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

Ane Izagirre, Urtzi Ayesta, Ina Maria Maaike Verloop

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

A. Izagirre[a,b,e], U. Ayesta[b,c,d,e], I.M. Verloop[a,e]
ane.izagirre@laas.fr, urtzi@laas.fr, maaike.verloop@enseeiht.fr

[a]CNRS ; IRIT ; 2 rue C. Camichel, F-31071 Toulouse, France
[b]CNRS ; LAAS ; 7 avenue du colonel Roche, F-31400 Toulouse, France
[c]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[d]UPV/EHU, Univ. of the Basque Country, 20018 Donostia, Spain
[e]Université de Toulouse ; INP, INSA ; *IRIT*, *LAAS* ; F-31400 Toulouse, France

Abstract

We study the steady-state queue-length vector in a multi-class queue with relative priorities. Upon service completion, the probability that the next served customer is from class k is controlled by class-dependent weights. Once a customer has started service, it is served without interruption until completion. We establish a state-space collapse for the scaled queue length vector in the heavy-traffic regime, 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. We observe that the scaled queue length reduces as classes with smaller mean service requirement obtain relatively larger weights. We finally show that the scaled waiting time of a class-k customer is distributed as the product of two exponentially distributed random variables.

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

$$\frac{n_k p_k}{\sum_j n_j p_j}, \tag{1}$$

where, $p_j > 0$, $j = 1, \ldots, K$, 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 non-anticipating. A non-anticipating policy does not use information of the actual service requirement of the customers.

The relative priority model is quite general, and it provides an appropriate framework to provide service differentiation in non-preemptive systems. In fact, following the analysis of Section 8.4.1 in [10] it could be shown that the family of relative priority policies as studied in this paper is complete, i.e., within this family of policies one can achieve any performance vector in the achievable region of the non-preemptive M/G/1 queue.

The relative priority model can have application in various domains, in particular in ATM networks [3], telecommunication networks [6], or genetic networks, where molecules are analogous to customers, the enzyme is analogous to the server and protein species correspond to classes, see [18]. In this paper we do not focus on any application in particular. Instead, our goal is to provide a thorough analysis in order to obtain insights into the performance of the relative priority model that could potentially be applied to different contexts. We also believe that our methodology can be of independent interest in the study of other queueing networks.

A special case of the model under study 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 [11] and it is referred to as discriminatory-random-order-of-service (DROS). In recent years several interesting studies have been published on DROS, [12, 13, 14]. Expressions for the mean waiting time of a customer given its class have been obtained in [12]. In [13, 14] 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 [16, 17, 19]. The Laplace transform for the waiting time distribution was obtained in [16]. In [16, 17, 21], 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 [7] 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 [7]. In [5] 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 and non-preemptive services, 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. We note that a similar state-space collapse result was observed in [20] 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 [10], 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 [17] 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.
We note that in this paper we consider the heavy-traffic limit of the steady-state metrics. In the literature there are state-space results available for the transient queue length processes, that is, when the heavy-traffic limit is directly taken of the queue length processes. See for example [8] for the heavy-traffic analysis of a multi-class system where all classes receive simultaneously service. In general, the heavy-traffic and steady-state limits cannot be interchanged, which explains the interest of our approach. Another important difference is that our approach allows to investigate the waiting time in the system, a metric that does not have a clear counterpart in the "process" world.
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.
An extended abstract version of this paper appeared in [2].

2 Model description

We consider a multi-class single-server queue with K classes of customers. Class-k customers, $k = 1, \ldots, K$, arrive according to independent Poisson processes with rate $\lambda_k > 0$. 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 $B_k(x)$ and Laplace-Stieltjes transform $B_k^*(s) = \int_0^\infty \mathrm{e}^{-sx} \mathrm{d}B_k(x)$, and we define $B_k^{*'}(s) = \frac{\mathrm{d}B_k^*(s)}{\mathrm{d}s}$. 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

$$\lambda \uparrow \hat{\lambda} := \frac{1}{\sum_{k=1}^{K} \alpha_k \mathbb{E}[B_k]}, \tag{2}$$

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, α_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. We note that, throughout the paper, we do not explicitly reflect the dependence of the random variables on the traffic load ρ, in order to keep notation compact. 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 next proposition states the main result of this section and shows that in

the limit, the queue length vector is the product of an exponentially distributed random variable and a deterministic vector. 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 $(\lambda_1,\ldots,\lambda_K)\to(\hat{\lambda}_1,\ldots,\hat{\lambda}_K)$, such that as $\rho\uparrow 1$,*

