



**HAL**  
open science

## Heavy-traffic analysis of a multi-class queue with relative priorities

Ane Izagirre, Urtzi Ayesta, Ina Maria Maaïke Verloop

► **To cite this version:**

Ane Izagirre, Urtzi Ayesta, Ina Maria Maaïke Verloop. Heavy-traffic analysis of a multi-class queue with relative priorities. 2013. hal-00790846v1

**HAL Id: hal-00790846**

**<https://hal.science/hal-00790846v1>**

Submitted on 21 Feb 2013 (v1), last revised 7 Jul 2014 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Heavy-traffic analysis of a multi-class queue with relative priorities

A. Izagirre<sup>b,e</sup>, U. Ayesta<sup>b,c,d,e</sup>, I.M. Verloop<sup>a,e</sup>

<sup>a</sup>CNRS, IRIT, 2 rue C. Camichel, 31071 Toulouse, France

<sup>b</sup>CNRS, LAAS, 7 avenue du colonel Roche, 31400 Toulouse, France

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

<sup>d</sup>Univ. of the Basque Country, Department of Computer Science and Artificial Intelligence, Manuel de Lardizabal, 1, 20018 Donostia, Spain

<sup>e</sup>Univ. de Toulouse, INP, INSA, LAAS, 31400 Toulouse, France

## Abstract

We study the steady-state queue-length vector in a multi-class single-server queue with relative priorities. Upon service completion, the probability that the next customer to be served is from class  $k$  is controlled by class-dependent weights. Once a customer has started service, it is served without interruption until completion. This is a generalization of the random-order-of-service discipline. We investigate the system in a heavy-traffic regime. We first establish a state-space collapse for the scaled queue length vector, that is, in the limit the scaled queue length vector is distributed as the product of an exponentially distributed random variable and a deterministic vector. As a direct consequence, we obtain that the scaled number of customers in the system reduces as classes with smaller mean service requirement obtain relatively larger weights. We then show that the scaled waiting time of a class- $k$  customer is distributed as the product of two exponentially distributed random variables. This allows us to determine the value of the weights that minimize the  $m$ -th moment of the scaled waiting time for a customer of arbitrary class. We simulate a system with two different classes of customers in order to numerically verify some of the analytical results.

## 1 Introduction

In this paper we study a multi-class  $M/G/1$  queue with relative priorities. Service is non-preemptive and upon service completion, the probability that the next customer to be served is from class  $k$  is

$$\frac{n_k p_k}{\sum_j n_j p_j}, \quad (1)$$

where,  $p_j$ ,  $j = 1, \dots, K$ , are given class-dependent weights, and  $n_j$  is the number of class- $j$  customers at the decision epoch. The intra-class scheduling discipline is non-preemptive and it does not use information of the actual service time requirements. The relative priority model can have application in various domains, in particular in telecommunication networks [5].

A special case of this model is when the intra-class scheduling discipline is uniform random, that is, within a class a customer is selected uniformly at random. This model was proposed in [9] and it is referred to as discriminatory-random-order-of-service (DROS). In recent years several interesting studies have been published on DROS, [10, 11, 12]. Expressions for the mean waiting time of a customer given its class have been obtained in [10]. In [11, 12] the authors derive differential equations that the transform of the joint queue lengths and the waiting time in steady-state must satisfy, respectively, and this allows the authors to find the moments of the queue lengths as a solution of linear equations.

In the single class case, DROS reduces to the well-studied random-order-of-service (ROS) discipline. Classical papers on ROS are for example [14, 15, 16]. The Laplace transform for the waiting time distribution was obtained in [14]. In [14, 15, 18], ROS is studied in a heavy-traffic setting and for

service requirements having finite variance it was shown that (i) the scaled queue length converges to an exponential distribution, and (ii) the scaled waiting time is equal in distribution to the product of two independent exponential random variables. More recently, the authors of [6] obtained the waiting time distribution in heavy traffic for certain service requirements having infinite variance. In addition, waiting time tail asymptotics have been obtained in [6]. In [4] the authors derive the relationship between the waiting time under ROS and the sojourn time under the processor-sharing discipline.

In the present study, we establish a state-space collapse for the scaled queue length vector in the heavy-traffic regime for a multi-class M/G/1 queue with relative priorities. We note that a similar state-space collapse result was observed in [17] for the discriminatory processor sharing model. The result shows that in the limit, the queue length vector is the product of an exponentially distributed random variable and a deterministic vector. In particular, this allows to show that the scaled number of customers in the system reduces as classes with higher value of  $c_k/\mathbb{E}[B_k]$  obtain a relatively larger weight, where  $c_k$  is the cost associated to class  $k$ , and  $\mathbb{E}[B_k]$  is the mean service requirement of a class- $k$  customer. This can be seen as an extension of the optimality result of the  $c\mu$ -rule [8], the strict priority discipline that gives priority in decreasing order of  $c_k/\mathbb{E}[B_k]$ .

For DROS, i.e., under the additional assumption that the intra-class discipline is uniform random we study in addition the waiting time in the heavy-traffic setting. Using the state-space collapse result, we obtain the distribution of the waiting time for a customer of a given class in heavy traffic and prove that it is distributed as the product of two exponentially distributed random variables. This generalizes [15] where this result was shown for the single-class ROS queue. Moreover, we also find the value of the weights that minimizes the  $m$ -th moment of the waiting time for a customer of arbitrary class.

Finally, we simulate a system with two different classes of customers under a DROS discipline and depict the queue length distribution and the first and second moments of the queue length and the waiting time in order to evaluate the analytical results outside the heavy-traffic regime.

The paper is organized as follows. In Section 2 the model is introduced and the heavy-traffic scaling is defined. In Section 3 and 4 the distribution of the scaled queue length vector at departure epochs and arbitrary epochs are presented, respectively. In Section 5 the distribution of the scaled waiting time of a given customer is presented. In Section 6 it is shown how the results presented in the previous sections can be used to optimize the scaled holding cost and the moments of the scaled waiting time of an arbitrary customer. In Section 7 we present our numerical results.

We note that an extended abstract version appeared in the proceedings of short papers in Performance Evaluation Review 2011, [2].

## 2 Model description

We consider a multi-class single-server queue with  $K$  classes of customers. Class- $k$  customers,  $k = 1, \dots, K$ , arrive according to independent Poisson processes with rate  $\lambda_k$ . We denote the overall arrival rate by  $\lambda = \sum_{k=1}^K \lambda_k$ . We assume that class- $k$  customers have i.i.d. generally distributed service requirements  $B_k$ , with distribution function  $B_k(t)$  and Laplace-Stieltjes transform  $B_k^*(s) = \int_0^\infty e^{-st} dB_k(t)$ , and we define  $B_k^{*'}(s) = \frac{dB_k^*(s)}{ds}$ . We assume  $\mathbb{E}[B_k^2] < \infty$ , for all  $k$ . The traffic intensity for class- $k$  customers is  $\rho_k = \lambda_k \mathbb{E}[B_k]$  and

$$\rho := \sum_{k=1}^K \rho_k = \sum_{k=1}^K \lambda_k \mathbb{E}[B_k] = \lambda \sum_{k=1}^K \alpha_k \mathbb{E}[B_k],$$

denotes the total traffic intensity, where  $\alpha_k = \lambda_k/\lambda$  denotes the probability that an arrival is of class  $k$ . Service is non-preemptive and upon service completion, the probability that the next customer to be served is of class  $k$  is given as in (1). Once a class is chosen to be served, an intra-class scheduling discipline determines which customer in this class will be served. We assume the intra-class discipline to be non-preemptive and not to make any use of information on the actual service requirements of the customers. We investigate the queue when it is near saturation, i.e.,  $\rho \uparrow 1$ , which is commonly referred to as the heavy-traffic regime. This regime can be obtained by letting

$$\lambda \uparrow \hat{\lambda} := \frac{1}{\sum_{i=1}^K \alpha_i \mathbb{E}[B_i]}, \quad (2)$$

since then  $\rho = \lambda \sum_{k=1}^K \alpha_k \mathbb{E}(B_k) \uparrow 1$ . When passing to the heavy-traffic regime we keep the fraction of class- $k$  arrivals,  $\alpha_k$ , fixed and we define  $\hat{\lambda}_k := \alpha_k \hat{\lambda}$ .

We denote the steady-state number of class- $k$  customers in the system at departure epochs by  $Q_k$  and at arbitrary epochs by  $N_k$ . We denote the waiting time of an arbitrary class- $k$  customer by  $W_k$ . In Section 3, Section 4 and Section 5 we will analyze  $Q_k, N_k$  and  $W_k$ , respectively, in the heavy-traffic setting.

### 3 Queue length at departure epochs

In this section we present the state-space collapse result for the steady-state queue length distribution at departure epochs. 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 next proposition states the main result of this section. The proof is provided in Section 3.2.

**Proposition 3.1.** *When scaled by  $1 - \rho$ , the queue length vector at departure epochs has a proper limiting distribution as  $(\lambda_1, \dots, \lambda_K) \rightarrow (\hat{\lambda}_1, \dots, \hat{\lambda}_K)$ , such that as  $\rho \uparrow 1$ ,*

$$(1 - \rho)(Q_1, \dots, Q_K) \xrightarrow{d} (\hat{Q}_1, \dots, \hat{Q}_K) \stackrel{d}{=} \left( \frac{\hat{\lambda}_1}{p_1}, \dots, \frac{\hat{\lambda}_K}{p_K} \right) \cdot Y,$$

where  $\xrightarrow{d}$  denotes convergence in distribution and  $Y$  is some one-dimensional random variable.

In Remark 2 of Section 4 we will show that, in fact,  $Y$  is exponentially distributed. To show this we require additional results presented in Section 4, and thus we refer the reader to Remark 2 for more details.

Before focusing on the heavy-traffic regime, we will introduce a system of equations that is satisfied by the probability generating function of the queue length distribution, as obtained by *Kim et al*, see [12]. Define

$$\pi(q_1, \dots, q_K) := \mathbb{P}((Q_1, \dots, Q_K) = (q_1, \dots, q_K)),$$

and let

$$p(\vec{z}) = \mathbb{E}[z_1^{Q_1} \dots z_K^{Q_K}] = \sum_{q_1=0}^{\infty} \dots \sum_{q_K=0}^{\infty} z_1^{q_1} \dots z_K^{q_K} \pi(\vec{q})$$

be its joint probability generating function. We define

$$r(\vec{z}) := \mathbb{E} \left[ \frac{z_1^{Q_1} \dots z_K^{Q_K}}{\sum_{j=1}^K Q_j p_j} \cdot \mathbf{1}_{(\sum_{j=1}^K Q_j > 0)} \right] = \sum_{(q_1, \dots, q_K) \neq (0, \dots, 0)} \frac{\pi(\vec{q})}{q_1 p_1 + \dots + q_K p_K} z_1^{q_1} \dots z_K^{q_K}.$$

In [12] the distribution of the queue length was studied assuming that the intra-class scheduling is uniform random. However, since the service discipline is non-preemptive and all class- $k$  customers in the queue are stochastically equivalent, the distribution of the queue length vector does not depend on the particular choice of the intra-class policy. Hence, for any arbitrary work-conserving intra-class policy we have the following result from [12].

**Theorem 3.2.** [12, Theorem 1 and 2] (a) *The probability generating function  $p(z_1, \dots, z_K)$  of the joint stationary queue lengths at departure epochs satisfies*

$$p(z_1, \dots, z_K) = 1 - \rho + \sum_{i=1}^K p_i z_i \frac{\partial}{\partial z_i} r(z_1, \dots, z_K). \quad (3)$$

(b) *The function  $r(z_1, \dots, z_K)$  satisfies*

$$\sum_{i=1}^K p_i \left( z_i - B_i^* \left( \lambda - \sum_{j=1}^K \lambda_j z_j \right) \right) \frac{\partial}{\partial z_i} r(z_1, \dots, z_K) = (\rho - 1) \left( 1 - \sum_{i=1}^K \frac{\lambda_i}{\lambda} B_i^* \left( \lambda - \sum_{j=1}^K \lambda_j z_j \right) \right). \quad (4)$$

In Section 3.1 we will show that Equations (3) and (4) simplify under the heavy-traffic scaling, which we will use in Section 3.2 to prove Proposition 3.1.

### 3.1 Heavy-traffic scaling

In this section we present three lemmas needed for the proof of Proposition 3.1. In the first lemma we show that the scaled queue length at departure epochs is tight. The proof may be found in Appendix A.

