Centrum Wiskunde & Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

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

I.M. Verloop, U. Ayesta, R. Núñez-Queija

REPORT PNA-E0905 MARCH 2009

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

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

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 i.i.d. 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.

2000 Mathematics Subject Classification: 68M20; 60K25
Keywords and Phrases: Discriminatory processor sharing; heavy traffic; phase-type service requirements; residual service requirements; scheduling
Note: 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)

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

I.M. Verloop[a], U. Ayesta[b,c,*], R. Núñez-Queija[a,d]

[a]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[b]BCAM - Basque Center for Applied Mathematics,
Bizkaia Technology Park, 48170 Zamudio, Spain
[c]LAAS-CNRS, Université de Toulouse, 7 Avenue Colonel Roche,
31077 Toulouse Cedex, France
[d]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 [26] 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 i.i.d. 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.

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

1 Introduction

The Discriminatory Processor Sharing (DPS) model, introduced in [24], is a versatile generalization of the celebrated (Egalitarian) Processor Sharing (PS) model. DPS allows class-based differentiation by assigning different weights to customers[1] of different classes. 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

*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).

[1]In this paper we adopt the traditional queueing theoretic terminology; often "customers" are abstract entities such as jobs, flows, packets, etc.

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 [4]. An extensive survey of the DPS literature can be found in [3]. The seminal paper [14] 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 [6]. This is further illustrated by the heavy-traffic bounds on the mean queue lengths reported in [2], 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 [6], and the tail index of the sojourn time distribution [9].

Several papers have analyzed the (discriminatory) processor sharing model assuming overload conditions with general service requirement distributions. In [4] the authors determine the queue length growth rates of the standard DPS model, generalizing the analogous result for egalitarian processor sharing [19]. Further extensions to bandwidth-sharing networks [13] and a network setting similar to ours [8], have been obtained more recently. In 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. [27]. Similar generalizations of DPS were previously considered by [8, 17, 18]. The analysis in [17] is particularly relevant for the present study. There, the generalized DPS model was investigated assuming heavy traffic conditions (and finite second moments of the service times), by using a direct relationship with critical Crump-Mode-Jagers branching processes. Through appropriate choices for a quite general functional of the queue length process, [17] determines 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-variate stochastic process under asymptotic (heavy-traffic) scaling has been demonstrated previously in other queueing models, see for example [7, 28, 20].

Our work is inspired by the heavy-traffic analysis of the traditional DPS model with exponential service requirement distributions in [26]. After developing a procedure to determine all moments of the queue length distributions from systems of linear equations, [26] 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 process. In [22] it was indicated that a similar approach as in [26] 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. Our analysis relies on heavy-traffic scaling techniques which typically require finiteness of second moments of the service requirements in many queueing models [7]. Since all phase-type distributions (with a finite number of phases) have a finite second moment, this restriction is implicit in our modeling approach. Indeed, our results show that the second moments appear in a natural fashion in the heavy-traffic limit. We believe that our results do extend to all distributions with a finite second moment, but we do not investigate this here.

Allowing the relative service weights of customers to change over time as they acquire service, effectively opens 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 disciplines and in a (ii) G/M/1-queue among all preemptive non-anticipating disciplines, see for example [16, 10, 25]. 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), 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 the queue, the remaining service requirements of the various customers are i.i.d., and distributed according to the forward recurrence time, a result that is well known for Egalitarian PS (see for example [12, 21]). In addition, we show that the 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 general Markovian framework and state the main result of the paper, 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 process 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 for the general model. Section 7 applies the state-space collapse result to the standard DPS queue with phase-type distributed service requirements. In addition, it discusses the optimal choice of the weights, and shows that residual service requirements are asymptotically i.i.d., and have the same distribution as the forward recurrence times. Concluding remarks can be found in Section 8.

2 General framework and main result

We consider a general 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}. Customers of type i have an exponentially distributed service requirement with mean $\frac{1}{\mu_i}$. After service completion, customers of type i become of type j with probability p_{ij}, and leave the system with probability $p_{i0} := 1-\sum_{j=1}^{J} p_{ij}$. We denote the number of type-j customers in the system by Q_j and the workload in type j by W_j. The J customer types share a common resource of capacity one. There are 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 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 $\mathbb{E}(\bar{R}) = (\mathbb{E}(R_1), \ldots, \mathbb{E}(R_J))^T$ and let P be a $J \times J$ matrix with $P = (p_{ij})$, $i, j = 1, \ldots, J$. Since P is a sub-stochastic matrix, $(I-P)^{-1}$ is well defined and we can write

$$\mathbb{E}(\bar{R}) = (I-P)^{-1}\bar{m}, \quad \text{with} \quad \bar{m} = (1/\mu_1, \ldots, 1/\mu_J)^T.$$