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

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 at departure epochs, as obtained by Kim et al, see [14]. Define

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

and let

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

be its joint probability generating function. We define

$$r(\vec{z}):=\mathbb{E}\left[\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{\sum_{k=1}^K Q_k p_k}\cdot\mathbf{1}_{(\sum_{k=1}^K Q_k>0)}\right]=\sum_{(q_1,\ldots,q_K)\neq(0,\ldots,0)}\frac{\pi(\vec{q})}{q_1p_1+\ldots+q_Kp_K}z_1^{q_1}\ldots z_K^{q_K}.$$

In [14] 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, non-anticipating 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 [14].

Theorem 3.2. *[14, Theorem 1 and 2] (a) The probability generating function $p(z_1,\ldots,z_K)$ of the joint stationary queue lengths at departure epochs satisfies*

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

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

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

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 $(1-\rho)(Q_1,\ldots,Q_K)$ is tight for ρ close enough to 1, that is, for all ϵ there is a $\bar{\rho}\in(0,1)$ and $M>0$ such that $\mathbb{P}((1-\rho)Q_k\geq M)<\epsilon$, for all $k=1,...K$, and $\rho>\bar{\rho}$.*

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

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

exists, then there is a (possibly defective) random vector $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K)$ such that $(1-\rho)(Q_1, Q_2, \ldots, Q_K)$ converges in distribution to $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K)$, and the distribution of $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K)$ is uniquely determined by the limit in (5) (cf.the Continuity theorem, see Feller (1971) [9]). For now, we assume that the limit exists; we come back to this assumption in the last part of the proof of Proposition 3.1. Below we give two lemmas that describe properties of $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$. 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

$$\hat{r}(\vec{s}) = \mathbb{E}\left[\frac{1 - \mathrm{e}^{-s_1\hat{Q}_1} \cdots \mathrm{e}^{-s_K\hat{Q}_K}}{\sum_{k=1}^K \hat{Q}_k p_k} 1_{(\sum_{k=1}^K \hat{Q}_k > 0)}\right]. \tag{6}$$

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

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

Proof: From (3) we have

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

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

$$\begin{aligned}
\lim_{\rho\uparrow 1} \frac{\partial r(\vec{z})}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} &= \lim_{\rho\uparrow 1} \frac{\partial \mathbb{E}\left[\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{\sum_{k=1}^K Q_k p_k} \cdot \mathbf{1}_{(\sum_{k=1}^K Q_k > 0)}\right]}{\partial z_i}\Bigg|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} \\
&= \lim_{\rho\uparrow 1} \mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^K Q_k p_k} \cdot \frac{\mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \ldots \cdot \mathrm{e}^{-(1-\rho)s_K Q_K}}{\mathrm{e}^{-(1-\rho)s_i}} \cdot \mathbf{1}_{(\sum_{k=1}^K Q_k > 0)}\right] \\
&= \mathbb{E}\left[\frac{\hat{Q}_i}{\sum_{k=1}^K \hat{Q}_k p_k} \cdot \mathrm{e}^{-s_1\hat{Q}_1} \cdot \ldots \cdot \mathrm{e}^{-s_K\hat{Q}_K} \cdot \mathbf{1}_{(\sum_{k=1}^K \hat{Q}_k > 0)}\right] \\
&= \frac{\partial \hat{r}(\vec{s})}{\partial s_i}.
\end{aligned} \tag{9}$$

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

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

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

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

$$F_i(\vec{s}) = p_i(-s_i + \mathbb{E}[B_i] \sum_{k=1}^{K} \hat{\lambda}_k s_k) \quad i = 1, \ldots, K, \tag{10}$$

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

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

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \frac{\sum_{i=1}^{K} p_i(\mathrm{e}^{-(1-\rho)s_i} - B_i^*(\lambda - \sum_{j=1}^{K} \lambda_j \mathrm{e}^{-(1-\rho)s_j}))}{1-\rho} \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\Big|_{z_i = \mathrm{e}^{-(1-\rho)s_i}} \\
&= \lim_{\rho\uparrow 1} -(1 - \sum_{i=1}^{K} \frac{\lambda_i}{\lambda} B_i^*(\lambda - \sum_{j=1}^{K} \lambda_j \mathrm{e}^{-(1-\rho)s_j})) = 0.
\end{aligned} \tag{11}$$

where the last equality follows by noting that $B_i^*(0) = 1$, $\forall i$. Making the change of variable $x_i = \mathrm{e}^{-s_i}$ we obtain