**Lemma 3.3.** *The random vector  $(1 - \rho)(Q_1, \dots, Q_K)$  is tight.*

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

$$\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}}) = \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)s_1 Q_1} \dots e^{-(1-\rho)s_K Q_K}] \quad (5)$$

exists, then there is a (possibly defective) random vector  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_K)$  such that  $(1-\rho)(Q_1, Q_2, \dots, Q_K)$  converges in distribution to  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_K)$ , and the distribution of  $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_K)$  is uniquely determined by the limit in (5) (cf. the Continuity theorem, see Feller (1971) [7]). For now, we assume that the limit exists; we come back to this assumption in Section 3.2.

Below we give two lemmas that describe properties of  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$ . In particular, in Lemma 3.5 we obtain a partial differential equation which will be the key element in the proof of Proposition 3.1. In order to describe the behaviour of the generating function, we define

$$\hat{r}(\vec{s}) = \mathbb{E} \left[ \frac{1 - e^{-s_1 \hat{Q}_1} \dots e^{-s_K \hat{Q}_K}}{\sum_{j=1}^K \hat{Q}_j p_j} \mathbf{1}_{(\sum_{j=1}^K \hat{Q}_j > 0)} \right]. \quad (6)$$

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.

**Lemma 3.4.** *If  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}})$  exists, then it satisfies*

$$\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}}) = \sum_{i=1}^K p_i \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \quad (7)$$

**Proof:** From (3) we have

$$\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\vec{s}}) = \lim_{\rho \uparrow 1} \sum_{i=1}^K p_i \frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}}. \quad (8)$$

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

$$\begin{aligned} \lim_{\rho \uparrow 1} \frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} &= \lim_{\rho \uparrow 1} \frac{\partial \mathbb{E} \left[ \frac{z_1^{Q_1} \dots z_K^{Q_K}}{\sum_{j=1}^K Q_j p_j} \cdot \mathbf{1}_{(\sum_{j=1}^K Q_j > 0)} \right]}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \\ &= \lim_{\rho \uparrow 1} \mathbb{E} \left[ \frac{Q_i}{\sum_{j=1}^K Q_j p_j} \cdot \frac{e^{-(1-\rho)s_1 Q_1} \dots e^{-(1-\rho)s_K Q_K}}{e^{-(1-\rho)s_i}} \cdot \mathbf{1}_{(\sum_{j=1}^K Q_j > 0)} \right] \\ &= \mathbb{E} \left[ \frac{\hat{Q}_i}{\sum_{j=1}^K \hat{Q}_j p_j} \cdot e^{-s_1 \hat{Q}_1} \dots e^{-s_K \hat{Q}_K} \cdot \mathbf{1}_{(\sum_{j=1}^K \hat{Q}_j > 0)} \right] \\ &= \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \end{aligned} \quad (9)$$

In the third step we used that  $\frac{Q_i}{\sum_{j=1}^K Q_j p_j} \cdot e^{-(1-\rho)s_1 Q_1} \dots e^{-(1-\rho)s_K Q_K} \cdot \mathbf{1}_{(\sum_{j=1}^K Q_j > 0)}$  is upper bounded by  $\frac{1}{\min_j(p_j)}$ , and, cf. the continuous mapping theorem, converges in distribution to  $\frac{\hat{Q}_i}{\sum_{j=1}^K \hat{Q}_j p_j} \cdot e^{-s_1 \hat{Q}_1} \dots e^{-s_K \hat{Q}_K} \cdot \mathbf{1}_{(\sum_{j=1}^K \hat{Q}_j > 0)}$ . From (8) and (9) we obtain (7).  $\square$

In the following lemma we show that the partial differential equation as given in (4) simplifies considerably in the heavy-traffic regime.

**Lemma 3.5.** *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:*

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

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

$$F_i(\vec{s}) = p_i(-s_i + \mathbb{E}[B_i] \sum_{j=1}^K \hat{\lambda}_j s_j) \quad i = 1, \dots, K, \quad (10)$$

with  $\hat{\lambda}_j = \alpha_j \hat{\lambda}$  and  $\hat{\lambda}$  as defined in (2).

**Proof:** Taking  $\vec{z}$  equal to  $e^{-(1-\rho)\vec{s}}$  in (4), dividing both sides by  $1 - \rho$  and taking the limit of  $\rho \uparrow 1$ , this gives

$$\begin{aligned} 0 &= \sum_{i=1}^K p_i \left( e^{-(1-\rho)s_i} - B_i^* (\lambda - \sum_{j=1}^K \lambda_j e^{-(1-\rho)s_j}) \right) \frac{\partial}{\partial z_i} r(z_1, \dots, z_K) \Big|_{\vec{z} = e^{-(1-\rho)\vec{s}}} \\ &= \sum_{i=1}^K p_i (-s_i + \mathbb{E}[B_i] \sum_{j=1}^K \hat{\lambda}_j s_j) \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \end{aligned}$$

In the second step we used (9) and the fact that  $\lim_{\rho \uparrow 1} \frac{x^{(1-\rho)} - 1}{1-\rho} = \ln(x)$ .  $\square$

## 3.2 Proof of Proposition 3.1

This subsection contains the proof of Proposition 3.1. The proof is based on the fact that the probability generating function satisfies the partial differential equation as described in Lemma 3.5. From this partial differential equation the following property for the function  $\hat{r}(\cdot)$  can be derived:

**Lemma 3.6.** *The function  $\hat{r}(s)$  is constant on the  $(K - 1)$ -dimensional hyperplane*

$$H_c := \{\vec{s} \geq \vec{0} : \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j = c\}, c > 0.$$

The proof of Lemma 3.6 may be found in Appendix B. We can now give the proof of Proposition 3.1.

**Proof of Proposition 3.1:** As  $\hat{r}(\vec{s})$  is constant on  $H_c$ , see Lemma 3.6,  $\hat{r}$  depends on  $\vec{s}$  only through  $\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j$ , so there exists a function  $\hat{r}^* : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j)$ . From Lemma 3.4 and  $\frac{\partial \hat{r}(s)}{\partial s_i} = \frac{\hat{\lambda}_i}{p_i} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}$  we obtain

$$\mathbb{E}[e^{-\sum_{i=1}^K s_i \hat{Q}_i}] = \lim_{\rho \rightarrow 1} p(e^{-(1-\rho)\vec{s}}) = \sum_{i=1}^K p_i \frac{\partial \hat{r}(s)}{\partial s_i} = \sum_{i=1}^K \hat{\lambda}_i \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j} = \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j},$$

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

$$\mathbb{E}[e^{-\sum_{i=1}^K s_i \hat{Q}_i}] = \mathbb{E} \left[ e^{-\frac{p_1}{\hat{\lambda}_1} \hat{Q}_1 \sum_{i=1}^K \frac{\hat{\lambda}_i}{p_i} s_i - s_2 \frac{\hat{\lambda}_2}{p_2} (\frac{p_2}{\hat{\lambda}_2} \hat{Q}_2 - \frac{p_1}{\hat{\lambda}_1} \hat{Q}_1) - \dots - s_K \frac{\hat{\lambda}_K}{p_K} (\frac{p_K}{\hat{\lambda}_K} \hat{Q}_K - \frac{p_1}{\hat{\lambda}_1} \hat{Q}_1)} \right].$$

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

$$(\hat{Q}_1, \dots, \hat{Q}_K) = \left( \frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, \dots, \frac{\hat{\lambda}_K}{p_K} \right) \frac{p_1}{\hat{\lambda}_1} \hat{Q}_1,$$

almost surely. Writing  $Y \stackrel{d}{=} \frac{p_1}{\lambda_1} \hat{Q}_1$  we get

$$(\hat{Q}_1, \dots, \hat{Q}_K) \stackrel{d}{=} \left( \frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, \dots, \frac{\hat{\lambda}_K}{p_K} \right) Y. \quad (11)$$

Recall that we assumed in Section 3.1 that  $\lim_{\rho \uparrow 1} p(e^{-(1-\rho)\bar{s}})$  exists, thereby showing that there is a unique limit (11). Since  $(1-\rho)(Q_1, \dots, Q_K)$  is tight, see Lemma 3.3, and since for any converging subsequence of  $\rho$  we obtain the same limit, we obtain that the limit itself exists (see corollary on page 59, Billingsley 1999). This concludes the proof.  $\square$

## 4 Queue length at arbitrary epochs

In this section we focus on the number of customers in the system at arbitrary epochs,  $(N_1, \dots, N_K)$ . The following result shows that in the limit the queue length vector is the product of an exponentially distributed random variable and a deterministic vector, referred to as a state-space collapse. The proof is presented in Section 4.2

**Remark 1.** *We note that a similar state-space collapse result was observed in [17] (Proposition 2.1.) for the discriminatory processor sharing model. In fact, the proof technique is similar to that of [17].*

**Proposition 4.1.** *When scaled by  $1-\rho$ , the queue length vector has a proper limiting distribution as  $(\lambda_1, \dots, \lambda_K) \rightarrow (\hat{\lambda}_1, \dots, \hat{\lambda}_K)$ , such that  $\rho \uparrow 1$ ,*

$$(1-\rho)(N_1, \dots, N_K) \xrightarrow{d} (\hat{N}_1, \dots, \hat{N}_K) \stackrel{d}{=} X \left( \frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, \dots, \frac{\hat{\lambda}_K}{p_K} \right), \quad (12)$$

where  $\xrightarrow{d}$  denotes convergence in distribution and  $X$  is an exponentially distributed random variable with mean  $1/\nu(\vec{p})$ , where

$$\nu(\vec{p}) := \frac{2 \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]}{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}. \quad (13)$$

Before focusing on the heavy-traffic regime, we will introduce a system of equations that is satisfied by the probability generating function of the queue length distribution, as obtained by *Kim et al*, see [12]. Let  $\psi(z_1, \dots, z_K)$  be the joint probability generating function of the stationary queue length, including the one in service if the system is nonempty, at arbitrary time, i.e,

$$\psi(z_1, \dots, z_K) := \mathbb{E}[z_1^{N_1} \dots z_K^{N_K}].$$

As mentioned in Section 3, the distribution of the queue length vector is independent of the particular choice of the intra-class scheduling discipline. We can therefore use the following result from [12].

**Theorem 4.2.** *[12, Theorem 3 and Theorem 4] The joint probability generating function  $\psi(z_1, \dots, z_K)$  of the stationary queue lengths at arbitrary time epochs is given by*

$$\psi(z_1, \dots, z_K) = 1 - \rho + \sum_{j=1}^K \lambda_j z_j \phi_j(z_1, \dots, z_K) \frac{1 - B_j^*(\lambda - \sum_{k=1}^K \lambda_k z_k)}{\lambda - \sum_{k=1}^K \lambda_k z_k}, \quad (14)$$

where  $\phi_j(z_1, \dots, z_K)$  (representing the joint probability generating function of the stationary queue lengths excluding the customer who has already started service, at service initiation epochs of class- $j$  customers) is given by

$$\phi_j(z_1, \dots, z_K) = 1 - \rho + \frac{\lambda p_j}{\lambda_j} \frac{\partial}{\partial z_j} r(z_1, \dots, z_K). \quad (15)$$

In Section 4.1 we will show that Equation (14) simplifies under the heavy-traffic scaling, and in Section 4.2 we will use this to characterize the distribution of the scaled queue length vector at arbitrary epochs.

## 4.1 Heavy-traffic scaling

In the next lemma we characterize Equation (14) in heavy traffic.

**Lemma 4.3.** *The limit of  $\psi(e^{-(1-\rho)\vec{s}})$  as  $\rho \uparrow 1$  exists and satisfies*

$$\lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) = \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}.$$

**Proof:** Since  $(1-\rho)(Q_1, \dots, Q_K)$  converges in distribution to  $(\hat{Q}_1, \dots, \hat{Q}_K)$ , we know that the limit of  $p(e^{-(1-\rho)\vec{s}})$  exists, and hence, by Equation (8), the limit of  $\frac{\partial r(\vec{z})}{\partial z_i} \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}}$  exists. It now follows directly from (15) that  $\lim_{\rho \uparrow 1} \phi_j(e^{-(1-\rho)\vec{s}})$  exists and it is given by  $\frac{\hat{\lambda}_j}{\hat{\lambda}_j} \frac{\partial \hat{r}(s)}{\partial s_j}$ .

As we have seen in Lemma 3.6,  $\hat{r}(\vec{s})$  is constant on  $H_c$ . Therefore, it depends on  $\vec{s}$  only through  $\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j$ , so there exists a function  $\hat{r}^* : R \rightarrow R$  such that  $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{j=1}^K \frac{\lambda_j}{p_j} s_j)$  and  $\frac{\partial \hat{r}(s)}{\partial s_i} = \frac{\hat{\lambda}_i}{p_i} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}$ .