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., $\bar{\gamma}^T = \bar{p}_0^T (I - P)^{-1}$, with $\bar{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$ and $\bar{p}_0 = (p_{01}, \ldots, p_{0J})^T$. 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 type-i customers by

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

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

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

Our main result shows that the steady-state distribution of the multi-dimensional queue length process 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 accomplished by fixing the $\bar{p}_0, P$ and $\bar{m}$, and letting

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

since then $\rho = \lambda \bar{p}_0^T (I - P)^{-1} \bar{m} \uparrow 1$. 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 type-i customers ($\sum_{j=1}^{J} \hat{\rho}_j = 1$).

We will now state our main result, which establishes a state-space collapse for the queue length vector in the heavy-traffic regime.

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

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

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

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

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. Note that, in general

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

since the expression within the expectation operator reflects the capacity share of class 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 indicates that

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

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

Numerical illustration of Proposition 2.1: We consider two types of customers and choose $g_1 = 2, g_2 = 1$, $\mu_1 = 2, \mu_2 = 5, p_{12} = 0.3, p_{21} = 0.1$, and we take different values for the loads. In Figure 1, the horizontal and vertical axes correspond to Q_1 and Q_2 respectively. We plot the joint queue length probabilities for loads $\rho = 0.8$ ($\rho_1 = 0.5872, \rho_2 = 0.2128$), $\rho = 0.90$ ($\rho_1 = 0.6605, \rho_2 = 0.2394$) and $\rho = 0.99$ ($\rho_1 = 0.7266, \rho_2 = 0.2634$), 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 $\frac{g_1}{\hat{\rho}_1}\frac{\hat{\rho}_2}{g_2}$, starting from the origin. In Figure 1 we see that as the load increases, the probable states indeed tend to concentrate more around this line. For load $\rho = 0.99$, this effect is clearly visible; the probable queue length states are strongly concentrated around the line with slope $\frac{g_1}{\rho_1}\frac{\rho_2}{g_2} \approx 0.73$.

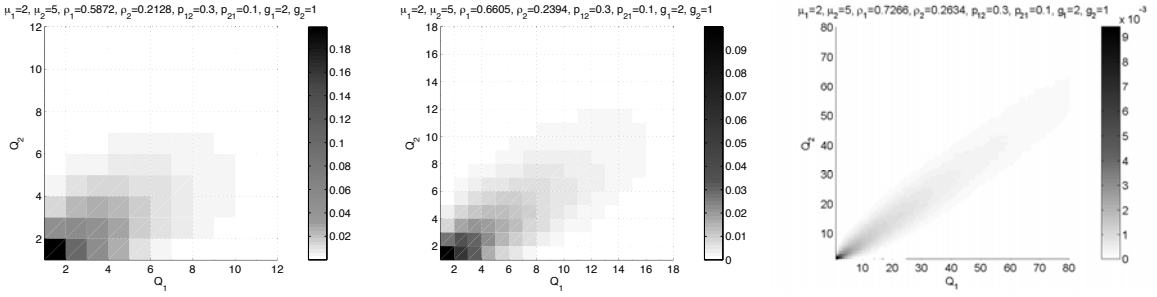

Figure 1: Joint queue length probabilities for load $\rho = 0.8$ (left), $\rho = 0.90$ (center) and $\rho = 0.99$ (right), respectively.

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

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

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

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

where $\delta_q = 1$ if $q > 0$, and $\delta_q = 0$ otherwise, and with $\bar{e}_i$ the i-th unit vector. It will be notationally

convenient to use the following transformation:

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

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

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

Note that

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

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

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

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

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

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

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

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 [26] 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 by $\bar{s} = (s_1, \ldots, s_J)$ and we will use the short hand notation $\mathrm{e}^{-(1-\rho)\bar{s}} = (\mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_J})$. If

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

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 [15]). For now, we assume that the limit exists and come back to this assumption in Section 5. In this section we give two lemma's that describe properties of $\lim_{\rho \uparrow 1} p(\mathrm{e}^{-(1-\rho)\bar{s}})$. 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

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

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.