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \frac{\sum_{i=1}^{K} p_i(\mathrm{e}^{-(1-\rho)s_i} - B_i^*(\lambda - \sum_{j=1}^{K} \lambda_j \mathrm{e}^{-(1-\rho)s_j}))}{1-\rho} \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\Big|_{z_i = \mathrm{e}^{-(1-\rho)s_i}} \\
&= \lim_{\rho\uparrow 1} \frac{\sum_{i=1}^{K} p_i(x_i^{1-\rho} - B_i^*(\lambda - \sum_{j=1}^{K} \lambda_j x_j^{1-\rho}))}{1-\rho} \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\Big|_{z_i = x_i^{1-\rho}} \\
&= \lim_{\rho\uparrow 1} \sum_{i=1}^{K} p_i(x_i^{1-\rho} \ln x_i \\
&\quad + \left(\frac{1}{\mathbb{E}(B)} - \left(\frac{1}{\mathbb{E}(B)} \sum_{j=1}^{K} \alpha_j x_j^{1-\rho} - \sum_{j=1}^{K} \lambda_j x_j^{1-\rho} \ln x_j \right) \right) (B_i^{*'}(\lambda - \sum_{j=1}^{K} \lambda_j x_j^{1-\rho}))) \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\Big|_{z_i = x_i^{1-\rho}} \\
&= \sum_{i=1}^{K} p_i(-s_i + \mathbb{E}(B_i) \sum_{j=1}^{K} \hat{\lambda}_j s_j) \frac{\partial \hat{r}(\vec{s})}{\partial s_i},
\end{aligned}$$

where in the second step we used l'Hopital's rule and in the third step we used (9) and that $B_i^{*'}(0) := \frac{\mathrm{d}B_i^*(s)}{\mathrm{d}s}\Big|_{s=0} = -\mathbb{E}[B_i]$ for all i.
Together with (11), we then obtain that

$$\sum_{i=1}^{K} p_i(-s_i + \mathbb{E}(B_i) \sum_{j=1}^{K} \hat{\lambda}_j s_j) \frac{\partial \hat{r}(\vec{s})}{\partial s_i} = 0.$$

□

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 function $\hat{r}(\vec{s})$ 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. *If* $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ *exists, then the function* $\hat{r}(s)$ *is constant on the* $(K-1)$*-dimensional hyperplane*

$$H_c := \{\vec{s} \geq \vec{0} \colon \sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k = c\}, c > 0.$$

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: Assume $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists. We come back to this assumption at the end of the proof. As $\hat{r}(\vec{s})$ is constant on H_c, see Lemma 3.6, $\hat{r}(\cdot)$ depends on $\vec{s}$ only through $\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k$, so there exists a function $\hat{r}^* : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{k=1}^{K} \frac{\lambda_k}{p_k} s_k)$. From Lemma 3.4 and $\frac{\partial \hat{r}(s)}{\partial s_i} = \frac{\hat{\lambda}_i}{p_i} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k}$ we obtain

$$\mathbb{E}[\mathrm{e}^{-\sum_{i=1}^{K} s_i \hat{Q}_i}] = \lim_{\rho\to 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{i=1}^{K} p_i \frac{\partial \hat{r}(s)}{\partial s_i} = \sum_{i=1}^{K} \hat{\lambda}_i \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k} = \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k}, \tag{12}$$

which again depends on $\vec{s}$ only through $\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k$. Equivalently, we can write

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

Since (by (12)) this only depends on $\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k$, 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

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

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

$$(\hat{Q}_1, ..., \hat{Q}_K) \stackrel{d}{=} (\frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, ..., \frac{\hat{\lambda}_K}{p_K}) Y. \tag{13}$$

Recall that we assumed that, for the sequence ρ, $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists, thereby showing that there is a unique limit (13). Since $(1-\rho)(Q_1, \ldots, Q_K)$ is tight, see Lemma 3.3, and since for any converging subsequence of ρ we obtain the same limit, we obtain that the limit itself exists (see corollary on page 59, Billingsley 1999). This concludes the proof. □

