

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

I.M. Verloop<sup>a,b</sup>, U. Ayesta<sup>a,c,\*</sup>, R. Núñez-Queija<sup>b,d</sup>

<sup>a</sup>BCAM - Basque Center for Applied Mathematics,  
Bizkaia Technology Park, 48170 Derio, Spain

<sup>b</sup>CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

<sup>c</sup>IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

<sup>d</sup>University of Amsterdam, Roetersstraat 11, 1018 WB, The Netherlands

## Abstract

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 (1996) 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  $d_k/E(B_k^{fw})$  obtain a larger share of the capacity, where  $d_k$  is the cost associated to class  $k$ , and  $E(B_k^{fw})$  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.

Keywords: Discriminatory processor sharing; heavy traffic; phase-type service requirements; residual service requirements; scheduling

## 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.

---

\*This work was initiated during a visit of Dr. U. Ayesta to The Netherlands, financially supported by Grant B 62-640 of NWO (Netherlands Organization for Scientific Research).

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 Fayolle et al. (1980) provided the first analysis of the mean sojourn time conditioned on the service requirement, by solving a system of integro-differential 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 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 only depend 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 to incorporate sophisticated scheduling techniques that give preferential treatment to customers that are close to service completion, thus reducing the numbers 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), 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 multi-dimensional stochastic process under heavy-traffic scaling has been demonstrated previously in other queueing models, see for example Bell and Williams (2001), Stolyar (2004), 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 multi-dimensional queue length vector. In Kessel et al. (2004) it was indicated that a similar approach as in Rege and Sengupta (1996) 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, which 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.

As 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, though. Since 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, effectively 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 non-preemptive work-conserving disciplines (Gelenbe and Mitrani (1980)) and in a (ii) G/M/1-queue among all preemptive non-anticipating 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  $\mu_k$  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), in addition to being the original mean service requirement,  $1/\mu_k$  also 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 a small weighted expected remaining service requirement.

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 multi-class queue close to saturation for the cases not covered by the  $c\mu$ -rule.

The paper is organized as follows. In Section 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 Section 3 and, under the heavy-traffic scaling, in Section 4. The proof of the main result is given in Section 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 Section 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  $\lambda$ , and an arriving customer is of type  $i$  with probability  $p_{0i}$ ,  $i = 1, \dots, J$ . Customers of type  $i$  have an exponentially distributed service requirement with mean  $\frac{1}{\mu_i}$ . After service completion, they become of type  $j$  with probability  $p_{ij}$ ,  $j = 1, \dots, J$ , and leave the system with probability  $p_{i0} := 1 - \sum_{j=1}^J p_{ij}$ . Let  $P$  be a  $J \times J$  matrix with  $P = (p_{ij})$ ,  $i, j = 1, \dots, J$ . We assume that all customers eventually leave the system. This implies  $\lim_{n \rightarrow \infty} P^n = 0$ , 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 one. There are strictly positive weights  $g_1, \dots, g_J$  associated with each of the types. Whenever there are  $q_i$  type- $i$  customers,  $i = 1, \dots, J$ , present in the system, each type- $j$  customer is served at rate

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

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 Section 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. Since 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:  $\mathbb{E}(R_i) = \frac{1}{\mu_i} + \sum_{j=1}^J p_{ij} \mathbb{E}(R_j)$ . Let  $\mathbb{E}(\vec{R}) = (\mathbb{E}(R_1), \dots, \mathbb{E}(R_J))^T$  and  $\vec{m} = (1/\mu_1, \dots, 1/\mu_J)^T$ , so that we can write

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

Denote the total traffic load by

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

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

$$\gamma_i = p_{0i} + \sum_{j=1}^J \gamma_j p_{ji}, \quad i = 1, \dots, J, \quad (1)$$

i.e.,  $\vec{\gamma}^T = \vec{p}_0^T (I - P)^{-1}$ , with  $\vec{\gamma} = (\gamma_1, \dots, \gamma_J)^T$  and  $\vec{p}_0 = (p_{01}, \dots, p_{0J})^T$ . Our assumption that none of the  $J$  types is redundant, entails that  $\vec{\gamma}$  is a vector with strictly positive elements. Note that  $\frac{\gamma_i}{\mu_i}$  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  $\rho$  we may equivalently write

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

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

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

since then  $\rho = \lambda \vec{p}_0^T (I - P)^{-1} \vec{m} \uparrow 1$ . 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  $\rho$ , in order to keep notation compact.

**Proposition 2.1** *Consider the general Markovian framework. When scaled by  $1 - \rho$ , the queue length vector has a proper limiting distribution as  $(\rho_1, \dots, \rho_J) \rightarrow (\hat{\rho}_1, \dots, \hat{\rho}_J)$ , such that  $\rho \uparrow 1$ ,*

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

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

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

The proof will be given in Section 5. Here we give some intuition for the result. Proposition 2.1 shows that in heavy traffic, the multi-dimensional 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

$$\rho_j = \mathbb{E} \left( \frac{g_j Q_j}{\sum_{i=1}^J g_i Q_i} \cdot \mathbf{1}_{(\sum_{i=1}^J Q_i > 0)} \right), \quad (6)$$

since 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

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

The pre-factor  $\sum_i g_i a_i$  is common to all  $a_j$ , which explains the appearance of the vector  $(\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \dots, \frac{\hat{\rho}_J}{g_J})$  in Proposition 2.1.

*Numerical illustration of Proposition 2.1:* We consider two types of customers and choose  $g_1 = 2, g_2 = 1, \mu_1 = 2, \mu_2 = 5, p_{01} = 0.6, p_{02} = 0.4, p_{12} = 0.3, p_{21} = 0.1$ . In Figure 1 we plot the joint queue length probabilities (obtained by simulation) for loads  $\rho = 0.8$  ( $\rho_1 \approx 0.59, \rho_2 \approx 0.21$ ),  $\rho = 0.90$  ( $\rho_1 \approx 0.66, \rho_2 \approx 0.24$ ) and  $\rho = 0.99$  ( $\rho_1 \approx 0.73, \rho_2 \approx 0.26$ ), 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 2.1, in heavy traffic the probabilities will lie on a straight line with slope  $\frac{g_1 \hat{\rho}_2}{\hat{\rho}_1 g_2} \approx 0.72$ , 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.

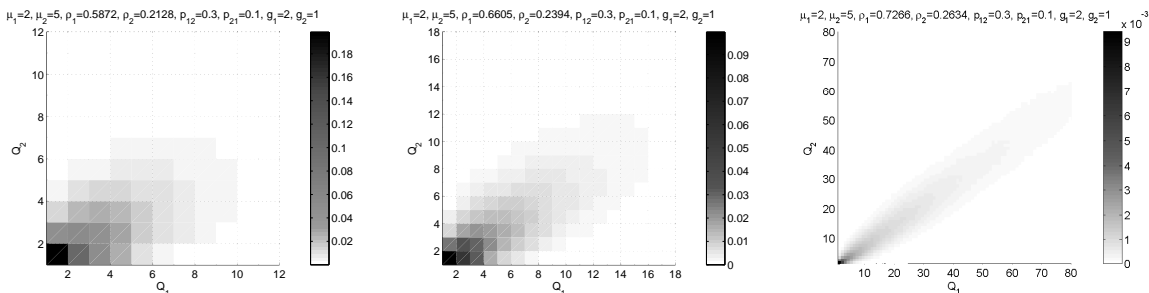


Figure 1: Joint queue length probabilities for load  $\rho = 0.8$  (left),  $\rho = 0.90$  (center) and  $\rho = 0.99$  (right), respectively.