This, together with (14) gives that

$$\begin{aligned} \lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) &= \lim_{\rho \uparrow 1} \left( 1 - \rho + \sum_{j=1}^K \lambda_j e^{-(1-\rho)s_j} \phi_j(e^{-(1-\rho)\vec{s}}) \frac{1 - B_j^*(\lambda - \sum_{k=1}^K \lambda_k e^{-(1-\rho)s_k})}{\lambda - \sum_{k=1}^K \lambda_k e^{-(1-\rho)s_k}} \right) \\ &= \sum_{j=1}^K \hat{\lambda}_j \frac{\hat{\lambda}_j p_j}{\hat{\lambda}_j} \frac{\partial \hat{r}(\vec{s})}{\partial s_j} (-B_j^{*'}(0)) = \sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j] \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j} \\ &= \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j} \sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j] = \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}, \end{aligned}$$

where in the first step we used l'Hopital's rule and  $B_j^{*'}(0) := \frac{dB_j^*(s)}{ds} \Big|_{s=0} = -\mathbb{E}[B_j]$  for all  $j$ .  $\square$

In particular, Lemma 4.3 implies that there exists a vector  $(\hat{N}_1, \dots, \hat{N}_K)$  such that the scaled queue length vector at arbitrary epochs converges in distribution to it.

## 4.2 Proof of Proposition 4.1

This subsection contains the proof of Proposition 4.1. It consists of two steps. Firstly, we show that the queue length vector is the product of a random variable and a deterministic vector, and secondly, we determine the distribution of the random variable  $X$ , concluding that it is exponentially distributed with mean as given in (13).

**Proof of Proposition 4.1:** Since  $\lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}})$  exists, see Lemma 4.3 we know there exists a random vector  $(\hat{N}_1, \dots, \hat{N}_K)$  such that

$$\mathbb{E}[e^{-\sum_{j=1}^K s_j \hat{N}_j}] = \lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) = \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \Big|_{v=\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}.$$

Using the same steps as in the proof of Proposition 3.1 we obtain that

$$(\hat{N}_1, \dots, \hat{N}_K) \stackrel{d}{=} \left( \frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, \dots, \frac{\hat{\lambda}_K}{p_K} \right) X, \quad (16)$$

with  $X$  distributed as  $\frac{p_1}{\hat{\lambda}_1} \hat{N}_1$ .

In order to determine the distribution of  $X$ , we consider the total workload in the queue at arbitrary epochs, denoted by  $V^{arb}$ . We first note that the total workload at the system is independent of the work-conserving scheduling discipline being used. In [13], Kingman considered a FCFS queue and showed that the scaled total workload in a  $M/G/1$  queue has a proper distribution as  $\rho \uparrow 1$ :

$$(1-\rho)V^{arb} \xrightarrow{d} \hat{V}^{arb},$$



where  $\hat{V}^{arb}$  is exponentially distributed with mean

$$\mathbb{E}[\hat{V}^{arb}] = \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2}. \quad (17)$$

Under the discipline DROS, the total workload at arbitrary epochs can equivalently be represented as

$$V^{arb} = \sum_{j=1}^K \sum_{h=1}^{N_j-1} B_{j,h} + \sum_{j=1}^K \tilde{B}_j,$$

with  $B_{j,h}$  the service requirement of the  $h$ -th class- $j$  customer and  $\tilde{B}_j$  the remaining service requirement of the first class- $j$  customer in line. On one hand, note that the service requirements of all class- $j$  customers are i.i.d., more precisely,  $B_{j,h} \stackrel{d}{=} B_j$  for all  $h$ . On the other hand,  $\tilde{B}_j$  is distributed as  $B_j$  if the  $N_j$ -th class- $j$  customer is not being served, and otherwise is given by the forward-recurrence time of  $B_j$ . Hence, for the scaled workload at arbitrary epochs we can write

$$\begin{aligned} \mathbb{E}[e^{-s\hat{V}^{arb}}] &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)sV^{arb}}] \\ &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)s(\sum_{j=1}^K \sum_{h=1}^{N_j-1} B_{j,h} + \sum_{j=1}^K \tilde{B}_j)}] \\ &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-s \sum_{j=1}^K (1-\rho)(N_j-1) \frac{\sum_{h=1}^{N_j-1} B_{j,h}}{(N_j-1)}} e^{-(1-\rho)s \sum_{j=1}^K \tilde{B}_j}] \\ &= \mathbb{E}[e^{-s \sum_{j=1}^K \mathbb{E}[B_j] \hat{N}_j}], \end{aligned} \quad (18)$$

where in the last step we used that  $e^{-s \sum_{j=1}^K (1-\rho)(N_j-1) \frac{\sum_{h=1}^{N_j-1} B_{j,h}}{(N_j-1)}}$  is bounded by 1 and converges in distribution to  $e^{-s \sum_{j=1}^K \mathbb{E}[B_j] \hat{N}_j}$ . From (18) we obtain that

$$\hat{V}^{arb} \stackrel{d}{=} \sum_{j=1}^K \mathbb{E}[B_j] \hat{N}_j, \quad (19)$$

and together with (16) this gives

$$\hat{V}^{arb} \stackrel{d}{=} X \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]. \quad (20)$$

Since  $\hat{V}^{arb}$  is exponentially distributed, the same is true for  $X$ . Hence, taking expectations in (20) and applying (17) we obtain

$$\mathbb{E}[X] = \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2 \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]},$$

which concludes the proof of Proposition 4.1.  $\square$

**Remark 2.** *In this remark we show that the two random variables that characterize the heavy traffic at departure and arbitrary epochs,  $Y$  and  $X$ , respectively, are equal in distribution.*

*Let us consider the arbitrary arrival of a class- $j$  customer. By the PASTA property, the number of class- $k$  customers in the system at this time is equal to  $N_k$ . The number of customers in the system after the first departure epoch is distributed as  $(Q_1, \dots, Q_K)$ . The number of customers that arrive in the time it takes for the customer in service to depart is of the order  $\rho$ , since it is distributed as the number of arrivals in a residual service requirement. It then follows that*

$$Q_j \stackrel{d}{=} N_j + \mathbf{O}(\rho).$$

*Multiplying the above equation by  $(1-\rho)$  and taking the limit  $\rho \uparrow 1$  we get that  $\hat{Q}_j \stackrel{d}{=} \hat{N}_j$  and hence  $X \stackrel{d}{=} Y$ .*

## 5 Waiting time

In this section we investigate the waiting time in the heavy-traffic setting. We focus on the random intra-class scheduling discipline, that is, we consider the specific model DROS. In order to study the scaled waiting time, we will need to assume throughout this section that  $(1 - \rho)Q_k^*$  is uniform integrable. As we mention in Section 7.2, numerics show arguments to believe that this is indeed satisfied.

**Assumption 1.** *For a random intra-class scheduling discipline, the random variable  $(1 - \rho)Q_k^*$  is uniform integrable for all  $k$ .*

We can now state our result, which shows that in the limit, the waiting time of a tagged class- $k$  customer, denoted by  $W_k$ , is the product of two exponentially distributed independent random variables.

**Proposition 5.1.** *Let Assumption 1 be satisfied and consider the random intra-class scheduling discipline (i.e., DROS). Then, as  $\rho \uparrow 1$ ,*

$$(1 - \rho)(W_k, Q_1, \dots, Q_K) \xrightarrow{d} X(Z_k, \frac{\hat{\lambda}_1}{p_1}, \dots, \frac{\hat{\lambda}_K}{p_K}),$$

where  $\xrightarrow{d}$  denotes convergence in distribution and  $X$  and  $Z_k$  are exponentially distributed independent random variables with  $\mathbb{E}[X] = 1/\nu(\vec{p})$  and  $\mathbb{E}[Z_k] = 1/p_k$ .

**Remark 3.** *Proposition 5.1 is a generalization of Kingman's result, see [15], where the asymptotic waiting time distribution is obtained for the single-class DROS queue (i.e., ROS).*

The proof of Proposition 5.1 is based on a result obtained in [12]. Let  $Q_k^*$  denote the number of class- $k$  customers in the system (excluding the tagged customer) immediately after service initiation of the tagged customer in case the tagged customer arrives while the server is busy, i.e.,  $W_k > 0$ . We now define the following joint transform:

$$T_k(u, z_1, \dots, z_K) := \mathbb{E}[e^{-uW_k} z_1^{Q_1^*} \dots z_K^{Q_K^*} \mathbf{1}_{\{W_k > 0\}}]. \quad (21)$$

Note that the joint transform of the waiting time  $W_k$  of an arbitrary class- $k$  customer and  $(Q_1^*, \dots, Q_K^*)$  is given by

$$\mathbb{E}[e^{-uW_k} z_1^{Q_1^*} \dots z_K^{Q_K^*}] = \mathbb{E}[e^{-u \cdot 0} z_1^0 \dots z_K^0 \mathbf{1}_{\{W_k = 0\}}] + e^{-u \cdot W_k} z_1^{Q_1^*} \dots z_K^{Q_K^*} \mathbf{1}_{\{W_k > 0\}}] = 1 - \rho + T_k(u, \vec{z}), \quad (22)$$

since  $1 - \rho$  is the probability that the tagged class- $k$  customer arrives in an idle period. For the random intra-class scheduling discipline we have from [12] the following result for the transform  $T_k(u, \vec{z})$ .

**Theorem 5.2.** *[12, Theorem 8] For the random intra-class scheduling discipline, the joint transform  $T_k(u, z_1, \dots, z_K)$  satisfies*

$$\sum_{j=1}^K \frac{p_j}{p_k} \left( \frac{\partial}{\partial z_j} T_k(u, z_1, \dots, z_K) \right) (z_j - B_j^*(u + \lambda - \sum_{r=1}^K \lambda_r z_r)) + T_k(u, z_1, \dots, z_K) = W_k^1(u, z_1, \dots, z_K), \quad (23)$$

where  $W_k^1(u, z_1, \dots, z_K)$  satisfies

$$\begin{aligned} & W_k^1(u, z_1, \dots, z_K) \\ &= \sum_{j=1}^K \left( (1 - \rho) \lambda_j + \lambda p_j \frac{\partial}{\partial z_j} r(z_1, \dots, z_K) \right) \frac{B_j^*(\lambda - \sum_{i=1}^K \lambda_i z_i) - B_j^*(u + \lambda - \sum_{i=1}^K \lambda_i z_i)}{u}. \end{aligned} \quad (24)$$

In order to prove Proposition 5.1, we need the following technical lemma that characterizes the value that the function  $W_k^1(u, z_1, \dots, z_K)$ , as defined in (24), takes in heavy traffic.

**Lemma 5.3.** *We consider the random intra-class scheduling discipline (i.e., DROS). Then, as  $\rho \uparrow 1$ , the limit  $W_k^1((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K})$  exists and satisfies*

$$\lim_{\rho \uparrow 1} W_k^1((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) = \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}},$$

with  $1/\nu(\vec{p})$  as given in (13).

The result of Lemma 5.3 implies that in heavy traffic the function  $W_k^1((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K})$  depends on  $s$  only through a linear combination of its components. The proof of Lemma 5.3 may be found in Appendix C. In the following lemma we show that the scaled waiting time of a class- $k$  customer has a proper limit.