4 Queue length at arbitrary epochs

In this section we focus on the number of customers in the system at arbitrary epochs, $(N_1, \ldots, N_K)$. The following result shows that in the limit the queue length vector at arbitrary epochs is the product of an exponentially distributed random variable and a deterministic vector. We refer to the latter 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 [20] (Proposition 2.1) for the discriminatory processor sharing model. In fact, the proof technique is similar to that of [20].*

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

$$(1-\rho)(N_1, \ldots, N_K) \stackrel{d}{\to} (\hat{N}_1, ..., \hat{N}_K) \stackrel{d}{=} (\frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, ..., \frac{\hat{\lambda}_K}{p_K}) X, \tag{14}$$

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

$$\nu(\vec{p}) := \frac{2\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k]}{\sum_{k=1}^{K} \hat{\lambda}_k \mathbb{E}[B_k^2]}. \tag{15}$$

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 [14]. Let $\psi(z_1, ..., z_K)$ be the joint probability generating function of $(N_1, \ldots, N_K)$, i.e,

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

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 [14].

Theorem 4.2. *[14, Theorem 3 and Theorem 4] The joint probability generating function $\psi(z_1, ..., z_K)$ of the joint stationary queue lengths at arbitrary time epochs is given by*

$$\psi(z_1, ..., z_K) = 1 - \rho + \sum_{i=1}^{K} \lambda_i z_i \phi_i(z_1, ..., z_K) \frac{1 - B_i^*(\lambda - \sum_{k=1}^{K} \lambda_k z_k)}{\lambda - \sum_{k=1}^{K} \lambda_k z_k}, \tag{16}$$

where $\phi_i(z_1, \ldots, z_K)$ (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-i customers) is given by

$$\phi_i(z_1, ...z_K) = 1 - \rho + \frac{\lambda p_i}{\lambda_i} \frac{\partial}{\partial z_i} r(z_1, ...z_K). \tag{17}$$

In Section 4.1 we will show that Equation (16) 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, that is, to prove Proposition 4.1.

4.1 Heavy-traffic scaling

In the next lemma we characterize Equation (16) in heavy traffic.

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

$$\lim_{\rho \uparrow 1} \psi(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{i=1}^{K} p_i \frac{\partial \hat{r}(s)}{\partial s_i} = \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k},$$

with $\hat{r}^*(\cdot)$ some function $\hat{r}^* : \mathbb{R} \longrightarrow \mathbb{R}$.

Proof: Since $(1-\rho)(Q_1, \ldots, Q_K)$ converges in distribution to $(\hat{Q}_1, \ldots, \hat{Q}_K)$, we know that the limit of $p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists, and hence, by Equation (8), the limit of $\frac{\partial r(\vec{z})}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}}$ exists. It now follows directly from (17) that $\lim_{\rho\uparrow 1} \phi_i(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists and it is given by $\frac{\hat{\lambda} p_i}{\hat{\lambda}_i} \frac{\partial \hat{r}(s)}{\partial s_i}$.
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 $\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k$, so there exists a function $\hat{r}^* : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{k=1}^{K} \frac{\lambda_k}{p_k} s_k)$ and $\frac{\partial \hat{r}(s)}{\partial s_i} = \frac{\hat{\lambda}_i}{p_i} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k}$.
This, together with (16) gives that

$$\begin{aligned}
\lim_{\rho \uparrow 1} \psi(\mathrm{e}^{-(1-\rho)\vec{s}}) \quad &= \lim_{\rho \uparrow 1} \left(1 - \rho + \sum_{i=1}^{K} \lambda_i \mathrm{e}^{-(1-\rho)s_i} \phi_i(\mathrm{e}^{-(1-\rho)\vec{s}}) \frac{1 - B_i^*(\lambda - \sum_{k=1}^{K} \lambda_k \mathrm{e}^{-(1-\rho)s_k})}{\lambda - \sum_{k=1}^{K} \lambda_k \mathrm{e}^{-(1-\rho)s_k}} \right) \\
&= \sum_{i=1}^{K} \hat{\lambda}_i \frac{\hat{\lambda} p_i}{\hat{\lambda}_i} \frac{\partial \hat{r}(\vec{s})}{\partial s_i} (-B_i^{*'}(0)) = \sum_{i=1}^{K} \hat{\lambda} \hat{\lambda}_i \mathbb{E}[B_i] \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k} \\
&= \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k} \sum_{i=1}^{K} \hat{\lambda}_i \mathbb{E}[B_i] = \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} s_k},
\end{aligned}$$

where in the first step we used l'Hopital's rule and $B_i^{*'}(0) := \frac{\mathrm{d}B_i^*(s)}{\mathrm{d}s}\Big|_{s=0} = -\mathbb{E}[B_i]$ for all i. □

In particular, Lemma 4.3 implies that there exists a vector $(\hat{N}_1, \ldots, \hat{N}_K)$ 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 (15).