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

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

Proof: From (11) we have

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

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

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

where in the third step we used that the function $\frac{Q_i}{\sum_{j=1}^{J} Q_j g_j} \cdot \mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \ldots \cdot \mathrm{e}^{-(1-\rho)s_1 Q_J} \cdot \mathbf{1}_{(\sum_{j=1}^{J} Q_j > 0)}$ is uniform integrable (since it is upper bounded by $\frac{1}{\min_j(g_j)}$), and converges in distribution to $\frac{\hat{Q}_i}{\sum_{j=1}^{J} \hat{Q}_j g_j} \cdot \mathrm{e}^{-s_1\hat{Q}_1} \cdot \ldots \cdot \mathrm{e}^{-s_1\hat{Q}_J} \cdot \mathbf{1}_{(\sum_{j=1}^{J} \hat{Q}_j > 0)}$. From (15) and (16) we obtain (14). □

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

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

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

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

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

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

gives

$$0 =$$

$$\lim_{\rho\uparrow 1}\sum_{i=1}^{J}\left(\mu_i g_i\frac{1-\mathrm{e}^{-(1-\rho)s_i}+\sum_{j=1}^{J}p_{ij}(\mathrm{e}^{-(1-\rho)s_j}-1)}{1-\rho}-\lambda g_i\mathrm{e}^{-(1-\rho)s_i}\sum_{j=1}^{J}p_{0j}\frac{1-\mathrm{e}^{-(1-\rho)s_j}}{1-\rho}\right)\frac{\partial r(\bar{z})}{\partial z_i}|_{\bar{z}=\mathrm{e}^{-(1-\rho)\bar{s}}}$$

$$=\sum_{i=1}^{J}g_i\left(\mu_i(s_i-\sum_{j=1}^{J}p_{ij}s_j)-\hat{\lambda}\sum_{j=1}^{J}p_{0j}s_j\right)\frac{\partial\hat{r}(\bar{s})}{\partial s_i}, \tag{18}$$

where in the second step we used equation (16) and the fact that $\lim_{\rho\uparrow 1}\frac{x^{1-\rho}-1}{1-\rho}=\ln(x)$. The result now follows. □

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

$$(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)\stackrel{d}{=}(\frac{\hat{\rho}_1}{g_1},\frac{\hat{\rho}_2}{g_2},\ldots,\frac{\hat{\rho}_J}{g_J})\cdot X, \tag{19}$$

for some random variable X. Second, we find 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 $\lim_{\rho\uparrow 1}p(\mathrm{e}^{-(1-\rho)\bar{s}})$ exists, thereby showing in Subsections 5.1 and 5.2 that there is a unique limit. For any converging subsequence this analysis can be performed, in particular for the lim sup and lim inf, which implies that the limit itself exists. This formally establishes the state-space collapse $(1-\rho)(Q_1,Q_2,\ldots,Q_J)\stackrel{d}{\to}(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)$ with $(\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_J)$ taking only values on the line described in (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 of Lemma 4.2. From this partial differential equation it can be derived that the function $\hat{r}(\bar{s})$ is constant on the $J-1$ dimensional set

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

as will be shown in Lemma 5.2. Hence, the function $\hat{r}(\bar{s})$ depends only on $\bar{s}$ through $\sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}s_j$, so there is a function $\hat{r}^*:\mathbb{R}\to\mathbb{R}$ such that $\hat{r}(\bar{s})=\hat{r}^*(\sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}s_j)$. From Lemma 4.1 and $\frac{\partial\hat{r}(\bar{s})}{\partial s_i}=\frac{\hat{\rho}_i}{g_i}\frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}s_j}$, we then obtain

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

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

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

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:

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

or equivalently

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

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

Before we proceed to prove that the generating function $\hat{r}(\bar{s})$ is constant on the plane H_c we first give a geometric interpretation for this fact in the particular case of $J = 3$. In Figure 2 (left) we depict the plane H_c for $J = 3$. For a given $s_0 \in H_c$, we draw a flow curve $\bar{f}(u)$, $u \geq 0$, defined such that the tangent at every point is precisely $\bar{f}'(u) := \bar{F}(\bar{f}(u))$ and $\bar{f}(0) = \bar{s}_0 \in H_c$. We will see in the proof of Lemma 5.2 that the vector $\bar{F}(\bar{s})$ is parallel to the plane H_c, for all $\bar{s} \in H_c$, thus the flow $\bar{f}(u)$ stays in the plane H_c for all $u \geq 0$. By Lemma 4.2, the vector $\bar{F}(\bar{s})$ and the gradient $\bigtriangledown\hat{r}(\bar{s})$ are perpendicular, for all $\bar{s}$, so $\bar{f}'(u) = \bar{F}(\bar{f}(u)) \perp \bigtriangledown\hat{r}(\bar{f}(u))$. Thus the function $\hat{r}$ has the same value in every point on a given flow $\bar{f}(u)$. In Figure 2 (right) we draw several flows in the plane H_c. In the proof of Lemma 5.2 we will see that all flows starting in the plane H_c go through one common point $c \cdot \bar{s}^*$. Since the function $\hat{r}$ is continuous and constant on each flow trajectory, it follows that $\hat{r}(\bar{s})$ is constant on the whole plane H_c, or equivalently, $\bigtriangledown\hat{r}(\bar{s}) \perp H_c$.