### 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  $(Q_1, Q_2, \dots, Q_J) \geq \vec{0}$  and  $(q_1, q_2, \dots, q_J) \geq \vec{0}$ , respectively. The equilibrium distribution  $\pi(\vec{q}) := \mathbb{P}(\vec{Q} = \vec{q})$  satisfies

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

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

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

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

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

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

$$\begin{aligned} p(\vec{z}) &= \mathbb{E}(z_1^{Q_1} \dots z_J^{Q_J}) = \sum_{q_1=0}^{\infty} \dots \sum_{q_J=0}^{\infty} z_1^{q_1} \dots z_J^{q_J} \pi(\vec{q}), \\ r(\vec{z}) &= \mathbb{E} \left( \frac{z_1^{Q_1} \dots z_J^{Q_J}}{\sum_{i=1}^J Q_i g_i} \cdot \mathbf{1}_{(\sum_{j=1}^J Q_j > 0)} \right) = \sum_{q_1=0}^{\infty} \dots \sum_{q_J=0}^{\infty} z_1^{q_1} \dots z_J^{q_J} R(\vec{q}). \end{aligned}$$

Note that

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

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

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

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

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

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

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

This equation turns out to be very useful to analyze the joint queue length distribution in heavy traffic, as 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  $z_i = e^{-s_i}$  with  $s_i > 0$ ,  $i = 1, \dots, J$ . Denote  $\vec{s} = (s_1, \dots, s_J)$  and  $e^{-(1-\rho)\vec{s}} = (e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_J})$ . If

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

exists, then there is a (possibly defective) random vector  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J)$  such that  $(1-\rho)(Q_1, Q_2, \dots, Q_J)$  converges in distribution to  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J)$ , and the distribution of  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J)$  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 Section 5. In this section we give two lemmas that describe properties of  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$ . In particular, in Lemma 4.2 we obtain a partial differential equation, which will be the key element in the proof of the main result stated in Proposition 2.1.

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

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

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.

**Lemma 4.1** *If  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$  exists, then it satisfies:*

$$\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}}) = \sum_{i=1}^J g_i \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \quad (14)$$

**Proof:** From (11) we have

$$\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}}) = \lim_{\rho \uparrow 1} \sum_{i=1}^J g_i \frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}}. \quad (15)$$

By definition of  $r(\vec{z})$  we can write

$$\begin{aligned} \lim_{\rho \uparrow 1} \frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} &= \lim_{\rho \uparrow 1} \frac{\partial \mathbb{E} \left( \frac{z_1^{Q_1} \cdot \dots \cdot z_J^{Q_J}}{\sum_{j=1}^J Q_j g_j} \cdot \mathbf{1}_{(\sum_{j=1}^J Q_j > 0)} \right)}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \\ &= \lim_{\rho \uparrow 1} \mathbb{E} \left( \frac{Q_i}{\sum_{j=1}^J Q_j g_j} \cdot \frac{e^{-(1-\rho)s_1 Q_1} \cdot \dots \cdot e^{-(1-\rho)s_J Q_J}}{e^{-(1-\rho)s_i}} \cdot \mathbf{1}_{(\sum_{j=1}^J Q_j > 0)} \right) \\ &= \mathbb{E} \left( \frac{\hat{Q}_i}{\sum_{j=1}^J \hat{Q}_j g_j} \cdot e^{-s_1 \hat{Q}_1} \cdot \dots \cdot e^{-s_J \hat{Q}_J} \cdot \mathbf{1}_{(\sum_{j=1}^J \hat{Q}_j > 0)} \right) \\ &= \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \end{aligned} \quad (16)$$

In the third step we used that  $\frac{Q_i}{\sum_{j=1}^J Q_j g_j} \cdot e^{-(1-\rho)s_1 Q_1} \cdot \dots \cdot e^{-(1-\rho)s_J Q_J} \cdot \mathbf{1}_{(\sum_{j=1}^J Q_j > 0)}$  is upper bounded by  $\frac{1}{\min_j(g_j)}$ , and, cf. the continuous mapping theorem (Billingsley (1999)), converges in distribution to  $\frac{\hat{Q}_i}{\sum_{j=1}^J \hat{Q}_j g_j} \cdot e^{-s_1 \hat{Q}_1} \cdot \dots \cdot e^{-s_J \hat{Q}_J} \cdot \mathbf{1}_{(\sum_{j=1}^J \hat{Q}_j > 0)}$ . From (15) and (16) we obtain (14).  $\square$

In the following lemma we show that the partial differential equation as given in (12) simplifies considerably in the heavy-traffic regime.

**Lemma 4.2** *If  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$  exists, then the function  $\hat{r}(\vec{s})$  satisfies the following partial differential equation:*

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

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

$$F_i(\vec{s}) = g_i \left( \mu_i(-s_i + \sum_{j=1}^J p_{ij}s_j) + \hat{\lambda} \sum_{j=1}^J p_{0j}s_j \right), \quad i = 1, \dots, J, \quad (17)$$

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

**Proof:** Taking  $\vec{z}$  equal to  $e^{-(1-\rho)\vec{s}}$  in (12), dividing both sides by  $1-\rho$  and taking the limit of  $\rho \uparrow 1$ , this gives

0 =

$$\begin{aligned} & \lim_{\rho \uparrow 1} \sum_{i=1}^J \left( \mu_i g_i \frac{1 - e^{-(1-\rho)s_i} + \sum_{j=1}^J p_{ij}(e^{-(1-\rho)s_j} - 1)}{1 - \rho} - \lambda g_i e^{-(1-\rho)s_i} \sum_{j=1}^J p_{0j} \frac{1 - e^{-(1-\rho)s_j}}{1 - \rho} \right) \cdot \frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \\ &= \sum_{i=1}^J g_i \left( \mu_i(s_i - \sum_{j=1}^J p_{ij}s_j) - \hat{\lambda} \sum_{j=1}^J p_{0j}s_j \right) \cdot \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \end{aligned} \quad (18)$$

In the second step we used (16) and the fact that  $\lim_{\rho \uparrow 1} \frac{x^{1-\rho} - 1}{1-\rho} = \ln(x)$ .  $\square$

## 5 Proof of the main result

This section contains the proof of the main result stated in Proposition 2.1. It consists of two steps, which will be treated separately. First we show in Subsection 5.1 that

$$(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J) \stackrel{d}{=} \left( \frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \dots, \frac{\hat{\rho}_J}{g_J} \right) \cdot X, \quad (19)$$

for some random variable  $X$ . Second, we demonstrate in Section 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 Section 4 we assumed that  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$  exists, thereby showing in Subsections 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 Section 5.2) we obtain that there exists a subsequence of  $\rho$  such that  $(1-\rho)Q_i$  converges in distribution, cf. Prohorov's theorem (Billingsley (1999)), and hence for this subsequence  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$  exists. Because for any converging subsequence we obtain the same limit, this implies that the limit itself exists, (see corollary Page 59 in Billingsley (1999)). This establishes the state-space collapse  $(1-\rho)(Q_1, Q_2, \dots, Q_J) \xrightarrow{d} (\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J)$  with  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J)$  taking only values on the line described in (19).