**Lemma 5.4.** *Let Assumption 1 be satisfied and consider the random intra-class scheduling discipline (i.e., DROS). Then, there exists a random vector  $(\hat{W}_k, \hat{Q}_1, \dots, \hat{Q}_K)$  such that  $(1-\rho)(W_k, Q_1, \dots, Q_K)$  converges in distribution to  $(\hat{W}_k, \hat{Q}_1, \dots, \hat{Q}_K)$  as  $\rho \uparrow 1$ .*

**Proof:** By definition, the following two equalities are satisfied:

$$T_k(u, 1, \dots, 1) = \mathbb{E}[e^{-uW_k} \mathbf{1}_{\{W_k > 0\}}], \quad (25)$$

and

$$\left. \frac{\partial}{\partial z_j} T_k(u, z_1, \dots, z_K) \right|_{\vec{z}=1} = \mathbb{E}[Q_j^* e^{-uW_k} \mathbf{1}_{\{W_k > 0\}}]. \quad (26)$$

Now, considering Equation (23) with  $\vec{z} = 1$  in heavy traffic we get:

$$\begin{aligned} & \lim_{\rho \uparrow 1} W_k^1((1-\rho)u, 1, \dots, 1) \quad (27) \\ &= \lim_{\rho \uparrow 1} \sum_{j=1}^K \frac{p_j}{p_k} (1-\rho) \left. \frac{\partial}{\partial z_j} T_k((1-\rho)u, z_1, \dots, z_K) \right|_{\vec{z}=1} \frac{1 - B_j^*((1-\rho)u)}{(1-\rho)} \\ &+ \lim_{\rho \uparrow 1} T_k((1-\rho)u, 1, \dots, 1) \\ &= \lim_{\rho \uparrow 1} \sum_{j=1}^K \frac{p_j}{p_k} u \mathbb{E}[B_j] \mathbb{E}[(1-\rho)Q_j^* e^{-(1-\rho)uW_k} \mathbf{1}_{\{W_k > 0\}}] + \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)uW_k} \mathbf{1}_{\{W_k > 0\}}] \\ &= \lim_{\rho \uparrow 1} \mathbb{E} \left[ \left( \sum_{j=1}^K \frac{p_j}{p_k} u \mathbb{E}[B_j] (1-\rho)Q_j^* + 1 \right) e^{-(1-\rho)uW_k} \mathbf{1}_{\{W_k > 0\}} \right] \\ &= \mathbb{E} \left[ \lim_{\rho \uparrow 1} \left( \sum_{j=1}^K \frac{p_j}{p_k} u \mathbb{E}[B_j] (1-\rho)Q_j^* + 1 \right) e^{-(1-\rho)uW_k} \mathbf{1}_{\{W_k > 0\}} \right] \quad (28) \end{aligned}$$

where in the second step we used (25) and (26) and in the fourth step we used the hypothesis that  $(1-\rho)Q_j^*$  is uniformly integrable (Assumption 1), [3, Theorem 3.5]. Note that  $W_k^1((1-\rho)u, e^{-(1-\rho)\vec{s}})$ , which is defined in Equation (24), has a proper limit when  $\rho \uparrow 1$ , see Lemma 5.3. Since (27) converges, the same must hold for (28). Besides,  $\sum_{j=1}^K \frac{p_j}{p_k} u \mathbb{E}[B_j] (1-\rho)Q_j^*$  converges in distribution to  $\sum_{j=1}^K \frac{p_j}{p_k} u \mathbb{E}[B_j] \hat{Q}_j$  and therefore, we conclude that the waiting time of an arbitrary class- $k$  customer in heavy traffic,  $\hat{W}_k$ , converges in distribution.  $\square$

From Lemma 5.3, we note that (28) should in fact be independent of  $u$ . It can be checked that in case  $(1-\rho)(W_k, Q_1, \dots, Q_K)$  is distributed as  $X(Z_k, \frac{\hat{\lambda}_1}{p_1}, \dots, \frac{\hat{\lambda}_K}{p_K})$ , as we want to show, see Proposition 5.1, this is indeed satisfied.

We can now prove Proposition 5.1 which consists in finding  $T_k(\cdot)$  by solving Equation (23) after the heavy-traffic scaling.

**Proof of Proposition 5.1:** We know by Lemma 5.4 that there is a random variable  $\hat{W}_k$  such that  $(1 - \rho)W_k$  converges in distribution to  $\hat{W}_k$ . Hence, we can define the function  $\hat{T}_k(u, \vec{s})$  as follows:

$$\begin{aligned}\hat{T}_k(u, \vec{s}) &:= \mathbb{E}[e^{-u\hat{W}_k} e^{-\sum_{i=1}^K s_i \hat{Q}_i^*}] \\ &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)uW_k} e^{-(1-\rho)s_1 Q_1^*} \dots e^{-(1-\rho)s_K Q_K^*}] \\ &= \lim_{\rho \uparrow 1} T_k((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}).\end{aligned}$$

We will evaluate Equation (23) in the point  $(u, \vec{z}) = (u(1-\rho), e^{-(1-\rho)\vec{s}})$  as  $\rho \uparrow 1$ . We have

$$\begin{aligned}& \lim_{\rho \uparrow 1} (1-\rho) \frac{\partial}{\partial z_j} T_k(u, \vec{z}) \Big|_{u=(1-\rho)u, \vec{z}=e^{-(1-\rho)\vec{s}}} \\ &= \lim_{\rho \uparrow 1} (1-\rho) \mathbb{E} \left( \frac{Q_j^* e^{-(1-\rho)uW_k} e^{-(1-\rho)s_1 Q_1^*} \dots e^{-(1-\rho)s_j (Q_j^* - 1)} \dots e^{-(1-\rho)s_K Q_K^*}}{e^{-(1-\rho)s_j}} \right) \\ &= \mathbb{E}[\hat{Q}_j^* e^{-u\hat{W}_k} e^{-s\hat{Q}_1^*} \dots e^{-s\hat{Q}_K^*}] = -\frac{\partial}{\partial s_j} \hat{T}_k(u, \vec{s}),\end{aligned}\tag{29}$$

where in the second step we used Assumption 1 and [3, Theorem 3.5].

Moreover, realize that  $\lim_{\rho \uparrow 1} T_k((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K})$  depends on  $\vec{s} = (s_1, \dots, s_K)$  only through  $y = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j$ . Thus, we will write  $\hat{T}_k(u, \vec{s}) = \hat{T}_k(u, y)$  and by the chain rule:

$$\frac{\partial}{\partial s_j} \hat{T}_k(u, \vec{s}) = \frac{\partial}{\partial s_j} \hat{T}_k(u, y) = \frac{\partial}{\partial y} \hat{T}_k(u, y) \frac{\partial y}{\partial s_j} = \frac{\hat{\lambda}_j}{p_j} \frac{\partial}{\partial y} \hat{T}_k(u, y) \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}}.\tag{30}$$

Then, taking the heavy-traffic limit in Equation (23) and using Equations (29), (30) and Lemma 5.3 we arrive to the following ordinary differential equation (ODE):

$$-\sum_{j=1}^K \frac{p_j}{p_k} \frac{\hat{\lambda}_j}{p_j} \frac{\partial \hat{T}_k(u, y)}{\partial y} \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} (-s_j + \mathbb{E}[B_j])(u + \sum_{i=1}^K \hat{\lambda}_i s_i) + \hat{T}_k(u, y) = \frac{\nu(\vec{p})}{\nu(\vec{p}) + y},$$

that is

$$-\frac{u}{p_k} \frac{\partial \hat{T}_k(u, y)}{\partial y} \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} + \hat{T}_k(u, y) = \frac{\nu(\vec{p})}{\nu(\vec{p}) + y}.\tag{31}$$

The solution of the ODE (31), see Appendix D for the details, is

$$\hat{T}_k(u, y) = \frac{p_k}{u} e^{\frac{p_k}{u} y} \int_y^\infty e^{-\frac{p_k}{u} x} \frac{\nu(\vec{p})}{\nu(\vec{p}) + x} dx.\tag{32}$$

Let  $Z_k$  and  $X$  be two exponentially distributed random variables with  $\mathbb{E}[Z_k] = 1/\eta$  and  $\mathbb{E}[X] = 1/\nu(\vec{p})$ . Then, the Laplace Transform of  $(Z_k \cdot X, \frac{\lambda_1}{p_1} X, \dots, \frac{\lambda_K}{p_K} X)$  is given by:

$$\begin{aligned}& \mathbb{E}[e^{-uZ_k \cdot X - s_1 \frac{\lambda_1}{p_1} X - \dots - s_K \frac{\lambda_K}{p_K} X}] = \mathbb{E}[e^{-uZ_k \cdot X - \sum_{j=1}^K s_j \frac{\lambda_j}{p_j} X}] = \mathbb{E}[e^{-uZ_k \cdot X - yX}] \\ &= \mathbb{E}[\mathbb{E}[e^{-uZ_k X - yX} | z_k = Z_k]]] = \mathbb{E}\left[\frac{\nu(\vec{p})}{\nu(\vec{p}) + uZ_k + y}\right] = \int_0^\infty \eta e^{-\eta z_k} \frac{\nu(\vec{p})}{\nu(\vec{p}) + uz_k + y} dz_k \\ &= \frac{1}{u} \int_{\nu(\vec{p})+y}^\infty \eta e^{-\eta \frac{x-\nu(\vec{p})-y}{u}} \frac{\nu(\vec{p})}{x} dx = \frac{\nu(\vec{p})\eta}{u} e^{\eta \frac{\nu(\vec{p})+y}{u}} \int_{\nu(\vec{p})+y}^\infty e^{-\eta \frac{x}{u}} \frac{1}{x} dx \\ &= \frac{\nu(\vec{p})\eta}{u} e^{\eta \frac{\nu(\vec{p})+y}{u}} \int_{\eta \frac{\nu(\vec{p})+y}{u}}^\infty \frac{e^{-l}}{l} dl.\end{aligned}\tag{33}$$

Making the change of variable  $l = p_k \frac{\nu(\vec{p})+x}{u}$  in Equation (32) we obtain

$$\hat{T}_k(u, y) = \nu(\vec{p}) \frac{p_k}{u} e^{p_k \frac{y}{u}} \int_y^\infty e^{-p_k \frac{x}{u}} \frac{1}{\nu(\vec{p}) + x} dx = p_k \frac{\nu(\vec{p})}{u} e^{p_k \frac{\nu(\vec{p})+y}{u}} \int_{p_k \frac{\nu(\vec{p})+y}{u}}^\infty e^{-l} \frac{1}{l} dl.$$

Hence, it coincides with the Laplace Transform of  $((1-\rho)W_k, (1-\rho)Q_1^*, \dots, (1-\rho)Q_K^*)$  obtained in (33) with  $p_k = \eta$ . Since the Laplace transform of a probability distribution is unique, (uniqueness theorem, [7]), we conclude that  $(1-\rho)(W_k, Q_1^*, \dots, Q_K^*)$  converges in distribution to  $(Z_k \cdot X, \frac{\lambda_1}{p_1} X, \dots, \frac{\lambda_K}{p_K} X)$ , where, as we have previously mentioned,  $Z_k$  and  $X$  are exponentially distributed independent random variables with  $\mathbb{E}(Z_k) = 1/p_k$  and  $\mathbb{E}(X) = 1/\nu(\vec{p})$ .

Moreover, from (22) we have

$$\begin{aligned} & \mathbb{E}[e^{-u(1-\rho)W_k} e^{-s_1(1-\rho)Q_1} \dots e^{-s_K(1-\rho)Q_K}] \\ &= \rho \mathbb{E}[e^{-u(1-\rho)W_k} e^{-s_1(1-\rho)Q_1^*} \dots e^{-s_K(1-\rho)Q_K^*}] + (1-\rho), \end{aligned}$$

where we used that  $\rho = \mathbb{P}[W_k > 0]$ . Hence

$$\begin{aligned} & \mathbb{E}[e^{-u\hat{W}_k} e^{-s_1\hat{Q}_1} \dots e^{-s_K\hat{Q}_K}] \\ &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-u(1-\rho)W_k} e^{-s_1(1-\rho)Q_1^*} \dots e^{-s_K(1-\rho)Q_K^*}] \\ &= \hat{T}(u, y) = \mathbb{E}[e^{-uXZ_k} e^{-s_1\frac{\lambda_1}{p_1}X} \dots e^{-s_K\frac{\lambda_K}{p_K}X}]. \end{aligned}$$

□

## 6 Optimal selection of the weights

In this section we show how the results of Proposition 4.1 and Proposition 5.1 can be used in order to optimize the performance. In particular, in Section 6.1 we focus on the holding cost and in Section 6.2 we find the weights that minimizes the moments of the waiting time of an arbitrary customer.

### 6.1 Holding Cost

With each class of customers we associate a cost  $c_j \geq 0, j = 1, \dots, K$ . As performance measure we take the holding cost  $\sum_{j=1}^K c_j N_j$ . In this section we will write  $N_j(\vec{p}), \hat{N}_j(\vec{p})$  instead of  $N_j, \hat{N}_j$  to emphasize the dependence on the weights  $\vec{p} := (p_1, \dots, p_K)$ . From Proposition 4.1 we obtain that the scaled holding cost,  $(1-\rho) \sum_{j=1}^K c_j N_j(\vec{p})$ , converges in distribution to an exponentially distributed random variable with mean

$$\sum_{j=1}^K c_j \mathbb{E}[\hat{N}_j(\vec{p})] = \frac{\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} c_j}{2 \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]} \sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2], \quad (34)$$

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 with a large value of  $\frac{c_j}{\mathbb{E}[B_j]}$ . This can be seen as an extension of the  $c\mu$ -rule for the heavy-traffic setting [8].

