=\hat{r}^*(\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j)$. From Lemma 1 and $\partial\hat{r}(\vec{s})/\partial s_i=(\hat{\rho}_i/g_i)(\mathrm{d}\hat{r}^*(v)/\mathrm{d}v)|_{v=\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j}$, we then obtain

$$\begin{aligned}\mathbb{E}\big(\mathrm{e}^{-\sum_{i=1}^{J}s_i\hat{Q}_i}\big)=\lim_{\rho\uparrow 1}p\big(\mathrm{e}^{-(1-\rho)\vec{s}}\big)&=\sum_{i=1}^{J}g_i\frac{\partial\hat{r}(\vec{s})}{\partial s_i}\\&=\sum_{i=1}^{J}\hat{\rho}_i\frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\bigg|_{v=\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j}\\&=\frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\bigg|_{v=\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j},\end{aligned}$$

which again depends on $\vec{s}$ only through $\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j$. Equivalently, we can write

$$\begin{aligned}\mathbb{E}\big(\mathrm{e}^{-\sum_{i=1}^{J}s_i\hat{Q}_i}\big)=\mathbb{E}\big(&\mathrm{e}^{-(g_1/\hat{\rho}_1)\hat{Q}_1\sum_{i=1}^{J}(\hat{\rho}_i/g_i)s_i}\\&\cdot\mathrm{e}^{-s_2(\hat{\rho}_2/g_2)((g_2/\hat{\rho}_2)\hat{Q}_2-(g_1/\hat{\rho}_1)\hat{Q}_1)}\\&\cdot\cdots\cdot\mathrm{e}^{-s_J(\hat{\rho}_J/g_J)((g_J/\hat{\rho}_J)\hat{Q}_J-(g_1/\hat{\rho}_1)\hat{Q}_1)}\big).\end{aligned}$$

Because this only depends on $\sum_{j=1}^{J}(\hat{\rho}_j/g_j)s_j$, it implies that $(g_i/\hat{\rho}_i)\hat{Q}_i=(g_j/\hat{\rho}_j)\hat{Q}_j$ almost surely for all i,j, and we obtain

$$(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)=\left(\frac{\hat{\rho}_1}{g_1},\frac{\hat{\rho}_2}{g_2},\ldots,\frac{\hat{\rho}_J}{g_J}\right)\cdot\frac{g_1}{\hat{\rho}_1}\hat{Q}_1,$$

almost surely,

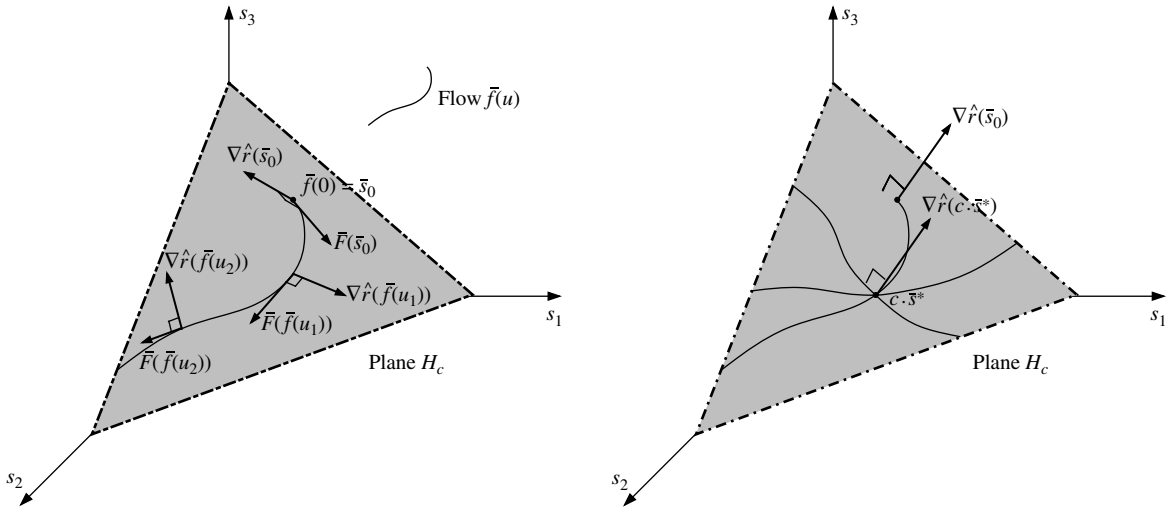
or equivalently

$$(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)\overset{d}{=}\left(\frac{\hat{\rho}_1}{g_1},\frac{\hat{\rho}_2}{g_2},\ldots,\frac{\hat{\rho}_J}{g_J}\right)\cdot X,$$

with X distributed as $(g_1/\hat{\rho}_1)\hat{Q}_1$.

The proof of Lemma 3 may be found in the e-companion. Here we give a geometric interpretation for the fact that the generating function $\hat{r}(\vec{s})$ is constant on the hyperplane H_c in the particular case of $J=3$. In Figure 2 (left) we depict the hyperplane $H_{\tilde{c}}$ for $J=3$. For a given $\vec{s}_0\in H_c$, we draw a flow curve $\vec{f}(u)$, $u\geqslant 0$, defined such that $\vec{f}(0)=\vec{s}_0\in H_c$ and the tangent at every point is precisely $\vec{f}'(u):=\vec{F}(\vec{f}(u))$, with $\vec{F}(\cdot)$ as defined in Lemma 2. In the proof of Lemma 3 (see the e-companion), it is derived that the vector $\vec{F}(\vec{s})$ is parallel to the hyperplane H_c, for all $\vec{s}\in H_c$; thus, the flow $\vec{f}(u)$ stays in the hyperplane H_c for all $u\geqslant 0$. By Lemma 2, the vector $\vec{F}(\vec{s})$ and the gradient $\nabla\hat{r}(\vec{s})$ are perpendicular for all $\vec{s}$, so $\vec{f}'(u)=\vec{F}(\vec{f}(u))\perp\nabla\hat{r}(\vec{f}(u))$. Thus, the function $\hat{r}$ has the same value in every point on a given flow $\vec{f}(u)$. In Figure 2 (right) we draw several flows in the hyperplane H_c. In the proof of Lemma 3 it is derived that all flows starting in the hyperplane H_c converge to one common point $c\cdot\vec{s}^*$. Because the function $\hat{r}(\cdot)$ is continuous and constant on each flow trajectory, it follows that $\hat{r}(\vec{s})$ is constant on the whole hyperplane H_c, or equivalently, $\nabla\hat{r}(\vec{s})\perp H_c$.