### 5.1 State-space collapse

In this section we give the proof of (19). The proof is based on the fact that the probability generating function satisfies the partial differential equation as described in Lemma 4.2. From this partial differential equation it can be derived that the function  $\hat{r}(\vec{s})$  is constant on the  $(J-1)$ -dimensional hyperplane

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

as will be shown in Lemma 5.2. From this it follows that the function  $\hat{r}(\vec{s})$  depends on  $\vec{s}$  only through  $\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} s_j$ , so there exists a function  $\hat{r}^* : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} s_j)$ . From Lemma 4.1 and



$\frac{\partial \hat{r}(\vec{s})}{\partial s_i} = \frac{\hat{\rho}_i}{g_i} \frac{d\hat{r}^*(v)}{dv} \Big|_{v=\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} s_j}$ , we then obtain

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

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

$$\mathbb{E}(e^{-\sum_{i=1}^J s_i \hat{Q}_i}) = \mathbb{E}(e^{-\frac{g_1}{\hat{\rho}_1} \hat{Q}_1 \sum_{i=1}^J \frac{\hat{\rho}_i}{g_i} s_i} \cdot e^{-s_2 \frac{\hat{\rho}_2}{g_2} (\frac{g_2}{\hat{\rho}_2} \hat{Q}_2 - \frac{g_1}{\hat{\rho}_1} \hat{Q}_1)} \dots \cdot e^{-s_J \frac{\hat{\rho}_J}{g_J} (\frac{g_J}{\hat{\rho}_J} \hat{Q}_J - \frac{g_1}{\hat{\rho}_1} \hat{Q}_1)}).$$

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

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

or equivalently

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

with  $X$  distributed as  $\frac{g_1}{\hat{\rho}_1} \hat{Q}_1$ .

Before we proceed to prove that the generating function  $\hat{r}(\vec{s})$  is constant on the hyperplane  $H_c$  we first give a geometrical interpretation for this fact in the particular case of  $J = 3$ . In Figure 2 (left) we depict the hyperplane  $H_c$  for  $J = 3$ . For a given  $\vec{s}_0 \in H_c$ , we draw a flow curve  $\vec{f}(u)$ ,  $u \geq 0$ , defined such that the tangent at every point is precisely  $\vec{f}'(u) := \vec{F}(\vec{f}(u))$  and  $\vec{f}(0) = \vec{s}_0 \in H_c$ . We will see in the proof of Lemma 5.2 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  $u \geq 0$ . By Lemma 4.2, 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 5.2 we will see that all flows starting in the hyperplane  $H_c$  converge to one common point  $c \cdot \vec{s}^*$ . Since 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,  $\nabla \hat{r}(\vec{s}) \perp H_c$ .

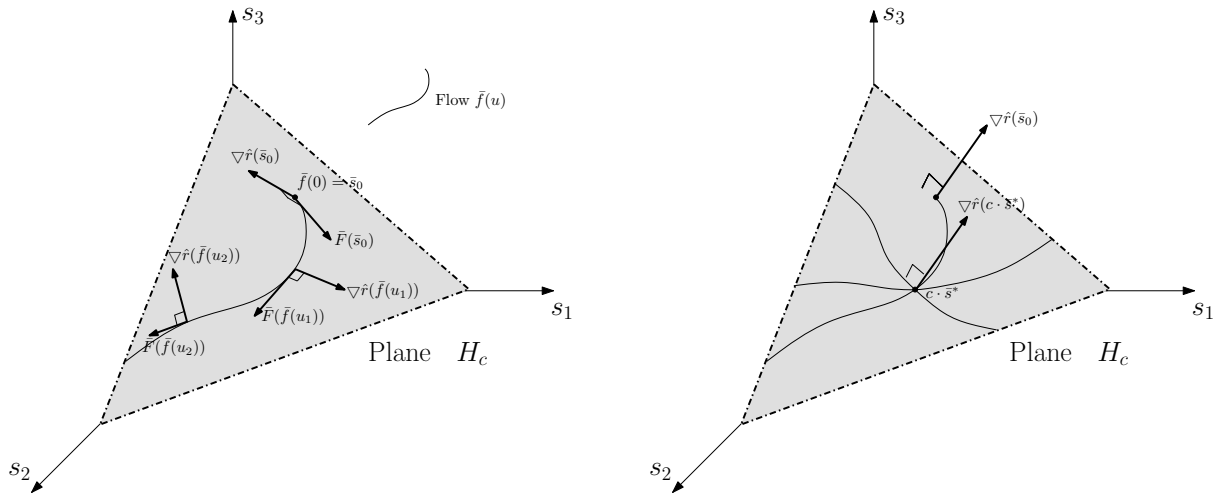


Figure 2: Geometrical interpretation of the proof of Lemma 5.2 for the case  $J = 3$ .

The following technical lemma is used in the proof of Lemma 5.2.

**Lemma 5.1** Consider the matrix

$$A = \begin{pmatrix} g_1(-\mu_1 + \mu_1 p_{11} + \hat{\lambda} p_{01}) & g_1(\mu_1 p_{12} + \hat{\lambda} p_{02}) & \dots & g_1(\mu_1 p_{1J} + \hat{\lambda} p_{0J}) \\ g_2(\mu_2 p_{21} + \hat{\lambda} p_{01}) & g_2(-\mu_2 + \mu_2 p_{22} + \hat{\lambda} p_{02}) & \dots & g_2(\mu_2 p_{2J} + \hat{\lambda} p_{0J}) \\ \vdots & \vdots & \ddots & \vdots \\ g_J(\mu_J p_{J1} + \hat{\lambda} p_{01}) & g_J(\mu_J p_{J2} + \hat{\lambda} p_{02}) & \dots & g_J(-\mu_J + \mu_J p_{JJ} + \hat{\lambda} p_{0J}) \end{pmatrix}, \quad (20)$$

where  $\hat{\lambda}$  is as defined in (3). One eigenvalue of  $A$  is 0 (with multiplicity 1), and all the other eigenvalues have a strictly negative real part. In addition, there exists a vector  $\vec{\eta} \geq \vec{0}$  with  $\sum_{j=1}^J \eta_j = 1$  such that  $\vec{s}^*$  with  $s_j^* := \frac{g_j}{\rho_j} \eta_j$  is an eigenvector of  $A$  corresponding to the eigenvalue 0, and  $\vec{s}^* \in H_1$ .

**Proof:** Define  $D$  as the diagonal matrix  $\text{diag}[d_1, d_2, \dots, d_J]$  with  $d_i = \frac{\hat{\rho}_i}{g_i}$ , and let  $S$  be the matrix

$$S := DAD^{-1} = \begin{pmatrix} g_1(-\mu_1 + \mu_1 p_{11} + \hat{\lambda} p_{01}) & \hat{\rho}_1 \frac{g_2}{\rho_2} (\mu_1 p_{12} + \hat{\lambda} p_{02}) & \dots & \hat{\rho}_1 \frac{g_J}{\rho_J} (\mu_1 p_{1J} + \hat{\lambda} p_{0J}) \\ \hat{\rho}_2 \frac{g_1}{\rho_1} (\mu_2 p_{21} + \hat{\lambda} p_{01}) & g_2(-\mu_2 + \mu_2 p_{22} + \hat{\lambda} p_{02}) & \dots & \hat{\rho}_2 \frac{g_J}{\rho_J} (\mu_2 p_{2J} + \hat{\lambda} p_{0J}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_J \frac{g_1}{\rho_1} (\mu_J p_{J1} + \hat{\lambda} p_{01}) & \hat{\rho}_J \frac{g_2}{\rho_2} (\mu_J p_{J2} + \hat{\lambda} p_{02}) & \dots & g_J(-\mu_J + \mu_J p_{JJ} + \hat{\lambda} p_{0J}) \end{pmatrix}.$$

The matrix  $A$  is similar to  $S$  and therefore  $A$ ,  $S$  and  $S^T$  have the same eigenvalues. Using Equation (1), it follows that

$$-\mu_i \hat{\rho}_i + \sum_{j=1}^J \hat{\rho}_j (\mu_j p_{ji} + \hat{\lambda} p_{0i}) = \hat{\lambda} (-\gamma_i + p_{0i} + \sum_{j=1}^J \gamma_j p_{ji}) = 0, \quad i = 1, \dots, J. \quad (21)$$

Hence, the sum of each row in  $S^T$  (sum of each column in  $S$ ) is equal to 0, and the off-diagonal elements in  $S^T$  are all positive. This implies that the matrix  $S^T$  is a generator corresponding to a finite-state continuous-time Markov chain. This Markov chain is irreducible, as will be shown at the end of the proof, and hence has a unique equilibrium distribution  $\vec{\eta}$ , i.e.,  $\vec{\eta} S^T = \vec{0}$  and  $\sum_{j=1}^J \eta_j = 1$ . This implies that 0 is an eigenvalue with multiplicity 1 of the matrix  $S^T$ , and, cf. (Asmussen 2003, Proposition 6.2), the real parts of all other eigenvalues are strictly negative. Since the eigenvalues of  $A$  and  $S^T$  coincide, the same holds for the matrix  $A$ . The eigenvector of  $A$  corresponding to the eigenvalue 0 is given by  $\vec{s}^* = D^{-1} \vec{\eta}$ , since  $A \vec{s}^* = D^{-1} D A D^{-1} \vec{\eta} = D^{-1} S \vec{\eta} = \vec{0}$ .