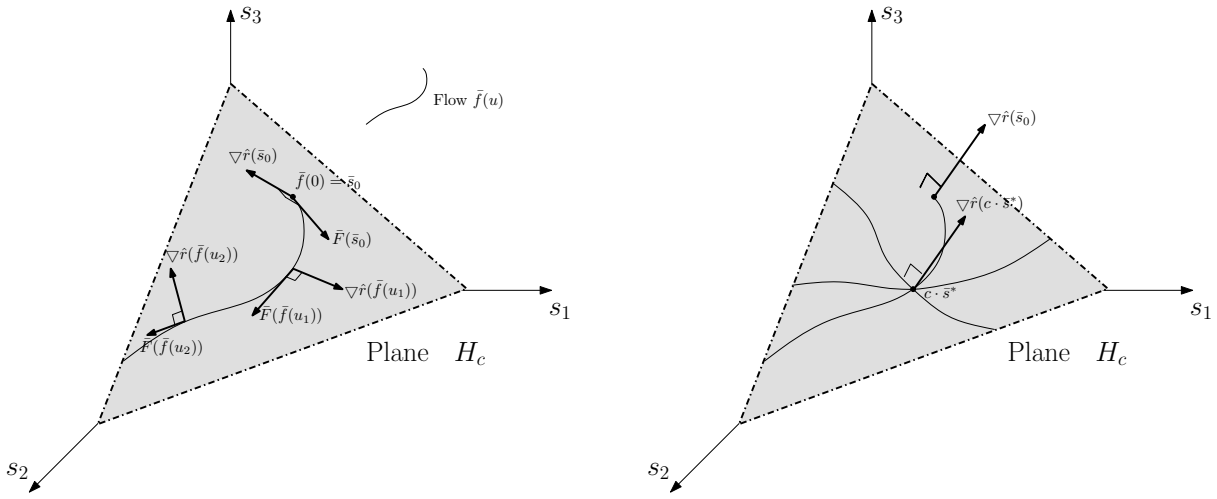

Figure 2: Geometrical interpretation of the proof of Lemma 5.2 for the case $J = 3$.

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

Lemma 5.1 *Consider the matrix*

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

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. The eigenvector corresponding to the eigenvalue 0 is equal to $\bar{s}^*$ with $s_j^* := \frac{g_j}{\hat{\rho}_j}\eta_j$, for a certain $\bar{\eta} \geq 0$ with $\sum_{j=1}^{J} \eta_j = 1$. In addition, $\bar{s}^* \in H_1$.

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

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

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

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

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 and hence has a unique equilibrium distribution $\bar{\eta}$, i.e., $\bar{\eta} S^T = 0$ and $\sum_{j=1}^{J} \eta_j = 1$. In addition, 0 is an eigenvalue of the matrix S^T with multiplicity 1, and the real parts of all other eigenvalues are strictly negative, cf. Perron-Frobenius theorem, [5]. By similarity the same holds for the matrix A. The eigenvector of A corresponding to the eigenvalue 0 is given by $\bar{s}^* = D^{-1}\bar{\eta}$, since $A\bar{s}^* = D^{-1}DAD^{-1}\bar{\eta} = D^{-1}S\bar{\eta} = 0$. □

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

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

Proof: From (21) it follows that

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

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

$$\bar{f}(0) = \bar{s} \quad \text{and} \quad \frac{\mathrm{d}f_i(u)}{\mathrm{d}u} = F_i(\bar{f}(u)), \quad \text{for all } i \text{ and } u \geq 0. \tag{22}$$

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

$$\frac{\mathrm{d}\hat{r}(\bar{f}(u))}{\mathrm{d}u} = \sum_{i=1}^{J} \frac{\mathrm{d}f_i(u)}{\mathrm{d}u} \cdot \frac{\partial \hat{r}(\bar{s})}{\partial s_i}\bigg|_{\bar{s}=\bar{f}(u)} = 0,$$