Figure 2. Geometrical interpretation of the proof of Lemma 3 for the case $J = 3$.



5.2. Determining the Common Factor

In the previous section we showed that $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) \stackrel{d}{=} (\hat{\rho}_1/g_1, \hat{\rho}_2/g_2, \ldots, \hat{\rho}_J/g_J) \cdot X$, with X some random variable. In this section we determine the distribution of X. In order to do so, we consider the total workload in the network, denoted by W. When scaled with $(1-\rho)$, the total workload has a proper distribution as $\rho \uparrow 1$; see Kingman (1961):

$$(1-\rho)W \stackrel{d}{\to} \widehat{W},$$

where $\widehat{W}$ is exponentially distributed with mean

$$\mathbb{E}(\widehat{W}) = \sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j). \tag{15}$$

The total workload can be represented as

$$W = \sum_{j=1}^{J} \sum_{h=1}^{Q_j} R_{j,h},$$

with $R_{j,h}$ the remaining service requirement of the hth type-j customer. Note that the remaining service requirements of all customers in phase j are i.i.d. and have the same phase-type distribution independent of $\vec{Q}$, more precisely, $R_{j,h} \stackrel{d}{=} R_j$ for all h. Hence,

$$\mathbb{E}(\mathrm{e}^{-sW}) = \mathbb{E}\Big(\mathrm{e}^{-s\sum_{j=1}^{J}\sum_{h=1}^{Q_j} R_{j,h}}\Big) = \mathbb{E}\bigg(\prod_{j=1}^{J} \mathbb{E}\Big(\mathrm{e}^{-s\sum_{h=1}^{Q_j} R_{j,h}} \mid \vec{Q}\Big)\bigg)$$

$$= \mathbb{E}\bigg(\prod_{j=1}^{J} (\mathbb{E}(\mathrm{e}^{-sR_j}))^{Q_j}\bigg) = \mathbb{E}\Big(\mathrm{e}^{\sum_{j=1}^{J} Q_j \ln(\mathbb{E}(\mathrm{e}^{-sR_j}))}\Big)$$

for $s > 0$. For the scaled workload we can therefore write

$$\mathbb{E}(\mathrm{e}^{-s\widehat{W}}) = \lim_{\rho\uparrow 1} \mathbb{E}(\mathrm{e}^{-(1-\rho)sW})$$

$$= \lim_{\rho\uparrow 1} \mathbb{E}\Big(\mathrm{e}^{\sum_{j=1}^{J} (\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))/((1-\rho)s))(1-\rho)sQ_j}\Big)$$

$$= \mathbb{E}\Big(\mathrm{e}^{-s\sum_{j=1}^{J}\mathbb{E}(R_j)\hat{Q}_j}\Big), \tag{16}$$

where in the last step we used that

$$\mathrm{e}^{\sum_{j=1}^{J} (\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))/(1-\rho)s)(1-\rho)sQ_j}$$

is bounded by 1 and converges in distribution to $\mathrm{e}^{-s\sum_{j=1}^{J}\mathbb{E}(R_j)\hat{Q}_j}$. The latter follows from $\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))/((1-\rho)s) \to -\mathbb{E}(R_j)$ as $\rho \uparrow 1$. From (16) we obtain that

$$\widehat{W} \stackrel{d}{=} \sum_{j=1}^{J} \mathbb{E}(R_j)\hat{Q}_j,$$

and together with (14) this gives

$$\widehat{W} \stackrel{d}{=} X \cdot \sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \mathbb{E}(R_j). \tag{17}$$

Because $\widehat{W}$ is exponentially distributed, the same is true for X. Taking expectations in (17), from (15) we obtain

$$\mathbb{E}(X) = \frac{\sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j)}{\sum_{j=1}^{J} (\hat{\rho}_j/g_j)\mathbb{E}(R_j)},$$

which concludes the proof of Proposition 1.

6. Size-Based Scheduling

Allowing the relative service weights of customers to change over time as they acquire service opens up a way to implement size-based scheduling by assigning relatively high weights in service phases that are more likely to lead to a quick service completion. In this section we investigate how the choice of the weights influences the performance of the system. With each type of customers we associate a cost $c_j \geqslant 0$, $j = 1, \ldots, J$. As a performance measure, we take the holding cost $\sum_{j=1}^{J} c_j Q_j$.

Recall that we consider the general Markovian framework where type-j customers have weight g_j. In this

section we will write $Q_j^{(g)}$ ($\hat{Q}_j^{(g)}$) instead of Q_j ($\hat{Q}_j$) to emphasize the dependence on the weights $g_1, \ldots, g_J$. From Proposition 1 we obtain that the scaled holding cost, $(1-\rho)\sum_{j=1}^J c_j Q_j^{(g)}$, converges in distribution to an exponentially distributed random variable with mean

$$\sum_{j=1}^J c_j \mathbb{E}(\hat{Q}_j^{(g)}) = \frac{\sum_{j=1}^J (\hat{\rho}_j/g_j)\cdot c_j}{\sum_{j=1}^J (\hat{\rho}_j/g_j)\cdot \mathbb{E}(R_j)} \cdot \sum_{j=1}^J \hat{\rho}_j \mathbb{E}(R_j), \tag{18}$$

as $\rho \uparrow 1$. Using this expression, we obtain the following monotonicity result in the heavy-traffic regime: the holding cost decreases "stochastically" as more preference is given to customers of types with a large value of $c_i/(\mathbb{E}(R_i))$.