**Proposition 6.1.** *Consider two policies with weights  $(p_1, \dots, p_K)$  and  $(q_1, \dots, q_K)$ , respectively. Let  $c_j \geq 0, j = 1, \dots, K$ . Without loss of generality we assume that the classes are ordered such that  $\frac{c_1}{\mathbb{E}[B_1]} \geq \frac{c_2}{\mathbb{E}[B_2]} \geq \dots \geq \frac{c_K}{\mathbb{E}[B_K]}$ . If  $\frac{p_j}{p_{j+1}} \leq \frac{q_j}{q_{j+1}}$ , for all  $j = 1, \dots, K-1$ , then*

$$\lim_{\rho \uparrow 1} (1-\rho) \sum_{j=1}^K c_j N_j(\vec{p}) \geq_{st} \lim_{\rho \uparrow 1} (1-\rho) \sum_{j=1}^K c_j N_j(\vec{q}),$$

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  $(1-\rho) \sum_{j=1}^K c_j N_j(\vec{p})$  converges in distribution to an exponentially distributed random variable with mean as stated in (34). Since exponentially distributed random variables are stochastically ordered according to their means, it only remains to check that

$$\frac{\sum_{j=1}^K \frac{c_j \hat{\lambda}_j}{p_j}}{\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]} \geq \frac{\sum_{j=1}^K \frac{c_j \hat{\lambda}_j}{q_j}}{\sum_{j=1}^K \frac{\hat{\lambda}_j}{q_j} \mathbb{E}[B_j]}.$$

This holds since

$$\begin{aligned} & \left( \sum_{j=1}^K \frac{c_j \hat{\lambda}_j}{p_j} \right) \left( \sum_{j=1}^K \frac{\hat{\lambda}_j}{q_j} \mathbb{E}[B_j] \right) = \sum_{j,i:j \neq i} \hat{\lambda}_j \hat{\lambda}_i \left( \frac{1}{p_j q_i} c_j \mathbb{E}[B_i] + \frac{1}{p_i q_j} c_i \mathbb{E}[B_j] \right) + \sum_{j=1}^K \hat{\lambda}_j^2 \frac{1}{p_j q_j} c_j \mathbb{E}[B_j] \\ & \geq \sum_{j,i:j \neq i} \hat{\lambda}_j \hat{\lambda}_i \left( \frac{1}{p_i q_j} c_j \mathbb{E}[B_i] + \frac{1}{p_j q_i} c_i \mathbb{E}[B_j] \right) + \sum_{j=1}^K \hat{\lambda}_j^2 \frac{1}{p_j q_j} c_j \mathbb{E}[B_j] = \left( \sum_{j=1}^K \frac{c_j \hat{\lambda}_j}{q_j} \right) \left( \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j] \right). \end{aligned}$$

Here we used that  $c_i \mathbb{E}[B_j] \left( \frac{1}{p_i q_j} - \frac{1}{p_j q_i} \right) \geq c_j \mathbb{E}[B_i] \left( \frac{1}{p_i q_j} - \frac{1}{p_j q_i} \right)$ , which follows from the fact that  $\frac{p_i}{p_j} \leq \frac{q_i}{q_j}$  and  $\frac{c_i}{\mathbb{E}[B_i]} \geq \frac{c_j}{\mathbb{E}[B_j]}$ , for  $i \leq j$ .  $\square$

## 6.2 Moments of the waiting time

In this section we will give the optimal values for the weights that minimize the  $m$ -th moment of the scaled waiting time. From Proposition 5.1 we know that

$$\lim_{\rho \uparrow 1} (1 - \rho) W_k \stackrel{d}{=} X \cdot Z_k, \quad (35)$$

where  $X$  and  $Z_k$  are exponentially distributed independent random variables with  $\mathbb{E}(Z_k) = 1/p_k$  and  $\mathbb{E}(X) = 1/\nu(\vec{p})$ . Now taking the expression in (35) and using that  $X$  and  $Z_k$  are independent random variables we observe that the  $m$ -th moment of  $\hat{W}_k$  is given by

$$\mathbb{E}[\hat{W}_k^m] = \mathbb{E}[X^m Z_k^m] = \mathbb{E}[X^m] \mathbb{E}[Z_k^m] = \frac{m!}{\nu(\vec{p})^m} \frac{m!}{p_k^m} = (m!)^2 \frac{1}{p_k^m} \left( \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2 \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]} \right)^m.$$

Hence the  $m$ -th moment of the waiting time for an arbitrary customer is given by

$$\mathbb{E}[\hat{W}^m] = \sum_{k=1}^K \frac{\hat{\lambda}_k}{\hat{\lambda}} \mathbb{E}[\hat{W}_k^m] = (m!)^2 \left( \sum_{k=1}^K \frac{\hat{\lambda}_k}{\hat{\lambda} p_k^m} \right) \left( \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2 \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]} \right)^m. \quad (36)$$

In what follows we will write  $\hat{W}(\vec{p})$  instead of  $\hat{W}$  to emphasize the dependence on the weights  $\vec{p}$ . Note that  $\mathbb{E}[\hat{W}(\vec{p})] = \frac{1}{\hat{\lambda}} \sum_{j=1}^K \mathbb{E}[\hat{N}_j(\vec{p})]$ . Hence, by applying Little's law to the result obtained in Proposition 6.1, we obtain the following corollary, which means that the mean waiting time decreases as more preference is given to customers with a small value of  $\mathbb{E}[B_i]$ ,  $i = 1, \dots, K$ .

**Corollary 6.2.** *Without loss of generality we assume that the classes are ordered such that  $\mathbb{E}[B_1] \leq \dots \leq \mathbb{E}[B_K]$ . If  $\frac{p_i}{p_{j+1}} \leq \frac{q_i}{q_{j+1}}$ , for all  $j = 1, \dots, K-1$ , then  $\mathbb{E}[\hat{W}(\vec{p})] \geq \mathbb{E}[\hat{W}(\vec{q})]$ .*

**Remark 4.** *The monotonicity result for the waiting time holds in the heavy-traffic setting. In the case of two classes,  $K=2$ , Corollary 6.2 is true for any stable system, i.e., for any value of  $\rho$ , not necessarily close to one. This can be seen as follows. The expression for the mean waiting time for  $K=2$  is the following:*

$$\begin{aligned} \mathbb{E}[W(\vec{p})] &= \sum_{i=1}^2 \frac{\lambda_i}{\lambda} \mathbb{E}[W_i] \\ &= \frac{\lambda_1 \mathbb{E}[B_1^2] + \lambda_2 \mathbb{E}[B_2^2]}{2\lambda} \frac{\lambda_1(1 - \rho p_1) + \lambda_2(1 - \rho p_2)}{(1 - \rho_1 - p_2 \rho_2)(1 - \rho_2 - p_1 \rho_1) - p_1 p_2 \rho_1 \rho_2}, \end{aligned} \quad (37)$$

where the expression of  $\mathbb{E}[W_i]$ ,  $i = 1, 2$ , was obtained in [12], Equation (38). Without loss of generality we assume that  $p_1 + p_2 = 1$ . Then, taking the derivative of (37) with respect to  $p_1$  we obtain the monotonicity result as stated in Corollary 6.2.

Moreover, we have written a code to calculate the mean waiting time as given in [12] for any value of  $K$ , i.e., for any number of classes of customers. We choose the weights such that  $\frac{p_i}{p_{j+1}} = \frac{1}{r}$ ,  $\forall j$ . In the figures we chose exponentially distributed service requirements, however, the monotonicity observed

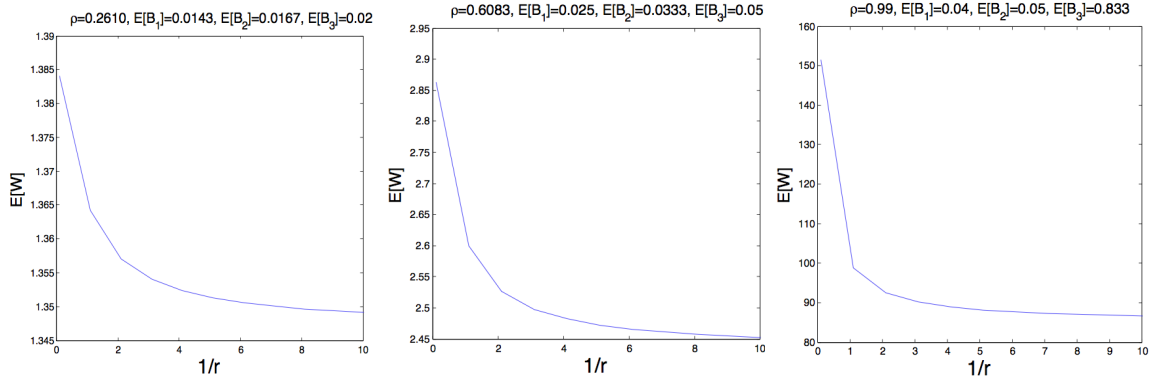


Figure 1: The mean waiting time for three classes of customers in the system,  $K=3$ , under DROS for the loads  $\rho = 0.2610$ ,  $\rho = 0.6083$  and  $\rho = 0.99$ , respectively. The horizontal axis corresponds to  $\frac{1}{r} = \frac{p_j}{p_{j+1}}$ ,  $j = 1, \dots, K-1$ .

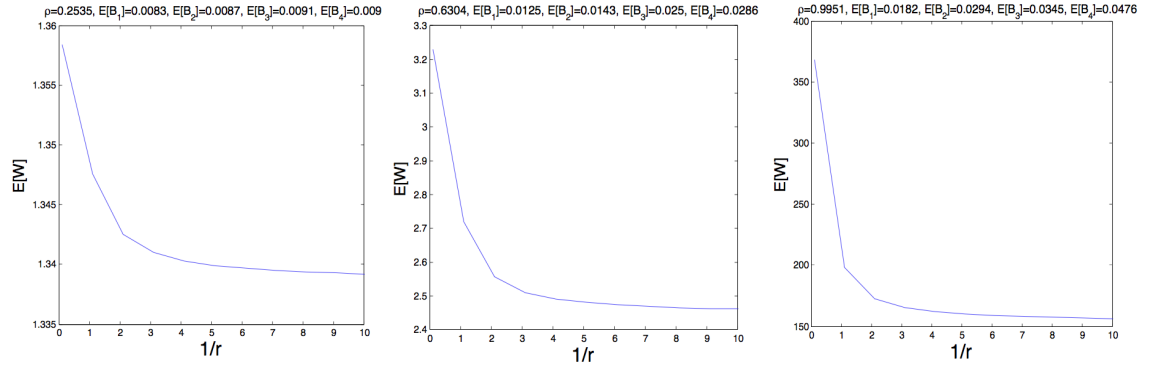


Figure 2: The mean waiting time for four classes of customers in the system,  $K=4$ , under DROS for the loads  $\rho = 0.2535$ ,  $\rho = 0.6304$  and  $\rho = 0.9951$ , respectively. The horizontal axis corresponds to  $\frac{1}{r} = \frac{p_j}{p_{j+1}}$ ,  $j = 1, \dots, K-1$ .

holds for any service requirement distribution (with the same first moment). The results obtained are shown in Figure 1 and Figure 2, for  $K=3$  and  $K=4$ , respectively, for different values of the load. It can be seen that under the assumptions of Corollary 6.2, as more priority is given to customers with small mean service requirement (i.e., as  $\frac{1}{r}$  becomes large), the mean waiting time decreases for any value of the load.

In Corollary 6.2 we considered the first moment of the scaled waiting time. In Proposition 6.3 we will investigate the  $m$ -th moment of the scaled waiting time and find the optimal value for the weights, which is non-trivial.