It remains to be shown that the Markov chain corresponding to the generator  $S^T$  is irreducible. Let  $\mathcal{J}_0 := \{j = 1, \dots, J : p_{0j} > 0\}$ , denote the non-empty set of types that receive external arrivals, and let

$$\mathcal{J}_n := \{j = 1, \dots, J : \text{there exist } j_0, \dots, j_{n-1} \text{ with } p_{0j_0} \cdot p_{j_0 j_1} \cdot \dots \cdot p_{j_{n-1} j} > 0\},$$

$n = 1, \dots, J-1$ , denote the set of types such that there is a strictly positive probability that a customer becomes of this type after  $n$  steps. Since  $J < \infty$  and eventually all types are observed, we have that  $\cup_{n=0}^{J-1} \mathcal{J}_n = \{1, \dots, J\}$ .

Since  $\gamma_i > 0$ , also  $\hat{\rho}_i > 0$  for all  $i = 1, \dots, J$ . Now consider the  $J \times J$  matrix  $S^T$ . If  $j \in \mathcal{J}_0$ , then the  $(j, i)$ -th element of  $S^T$ ,  $\hat{\rho}_i \frac{g_j}{\rho_j} (\mu_i p_{ij} + \hat{\lambda} p_{0j})$ , is strictly positive for all  $i \neq j$ . Thus, in the Markov chain corresponding to the generator  $S^T$ , from any state in  $\mathcal{J}_0$  one can reach all other states. In order to prove irreducibility it is now sufficient to show that from any state in  $\{1, \dots, J\} \setminus \mathcal{J}_0$ , some state in  $\mathcal{J}_0$  can be reached.

Assume  $j \in \mathcal{J}_1$ . By definition, there exists an  $i \in \mathcal{J}_0$  such that  $p_{ij} > 0$ . Hence, the  $(j, i)$ -th element of the matrix  $S^T$ ,  $\hat{\rho}_i \frac{g_j}{\rho_j} (\mu_i p_{ij} + \hat{\lambda} p_{0j})$ , is strictly positive. This implies that from every state in  $\mathcal{J}_1$ , a state in  $\mathcal{J}_0$  can be reached. Now consider a state  $j \in \mathcal{J}_2$ . By definition, there exists a state  $i \in \mathcal{J}_1$  such that  $p_{ij} > 0$ . Similarly to the previous case, this implies that the  $(j, i)$ -th element of  $S^T$  is strictly positive, and we can conclude that from every state in  $\mathcal{J}_2$ , a state in  $\mathcal{J}_1$  can be reached. Proceeding along these lines, it can be shown that from every state in  $\mathcal{J}_n$ , a state in  $\mathcal{J}_{n-1}$  can be reached,  $n = 1, \dots, J-1$ . Since  $\cup_{n=0}^{J-1} \mathcal{J}_n = \{1, \dots, J\}$ , we obtain that from any state outside  $\mathcal{J}_0$ , some state in  $\mathcal{J}_0$  can be reached, which concludes the proof.  $\square$

The following lemma shows that the generating function  $\hat{r}(\vec{s})$  is constant on the hyperplane  $H_c$ .

**Lemma 5.2** For any  $c > 0$ , the function  $\hat{r}(\vec{s})$  is constant on  $H_c$ .

**Proof:** From (21) it follows that

$$\begin{aligned} \sum_{i=1}^J \frac{\hat{\rho}_i}{g_i} \cdot F_i(\vec{s}) &= \sum_{i=1}^J \hat{\rho}_i \cdot \left( \mu_i(-s_i + \sum_{j=1}^J p_{ij}s_j) + \hat{\lambda} \sum_{j=1}^J p_{0j}s_j \right) \\ &= \sum_{i=1}^J (-\mu_i \hat{\rho}_i + \sum_{j=1}^J \hat{\rho}_j (\mu_j p_{ji} + \hat{\lambda} p_{0i})) s_i \\ &= 0. \end{aligned}$$

This implies that for all  $\vec{s} \in H_c$ , the vector  $\vec{F}(\vec{s})$  is parallel to the hyperplane  $H_c$ . Since  $\vec{F}$  is  $C^1$ , for each state  $\vec{s} \geq \vec{0}$  there exists a unique flow  $\vec{f}(u) = (f_1(u), \dots, f_J(u))$ , parametrized by  $u \geq 0$ , such that

$$\vec{f}(0) = \vec{s} \quad \text{and} \quad \frac{df_i(u)}{du} = F_i(\vec{f}(u)), \quad \text{for all } i \text{ and } u \geq 0. \quad (22)$$

Since  $\vec{F}(\vec{s})$  is parallel to  $H_c$  for all  $\vec{s} \in H_c$ , when started in  $H_c$ , the flow  $\vec{f}(u)$  will stay in  $H_c$ . Another important property of this flow  $\vec{f}(u)$  is that

$$\frac{d\hat{r}(\vec{f}(u))}{du} = \sum_{i=1}^J \frac{df_i(u)}{du} \cdot \left. \frac{\partial \hat{r}(\vec{s})}{\partial s_i} \right|_{\vec{s}=\vec{f}(u)} = 0,$$

which follows from the chain rule, Lemma 4.2, and Equation (22). Hence, along each flow  $\vec{f}(u)$ , which lies in  $H_c$ , the function  $\hat{r}(\vec{f}(u))$  is constant. We will now show that each flow in  $H_c$  converges to a certain point  $c \cdot \vec{s}^* \geq \vec{0}$  as  $u \rightarrow \infty$ .

Relation (22) can be written as  $\vec{f}'(u) = A\vec{f}(u)$ , with  $A$  as defined in (20), see (17). In Lemma 5.1 it is proved that one eigenvalue of  $A$  is 0 with eigenvector  $\vec{s}^* \geq \vec{0}$ ,  $\vec{s}^* \in H_1$ , and all the other eigenvalues have a strictly negative real part. Hence, the solution of  $\vec{f}'(u) = A\vec{f}(u)$  with  $\vec{f}(0) \in H_c$  can be written as  $\vec{f}(u) = c \cdot \vec{s}^* + \vec{g}(u)$ , where  $\lim_{u \rightarrow \infty} \vec{g}(u) = \vec{0}$  and  $\vec{s}^* \geq \vec{0}$ . This implies that all the flows in the hyperplane  $H_c$  converge to one common point  $c \cdot \vec{s}^* \geq \vec{0}$ .