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

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

Since the function $\hat{r}(\bar{s})$ is constant along one flow, and all flows in the plane H_c go through the common point $c \cdot \bar{s}^* \geq 0$, we obtain that the function $\hat{r}(\bar{s})$ is constant on the plane H_c. □

5.2 Determining the common factor

In the previous section we showed that $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) \stackrel{d}{=} (\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J}) \cdot X$, 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, W. When scaled with $(1-\rho)$ the total workload has a proper distribution as $\rho \uparrow 1$, see [23]:

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

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

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

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 residual 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, i.e., $R_{j,h} \stackrel{d}{=} R_j$ for all h. Hence,

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

For the scaled workload we can write

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

where in the last step we used that $\mathrm{e}^{\sum_{j=1}^{J} \frac{\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))}{(1-\rho)s}(1-\rho)sQ_j}$ 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 $\frac{\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))}{(1-\rho)s} \to -\mathbb{E}(R_j)$ 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 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 for the weights influences the performance of the system. To each type of customers we associate a cost $c_j \geq 0$, $j = 1, \ldots, J$. 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 $g_1, \ldots, g_J$. 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 for $\frac{c_i}{\mathbb{E}(R_i)}$.

Proposition 6.1 *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 $c_j \geq 0$, $j = 1, \ldots, J$. Without loss of generality we assume that the types are ordered such that $\frac{c_1}{\mathbb{E}(R_1)} \geq \frac{c_2}{\mathbb{E}(R_2)} \geq \ldots \geq \frac{c_J}{\mathbb{E}(R_J)}$.*

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

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

Here we used that $c_i\mathbb{E}(R_j)(\frac{1}{g_i\tilde{g}_j} - \frac{1}{g_j\tilde{g}_i}) \geq c_j\mathbb{E}(R_i)(\frac{1}{g_i\tilde{g}_j} - \frac{1}{g_j\tilde{g}_i})$, 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 $i \leq j$. □

As mentioned in the Introduction, the so-called $c\mu$-rule minimizes the mean holding cost in an (i) M/G/1-queue among all non-preemptive 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 more preference is given according to the values of $\frac{c_i}{\mathbb{E}(R_i)}$, $i = 1, \ldots, J$.

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 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 to class-k customers. The capacity is shared among the customers of the various classes in accordance with the Discriminatory Processor Sharing (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 to 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.

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^{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 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

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

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 γ_i's can be found from the following set of equations:

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

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

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

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 characterize 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

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

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

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

see [5, 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

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

We now show the state-space collapse for the standard DPS queue with phase-type distributed service requirements. When passing $\rho \to 1$ 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 will 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 7.1 *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights $w_1, \ldots, w_K$. When scaled with $1-\rho$, the queue length vector has a proper distribution as $\rho \to 1$,*

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

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

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

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

Proof: 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

$$\sum_{j=1}^J \frac{\rho_j}{g_j}\mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l} \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l}\mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l}\frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^K \frac{\lambda p_l}{w_l}\frac{\mathbb{E}(B_l^2)}{2}. \tag{34}$$

Similarly, we have that

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

Obviously, Equations (34) and (35) remain valid in heavy traffic. Equation (33) follows after substituting (34) and (35) into (5). □

Note that, although the limiting distribution depends on the second moment 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. [2]. Similar partial insensitivity results have also been proved for the mean sojourn time conditioned on the service requirement, [6], and the tail index of the sojourn time distribution, [9].

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 on the residual service requirements of the customers in the various classes. In Section 7.2, monotonicity in the weights of the DPS policy is investigated.

7.1 Residual service requirements

The distribution of the residual service requirement of a customer 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 h-th class-k customer in the PS queue, $k = 1, \ldots, K$, $h = 1, \ldots, n_k$. The following result is known for PS:

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

with $x_{k,h} \geq 0$. The joint distribution of the numbers of customers is of product form: $\mathbb{P}(N_k = n_k, k = 1, \ldots, K) = (1-\rho)\frac{(n_1+\ldots+n_K)!}{n_1!\cdot\ldots\cdot n_K!} \prod_{k=1}^{K} \varrho_k^{n_k}$, see for example [12, 21]. In this section we show that in a heavy-traffic setting a similar result as for the PS queue is true 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 number 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 define $B^r_{k,h} = 0$ whenever $h > N_k$, $k = 1, \ldots, K$.