PROPOSITION 2. *Consider the general Markovian framework and consider two policies with weights* $(g_1, \ldots, g_J)$ *and* $(\tilde{g}_1, \ldots, \tilde{g}_J)$, *respectively. Let* $c_j \geqslant 0$, $j = 1, \ldots, J$. *Without loss of generality, we assume that the types are ordered such that* $c_1/(\mathbb{E}(R_1)) \geqslant c_2/(\mathbb{E}(R_2)) \geqslant \cdots \geqslant c_J/(\mathbb{E}(R_J))$.

If $g_j/g_{j+1} \leqslant \tilde{g}_j/\tilde{g}_{j+1}$, *for all* $j = 1, \ldots, J-1$, *then*

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^J c_j Q_j^{(g)} \geqslant_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^J c_j Q_j^{(\tilde{g})},$$

where $\geqslant_{st}$ *denotes the usual stochastic ordering, i.e.,* $X \geqslant_{st} Y$ *if and only if* $\mathbb{P}(X \geqslant z) \geqslant \mathbb{P}(Y \geqslant z)$ *for all* z.

PROOF. We have that $(1-\rho)\sum_{j=1}^J c_j Q_j^{(g)}$ converges in distribution to an exponentially distributed random variable with mean as stated in (18). Because exponentially distributed random variables are stochastically ordered according to their means, it only remains to check that

$$\frac{\sum_{j=1}^J c_j \hat{\rho}_j/g_j}{\sum_{j=1}^J (\hat{\rho}_j/g_j)\mathbb{E}(R_j)} \geqslant \frac{\sum_{j=1}^J c_j \hat{\rho}_j/\tilde{g}_j}{\sum_{j=1}^J (\hat{\rho}_j/\tilde{g}_j)\mathbb{E}(R_j)}.$$

This holds because

$$\left(\sum_{j=1}^J \frac{c_j\hat{\rho}_j}{g_j}\right)\cdot\left(\sum_{j=1}^J \frac{\hat{\rho}_j}{\tilde{g}_j}\mathbb{E}(R_j)\right)$$

$$= \sum_{j,i:j\neq i} \hat{\rho}_j\hat{\rho}_i\left(\frac{1}{g_j\tilde{g}_i}c_j\mathbb{E}(R_i) + \frac{1}{g_i\tilde{g}_j}c_i\mathbb{E}(R_j)\right) + \sum_{j=1}^J \hat{\rho}_j^2 \frac{1}{g_j\tilde{g}_j}c_j\mathbb{E}(R_j)$$

$$\geqslant \sum_{j,i:j\neq i} \hat{\rho}_j\hat{\rho}_i\left(\frac{1}{g_i\tilde{g}_j}c_j\mathbb{E}(R_i) + \frac{1}{g_j\tilde{g}_i}c_i\mathbb{E}(R_j)\right) + \sum_{j=1}^J \hat{\rho}_j^2 \frac{1}{g_j\tilde{g}_j}c_j\mathbb{E}(R_j)$$

$$= \left(\sum_{j=1}^J \frac{c_j\hat{\rho}_j}{\tilde{g}_j}\right)\cdot\left(\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j}\mathbb{E}(R_j)\right).$$

Here we used that $c_i\mathbb{E}(R_j)(1/(g_i\tilde{g}_j) - 1/(g_j\tilde{g}_i)) \geqslant c_j\mathbb{E}(R_i)\cdot(1/(g_i\tilde{g}_j) - 1/(g_j\tilde{g}_i))$, which follows from the fact that $g_i/g_j \leqslant \tilde{g}_i/\tilde{g}_j$ and $c_i/(\mathbb{E}(R_i)) \geqslant c_j/(\mathbb{E}(R_j))$, for $i \leqslant j$. □

As mentioned in the introduction, the $c\mu$-rule minimizes the mean holding cost in an (i) $M/G/1$-queue among all nonpreemptive work-conserving disciplines as well as in an (ii) $G/M/1$-queue among all preemptive nonanticipating disciplines. In both systems the *expected remaining service requirement* of a class-k customer at a scheduling decision epoch is precisely $1/\mu_k$. Hence, the $c\mu$-rule gives priority according to the cost c_k divided by the expected remaining service requirement of a class-k customer. Proposition 2 can be seen as an extension of the $c\mu$-rule for DPS-based disciplines in the heavy-traffic regime: the performance improves as larger weights are assigned according to the values of $c_j/(\mathbb{E}(R_j))$, $j = 1, \ldots, J$. In particular, we obtain that a policy that gives lowest priority to type $i = \arg\min_{j=1,\ldots,J} c_j/(\mathbb{E}(R_j))$ minimizes the scaled holding cost in heavy traffic among all DPS-based policies.

7. The Standard DPS Queue in Heavy Traffic

In this section we specialize the results obtained so far to the standard DPS queue with phase-type distributed service requirements. In order to show how this queue fits into the Markovian framework of §2, let us give a brief description of the standard DPS queue.

We consider a single-server system with capacity 1 and Poisson arrivals with rate λ. With probability p_k an arrival is a class-k customer. Class-k customers have phase-type distributed service requirements, B_k, with a finite number of phases. In particular, this implies that the second moment of B_k is finite. Let

$$\varrho_k := \lambda p_k \mathbb{E}(B_k)$$

be the load associated with class-k customers. The capacity is shared among the customers of the various classes in accordance with the DPS discipline. When there are n_k class-k customers present in the system, $k = 1, \ldots, K$, each class-k customer is served at rate

$$\frac{w_k}{\sum_{l=1}^K w_l n_l},$$

where w_k is the weight associated with class k. It is important to note that the weight for a class-k customer is *independent* of the current phase of its service requirement. Denote by N_k the number of class-k customers in the DPS queue in steady state.