**Proposition 6.3.** *The  $m$ -th moment of the scaled waiting time,  $\mathbb{E}[\hat{W}^m(\vec{p})]$ , is minimized in  $\vec{p}^* = (p_1^*, \dots, p_K^*)$ , with*

$$p_k^* := \left( \frac{1/\mathbb{E}[B_k]}{\sum_{j=1}^K \frac{1}{\mathbb{E}[B_j]}} \right)^{1/m-1}, \quad (38)$$

for each  $k \in \{1, \dots, K\}$ ,  $m = 2, 3, \dots$

**Proof :** We need to show that  $\mathbb{E}[\hat{W}^k(p^*)] \leq \mathbb{E}[\hat{W}^k(p)]$ . This holds if and only if

$$\frac{\sum_{i=1}^k \hat{\lambda}_i \mathbb{E}[B_i]^{k/k-1}}{\hat{\lambda} (\sum_{j=1}^k \hat{\lambda}_j \mathbb{E}[B_j]^{k/k-1})^k} \cdot \left( \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2} \right)^k \leq \frac{\hat{\lambda}_i}{\hat{\lambda} p_i^k} \frac{1}{(\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j])^k} \cdot \left( \frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2} \right)^k$$

which follows by definition. Now, simplifying the terms which do not depend on the weights we obtain

$$\frac{1}{\left(\sum_{j=1}^k \hat{\lambda}_j \mathbb{E}[B_j]^{k/k-1}\right)^{k-1}} \leq \frac{\sum_{i=1}^k \frac{\hat{\lambda}_i}{p_i^k}}{\left(\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]\right)^k}$$

and rewriting it

$$\left(\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} \mathbb{E}[B_j]\right)^k \leq \sum_{i=1}^k \frac{\hat{\lambda}_i}{p_i^k} \left(\sum_{j=1}^k \hat{\lambda}_j \mathbb{E}[B_j]^{k/k-1}\right)^{k-1}, \quad (39)$$

which holds by Hölder's inequality.  $\square$

**Remark 5.** *By Equation (38) we get that the ratio of two optimal weights is the following:*

$$\frac{p_k^*}{p_j^*} = \left(\frac{\mathbb{E}[B_j]}{\mathbb{E}[B_k]}\right)^{1/(m-1)}. \quad (40)$$

*In general, the optimal choice for the weights is non trivial. However, note that when  $m \rightarrow 1$  we deduce that under the optimal weights a class- $k$  customer has strict priority over a class- $j$  customer if  $\mathbb{E}[B_k] < \mathbb{E}[B_j]$ . This is exactly the result that the  $c\mu$ -rule states. In addition, when  $m \rightarrow \infty$ , from (40) we see that the ratio of the optimal weights converges to 1. This implies that as  $m$  gets larger, it becomes optimal to treat all classes equal.*

## 7 Numerical results

In this section we present numerical experiments related to the results obtained in this paper. We consider a system under the discipline DROS with two classes of customers ( $K = 2$ ) and assume exponentially distributed service requirements. For each experiment in the order of  $10^5$  busy periods are simulated. A busy period refers to the period of time between two consecutive time epochs in which the system is empty, and every busy period is a regenerative point of the stochastic process. In Section 7.1 we present the numerical results corresponding to the distribution of the number of customers in the queue, in Sections 7.2 we focus on the moments of the queue length and waiting time and in Section 7.3 we investigate the optimal weights.

### 7.1 State-space collapse for the queue lengths

In this section we simulate the distribution of the joint queue length vector. As parameters we chose  $\lambda_1 = 2.15, \lambda_2 = 2.85, \mathbb{E}[B_1] = 1/4$  and  $\mathbb{E}[B_2] = 1/6$ , so that  $\rho = 0.9994$ . In Figure 3 we plot the joint queue length probabilities (obtained by simulation) for the weights  $p_1 = 0.7, p_2 = 0.3$ . The horizontal and vertical axis correspond to  $N_1$  and  $N_2$ , respectively. As a consequence of the state-space collapse stated in Proposition 4.1, in heavy traffic the probabilities will lie on a straight line with slope  $\frac{\hat{N}_2}{\hat{N}_1} = \frac{p_1 \hat{\lambda}_2}{\hat{\lambda}_1 p_2} \approx 3.1$ , starting from the origin. This result coincides with the slope of the figure obtained.

### 7.2 Moments of waiting time and queue length

In Figure 4 we plot  $(1-\rho)\mathbb{E}[N]$  (using Little's Law and Equation (37)) and  $(1-\rho)^2\mathbb{E}[N^2]$  (obtained by simulation) for different values of the load  $\rho$ . When doing so, we keep the mean service requirements fixed,  $\mathbb{E}[B_1] = 1/4$  and  $\mathbb{E}[B_2] = 1/6$ , and take  $\lambda_2 = 1.5\lambda_1$ . Moreover, we calculate the first and second moment of the scaled queue length as  $\rho \uparrow 1$ , i.e.,  $\mathbb{E}[\lim_{\rho \uparrow 1} (1-\rho)^m N^m] = \mathbb{E}[\hat{N}^m], m = 1, 2$ , using (12), giving as a result the values indicated with a dot in Figure 4, which are  $\mathbb{E}[\hat{N}] = 0.9589$  and  $\mathbb{E}[\hat{N}^2] = 1.9510$ . As it can be seen in Figure 4, in both cases, as the load gets close to one the functions  $\mathbb{E}[(1-\rho)^m N^m], m = 1, 2$ , converge to the values indicated with the dot. This would imply that an interchange of the limit and expectation holds for the random variable  $(1-\rho)N_k$ , i.e.,  $\lim_{\rho \uparrow 1} \mathbb{E}[(1-\rho)^m N_k^m] = \mathbb{E}[\lim_{\rho \uparrow 1} (1-\rho)^m N_k^m], m = 1, 2$ .



$\rho=0.9994, \lambda_1=2.15, \lambda_2=2.85, E[B_1]=0.24, E[B_2]=0.16, p_1=0.7, p_2=0.3$

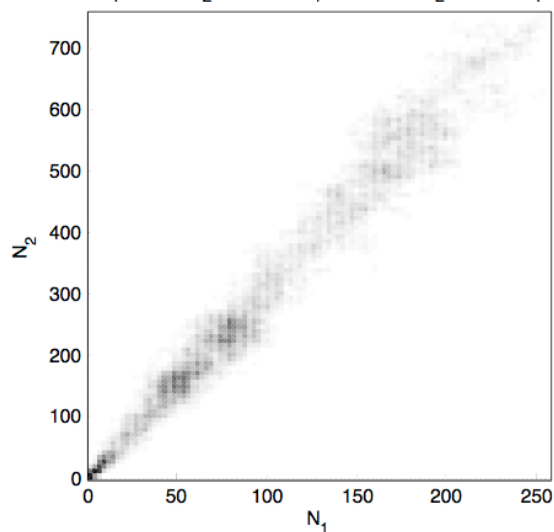


Figure 3: Joint queue length probability. The darkness of the points specifies the probability of being into a particular state. The darker the point is, the higher the probability of being in that state.

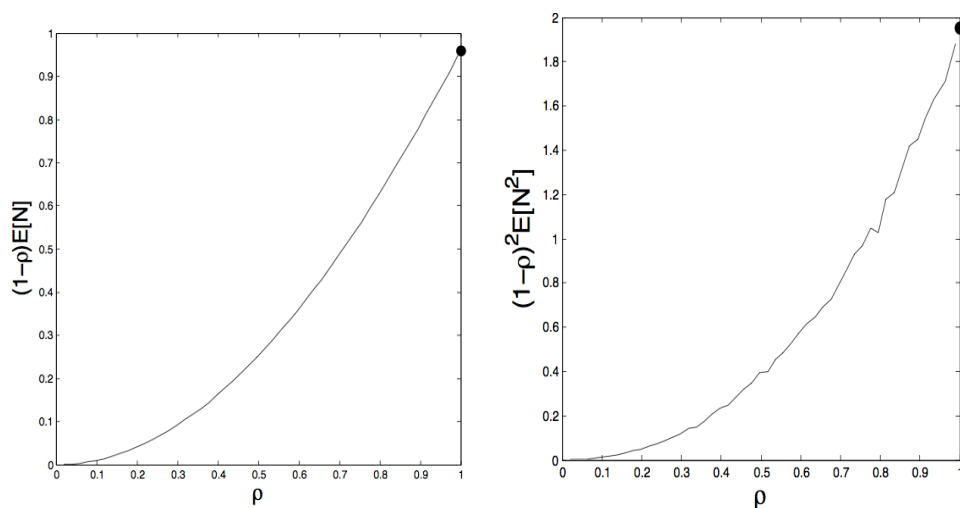


Figure 4: First and second moment of the scaled queue length obtained for different values of the load  $\rho$ . The dots in both pictures are calculated by using (12) giving as a result  $\mathbb{E}[\hat{N}] = 0.9589$  and  $\mathbb{E}[\hat{N}^2] = 1.9510$ .

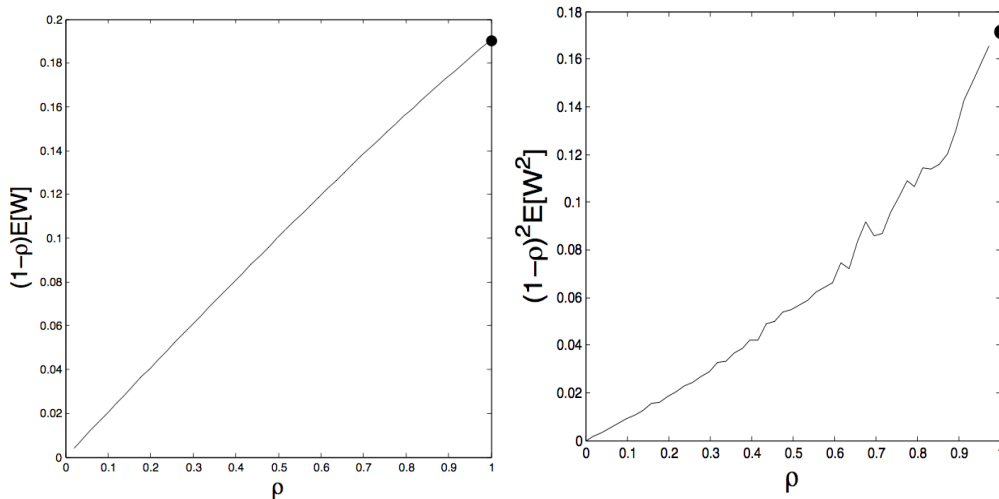


Figure 5: First and second moment of the scaled waiting time obtained for different values of the load  $\rho$ . The dots in both pictures are calculated by using Equation (36) giving as a result  $\mathbb{E}[\hat{W}] = 0.1906$  and  $\mathbb{E}[\hat{W}^2] = 0.1713$ .

We note that if the limits are indeed interchangeable, together with the convergence in distribution of the scaled queue lengths this would imply the uniform integrability of the scaled queue length (see [3, Theorem 3.6]), as assumed in Assumption 1.

In Figure 5 we plot  $(1 - \rho)\mathbb{E}[W]$  (using Equation (37)) and  $(1 - \rho)^2\mathbb{E}[W^2]$  (obtained by simulation) for different values of the load  $\rho$ . The simulation setting is the same as the one used for the queue length. We calculate the value of Equation (36) for the cases  $m = 1$  and  $m = 2$  giving as a result the values indicated with a dot in Figure 5, which are  $\mathbb{E}[\hat{W}] = 0.1906$  and  $\mathbb{E}[\hat{W}^2] = 0.1713$ . In both cases, as the load gets close to one the functions converge to the value obtained in Equation (36), which would imply again that an interchange of the limit and expectation holds for the random variable  $(1 - \rho)W_k$ , i.e.,  $\lim_{\rho \uparrow 1} \mathbb{E}[(1 - \rho)^m W_k^m] = \mathbb{E}[\lim_{\rho \uparrow 1} (1 - \rho)^m W_k^m]$ ,  $m = 1, 2$ . In fact, for the first moment, taking the limit as  $\rho \uparrow 1$  in Equation (37) it is easy to see that it indeed converges to the heavy-traffic limit as characterized by Equation (36) when  $K = 2$ .

### 7.3 Optimal values for the weights

In Proposition 6.3 we presented the optimal choices for the weights  $p^*$  in order to minimize the moments of the scaled waiting time  $\hat{W}$ . In this section we numerically evaluate the validity of the optimal weights outside the heavy-traffic regime. We set  $\mathbb{E}[B_1] = 0.2439$  and  $\mathbb{E}[B_2] = 0.1667$  and plot  $(1 - \rho)^2\mathbb{E}[W^2(p_1, 1 - p_1)]$  for three different values of the load,  $\rho = 0.7$ ,  $\rho = 0.8$  and  $\rho = 0.9$ , see Figure 6. The value of  $p_1^*$  is in this particular case equal to  $p_1^* = 0.4059$  (see (38)). It can be seen that the weight  $p_1^* = 0.4059$  is a good approximation for the minimizer of  $(1 - \rho)^2\mathbb{E}[W^2(p_1, 1 - p_1)]$  for load equal to  $\rho = 0.9$ . As the load decreases the approximation becomes worse, but it is still close to the minimum of the function. We also plot  $\mathbb{E}[\lim_{\rho \uparrow 1} (1 - \rho)^2 W^2(p_1, 1 - p_1)] = \mathbb{E}[\hat{W}^2(p_1, 1 - p_1)]$ , which is seen to be a good approximation for  $(1 - \rho)^2\mathbb{E}[W^2(p_1, 1 - p_1)]$  as the load gets close to 1.

## Acknowledgements

The authors wish to express their gratitude to A.M. Makowski and P. Robert for useful discussions.

## References

- [1] S. Asmussen. *Applied Probability and Queues*. Springer, 2003.