Proposition 7.2 *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights* $w_1, \ldots, w_K$. *Then,*

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

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

Recall that $(\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_K) \stackrel{d}{=} X \cdot \left(\frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \ldots, \frac{\hat{\varrho}_K}{w_K}\right)$, where X is an exponentially distributed random variable 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 7.2: It will be convenient to first analyze the conditional expectation $\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K}\sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \bar{Q}\right)$. 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 interpretation. Let $\bar{I} = (I_{1,1}, \ldots, I_{1,y_1}, \ldots, I_{K,1}, \ldots, I_{K,y_K})$, which takes values in the set

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

Conditioning on the types of the customers, we can write

$$\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K}\sum_{h=1}^{y_l} s_{l,h}B^r_{l,h}}\middle|\bar{Q}\right) = \sum_{\bar{i}\in\mathcal{I}} \mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K}\sum_{h=1}^{y_l} s_{l,h}B^r_{l,h}}\middle|\bar{I}=\bar{i},\ \bar{Q}\right)\cdot\mathbb{P}(\bar{I}=\bar{i}|\bar{Q}). \tag{36}$$

Define the random variable Y_l as

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

and note that $\mathbb{P}(Y_l = y_l) = \mathbb{P}(\sum_{j:k(j)=l} Q_j > y_l) \to 1$ as $\rho\uparrow 1$, cf. Proposition 2.1. 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 $R_{i_{l,h}}$, $h=1,\ldots,y_l$. Hence,

$$\begin{aligned}\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K}\sum_{h=1}^{y_l} s_{l,h}B^r_{l,h}}\middle|\bar{I}=\bar{i},\ \bar{Q}\right) &= \prod_{l=1}^{K}\prod_{h=1}^{Y_l}\mathbb{E}\left(\mathrm{e}^{-s_{l,h}R_{i_{l,h}}}\right),\\ &\to \prod_{l=1}^{K}\prod_{h=1}^{y_l}\mathbb{E}\left(\mathrm{e}^{-s_{l,h}R_{i_{l,h}}}\right), \quad \text{as } \rho\uparrow 1,\end{aligned} \tag{37}$$

where the convergence holds in probability.

Given the population vector $\bar{Q}$, 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 $\frac{Q_j - 1_{(i=j)}}{\sum_{j:k(j)=l}Q_j - 1}$, etc. So we obtain

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

The latter converges in probability to

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

where we used that $(1-\rho)(Q_1,\ldots,Q_J)\xrightarrow{d} X\cdot(\hat{\rho}_1/g_1,\ldots,\hat{\rho}_J/g_J)$ (see Proposition 2.1), the fact that Y_l converges in probability to y_l, and $g_{i_{l,h}} = w_{k(i_{l,h})} = w_l$. Together with (30), (36) and (37) this gives that

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

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

$$\begin{aligned}\lim_{\rho\uparrow 1}\mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J}s_j(1-\rho)Q_j-\sum_{l=1}^{K}\sum_{h=1}^{y_l}s_{l,h}B^r_{l,h}}\right) &= \lim_{\rho\uparrow 1}\mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J}s_j(1-\rho)Q_j}\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K}\sum_{h=1}^{y_l}s_{l,h}B^r_{l,h}}\middle|\bar{Q}\right)\right)\\ &= \mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J}s_j\hat{Q}_j}\right)\cdot\prod_{l=1}^{K}\prod_{h=1}^{y_l}\mathbb{E}\left(\mathrm{e}^{-s_{l,h}B^{fwd}_l}\right).\end{aligned} \tag{38}$$

The result now follows by setting $s_j = \tilde{s}_{k(j)}$ in Equation (38) and noting that $\sum_{j:k(j)=l}\hat{Q}_j = \hat{N}_l$. □

7.2 Monotonicity in the weights

In Section 6 we showed monotonicity in the holding cost. In this section, we investigate the implications for the standard DPS queue, assigning a cost per class instead of per type. So the cost $d_k \geq 0$ represents the cost associated to a class-k customer. As we will see in the proposition below, the scaled holding cost stochastically decreases when relatively larger weights are assigned to customers of classes according to the values of $d_k/\mathbb{E}(B_k^{fwd})$, $k = 1, \ldots, K$. From Proposition 7.2 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, so as to decrease queue lengths 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. Indeed, in the particular case of exponentially distributed service requirements it holds that $d_k/\mathbb{E}(B_k^{fwd}) = d_k\mu_k$.