We now describe how the DPS queue with phase-type distributed service requirements fits into the Markovian framework as described in §2. Within each customer *class* of the DPS queue, we distinguish between customers residing in different service phases and represent them in the general framework as different customer *types*. Denoting the number of phases of the class-k phase-type distribution with J_k, the total number of types is $J := \sum_{k=1}^K J_k$. With slight abuse of terminology, we also refer to a class-k customer in the jth service phase as being of type $\sum_{l=1}^{k-1} J_l + j$. We use $k(j)$ to denote the customer class to which type-j customers belong. If types i and j belong to the same customer class, then they are associated with the same weight,

i.e., $g_i = g_j = w_{k(j)}$ when $k(i) = k(j)$. The p_{0j} in the general framework is taken such that for $l = k(j)$, p_{0j}/p_l is the probability that a class-l customer starts with service phase j. In the DPS queue, no transitions are possible between types belonging to different customer classes; hence, for the general framework this implies that $p_{ij} = 0$ if $k(i) \neq k(j)$. If a class-$k(i)$ customer finishes phase i, then p_{ij} is the probability that it continues in phase j (with $k(i) = k(j)$). The number of class-l customers in the DPS model can be written as $N_l = \sum_{j:k(j)=l} Q_j$.

The mean service requirement of a class-l customer may be written as $\mathbb{E}(B_l) = \sum_{j:k(j)=l}(p_{0j}/p_l)\mathbb{E}(R_j)$. Hence, the load in class l can be expressed by

$$\varrho_l = \lambda p_l \mathbb{E}(B_l) = \lambda \sum_{j:k(j)=l} p_{0j}\mathbb{E}(R_j). \tag{19}$$

For the DPS queue, the set of equations as given in (1) simplify: per class there is a set of equations that can be solved independently. For class l, the corresponding γ_is can be found from the following set of equations:

$$\gamma_i = p_{0i} + \sum_{j:k(j)=l} \gamma_j p_{ji}, \quad \text{for all } i \text{ s.t. } k(i) = l.$$

Applying the same reasoning as we followed to obtain Equation (2), it follows that an equivalent representation of ϱ_l is

$$\varrho_l = \lambda \sum_{j:k(j)=l} \frac{\gamma_j}{\mu_j} = \sum_{j:k(j)=l} \rho_j. \tag{20}$$

Note that the total load in the DPS queue equals $\sum_{l=1}^K \varrho_l = \sum_{l=1}^K \sum_{j:k(j)=l} \rho_j =: \rho$, i.e., it coincides indeed with the total load in the general framework.

Before proceeding with the main result of this section, we first give expressions for the forward recurrence time of the service requirements. For class l, we denote this random variable by B_l^{fwd}. From renewal theory we know that the associated distribution is

$$\mathbb{P}(B_l^{fwd} \leqslant x) := \frac{1}{\mathbb{E}B_l} \int_{y=0}^{x} \mathbb{P}(B_l > y)\, dy, \tag{21}$$

and hence $\mathbb{E}(B_l^{fwd}) = \mathbb{E}((B_l)^2)/(2\mathbb{E}(B_l))$. Alternatively, we can write

$$\mathbb{P}(B_l^{fwd} \leqslant x) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{P}(R_j \leqslant x); \tag{22}$$

see Asmussen (2003, Chapter III, Corollary 5.3). Intuitively, Relation (22) can be explained as follows: Note that γ_j/p_l represents the expected number of visits to phase j during the lifetime of the random variable B_l, with $k(j) = l$. As a consequence, $\gamma_j/(p_l\mu_j)$ is the expected time spent in phase j. Thus, with probability

$$\frac{\gamma_j/(p_l\mu_j)}{\sum_{i:k(i)=l} \gamma_i/(p_l\mu_i)} = \frac{\rho_j}{\sum_{i:k(i)=l} \rho_i} = \frac{\rho_j}{\varrho_l},$$

the residual lifetime equals the residual service requirement starting in phase j, and this gives Relation (22). Combining (21) and (22), we obtain that the mean forward recurrence time of B_l satisfies

$$\frac{\mathbb{E}((B_l)^2)}{2\mathbb{E}(B_l)} = \mathbb{E}(B_l^{fwd}) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{E}(R_j). \tag{23}$$

We now show the state-space collapse for the standard DPS queue with phase-type distributed service requirements. When passing $\rho \uparrow 1$ as described in §2, we actually fix the service requirement distributions and the class probabilities p_k, while increasing the arrival rate. In particular, the heavy-traffic scaling as considered in §2, $\lambda \uparrow \hat{\lambda} = (\vec{p}_0(I-P)^{-1}\vec{m}^T)^{-1}$, is equivalent with $\lambda \uparrow (\sum_l p_l\mathbb{E}(B_l))^{-1}$, because $\sum_{l=1}^K p_l\mathbb{E}(B_l) = \sum_{j=1}^J p_{0j}\mathbb{E}(R_j) = \vec{p}_0(I-P)^{-1}\vec{m}^T$. We denote the limiting loads of all classes by $\hat{\varrho}_l = \hat{\lambda} p_l \mathbb{E}(B_l)$, $l = 1, \ldots, K$ (or equivalently, $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$).

Proposition 3. *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights $w_1, \ldots, w_K$. When scaled by $1-\rho$, the queue-length vector has a proper distribution as $\rho \uparrow 1$,*

$$(1-\rho)(N_1, N_2, \ldots, N_K) \xrightarrow{d} (\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_K) \stackrel{d}{=} X \cdot \left(\frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \ldots, \frac{\hat{\varrho}_K}{w_K}\right), \tag{24}$$

where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean

$$\mathbb{E}(X) = \frac{\sum_k p_k \mathbb{E}((B_k)^2)}{\sum_k p_k \mathbb{E}((B_k)^2)/w_k}, \tag{25}$$

which is equal to 1 when $w_k = 1$ for all k, i.e., in the case of a standard PS queue.