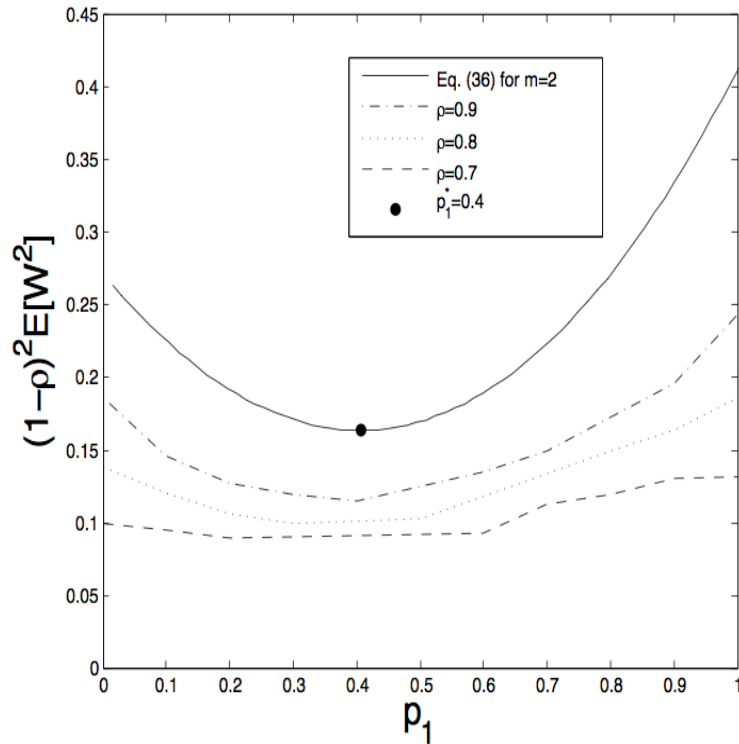


Figure 6: The second moment  $\mathbb{E}[(1 - \rho)^2 W^2]$  for different values of the weights  $(p_1, p_2) = (p_1, 1 - p_1)$ .

- [2] U. Ayesta, A. Izagirre, and I.M. Verloop. Heavy-traffic analysis of the discriminatory random-order-of-service discipline. *Performance Evaluation Review*, 32(2):41–43, 2011.
- [3] P. Billingsley. *Convergence of Probability Measures*. Wiley, 1999.
- [4] S.C. Borst, O.J. Boxma, J.A. Morrison, and R. Núñez-Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31:254–262, 2003.
- [5] O.J. Boxma, D. Denteneer, and J.A.C. Resing. Some models for contention resolution in cable networks. *Lecture Notes in Computer Science*, 2345:117–128, 2002.
- [6] O.J. Boxma, S.G. Foss, J.-M. Lasgouttes, and R. Núñez-Queija. Waiting time asymptotics in the single server queue with service in random order. *Queueing Systems*, 46:35–73, 2004.
- [7] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, New York, 1971.
- [8] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. London: Academic Press, 1980.
- [9] M. Haviv and J. van der Wal. Equilibrium strategies for processor sharing and random queues with relative priorities. *Probability in the Engineering and Informational Sciences*, 11:403–412, 1997.
- [10] M. Haviv and J. van der Wal. Waiting times in queues with relative priorities. *Operations Research Letters*, 35:591–594, 2007.
- [11] J. Kim. Queue length distribution in a queue with relative priorities. *Bull. Korean Math. Soc.*, 46:107–116, 2009.
- [12] J. Kim, J. Kim, and B. Kim. Analysis of the M/G/1 queue with discriminatory random order service policy. *Performance Evaluation*, 68(3):256–270, 2011.

- [13] J.F.C. Kingman. The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.*, 57:902–904, 1961.
- [14] J.F.C. Kingman. On queues in which customers are served in random order. *Proc. Cambridge Philos. Soc.*, 58:79–91, 1962.
- [15] J.F.C. Kingman. Queue disciplines in heavy traffic. *Mathematics of Operations Research*, 7(2):262–271, 1982.
- [16] C. Palm. Waiting times with random served queue. *Tele1 (English edition; original 1938)*, 1–107, 1957.
- [17] I.M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Operations Research*, 59:648–660, 2011.
- [18] A.P. Zwart. Heavy-traffic asymptotics for the single-server queue with random order of service. *Operations Research Letters*, 33:511–518, 2005.

## Appendix A: Proof of Lemma 3.3

The total workload at departure epochs can be represented as

$$V^{\text{dep}} = \sum_{j=1}^K \sum_{h=1}^{Q_j} B_{j,h},$$

with  $B_{j,h}$  the service requirement of the  $h$ -th class- $j$  customer. Note that the service requirements of all class- $j$  customers are i.i.d., and  $B_{j,h} \stackrel{d}{=} B_j$  for all  $h$ .

For  $\epsilon' > 0$  we have

$$\begin{aligned}
\mathbb{P}((1-\rho)Q_j \geq M) &= \mathbb{P}\left(Q_j \geq \frac{M}{(1-\rho)}\right) \\
&\leq \mathbb{P}\left(\sum_{h=1}^{Q_j} B_{j,h} \geq \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h}\right) \\
&\leq \mathbb{P}\left((1-\rho) \sum_{j=1}^K \sum_{h=1}^{Q_j} B_{j,h} \geq M \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h}\right) \\
&= \mathbb{P}\left(\frac{(1-\rho)V^{\text{dep}}}{M} - \mathbb{E}[B_j] \geq \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j]\right) \\
&= \mathbb{P}\left(\frac{(1-\rho)V^{\text{dep}}}{M} - \mathbb{E}[B_j] \geq \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j]\right) \\
&\quad \left| \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j] > -\epsilon \right) \mathbb{P}\left(\frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j] > -\epsilon\right) \\
&+ \mathbb{P}\left(\frac{(1-\rho)V^{\text{dep}}}{M} - \mathbb{E}[B_j] \geq \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j]\right) \\
&\quad \left| \frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j] \leq -\epsilon \right) \mathbb{P}\left(\frac{(1-\rho)}{M} \sum_{h=1}^{\lfloor M/(1-\rho) \rfloor} B_{j,h} - \mathbb{E}[B_j] \leq -\epsilon\right) \\
&\leq \mathbb{P}\left(\frac{(1-\rho)V^{\text{dep}}}{M} - \mathbb{E}[B_j] > -\epsilon\right) + \tilde{\epsilon} \\
&= \mathbb{P}\left((1-\rho)V^{\text{dep}} \geq M(\mathbb{E}[B_j] - \epsilon)\right) + \tilde{\epsilon} \\
&< \bar{\epsilon} + \tilde{\epsilon} = \epsilon', \text{ for } \rho \text{ close enough to 1 and } M \text{ large enough.} \tag{41}
\end{aligned}$$

In the last step we used the fact that the workload, independently of the work-conserving scheduling discipline being used, is tight in heavy traffic, see Kingman [13], that is  $\forall \epsilon' \exists M'$  such that  $\mathbb{P}((1 - \rho)V^{\text{dep}} \geq M') < \epsilon'$ . From (41) we conclude that  $(1 - \rho)(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_K)$  is tight.  $\square$

## Appendix B: Proof of Lemma 3.6

The proof of Lemma 3.6 is based on the proof of Lemma 3 in [17]. We have

$$\begin{aligned} \sum_{i=1}^K \frac{\hat{\lambda}_i}{p_i} F_i(\vec{s}) &= \sum_{i=1}^K \frac{\hat{\lambda}_i}{p_i} p_i (-s_i + \mathbb{E}[B_i] \sum_{j=1}^K \hat{\lambda}_j s_j) \\ &= - \sum_{i=1}^K \hat{\lambda}_i s_i + \sum_{i=1}^K \hat{\lambda}_i \mathbb{E}[B_i] \sum_{j=1}^K \hat{\lambda}_j s_j \\ &= - \sum_{i=1}^K \hat{\lambda}_i s_i + \sum_{j=1}^K \hat{\lambda}_j s_j \\ &= 0. \end{aligned}$$

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  $\vec{f}(u) = (f_1(u), \dots, f_K(u))$ , parametrized by  $u \geq 0$ , such that

$$\vec{f}(0) = \vec{s} \quad \text{and} \quad \frac{\partial f_i(u)}{\partial u} = F_i(\vec{f}(u)), \quad \text{for all } i \text{ and } u \geq 0. \quad (42)$$

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

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

which follows from the chain rule, Lemma 3.5, and equation (42). 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  $u \rightarrow \infty$ .

From (10) we get that (42) can be written as  $\vec{f}'(u) = \vec{s}$  and  $\vec{f}'(u)^T = A \vec{f}(u)^T$  with

$$A = \begin{pmatrix} p_1(-1 + \mathbb{E}[B_1]\hat{\lambda}_1) & p_1\mathbb{E}[B_1]\hat{\lambda}_2 & \cdots & p_1\mathbb{E}[B_1]\hat{\lambda}_K \\ p_2\mathbb{E}[B_2]\hat{\lambda}_1 & p_2(-1 + \mathbb{E}[B_2]\hat{\lambda}_2) & \cdots & p_2\mathbb{E}[B_2]\hat{\lambda}_K \\ \vdots & \vdots & \ddots & \vdots \\ p_K\mathbb{E}[B_K]\hat{\lambda}_1 & p_K\mathbb{E}[B_K]\hat{\lambda}_2 & \cdots & p_K(-1 + \mathbb{E}[B_K]\hat{\lambda}_K) \end{pmatrix}. \quad (43)$$

In Lemma 7.1 below 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  $\vec{f}'(u)^T = A \vec{f}(u)^T$  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  $H_c$ .  $\square$

The following technical lemma is used in the proof of Lemma 3.6.

**Lemma 7.1.** *Consider the matrix  $A$  as defined in (43). 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  $\vec{\eta} = (\eta_1, \dots, \eta_K) \geq \vec{0}$  with  $\sum_{j=1}^K \eta_j = 1$  such that  $\vec{s}^* = (s_1^*, \dots, s_K^*)$  with  $s_j^* := \frac{p_j}{\lambda_j} \eta_j$  is an eigenvector of  $A$  corresponding to the eigenvalue 0, and  $\vec{s}^* \in H_1$ .*

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

$$S = DAD^{-1} = \begin{pmatrix} p_1(-1 + \mathbb{E}[B_1]\hat{\lambda}_1) & p_2\hat{\lambda}_1\mathbb{E}[B_1] & \cdots & p_K\hat{\lambda}_1\mathbb{E}[B_1] \\ p_1\hat{\lambda}_2\mathbb{E}[B_2] & p_2(-1 + \mathbb{E}[B_2]\hat{\lambda}_2) & \cdots & p_K\hat{\lambda}_2\mathbb{E}[B_2] \\ \vdots & \vdots & \ddots & \vdots \\ p_1\hat{\lambda}_K\mathbb{E}[B_K] & p_2\hat{\lambda}_K\mathbb{E}[B_K] & \cdots & p_K(-1 + \mathbb{E}[B_K]\hat{\lambda}_K) \end{pmatrix}. \quad (44)$$

The matrix  $A$  is similar to  $S$  and therefore  $A$ ,  $S$  and  $S^T$  have the same eigenvalues. The sum of each row of  $S^T$  is 0 because  $\sum_{i=1}^K \mathbb{E}[B_i]\hat{\lambda}_i = 1$ , and the off-diagonal elements in  $S^T$  are all strictly positive. This implies that the matrix  $S^T$  is a generator corresponding to a finite-state continuous-time irreducible Markov chain. Hence, it has a unique equilibrium distribution  $\vec{\eta}$ , i.e.,  $\vec{\eta}S^T = \vec{0}$  and  $\sum_{j=1}^K \eta_j = 1$ . In particular, 0 is an eigenvalue of the matrix  $S^T$ , with multiplicity 1 and corresponding to the left eigenvector  $\vec{\eta}$ , and, cf.(Proposition 6.2,[1]), 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  $\vec{s}^{*T} = D^{-1}\vec{\eta}^T$ , since  $A\vec{s}^{*T} = D^{-1}DAD^{-1}\vec{\eta}^T = D^{-1}S\vec{\eta}^T = \vec{0}^T$ .  $\square$

## Appendix C: Proof of Lemma 5.3

Taking  $(u, z_1, \dots, z_K) = ((1 - \rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K})$  in (24) we get

$$\begin{aligned} W_i^1((1 - \rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) &= \\ \sum_{j=1}^K ((1 - \rho)\lambda_j + \lambda p_j \frac{\partial}{\partial z_j} r(z_1, \dots, z_K) \Big|_{z_j = e^{-(1-\rho)s_j}}) &\cdot \\ \frac{B_j^*(\lambda - \sum_{k=1}^K \lambda_k e^{-(1-\rho)s_k}) - B_j^*((1 - \rho)u + \lambda - \sum_{k=1}^K \lambda_k e^{-(1-\rho)s_k})}{(1 - \rho)u} &. \end{aligned} \quad (45)$$

By applying l'Hopital's rule we get the expression below:

$$\begin{aligned} \lim_{\rho \uparrow 1} W_i^1((1 - \rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) &= \\ = -\frac{1}{u} \sum_{j=1}^K \left( \hat{\lambda}_j \cdot 0 + \hat{\lambda} p_j \frac{\partial}{\partial s_j} \hat{r}(s_1, \dots, s_K) (B_j^{*'}(0) (-\sum_{k=1}^K \hat{\lambda}_k s_k) - B_j^{*'}(0) (-u - \sum_{k=1}^K \hat{\lambda}_k s_k)) \right) & \\ = \hat{\lambda} \sum_{j=1}^K \mathbb{E}[B_j] p_j \frac{\hat{\lambda}_j}{p_j} \frac{\partial}{\partial v} \hat{r}^*(v) \Big|_{v = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j} = \hat{\lambda} \frac{\partial}{\partial v} \hat{r}^*(v) \Big|_{v = \sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} s_j}, & \end{aligned}$$

where  $\hat{r}(\vec{s})$  is defined in (6) and satisfies (9). Hence,  $W_i^1$  does not depend on  $u$  in heavy traffic. In order to calculate  $\lim_{\rho \uparrow 1} W_i^1((1 - \rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K})$  it is therefore sufficient to study the left hand side of Equation (24) in the point  $u = 0$ . This is what we will do in the remainder of the proof.