Proposition 7.3 *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 $d_k \geq 0$, $k = 1, \ldots, K$. Without loss of generality we assume that the classes are ordered such that $d_1/\mathbb{E}(B_1^{fwd}) \geq \ldots \geq d_K/\mathbb{E}(B_K^{fwd})$.*
If $\frac{w_k}{w_{k+1}} \leq \frac{\tilde{w}_k}{\tilde{w}_{k+1}}$, for all $k = 1, \ldots, K-1$, 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 $w_1, \ldots, w_K$.

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

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

hence we need to check that

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

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

Remark 7.4 *In [11, p. 188–199] it was conjectured that $\frac{Var(B)}{\mathbb{E}((B)^2)} < 1$ 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 [29] the authors found a counterexample to this conjecture, and it was later shown in [1] 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 7.3 which are based on $\mathbb{E}(B_k^{fwd}) = \frac{\mathbb{E}((B_k)^2)}{2\mathbb{E}(B_k)} = \frac{1}{2}\left(\frac{Var(B_k)}{\mathbb{E}(B_k)} + \mathbb{E}(B_k)\right)$. 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 [11].*

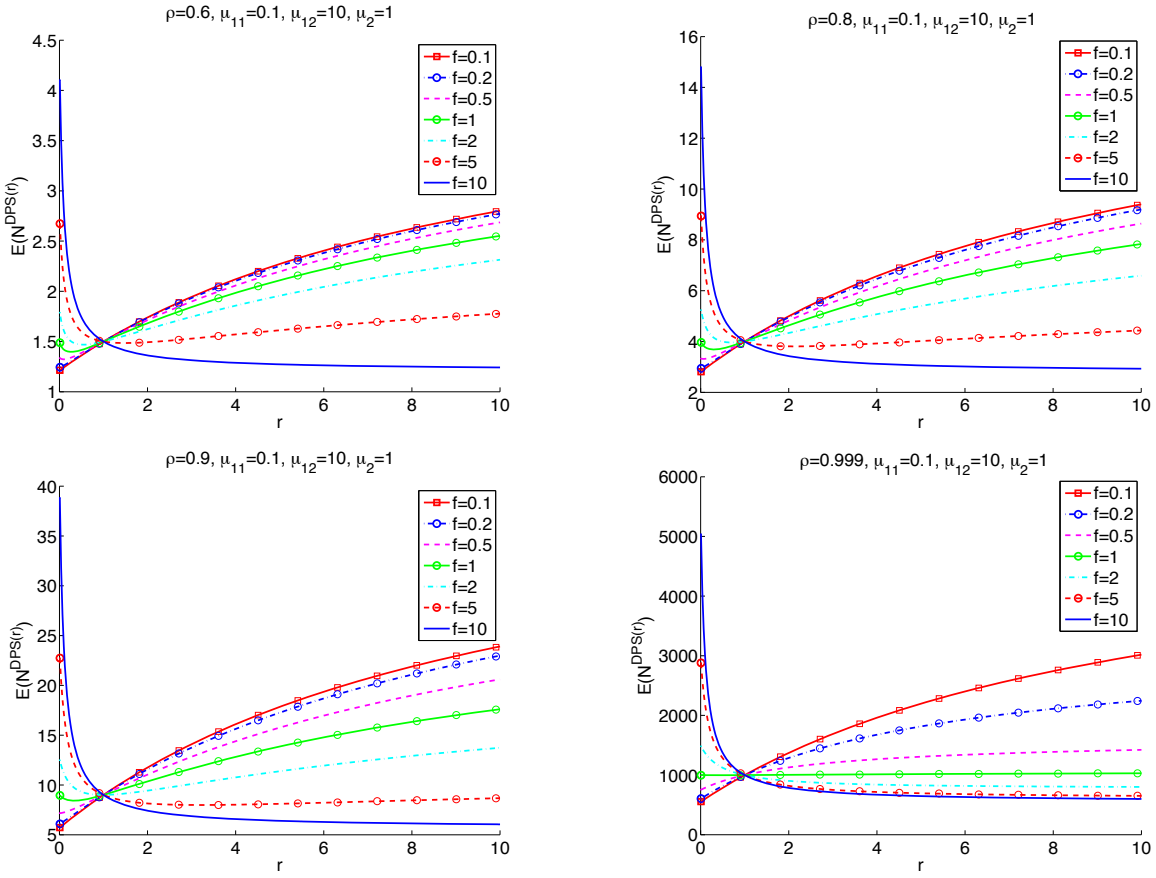

Figure 3: Total mean number of customers under a DPS policy with weights $w_1 = 1$ and $w_2 = r$. Class-1 service requirements are hyper-exponentially distributed (with parameters $\mu_{11} = 0.1, \mu_{12} = 10$) and class-2 service requirements are exponentially distributed (with $\mu_2 = 1$). The load $\rho = \varrho_1 + \varrho_2$ equals 0.6, 0.8, 0.9 and 0.999, respectively.