Since the continuous function  $\hat{r}(\vec{s})$  is constant along each flow, and all flows in the hyperplane  $H_c$  converge to  $c \cdot \vec{s}^* \in H_c$ , we obtain that the function  $\hat{r}(\vec{s})$  is constant on  $H_c$ .  $\square$

## 5.2 Determining the common factor

In the previous section we showed that  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J) \stackrel{d}{=} (\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \dots, \frac{\hat{\rho}_J}{g_J}) \cdot X$ , 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):

$$(1 - \rho)W \xrightarrow{d} \hat{W},$$

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

$$\mathbb{E}(\hat{W}) = \sum_{j=1}^J \hat{\rho}_j \mathbb{E}(R_j). \quad (23)$$

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  $h$ -th 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,

$$\begin{aligned}\mathbb{E}(e^{-sW}) &= \mathbb{E}(e^{-s \sum_{j=1}^J \sum_{h=1}^{Q_j} R_{j,h}}) = \mathbb{E}\left(\prod_{j=1}^J \mathbb{E}(e^{-s \sum_{h=1}^{Q_j} R_{j,h}} | \vec{Q})\right) \\ &= \mathbb{E}\left(\prod_{j=1}^J (\mathbb{E}(e^{-sR_j}))^{Q_j}\right) = \mathbb{E}(e^{\sum_{j=1}^J Q_j \ln(\mathbb{E}(e^{-sR_j}))}),\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}(e^{-s\hat{W}}) &= \lim_{\rho \uparrow 1} \mathbb{E}(e^{-(1-\rho)sW}) = \lim_{\rho \uparrow 1} \mathbb{E}(e^{\sum_{j=1}^J \frac{\ln(\mathbb{E}(e^{-(1-\rho)sR_j}))}{(1-\rho)s} (1-\rho)sQ_j}) \\ &= \mathbb{E}(e^{-s \sum_{j=1}^J \mathbb{E}(R_j) \hat{Q}_j}),\end{aligned}\tag{24}$$

where in the last step we used that  $e^{\sum_{j=1}^J \frac{\ln(\mathbb{E}(e^{-(1-\rho)sR_j}))}{(1-\rho)s} (1-\rho)sQ_j}$  is bounded by 1 and converges in distribution to  $e^{-s \sum_{j=1}^J \mathbb{E}(R_j) \hat{Q}_j}$ . The latter follows from  $\frac{\ln(\mathbb{E}(e^{-(1-\rho)sR_j}))}{(1-\rho)s} \rightarrow -\mathbb{E}(R_j)$  as  $\rho \uparrow 1$ . From (24) we obtain that

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

and together with (19) this gives

$$\hat{W} \stackrel{d}{=} X \cdot \sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} \mathbb{E}(R_j).\tag{25}$$

Since  $\hat{W}$  is exponentially distributed, the same is true for  $X$ . Taking expectations in (25), from (23) we obtain

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

which concludes the proof of Proposition 2.1.

## 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  $c_j \geq 0$ ,  $j = 1, \dots, J$ . As 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  $g_1, \dots, g_J$ . From Proposition 2.1 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

$$\sum_{j=1}^J c_j \mathbb{E}(\hat{Q}_j^{(g)}) = \frac{\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} \cdot c_j}{\sum_{j=1}^J \frac{\hat{\rho}_j}{g_j} \cdot \mathbb{E}(R_j)} \cdot \sum_{j=1}^J \hat{\rho}_j \mathbb{E}(R_j),\tag{26}$$

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  $\frac{c_j}{\mathbb{E}(R_j)}$ .

**Proposition 6.1** *Consider the general Markovian framework and consider two policies with weights  $(g_1, \dots, g_J)$  and  $(\tilde{g}_1, \dots, \tilde{g}_J)$ , respectively. Let  $c_j \geq 0$ ,  $j = 1, \dots, J$ . Without loss of generality we assume that the types are ordered such that  $\frac{c_1}{\mathbb{E}(R_1)} \geq \frac{c_2}{\mathbb{E}(R_2)} \geq \dots \geq \frac{c_J}{\mathbb{E}(R_J)}$ .*

If  $\frac{g_j}{g_{j+1}} \leq \frac{\tilde{g}_j}{\tilde{g}_{j+1}}$ , for all  $j = 1, \dots, J-1$ , then

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

where  $\geq_{st}$  denotes the usual stochastic ordering, i.e.,  $X \geq_{st} Y$  if and only if  $\mathbb{P}(X \geq z) \geq \mathbb{P}(Y \geq z)$  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 (26). Hence, it only remains to check that

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

This holds since

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

Here we used that  $c_i \mathbb{E}(R_j) \left( \frac{1}{g_i \tilde{g}_j} - \frac{1}{g_j \tilde{g}_i} \right) \geq c_j \mathbb{E}(R_i) \left( \frac{1}{g_i \tilde{g}_j} - \frac{1}{g_j \tilde{g}_i} \right)$ , which follows from the fact that  $\frac{g_i}{g_j} \leq \frac{\tilde{g}_i}{\tilde{g}_j}$  and  $\frac{c_i}{\mathbb{E}(R_i)} \geq \frac{c_j}{\mathbb{E}(R_j)}$ , for  $i \leq j$ .  $\square$

As mentioned in the Introduction, the  $c\mu$ -rule minimizes the mean holding cost in an (i) M/G/1-queue among all non-preemptive work-conserving disciplines as well as in an (ii) G/M/1-queue among all preemptive non-anticipating 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 6.1 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  $\frac{c_j}{\mathbb{E}(R_j)}$ ,  $j = 1, \dots, J$ . In particular, we obtain that a policy that gives lowest priority to type  $i = \arg \min_{j=1, \dots, J} \frac{c_j}{\mathbb{E}(R_j)}$  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 Section 2, let us give a brief description of the standard DPS queue.

We consider a single-server system with capacity one and Poisson arrivals with rate  $\lambda$ . 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,  $k = 1, \dots, K$ , 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 Section 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  $j^{\text{th}}$  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 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  $\mathbb{E}(B_l) = \sum_{j:k(j)=l} \frac{p_{0j}}{p_l} \mathbb{E}(R_j)$ . Hence, the load in class  $l$  can be expressed by

$$\varrho_l = \lambda p_l \mathbb{E}(B_l) = \lambda \sum_{j:k(j)=l} p_{0j} \mathbb{E}(R_j). \quad (27)$$

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  $\gamma_i$ 's 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  $\varrho_l$  is

$$\varrho_l = \lambda \sum_{j:k(j)=l} \frac{\gamma_j}{\mu_j} = \sum_{j:k(j)=l} \rho_j. \quad (28)$$

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  $B_l^{\text{fwd}}$ . From renewal theory we know that the associated distribution is

$$\mathbb{P}(B_l^{\text{fwd}} \leq x) := \frac{1}{\mathbb{E}B_l} \int_{y=0}^x \mathbb{P}(B_l > y) dy, \quad (29)$$

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

$$\mathbb{P}(B_l^{\text{fwd}} \leq x) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{P}(R_j \leq x), \quad (30)$$

see (Asmussen 2003, Chapter III, Corollary 5.3). Intuitively Relation (30) can be explained as follows: Note that  $\frac{\gamma_j}{p_l}$  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

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

the residual life time equals the residual service requirement starting in phase  $j$ , and this gives Relation (30). Combining (29) and (30), we obtain that the mean forward recurrence time of  $B_l$  satisfies

$$\frac{\mathbb{E}((B_l)^2)}{2\mathbb{E}(B_l)} = \mathbb{E}(B_l^{fwd}) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{E}(R_j). \quad (31)$$

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 Section 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 Section 2,  $\lambda \uparrow \hat{\lambda} = \left( \bar{p}_0^T (I - P)^{-1} \bar{m} \right)^{-1}$ , is equivalent with  $\lambda \uparrow (\sum_l p_l \mathbb{E}(B_l))^{-1}$ , since  $\sum_{l=1}^K p_l \mathbb{E}(B_l) = \sum_{j=1}^J p_{0j} \mathbb{E}(R_j) = \bar{p}_0^T (I - P)^{-1} \bar{m}$ . We denote the limiting loads of all classes by  $\hat{\varrho}_l = \hat{\lambda} p_l \mathbb{E}(B_l)$ ,  $l = 1, \dots, K$  (or equivalently,  $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$ ).