Remark 1. In the case of exponentially distributed service requirements, in Kang et al. (2009) a related result is proved. The authors consider a sequence of systems indexed by r such that $\varrho_k^r \to \hat{\varrho}_k$, $\rho^r = \sum_{k=1}^K \varrho_k^r \uparrow 1$, and $\sqrt{r}(1-\rho^r) \to 1$, as $r \to \infty$, and obtain that $(1-\rho^r)\vec{N}^r(rt)$ converges in distribution to

$$\frac{\widehat{W}(t)}{\sum_{k=1}^K \hat{\varrho}_k/(w_k\mu_k)} \cdot \left(\frac{\hat{\varrho}_1}{w_1}, \ldots, \frac{\hat{\varrho}_K}{w_K}\right), \tag{26}$$

with $\widehat{W}(t)$ the diffusion-scaled workload process, being equal to a reflected Brownian motion with negative drift. The stationary distribution of the latter process is exponential with mean $\sum_{k=1}^K \hat{\varrho}_k/\mu_k$. Hence, for exponentially distributed service requirements, the stationary limit of (26) coincides with the heavy-traffic limit of the steady-state queue lengths (24) as derived in Proposition 3. Interestingly, this shows that the heavy-traffic limit and the steady-state limit commute in the case of exponentially distributed service requirements.

Proof of Proposition 3. Recall that the DPS queue with phase-type distributed service requirements is a special case of the general framework of §2 when the parameters are chosen as described in the beginning of this section. In particular, recall that $g_i = g_j = w_l$ when $k(i) = k(j) = l$. Because $N_l = \sum_{j:k(j)=l} Q_j$, $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$ (see (20)), and because for the general framework Relation (4) holds, Relation (24) follows directly where X is an exponentially distributed random variable with mean as given in (5). We are left with showing that (5) reduces to (25).

From (19) and (23), and because type-j customers belong to class $k(j)$ and have weight $g_j = w_{k(j)}$, we obtain that

$$\sum_{j=1}^{J} \frac{\rho_j}{g_j}\mathbb{E}(R_j) = \sum_{l=1}^{K} \frac{\varrho_l}{w_l} \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l}\mathbb{E}(R_j)$$
$$= \sum_{l=1}^{K} \frac{\varrho_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^{K} \frac{\lambda p_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2}. \tag{27}$$

Similarly, we have that

$$\sum_{j=1}^{J} \rho_j \mathbb{E}(R_j) = \sum_{l=1}^{K} \varrho_l \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l}\mathbb{E}(R_j)$$
$$= \sum_{l=1}^{K} \varrho_l \mathbb{E}((B_l)^2)/(2\mathbb{E}(B_l))$$
$$= \sum_{l=1}^{K} \lambda p_l \mathbb{E}((B_l)^2)/2. \tag{28}$$

Obviously, Equations (27) and (28) remain valid in heavy traffic. Equation (25) follows after substituting (27) and (28) into (5). □

Note that although the limiting distribution depends on the second moments of the service requirement distributions through $\mathbb{E}(X)$, the impact of the second moment on $\mathbb{E}(X)$ is uniformly bounded, and in particular

$$\min_k w_k \leqslant \mathbb{E}(X) \leqslant \max_k w_k,$$

(cf. Aalto et al. 2007).

The state-space collapse, as demonstrated above, allows us to show further interesting properties for the DPS queue in heavy traffic. In §7.1 we obtain heavy-traffic results for the residual service requirements of the customers in the various classes. In §7.2, monotonicity in the weights of the standard DPS policy is investigated.

7.1. Residual Service Requirements

The distribution of the residual service requirement of a customer, without having knowledge of the current phase of its service requirement, depends on the used scheduling discipline. For example, in a first-come first-served queue the residual service requirement for customers waiting to be served is given by their original service requirement. In case of a standard PS queue, the residual service requirements are independent random variables distributed according to the forward recurrence times of the service requirements. Given that there are n_k class-k customers in the system, let $B^r_{k,h}$ denote the remaining service requirement of the hth class-k customer, $k = 1, \ldots, K$, $h = 1, \ldots, n_k$. The following result is known for PS:

$$\mathbb{P}(B^r_{k,h} \leqslant x_{k,h}, N_k = n_k,\ k = 1, \ldots, K,\ h = 1, \ldots, n_k)$$
$$= \mathbb{P}(N_k = n_k,\ k = 1, \ldots, K) \prod_{k=1}^{K} \prod_{h=1}^{n_k} \mathbb{P}(B_k^{fwd} \leqslant x_{k,h}),$$

with $x_{k,h} \geqslant 0$. The joint distribution of the numbers of customers is of product form: $\mathbb{P}(N_k = n_k, k = 1, \ldots, K) = (1-\rho)((n_1 + \cdots + n_K)!/n_1! \cdot \cdots \cdot n_K!) \prod_{k=1}^{K} \varrho_k^{n_k}$, see for example Cohen (1979), Kelly (1979). In this section we show that in a heavy-traffic setting a similar result holds for the DPS queue.

Obviously, in the heavy-traffic limit, there will be an infinite number of customers present in the system. Therefore, we concentrate on the first $y_k < \infty$ class-k customers, $k = 1, \ldots, K$. In the following proposition we show that the scaled numbers of customers in the various classes and the remaining service requirements of any finite subset of customers are independent in a heavy-traffic setting. In particular, the remaining service requirement of a class-k customer is distributed according to the forward recurrence time of its service requirement B_k. It will be convenient to set $B^r_{k,h} = 0$ whenever $h > N_k$, $k = 1, \ldots, K$.