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

$$\mathbb{E}\big[\mathrm{e}^{-\sum_{k=1}^K s_k\hat{N}_k}\big]=\lim_{\rho\uparrow 1}\psi(\mathrm{e}^{-(1-\rho)\vec{s}})=\hat{\lambda}\frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^K\frac{\hat{\lambda}_k}{p_k}s_k}. \tag{18}$$

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

$$(\hat{N}_1,...,\hat{N}_K)\overset{d}{=}(\frac{\hat{\lambda}_1}{p_1},\frac{\hat{\lambda}_2}{p_2},...,\frac{\hat{\lambda}_K}{p_K})X, \tag{19}$$

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 [15], Kingman considered a FCFS queue and showed that the scaled total workload in a $M/G/1$ queue has a proper distribution as $\rho\uparrow 1$:

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

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

$$\mathbb{E}[\hat{V}^{arb}]=\frac{\sum_{k=1}^K\hat{\lambda}_k\mathbb{E}[B_k^2]}{2}. \tag{20}$$

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

$$V^{arb}=\sum_{k=1}^K\sum_{h=1}^{N_k-1}B_{k,h}+\sum_{k=1}^K\tilde{B}_k,$$

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

$$\begin{aligned}\mathbb{E}\big[\mathrm{e}^{-s\hat{V}^{arb}}\big] &= \lim_{\rho\uparrow 1}\mathbb{E}\big[\mathrm{e}^{-(1-\rho)sV^{arb}}\big]\\ &= \lim_{\rho\uparrow 1}\mathbb{E}\big[\mathrm{e}^{-(1-\rho)s(\sum_{k=1}^K\sum_{h=1}^{N_k-1}B_{k,h}+\sum_{k=1}^K\tilde{B}_k)}\big]\\ &= \lim_{\rho\uparrow 1}\mathbb{E}\big[\mathrm{e}^{-s\sum_{k=1}^K(1-\rho)(N_k-1)\frac{\sum_{h=1}^{N_k-1}B_{k,h}}{(N_k-1)}}\mathrm{e}^{-(1-\rho)s\sum_{k=1}^K\tilde{B}_k}\big]\\ &= \mathbb{E}\big[\mathrm{e}^{-s\sum_{k=1}^K\mathbb{E}[B_k]\hat{N}_k}\big],\end{aligned} \tag{21}$$

where in the last step we used that $\mathrm{e}^{-s\sum_{k=1}^K(1-\rho)(N_k-1)\frac{\sum_{h=1}^{N_k-1}B_{k,h}}{(N_k-1)}}$ is bounded by 1 and converges in distribution to $\mathrm{e}^{-s\sum_{k=1}^K\mathbb{E}[B_k]\hat{N}_k}$. From (21) we obtain that

$$\hat{V}^{arb}\overset{d}{=}\sum_{k=1}^K\mathbb{E}[B_k]\hat{N}_k, \tag{22}$$

and together with (19) this gives

$$\hat{V}^{arb} \stackrel{d}{=} X \sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k]. \tag{23}$$

Since $\hat{V}^{arb}$ is exponentially distributed, the same is true for X. Hence, taking expectations in (23) and applying (20) we obtain

$$\mathbb{E}[X] = \frac{\sum_{k=1}^{K} \hat{\lambda}_k \mathbb{E}[B_k^2]}{2 \sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k]},$$

which concludes the proof of Proposition 4.1. □

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-k 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 $(Q_1, \ldots, Q_K)$. The number of customers that arrive in the time it takes for the customer in service to depart is of the order ρ, since it is distributed as the number of arrivals in a residual service requirement. It then follows that*

$$Q_k \stackrel{d}{=} N_k + \mathbf{O}(\rho).$$

Multiplying the above equation by $(1-\rho)$ and taking the limit $\rho \uparrow 1$ we get that $\hat{Q}_k \stackrel{d}{=} \hat{N}_k$ 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.
Let W_l denote a generic random variable for the waiting time of an arbitrary class-l customer. We refer to this customer as the tagged class-l customer. 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., $W_l > 0$. We now define the following joint transform:

$$T_l(u, z_1, \ldots, z_K) := \mathbb{E}[\mathrm{e}^{-uW_l} z_1^{Q_1^*} \cdots z_K^{Q_K^*} 1_{\{W_l > 0\}}]. \tag{24}$$