**Proposition 7.1** *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights  $w_1, \dots, w_K$ . When scaled by  $1 - \rho$ , the queue length vector has a proper distribution as  $\rho \uparrow 1$ ,*

$$(1 - \rho)(N_1, N_2, \dots, N_K) \xrightarrow{d} (\hat{N}_1, \hat{N}_2, \dots, \hat{N}_K) \stackrel{d}{=} X \cdot \left( \frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \dots, \frac{\hat{\varrho}_K}{w_K} \right), \quad (32)$$

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

$$\mathbb{E}(X) = \frac{\sum_k p_k \mathbb{E}((B_k)^2)}{\sum_k p_k \mathbb{E}((B_k)^2)/w_k}, \quad (33)$$

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

**Remark 7.2** *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  $\varrho_k^r \rightarrow \hat{\varrho}_k$ ,  $\rho^r = \sum_{k=1}^K \varrho_k^r \uparrow 1$ , and  $\sqrt{r}(1 - \rho^r) \rightarrow 1$ , as  $r \rightarrow \infty$ , and obtain that  $(1 - \rho^r) \vec{N}^r(rt)$  converges in distribution to*

$$\frac{\hat{W}(t)}{\sum_{k=1}^K \frac{\hat{\varrho}_k}{w_k \mu_k}} \cdot \left( \frac{\hat{\varrho}_1}{w_1}, \dots, \frac{\hat{\varrho}_K}{w_K} \right), \quad (34)$$

with  $\hat{W}(t)$  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  $\sum_{k=1}^K \frac{\hat{\varrho}_k}{\mu_k}$ . Hence, for exponentially distributed service requirements, the stationary limit of (34) coincides with the heavy-traffic limit of the steady-state queue lengths (32) as derived in Proposition 7.1. 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 7.1.** Recall that the DPS queue with phase-type distributed service requirements is a special case of the general framework of Section 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  $k(i) = k(j) = l$ . Since  $N_l = \sum_{j:k(j)=l} Q_j$ ,  $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$  (see (28)), and since for the general framework Relation (4) holds, Relation (32) 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 (33).

From (27) and (31), and since type- $j$  customers belong to class  $k(j)$  and have weight  $g_j = w_{k(j)}$ , we obtain that

$$\sum_{j=1}^J \frac{\rho_j}{g_j} \mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l} \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^K \frac{\lambda p_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2}. \quad (35)$$

Similarly, we have that

$$\sum_{j=1}^J \rho_j \mathbb{E}(R_j) = \sum_{l=1}^K \varrho_l \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \mathbb{E}(R_j) = \sum_{l=1}^K \varrho_l \frac{\mathbb{E}((B_l)^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^K \lambda p_l \frac{\mathbb{E}((B_l)^2)}{2}. \quad (36)$$

Obviously, Equations (35) and (36) remain valid in heavy traffic. Equation (33) follows after substituting (35) and (36) into (5).  $\square$

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

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

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 Section 7.1 we obtain heavy-traffic results for the residual service requirements of the customers in the various classes. In Section 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  $B_{k,h}^r$  denote the remaining service requirement of the  $h$ -th class- $k$  customer,  $k = 1, \dots, K$ ,  $h = 1, \dots, n_k$ . The following result is known for PS:

$$\mathbb{P}(B_{k,h}^r \leq x_{k,h}, N_k = n_k, k = 1, \dots, K, h = 1, \dots, n_k) = \mathbb{P}(N_k = n_k, k = 1, \dots, K) \prod_{k=1}^K \prod_{h=1}^{n_k} \mathbb{P}(B_k^{fwd} \leq x_{k,h}),$$

with  $x_{k,h} \geq 0$ . The joint distribution of the numbers of customers is of product form:  $\mathbb{P}(N_k = n_k, k = 1, \dots, K) = (1 - \rho) \frac{(n_1 + \dots + n_K)!}{n_1! \dots n_K!} \prod_{k=1}^K \varrho_k^{n_k}$ , 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,  $k = 1, \dots, K$ . 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  $B_{k,h}^r = 0$  whenever  $h > N_k$ ,  $k = 1, \dots, K$ .

**Proposition 7.3** *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights  $w_1, \dots, w_K$ . Then,*

$$\lim_{\rho \uparrow 1} \mathbb{E} \left( e^{-\sum_{l=1}^K s_l (1-\rho) N_l - \sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \right) = \mathbb{E} \left( e^{-\sum_{l=1}^K s_l \hat{N}_l} \right) \cdot \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E} \left( e^{-s_{l,h} B_l^{fwd}} \right),$$

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

Recall that  $(\hat{N}_1, \hat{N}_2, \dots, \hat{N}_K) \stackrel{d}{=} X \cdot \left( \frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \dots, \frac{\hat{\varrho}_K}{w_K} \right)$ , where  $X$  is exponentially distributed with mean  $\mathbb{E}(X) = \frac{\sum_{l=1}^K p_l \mathbb{E}((B_l)^2)}{\sum_{l=1}^K p_l \mathbb{E}((B_l)^2)/w_l}$ , cf. Proposition 7.1.

**Proof of Proposition 7.3:** It will be convenient to first analyze the conditional expectation  $\mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{Q} \right)$ . In order to do so, we condition on the type of the  $h$ -th class- $l$  customer, which we denote by  $I_{l,h}$  and takes values in  $\{i : k(i) = l\}$ . For convenience, if  $h > \sum_{j:k(j)=l} Q_j$ , then  $I_{l,h}$  has no significance. Let  $\vec{I} = (I_{1,1}, \dots, I_{1,y_1}, \dots, I_{K,1}, \dots, I_{K,y_K})$ , which takes values in the set

$$\mathcal{I} := \{\vec{i} : k(i_{1,1}) = 1, \dots, k(i_{1,y_1}) = 1, \dots, k(i_{K,1}) = K, \dots, k(i_{K,y_K}) = K\}.$$



Conditioning on the types of the customers, we can write

$$\mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{Q} \right) = \sum_{\vec{i} \in \mathcal{I}} \mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{I} = \vec{i}, \vec{Q} \right) \cdot \mathbb{P}(\vec{I} = \vec{i} | \vec{Q}). \quad (37)$$

Define the random variable  $Y_l$  as

$$Y_l = \min(y_l, \sum_{j:k(j)=l} Q_j) = \min(y_l, N_l), \quad l = 1, \dots, K,$$

and note that  $\mathbb{P}(Y_l = y_l) = \mathbb{P}(\sum_{j:k(j)=l} Q_j > y_l) \rightarrow 1$  as  $\rho \uparrow 1$ , cf. Proposition 2.1. Since  $y_l$  is a deterministic value, this implies convergence of  $Y_l$  to  $y_l$  in probability. By definition, if the  $h$ -th class- $l$  customer is of type  $i_{l,h}$ , then the corresponding residual service requirement has the same distribution as  $R_{i_{l,h}}$ ,  $h = 1, \dots, y_l$ . Hence,

$$\begin{aligned} \mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{I} = \vec{i}, \vec{Q} \right) &= \prod_{l=1}^K \prod_{h=1}^{Y_l} \mathbb{E} \left( e^{-s_{l,h} R_{i_{l,h}}} \right), \\ &\rightarrow \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E} \left( e^{-s_{l,h} R_{i_{l,h}}} \right), \quad \text{as } \rho \uparrow 1, \end{aligned} \quad (38)$$

where the convergence holds in probability (since the conditional expectation on the left-hand side converges to a deterministic value).