Proposition 4. *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights $w_1, \ldots, w_K$. Then,*

$$\lim_{\rho \uparrow 1} \mathbb{E}\big(\mathrm{e}^{-\sum_{l=1}^{K} s_l (1-\rho) N_l - \sum_{l=1}^{K} \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}}\big)$$
$$= \mathbb{E}\big(\mathrm{e}^{-\sum_{l=1}^{K} s_l \widehat{N}_l}\big) \cdot \prod_{l=1}^{K} \prod_{h=1}^{y_l} \mathbb{E}\big(\mathrm{e}^{-s_{l,h} B_l^{fwd}}\big),$$

for $y_l \in \{0, 1, \ldots\}$ and $s_{l,h}, s_l > 0$, $l = 1, \ldots, K$, $h = 1, \ldots, y_l$.

The proof may be found in the e-companion. Recall that $(\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_K) \stackrel{d}{=} X \cdot (\hat{\varrho}_1/w_1, \hat{\varrho}_2/w_2, \ldots, \hat{\varrho}_K/w_K)$, where X is exponentially distributed with mean $\mathbb{E}(X) = \sum_{l=1}^{K} p_l \mathbb{E}((B_l)^2)/(\sum_{l=1}^{K} p_l \mathbb{E}((B_l)^2)/w_l)$, cf. Proposition 3.

7.2. Monotonicity in the Weights

In this section, we investigate how the choice of the weights influences the holding cost for the standard DPS queue. We denote by d_k the cost associated with a class-k customer. Note that this is a different setting compared to §6, where a cost was assigned per type. As we will see in the proposition below, the scaled holding cost stochastically decreases when relatively larger weights are assigned to classes according to the values of $d_k/\mathbb{E}(B_k^{fwd})$, $k = 1, \ldots, K$.

From Proposition 4 it follows that the expected residual service requirement of a class-k customer is $\mathbb{E}(B_k^r) = \mathbb{E}(B_k^{fwd})$ in heavy traffic. Hence, in order to decrease the holding cost in heavy traffic, priority should be given according to the cost d_k divided by the expected residual service requirement of a class-k customer. This agrees with the celebrated $c\mu$-rule; see also §6.

PROPOSITION 5. *Assume phase-type distributed service requirements and consider two standard DPS queues with weights* $(w_1, \ldots, w_K)$ *and* $(\tilde{w}_1, \ldots, \tilde{w}_K)$. *Let* $d_k \geqslant 0$, $k = 1, \ldots, K$. *Without loss of generality, we assume that the classes are ordered such that* $d_1/\mathbb{E}(B_1^{fwd}) \geqslant \cdots \geqslant d_K/\mathbb{E}(B_K^{fwd})$.

If $w_k/w_{k+1} \leqslant \tilde{w}_k/\tilde{w}_{k+1}$, *for all* $k = 1, \ldots, K-1$, *then*

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^{K} d_k N_k^{DPS(w)} \geqslant_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^{K} d_k N_k^{DPS(\tilde{w})},$$

where $\geqslant_{st}$ *denotes the usual stochastic ordering, and* $N_k^{DPS(w)}$ *denotes the number of class-k customers in the DPS queue with weights* $w_1, \ldots, w_K$.

The proof is similar to the proof of Proposition 2 and may be found in the e-companion.

When the service requirements are exponentially distributed, it holds that $d_k/\mathbb{E}(B_k^{fwd}) = d_k\mu_k$. Hence, the $c\mu$-rule can be obtained in the limit from a DPS policy by letting the ratios w_k/w_{k+1}, $k = 1, \ldots, K-1$, all go to ∞.

REMARK 2. In Kim and Kim (2006) it was conjectured that, in the case of exponentially distributed service requirements, a result similar to Proposition 5 holds outside heavy traffic; see also Verloop et al. (2010, §6.1).

REMARK 3. In Coffman and Denning (1973, pp. 188–199) it was conjectured that $Var(B)/(\mathbb{E}((B)^2)) < 1$ is a sufficient condition to ensure that the queue length under PS has a smaller mean than under the least attained service discipline (denoted by LAS or FB), which gives service to the customers that have received the least amount of service. In Wierman et al. (2004), the authors found a counterexample to this conjecture, and it was later shown in Aalto and Ayesta (2006) that a stronger condition is needed in order to compare the performance of LAS and PS; to be specific, the distribution needs to have an "increasing mean residual life." This result is in concordance with the intuition behind size-based scheduling: queue lengths can be reduced by prioritizing customers that (are likely to) have smaller residual service requirements. The same intuition also explains the conditions in Proposition 5 which are based on $\mathbb{E}(B_k^{fwd}) = \mathbb{E}((B_k)^2)/(2\mathbb{E}(B_k)) = \frac{1}{2}((Var(B_k)/(\mathbb{E}(B_k))) + \mathbb{E}(B_k))$. Customers belonging to classes with highly variable service distributions are likely to have longer service requirements.

Although the monotonicity of the weight structure in Proposition 5 is only proved in the heavy-traffic limit, it is actually a good rule of thumb for systems operating close to saturation as well. We conclude this section with a numerical example where the behavior of the DPS queue is numerically investigated for different values of the total load.