Although the monotonicity of the weight structure in Proposition 7.3 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 holding cost is numerically investigated for different values of the total load.

Numerical evaluation of Proposition 7.3: 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., $\varrho_1 = \varrho_2$. We will be 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 take the weight for class 1 as $w_1 = 1$, and that of class 2 as $w_2 = r$, with $r > 0$. Proposition 7.3 states that in a heavily-loaded system the steady-state 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 users is given by $\frac{\rho}{1-\rho}$.

In Figure 3 we plot the mean total number of customers (denoted by $\mathbb{E}(N^{DPS(r)})$) as a function of the weight parameter 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 was obtained by solving a system of linear equations as described in [14]. 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 mean total number of users 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 Discriminatory Processor Sharing (DPS) queue with phase-type distributed service requirements is a special case. In our main result we have shown that, under heavy traffic conditions, the queue length process 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 have seen that the DPS model in heavy traffic inherits several well known properties of the standard PS queue (not necessarily in heavy traffic). For example, in the limit, the (scaled) number of customers present in a DPS queue is exponentially distributed, which is the continuous analogue of the geometric queue length distribution of the PS queue. In addition, we showed that (again, in a heavy-traffic regime) the residual service requirements are i.i.d. 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 performance improves as customers with lower variability in their service requirements get larger weights. This property can be understood from the standard intuition of size-based scheduling: Customers belonging to classes with highly variable service distributions are likely to have longer residual service requirements and should therefore be given lower priority.

References

[1] S. Aalto and U. Ayesta. On the nonoptimality of the foreground-background discipline for IMRL service times. *Journal of Applied Probability*, 43:523–534, 2006.

[2] S. Aalto, U. Ayesta, S.C. Borst, V. Misra, and R. Núñez-Queija. Beyond processor sharing. *Performance Evaluation Review*, 34:36–43, 2007.

[3] E. Altman, K.E. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53:53–63, 2006.

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

[5] S. Asmussen. *Applied Probability and Queues*. Springer, New York, 2003.

[6] K.E. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM*, Miami, FL, USA, 2005.

[7] S.L. Bell and R.J. Williams. 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, 2001.

[8] A. Ben Tahar and A. Jean-Marie. The fluid limit of the multiclass processor sharing queue. *INRIA Research Report*, RR-6867, 2009.

[9] S.C. Borst, R. Núñez-Queija, and A.P. Zwart. Sojourn time asymptotics in processor-sharing queues. *Queueing Systems*, 53:31–51, 2006.

[10] C. Buyukkoc, P. Varaya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[11] E.G. Coffman and P. Denning. *Operating system theory*. Prentice-Hall, 1973.

[12] J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12:245–284, 1979.

[13] R. Egorova, S.C. Borst, and A.P. Zwart. Bandwidth-sharing networks in overload. *Performance Evaluation*, 64:978–993, 2007.

[14] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27:519–532, 1980.

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

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

[17] G. Grishechkin. On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability*, 24:653–698, 1992.

[18] M. Haviv and J. van der Wal. Mean sojourn times for phase-type discriminatory processor sharing systems. *European Journal of Operational Research*, 189:375–386, 2008.

[19] A. Jean-Marie and P. Robert. On the transient behavior of the processor-sharing queue. *Queueing Systems*, 17:129–136, 1994.

[20] W.N. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Annals of Applied Probability*, 2009. To appear.

[21] F.P. Kelly. *Stochastic Networks and Reversibility*. Wiley, Chichester, 1979.

[22] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Asymptotic regimes and approximations for discriminatory processor sharing. *Performance Evaluation Review*, 32:44–46, 2004.

[23] J.F.C. Kingman. The single server queue in heavy traffic. *Proc. Cambridge Philos.*, 57:902–904, 1961.

[24] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14:242–261, 1967.

[25] P. Nain and D. Towsley. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Transactions on Automatic Control*, 39:1853–1855, 1994.

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

[27] R. Righter and J.G. Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, 3:323–333, 1989.

[28] A.L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, 14:1–53, 2004.

[29] A. Wierman, N. Bansal, and M. Harchol-Balter. A note comparing response times in the M/GI/1/FB and M/GI/1/PS queues. *Operations Research Letters*, 32:73–76, 2004.