Given the population vector  $\vec{Q}$ , the first chosen class- $l$  customer is of type  $i$ ,  $k(i) = l$ , with probability  $\frac{Q_i}{\sum_{j:k(j)=l} Q_j}$ . The next chosen class- $l$  customer is of type  $j$ ,  $k(j) = l$ , with probability  $\frac{Q_j - \mathbf{1}_{(i=j)}}{\sum_{j:k(j)=l} Q_j - 1}$ , etc. So we obtain

$$\begin{aligned} \mathbb{P}(\vec{I} = \vec{i} | \vec{Q}) &= \frac{Q_{i_{1,1}}}{\sum_{j:k(j)=1} Q_j} \cdot \frac{Q_{i_{1,2}} - \mathbf{1}_{(i_{1,1}=i_{1,2})}}{\sum_{j:k(j)=1} Q_j - 1} \cdots \frac{Q_{i_{1,Y_1}} - \sum_{h=1}^{Y_1-1} \mathbf{1}_{(i_{1,h}=i_{1,Y_1})}}{\sum_{j:k(j)=1} Q_j - (Y_1 - 1)} \cdots \\ &\quad \frac{Q_{i_{K,1}}}{\sum_{j:k(j)=K} Q_j} \cdot \frac{Q_{i_{K,2}} - \mathbf{1}_{(i_{K,1}=i_{K,2})}}{\sum_{j:k(j)=K} Q_j - 1} \cdots \frac{Q_{i_{K,Y_K}} - \sum_{h=1}^{Y_K-1} \mathbf{1}_{(i_{K,h}=i_{K,Y_K})}}{\sum_{j:k(j)=K} Q_j - (Y_K - 1)}. \end{aligned}$$

The latter converges in probability to

$$\prod_{l=1}^K \prod_{h=1}^{y_l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{Q}_l}, \quad \text{as } \rho \uparrow 1,$$

where we used that  $(1 - \rho)(Q_1, \dots, Q_J) \xrightarrow{d} X \cdot (\hat{\rho}_1/g_1, \dots, \hat{\rho}_J/g_J)$  (see Proposition 2.1), the fact that  $Y_l$  converges in probability to  $y_l$ , the continuous mapping theorem,  $g_{i_{l,h}} = w_{k(i_{l,h})} = w_l$ , and (28). Together with (30), (37) and (38) we now obtain

$$\begin{aligned} \mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{Q} \right) &\rightarrow \sum_{\vec{i} \in \mathcal{I}} \prod_{l=1}^K \prod_{h=1}^{y_l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{Q}_l} \cdot \mathbb{E} \left( e^{-s_{l,h} R_{i_{l,h}}} \right) = \prod_{l=1}^K \prod_{h=1}^{y_l} \sum_{i_{l,h}:k(i_{l,h})=l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{Q}_l} \mathbb{E} \left( e^{-s_{l,h} R_{i_{l,h}}} \right) \\ &= \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E} \left( e^{-s_{l,h} B_l^{fwd}} \right), \end{aligned}$$

in probability as  $\rho \uparrow 1$ . By the law of total expectation we therefore have

$$\begin{aligned} \lim_{\rho \uparrow 1} \mathbb{E} \left( e^{-\sum_{j=1}^J s_j (1-\rho) Q_j - \sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \right) &= \lim_{\rho \uparrow 1} \mathbb{E} \left( e^{-\sum_{j=1}^J s_j (1-\rho) Q_j} \mathbb{E} \left( e^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B_{l,h}^r} \middle| \vec{Q} \right) \right) \\ &= \mathbb{E} \left( e^{-\sum_{j=1}^J s_j \hat{Q}_j} \right) \cdot \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E} \left( e^{-s_{l,h} B_l^{fwd}} \right). \end{aligned} \quad (39)$$

The result now follows by setting  $s_j = \tilde{s}_{k(j)}$ ,  $j = 1, \dots, J$ , in Equation (39) and noting that  $\sum_{j:k(j)=l} \hat{Q}_j = \hat{N}_l$ .  $\square$

## 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 Section 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  $d_k/\mathbb{E}(B_k^{fwd})$ ,  $k = 1, \dots, K$ . From Proposition 7.3 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 Section 6.

**Proposition 7.4** *Assume phase-type distributed service requirements and consider two standard DPS queues with weights  $(w_1, \dots, w_K)$  and  $(\tilde{w}_1, \dots, \tilde{w}_K)$ . Let  $d_k \geq 0$ ,  $k = 1, \dots, K$ . Without loss of generality we assume that the classes are ordered such that  $d_1/\mathbb{E}(B_1^{fwd}) \geq \dots \geq d_K/\mathbb{E}(B_K^{fwd})$ .*

*If  $\frac{w_k}{w_{k+1}} \leq \frac{\tilde{w}_k}{\tilde{w}_{k+1}}$ , for all  $k = 1, \dots, K-1$ , then*

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

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

**Proof:** From Proposition 7.1 we obtain that  $(1 - \rho) \sum_{k=1}^K d_k N_k^{DPS(w)}$  converges in distribution to an exponentially distributed random variable with mean

$$\frac{\sum_{k=1}^K \frac{d_k \hat{w}_k}{w_k}}{\sum_k p_k \mathbb{E}((B_k)^2)/w_k} \cdot \sum_k p_k \mathbb{E}((B_k)^2),$$

hence we need to check that

$$\frac{\sum_{k=1}^K \frac{d_k \hat{w}_k}{w_k}}{\sum_{k=1}^K \frac{\hat{w}_k}{w_k} \frac{\mathbb{E}((B_k)^2)}{\mathbb{E}(B_k)}} \geq \frac{\sum_{k=1}^K \frac{d_k \hat{w}_k}{\tilde{w}_k}}{\sum_{k=1}^K \frac{\hat{w}_k}{\tilde{w}_k} \frac{\mathbb{E}((B_k)^2)}{\mathbb{E}(B_k)}}.$$

This follows using similar arguments as in the proof of Proposition 6.1 and noting that  $\frac{\mathbb{E}((B_k)^2)}{2\mathbb{E}(B_k)} = \mathbb{E}(B_k^{fwd})$ .  $\square$

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  $w_k/w_{k+1}$ ,  $k = 1, \dots, K-1$ , all go to  $\infty$ .

**Remark 7.5** *In Kim and Kim (2006) it was conjectured that, in the case of exponentially distributed service requirements, a result similar to Proposition 7.4 holds outside heavy traffic, see also (Verloop et al. 2009, Section 6.1).*

**Remark 7.6** *In (Coffman and Denning 1973, p. 188–199) it was conjectured that  $\frac{\text{Var}(B)}{\mathbb{E}(B)^2} < 1$  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 7.4 which are based on  $\mathbb{E}(B_k^{fwd}) = \frac{\mathbb{E}((B_k)^2)}{2\mathbb{E}(B_k)} = \frac{1}{2} \left( \frac{\text{Var}(B_k)}{\mathbb{E}(B_k)} + \mathbb{E}(B_k) \right)$ . Customers belonging to classes with highly variable service distributions are likely to have longer service requirements. The variance also appears in the criteria conjectured in Coffman and Denning (1973).*