Numerical evaluation of Proposition 5: We consider a DPS queue with two classes. Class-1 customers have hyperexponentially distributed service requirements, i.e., with a certain probability p a class-1 customer has an exponentially distributed service requirement with mean $1/\mu_{11}$, and with probability $1-p$ it has an exponentially distributed service requirement with mean $1/\mu_{12}$. Class-2 customers have exponentially distributed service requirements with mean $1/\mu_2$. Furthermore, we assume the load is equally distributed between classes 1 and 2, i.e., $\varrho_1 = \varrho_2$. We are interested in the total number of customers in the system; hence, we set $d_1 = d_2 = 1$. Note that

$$\mathbb{E}(B_1^{fwd}) = \frac{p/\mu_{11}^2 + (1-p)/\mu_{12}^2}{p/\mu_{11} + (1-p)/\mu_{12}} \quad \text{and} \quad \mathbb{E}(B_2^{fwd}) = 1/\mu_2.$$

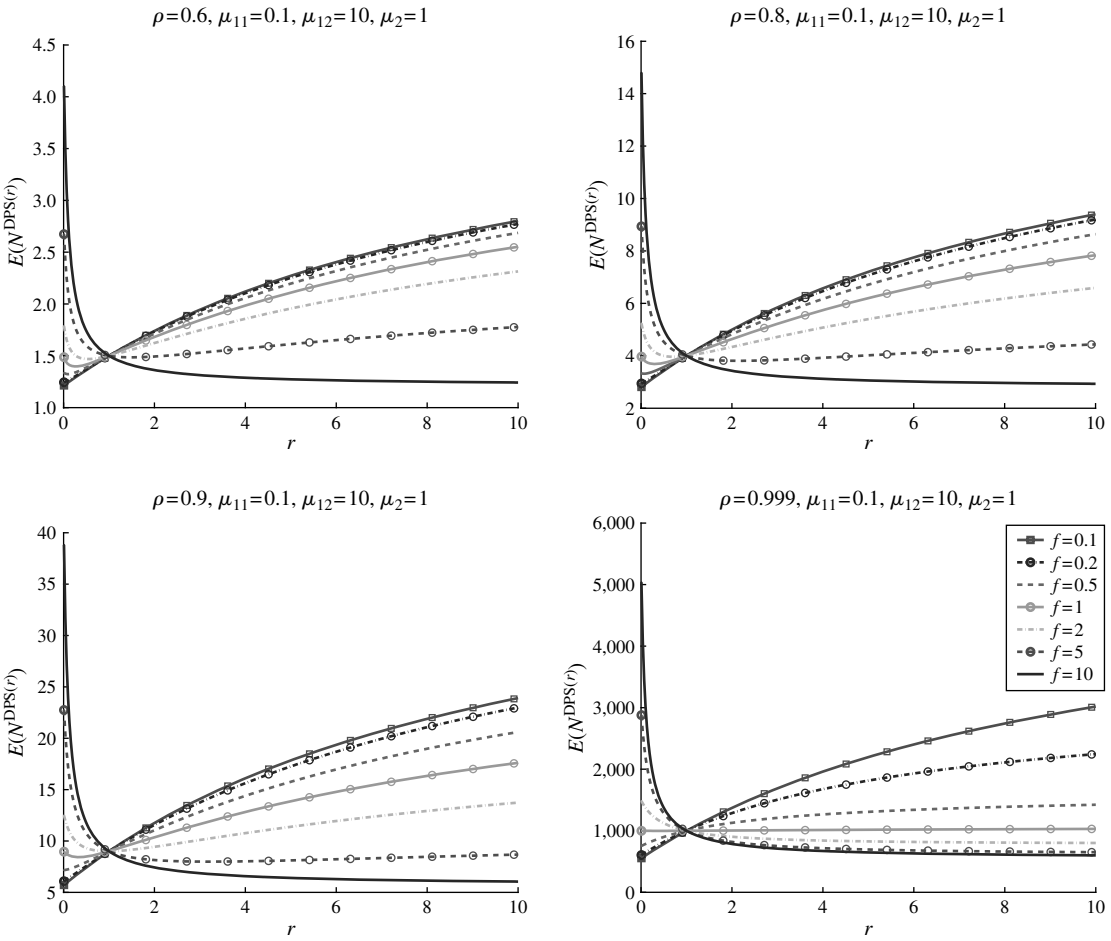
Without loss of generality we set $w_1 = 1$ and $w_2 = r$, with $r > 0$. Proposition 5 states that in a heavily loaded system the steady-state scaled total number of customers is stochastically increasing in r when $\mathbb{E}(B_1^{fwd}) < \mathbb{E}(B_2^{fwd})$, is constant in r when $\mathbb{E}(B_1^{fwd}) = \mathbb{E}(B_2^{fwd})$, and is stochastically decreasing in r when $\mathbb{E}(B_1^{fwd}) > \mathbb{E}(B_2^{fwd})$. Note that when $r = 1$, the policy reduces to standard PS, and in that case the total mean number of customers is given by $\rho/(1-\rho)$.

In Figure 3 we plot the total mean number of customers as a function of the weight parameter r (denoted by $\mathbb{E}(N^{DPS(r)})$). We consider the case $\mu_{11} = 0.1$, $\mu_{12} = 10$, and $\mu_2 = 1$, while choosing several values for $f := \mathbb{E}(B_1^{fwd})/\mathbb{E}(B_2^{fwd})$. The total mean number of customers is obtained by solving a system of linear equations as described in Fayolle et al. (1980). For $\rho = \varrho_1 + \varrho_2$ we chose the following values: 0.6, 0.8, 0.9, and 0.999. We see that in the latter case, a heavily loaded system, the total mean number of customers indeed exhibits the above-described phenomena depending on whether $f < 1$ (increasing), $f = 1$ (constant), or $f > 1$ (decreasing). As the total load decreases, the monotonicity no longer necessarily holds. This can be explained as follows. Because $\mu_{11} < \mu_2 < \mu_{12}$, the $c\mu$-rule suggests to prioritize class-1 customers in phase 2, whereas the class-1 customers in phase 1 should receive lowest priority. In the DPS queue no differentiation can be made between customers residing in different phases. Therefore, the way the weight r affects the total mean number of customers depends on the typical mix of numbers of class-1 customers residing in the two phases. In heavy traffic, this mix is characterized by the loads corresponding to the work of class 1 residing in phases 1 and 2, cf. Proposition 1, and is hence independent of r. However, away from heavy traffic, this mix may itself be influenced by r, leading to the observed nonmonotonic behavior in the figures.

8. Conclusion

We have studied a multiple-phase network of which the DPS queue with phase-type distributed service requirements is a special case. In our main result we have

Figure 3. Total mean number of customers under a DPS policy with weights $w_1 = 1$ and $w_2 = r$.