We also remark that  $\lim_{\rho \uparrow 1} (1 - \rho)Q_i^*$  is distributed as  $\hat{Q}_i$ . This follows by setting  $u = 0$  in Equation (22):

$$\begin{aligned} \mathbb{E}[e^{-(1-\rho)\sum_{i=1}^K s_i Q_i}] &= \mathbb{E}[\mathbf{1}_{\{W_k=0\}}] + \mathbb{E}[e^{-(1-\rho)\sum_{i=1}^K s_i Q_i} \mathbf{1}_{\{W_k>0\}}] \cdot \mathbb{E}[\mathbf{1}_{\{W_k>0\}}] \\ &= (1 - \rho) + \rho \mathbb{E}[e^{-(1-\rho)\sum_{i=1}^K s_i Q_i^*}] \end{aligned} \quad (46)$$

Taking the limit  $\rho \uparrow 1$  in equation (46) we immediately obtain that  $\lim_{\rho \uparrow 1} (1 - \rho)Q_i^* \stackrel{d}{=} \hat{Q}_i \stackrel{d}{=} Y \frac{\lambda_i}{p_i}$ , where the last step follows from Proposition 3.1.

Using this result and taking  $u = 0$  in (21) we get:

$$\begin{aligned} \lim_{\rho \uparrow 1} T_k(0, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) &= \lim_{\rho \uparrow 1} \mathbb{E}[e^{-(1-\rho)s_1 Q_1^*} \dots e^{-(1-\rho)s_K Q_K^*}] \\ &= \mathbb{E}[e^{-Y(\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j})}] = \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}}. \end{aligned} \quad (47)$$

Let us now calculate the derivative of  $T_k$  with respect to  $z_j$ . By definition (see (21)) we have

$$T(0, \vec{z}) = \mathbb{E}[z_1^{Q_1^*} \dots z_K^{Q_K^*} \mathbf{1}_{\{W_k > 0\}}].$$

Hence, the derivative is

$$\frac{\partial T(0, \vec{z})}{\partial z_j} = \mathbb{E}\left[\frac{Q_j^* z_1^{Q_1^*} \dots z_K^{Q_K^*}}{z_j} \mathbf{1}_{\{W_k > 0\}}\right].$$

Taking the limit when  $\rho \uparrow 1$  we get

$$\begin{aligned} \lim_{\rho \uparrow 1} (1-\rho) \frac{\partial T(0, \vec{z})}{\partial z_j} \Big|_{\vec{z}=e^{(1-\rho)\vec{s}}} &= \lim_{\rho \uparrow 1} \mathbb{E}\left[\frac{-(1-\rho)Q_j^* e^{-(1-\rho)s_1 Q_1^*} \dots e^{-(1-\rho)s_K Q_K^*}}{e^{-(1-\rho)s_j}}\right] \\ &= \mathbb{E}[\hat{Q}_j^* e^{-s_1 \hat{Q}_1^*} \dots e^{-s_K \hat{Q}_K^*}] = \frac{\lambda_j}{p_j} \frac{\nu(\vec{p})}{\left(\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}\right)^2}, \end{aligned} \quad (48)$$

where in the second step we used Assumption 1 and [3, Theorem 3.5], and in the last step the fact that  $\hat{Q}_j \stackrel{d}{=} \frac{\hat{\lambda}_j}{p_j} Y$  with  $Y$  exponentially distributed with rate  $\nu(\vec{p})$ . Let us consider the term which is multiplying  $\frac{\partial}{\partial z_j} T_k(u, z_1, \dots, z_K)$  in Equation (23):

$$\begin{aligned} &\lim_{\rho \uparrow 1} \frac{e^{-(1-\rho)s_j} - B_j^*((1-\rho)u + \lambda - \sum_{r=1}^K \lambda_r e^{-(1-\rho)s_r})}{1-\rho} \\ &= \lim_{\rho \uparrow 1} \frac{s_j e^{-(1-\rho)s_j} - B_j^*((1-\rho)u + \lambda - \sum_{r=1}^K \lambda_r e^{-(1-\rho)s_r})}{-1} \frac{(-u - \sum_{r=1}^K \lambda_r e^{-(1-\rho)s_r} s_r)}{-1} \\ &= -(s_j - B_j^*(0))(-u - \sum_{r=1}^K \hat{\lambda}_r s_r) = -(s_j - \mathbb{E}[X_j])(u + \sum_{r=1}^K \hat{\lambda}_r s_r), \end{aligned} \quad (49)$$

where in the first step we used l'Hopital's rule.

Due to the fact that the limit of a sum is the sum of the limits and it satisfies the same property for the product, substituting (47), (48) and (49) in equation (23) and since we know that  $W_i^1$  does not

depend on  $s$  in heavy-traffic we get what we desired, namely,

$$\begin{aligned}
& \lim_{\rho \uparrow 1} W_i^1((1-\rho)u, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) = \lim_{\rho \uparrow 1} W_i^1(0, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) \\
&= \lim_{\rho \uparrow 1} \sum_{j=1}^K \frac{p_j}{p_i} \left( \frac{\partial}{\partial z_j} T_k(0, z_1, \dots, z_K) \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \right) (z_j - B_j^*(\lambda - \sum_{r=1}^K \lambda_r z_r)) + T_k(0, z_1, \dots, z_K) \\
&= \lim_{\rho \uparrow 1} \sum_{j=1}^K \frac{p_j}{p_i} \left( (1-\rho) \frac{\partial}{\partial z_j} T_k(0, z_1, \dots, z_K) \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \right) \frac{(z_j - B_j^*(\lambda - \sum_{r=1}^K \lambda_r z_r))}{(1-\rho)} \\
&\quad + T_k(0, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) \\
&= \lim_{\rho \uparrow 1} \sum_{j=1}^K \frac{p_j}{p_i} \left( (1-\rho) \frac{\partial}{\partial z_j} T_k(0, z_1, \dots, z_K) \Big|_{\vec{z}=e^{-(1-\rho)\vec{s}}} \right) \lim_{\rho \uparrow 1} \frac{(z_j - B_j^*(\lambda - \sum_{r=1}^K \lambda_r z_r))}{(1-\rho)} \\
&\quad + \lim_{\rho \uparrow 1} T_k(0, e^{-(1-\rho)s_1}, \dots, e^{-(1-\rho)s_K}) \\
&= \sum_{j=1}^K \frac{p_j}{p_i} \left( \frac{\hat{\lambda}_j}{p_j} \frac{\nu(\vec{p})}{(\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j})^2} \right) (-s_j + \mathbb{E}[B_j] \sum_{r=1}^K \hat{\lambda}_r s_r) + \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} \\
&= \frac{1}{p_i} \sum_{j=1}^K \frac{\nu(\vec{p})}{(\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j})^2} (-\hat{\lambda}_j s_j + \sum_{r=1}^K \hat{\lambda}_r s_r) + \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} \\
&= \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}}.
\end{aligned}$$

□

## Appendix D: Solution of the ODE (31)

The solution of (31) is given by the sum of the solution of the homogeneous case,  $T_k^H(u, y)$ , and a particular solution,  $T_k^P(u, y)$ . The homogeneous solution is given by:

$$T_k^H(u, y) = C(u) e^{\frac{p_k}{u} y}, \quad (50)$$

where  $C(u)$  is an arbitrary function of  $u$ . In order to find the particular solution we rewrite (31) as

$$\frac{\partial \hat{T}_k^P(u, y)}{\partial y} \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} - \frac{p_k}{u} \hat{T}_k^P(u, y) = -\frac{p_k}{u} \frac{\nu(\vec{p})}{\nu(\vec{p}) + y}. \quad (51)$$

Let us solve the new equation using the integrating factor technique. In order to do so, let define the function  $\mu(y) = e^{-\frac{p_k}{u} y}$  and multiply (51) by it. The derivative of  $\mu(y)$  satisfies  $\frac{d\mu(y)}{dy} = -\mu(y) \frac{p_k}{u}$ . Then, our equation becomes

$$\begin{aligned}
-\mu(y) \frac{p_k}{u} \frac{\nu(\vec{p})}{\nu(\vec{p}) + y} &= \mu(y) \frac{\partial \hat{T}_k^P(u, y)}{\partial y} \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} - \mu(y) \frac{p_k}{u} \hat{T}_k^P(u, y) \\
&= \mu(y) \frac{\partial \hat{T}_k^P(u, y)}{\partial y} \Big|_{y=\sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j}} + \hat{T}_k^P(u, y) \frac{d\mu(y)}{dy} \\
&= \frac{\partial}{\partial y} (\mu(y) \hat{T}_k^P(u, y))
\end{aligned}$$



which can be solved by integration. Integrating each side with respect to  $y$  gives us a particular solution for (31), which is,

$$\hat{T}_k^P(u, y) = -\frac{p_k}{u} e^{\frac{p_k}{u} y} \int_0^y e^{-\frac{p_k}{u} x} \frac{\nu(\vec{p})}{\nu(\vec{p}) + x} dx,$$

or which is the same

$$\hat{T}_k^P(u, y) = \frac{p_k}{u} e^{\frac{p_k}{u} y} \int_y^\infty e^{-\frac{p_k}{u} x} \frac{\nu(\vec{p})}{\nu(\vec{p}) + x} dx. \quad (52)$$

In conclusion, the general solution of the ODE (31) is given by

$$\hat{T}_k(u, y) = T_k^H(u, y) + \hat{T}_k^P(u, y) = C(u) e^{\frac{p_k}{u} y} + \frac{p_k}{u} e^{\frac{p_k}{u} y} \int_y^\infty e^{-\frac{p_k}{u} x} \frac{\nu(\vec{p})}{\nu(\vec{p}) + x} dx. \quad (53)$$

We will now show that the constant  $C(u)$  is equal to zero. First, note that the solution for the homogeneous case (50) goes to infinity as  $y \rightarrow \infty$ . Second, from (47) it is immediate that  $T_k(u, y)$  converges to 0 when  $y = \sum_{j=1}^K s_j \frac{\hat{\lambda}_j}{p_j} \rightarrow \infty$ . Moreover, if we take the particular solution (52), applying l'Hopital's rule for  $y \rightarrow \infty$  we obtain that it also converges to zero, namely:

$$\lim_{y \rightarrow \infty} \hat{T}_k(u, y) = \lim_{y \rightarrow \infty} \frac{\nu(\vec{p}) p_k \int_y^\infty e^{-\frac{p_k}{u} x} \frac{1}{\nu(\vec{p}) + x} dx}{e^{-\frac{p_k}{u} y}} = \lim_{y \rightarrow \infty} \frac{\nu(\vec{p}) p_k}{u} \frac{-e^{-\frac{p_k}{u} y} \frac{1}{\nu(\vec{p}) + y}}{-\frac{p_k}{u} e^{-\frac{p_k}{u} y}} = \lim_{y \rightarrow \infty} \frac{\nu(\vec{p})}{\nu(\vec{p}) + y} = 0.$$

Then, the necessary condition for  $C(u) e^{\frac{p_k}{u} y}$  to converge to zero as  $y \rightarrow \infty$  is  $C(u) = 0$ . As a consequence, we conclude that the solution of (31) is

$$\hat{T}_k(u, y) = \frac{p_k}{u} e^{\frac{p_k}{u} y} \int_y^\infty e^{-\frac{p_k}{u} x} \frac{\nu(\vec{p})}{\nu(\vec{p}) + x} dx.$$