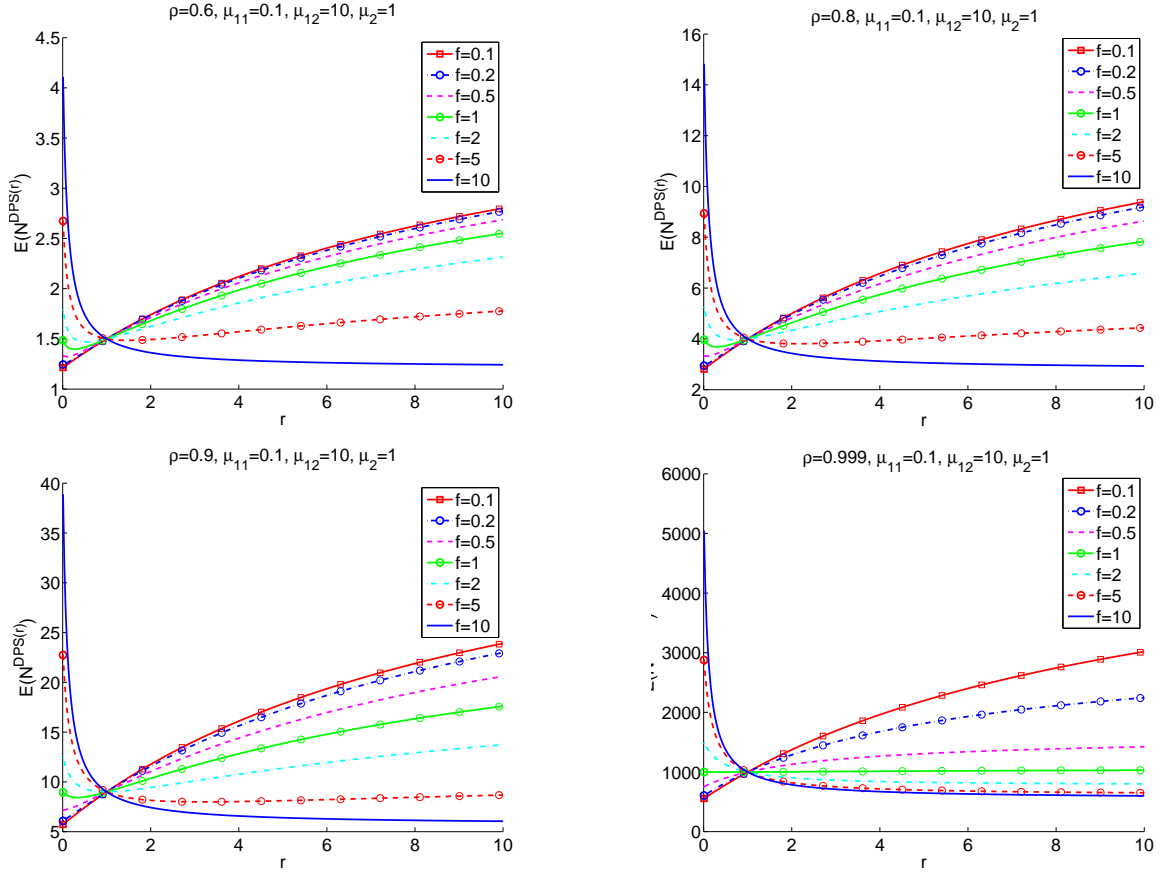


Figure 3: Total mean number of customers under a DPS policy with weights  $w_1 = 1$  and  $w_2 = r$ . Class-1 service requirements are hyper-exponentially distributed (with parameters  $\mu_{11} = 0.1, \mu_{12} = 10$ ) and class-2 service requirements are exponentially distributed (with  $\mu_2 = 1$ ). The load  $\rho = \rho_1 + \rho_2$  equals 0.6, 0.8, 0.9 and 0.999, respectively.

Although the monotonicity of the weight structure in Proposition 7.4 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 7.4:* We consider a DPS queue with two classes. Class-1 customers have hyper-exponentially 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.,  $\rho_1 = \rho_2$ . 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 7.4 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  $\frac{\rho}{1-\rho}$ .

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. Since  $\mu_{11} < \mu_2 < \mu_{12}$ , the  $c\mu$ -rule suggests to prioritize class-1 customers in phase 2, while 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 2.1, and is hence independent of  $r$ . However, away from heavy traffic, this mix may itself be influenced by  $r$ , leading to the observed non-monotonic 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 shown that, in heavy-traffic conditions, the queue length vector exhibits a so-called state-space collapse: The multi-dimensional 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.

## References

- Aalto, S., U. Ayesta. 2006. On the nonoptimality of the foreground-background discipline for IMRL service times. *Journal of Applied Probability* **43** 523–534.
- Aalto, S., U. Ayesta, S.C. Borst, V. Misra, R. Núñez-Queija. 2007. Beyond processor sharing. *Performance Evaluation Review* **34(4)** 36–43.
- Altman, E., K.E. Avrachenkov, U. Ayesta. 2006. A survey on discriminatory processor sharing. *Queueing Systems* **53** 53–63.
- Altman, E., T. Jimenez, D. Kofman. 2004. DPS queues with stationary ergodic service times and the performance of TCP in overload. *Proceedings of 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. *Proceedings of IEEE INFOCOM*. Miami FL, USA.
- 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. *Annals of Applied Probability* **11** 608–649.
- Ben Tahar, A., A. Jean-Marie. 2009. The fluid limit of the multiclass processor sharing queue. *INRIA Research Report RR-6867*.
- 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** 31–51.
- Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The  $c\mu$  rule revisited. *Advances in Applied Probability* **17** 237–238.
- Coffman, E.G., P. Denning. 1973. *Operating System Theory*. Prentice-Hall.

- Cohen, J.W. 1979. The multiple phase service network with generalized processor sharing. *Acta Informatica* **12** 245–284.
- Egorova, R., S.C. Borst, A.P. Zwart. 2007. Bandwidth-sharing networks in overload. *Performance Evaluation* **64** 978–993.
- Fayolle, G., I. Mitrani, R. Iasnogorodski. 1980. Sharing a processor among many job classes. *Journal of the ACM* **27** 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. *Advances in Applied Probability* **24** 653–698.
- Haviv, M., J. van der Wal. 2008. Mean sojourn times for phase-type discriminatory processor sharing systems. *European Journal of Operational Research* **189** 375–386.
- Jean-Marie, A., P. Robert. 1994. On the transient behavior of the processor-sharing queue. *Queueing Systems* **17** 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. *Annals of Applied Probability* **19** 1719–1780.
- Kelly, F.P. 1979. *Stochastic Networks and Reversibility*. Wiley, Chichester.
- Kessel, G. van, R. Núñez-Queija, S.C. Borst. 2004. Asymptotic regimes and approximations for discriminatory processor sharing. *Performance Evaluation Review* **32(2)** 44–46.
- Kim, B., J. Kim. 2006. Comparison of DPS and PS systems according to DPS weights. *IEEE Communications Letters* **10(7)** 558–560.
- Kingman, J.F.C. 1961. The single server queue in heavy traffic. *Proc. Cambridge Philos.* **57** 902–904.
- Kleinrock, L. 1967. Time-shared systems: A theoretical treatment. *Journal of the ACM* **14** 242–261.
- Nain, P., D. Towsley. 1994. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Transactions on Automatic Control* **39** 1853–1855.
- Rege, K.M., B. Sengupta. 1996. Queue length distribution for the discriminatory processor-sharing queue. *Operations Research* **44** 653–657.
- Righter, R., J.G. Shanthikumar. 1989. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences* **3** 323–333.
- Sigman, K. 1991. A note on a sample-path rate conservation law and its relationship with  $H = \lambda G$ . *Advances in Applied Probability* **23** 662–665.
- Stolyar, A.L. 2004. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability* **14** 1–53.
- Verloop, I.M., U. Ayesta, S.C. Borst. 2009. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems* DOI: 10.1007/s10626-009-0069-4.
- 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. *Operations Research Letters* **32** 73–76.