Notes. Class-1 service requirements are hyper-exponentially distributed (with parameters $\mu_{11} = 0.1$, $\mu_{12} = 10$), and class-2 service requirements are exponentially distributed (with $\mu_2 = 1$). The load $\rho = \varrho_1 + \varrho_2$ equals 0.6, 0.8, 0.9, and 0.999, respectively.

shown that in heavy-traffic conditions the queue-length vector exhibits a so-called state-space collapse: The multidimensional vector describing the numbers of customers in the various classes converges in distribution to a one-dimensional random vector. Based on this result, we found that the DPS model in heavy traffic inherits several well-known properties of PS (not necessarily in heavy traffic). For example, in the limit, the (scaled) number of customers present in a DPS model is exponentially distributed, which is the continuous analogue of the geometric queue-length distribution of the PS queue. In addition, in a heavy-traffic regime the residual service requirements are independent and distributed according to the forward recurrence times, which is true for PS as well.

We have investigated the performance of a DPS queue as a function of the weights and showed that the scaled holding cost reduces as customers with smaller weighted residual service requirements get larger weights. This property can be understood from the standard intuition of the $c\mu$-rule.

9. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgments

This work was supported by two visiting grants of NWO (Netherlands Organization for Scientific Research): grant B 62-640 (U. Ayesta) and grant PHC 20457QC (R. Núñez-Queija). The authors are thankful to Florian Simatos for useful discussions on tightness of random variables.

References

Aalto, S., U. Ayesta. 2006. On the nonoptimality of the foreground-background discipline for IMRL service times. *J. Appl. Probab.* **43**(2) 523–534.

Aalto, S., U. Ayesta, S. C. Borst, V. Misra, R. Núñez-Queija. 2007. Beyond processor sharing. *Performance Evaluation Rev.* **34**(4) 36–43.

Altman, E., K. E. Avrachenkov, U. Ayesta. 2006. A survey on discriminatory processor sharing. *Queueing Systems* **53**(1–2) 53–63.

Altman, E., T. Jimenez, D. Kofman. 2004. DPS queues with stationary ergodic service times and the performance of TCP in overload. *Proc. IEEE INFOCOM, Hong Kong*.

Asmussen, S. 2003. *Applied Probability and Queues*. Springer-Verlag, New York.

Avrachenkov, K. E., U. Ayesta, P. Brown, R. Núñez-Queija. 2005. Discriminatory processor sharing revisited. *Proc. IEEE INFOCOM, Miami*.

Bell, S. L., R. J. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* **11**(3) 608–649.

Ben Tahar, A., A. Jean-Marie. 2009. The fluid limit of the multiclass processor sharing queue. INRIA Research Report RR-6867, INRIA, Rocquencourt, France.

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Borst, S. C., R. Núñez-Queija, A. P. Zwart. 2006. Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* **53**(1–2) 31–51.

Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ rule revisited. *Adv. Appl. Probab.* **17** 237–238.

Coffman, E. G., P. Denning. 1973. *Operating System Theory*. Prentice-Hall, Englewood Cliffs, NJ.

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

Egorova, R., S. C. Borst, A. P. Zwart. 2007. Bandwidth-sharing networks in overload. *Performance Evaluation* **64**(9–12) 978–993.

Fayolle, G., I. Mitrani, R. Iasnogorodski. 1980. Sharing a processor among many job classes. *J. ACM* **27**(3) 519–532.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*, Vol. II. Wiley, New York.

Gelenbe, E., I. Mitrani. 1980. *Analysis and Synthesis of Computer Systems*. Academic Press, London.

Grishechkin, G. 1992. On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Probab.* **24**(3) 653–698.

Haviv, M., J. van der Wal. 2008. Mean sojourn times for phase-type discriminatory processor sharing systems. *Eur. J. Oper. Res.* **189**(2) 375–386.

Jean-Marie, A., P. Robert. 1994. On the transient behavior of the processor-sharing queue. *Queueing Systems* **17**(1–2) 129–136.

Kang, W. N., F. P. Kelly, N. H. Lee, R. J. Williams. 2009. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* **19**(5) 1719–1780.

Kelly, F. P. 1979. *Stochastic Networks and Reversibility*. Wiley, Chichester, UK.

Kim, B., J. Kim. 2006. Comparison of DPS and PS systems according to DPS weights. *IEEE Comm. Lett.* **10**(7) 558–560.

Kingman, J. F. C. 1961. The single server queue in heavy traffic. *Proc. Cambridge Philos.* **57**(4) 902–904.

Kleinrock, L. 1967. Time-shared systems: A theoretical treatment. *J. ACM* **14**(2) 242–261.

Nain, P., D. Towsley. 1994. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Trans. Automatic Control* **39**(9) 1853–1855.

Rege, K. M., B. Sengupta. 1996. Queue length distribution for the discriminatory processor-sharing queue. *Oper. Res.* **44**(4) 653–657.

Righter, R., J. G. Shanthikumar. 1989. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probab. Engrg. Inform. Sci.* **3** 323–333.

Sigman, K. 1991. A note on a sample-path rate conservation law and its relationship with $H = \lambda G$. *Adv. Appl. Probab.* **23**(3) 662–665.

Stolyar, A. L. 2004. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* **14**(1) 1–53.

van Kessel, G., R. Núñez-Queija, S. C. Borst. 2004. Asymptotic regimes and approximations for discriminatory processor sharing. *Performance Evaluation Rev.* **32**(2) 44–46.

Verloop, I. M., U. Ayesta, S. C. Borst. 2010. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynam. Systems* **20**(4) 473–509.

Wierman, A., N. Bansal, M. Harchol-Balter. 2004. A note comparing response times in the $M/GI/1/FB$ and $M/GI/1/PS$ queues. *Oper. Res. Lett.* **32**(1) 73–76.