Note that the transform of the waiting time W_l of the tagged class-l customer is given by

$$\mathbb{E}[\mathrm{e}^{-uW_l}] = \mathbb{E}[\mathrm{e}^{-u\cdot 0} \mathbf{1}_{\{W_l=0\}} + \mathrm{e}^{-u\cdot W_l} \mathbf{1}_{\{W_l>0\}}] = 1 - \rho + T_l(u, \vec{0}), \tag{25}$$

since $1-\rho$ is the probability that the tagged class-l customer arrives in an idle period. For the random intra-class scheduling discipline we have from [14] the following result for the transform $T_l(u, \vec{z})$.

Theorem 5.1. *[14, Theorem 8] For the random intra-class scheduling discipline, the joint transform $T_l(u, z_1, \ldots, z_K)$ satisfies*

$$\sum_{i=1}^{K} \frac{p_i}{p_l} \left(\frac{\partial}{\partial z_i} T_l(u, z_1, \ldots, z_K)\right)\left(z_i - B_i^*\left(u + \lambda - \sum_{k=1}^{K} \lambda_k z_k\right)\right) + T_l(u, z_1, \ldots, z_K) = W_l^1(u, z_1, \ldots, z_K), \tag{26}$$

where $W_l^1(u, z_1, \ldots, z_K)$ satisfies

$$\begin{aligned} &W_l^1(u, z_1, \ldots, z_K) \\ &= \sum_{i=1}^{K} \left((1-\rho)\lambda_i + \lambda p_i \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\right) \frac{B_i^*\left(\lambda - \sum_{k=1}^{K} \lambda_k z_k\right) - B_i^*\left(u + \lambda - \sum_{k=1}^{K} \lambda_k z_k\right)}{u}. \end{aligned} \tag{27}$$

In order to study the scaled waiting time, we will need to assume throughout this section that $(1-\rho)Q_k^*$ is uniform integrable, for all k. 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 family of random variables* $\{(1-\rho)Q_k^*\}$ *is uniform integrable for all* k.

We can now state our result that shows that in the limit the waiting time of a tagged class-l customer, W_l, is the product of two exponentially distributed independent random variables.

Proposition 5.2. *Let Assumption 1 be satisfied and consider the random intra-class scheduling discipline (i.e., DROS). Then, as* $\rho \uparrow 1$,

$$(1-\rho)(W_l, Q_1^*, \ldots, Q_K^*) \xrightarrow{d} (\hat{W}_l, \hat{Q}_1^*, ..., \hat{Q}_K^*) \stackrel{d}{=} (Z_l, \frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, ..., \frac{\hat{\lambda}_K}{p_K})X,$$

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

Remark 3. *Proposition 5.2 is a generalization of Kingman's result, see [17], where the asymptotic waiting time distribution is obtained for the single-class DROS queue (i.e., ROS).*

In order to prove Proposition 5.2, we will need the following three technical lemmas. The first lemma states that the scaled vector $(Q_1^*, \ldots, Q_K^*)$ has a proper limit.

Lemma 5.3. *When scaled by* $1-\rho$, *the queue length vector* $(Q_1^*, \ldots, Q_K^*)$ *has a proper limiting distribution as* $(\lambda_1, \ldots, \lambda_K) \to (\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$, *such that as* $\rho \uparrow 1$,

$$(1-\rho)(Q_1^*, \ldots, Q_K^*) \xrightarrow{d} (\hat{Q}_1^*, ..., \hat{Q}_K^*) \stackrel{d}{=} (\frac{\hat{\lambda}_1}{p_1}, \ldots, \frac{\hat{\lambda}_K}{p_K}) \cdot X,$$

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

Proof: Denote by $\tilde{Q}_i$ the class-i queue length at a service initiation epoch of a tagged class-i customers (excluding the tagged customer). By definition the following equality is satisfied:

$$\begin{aligned}
\phi_l(\mathrm{e}^{-s_1}, \ldots, \mathrm{e}^{-s_K}) &= \mathbb{E}[\mathrm{e}^{-\sum_{i=1}^K s_i \tilde{Q}_i}] \\
&= \mathbb{E}\left[\mathrm{e}^{-\sum_{i=1}^K s_i \tilde{Q}_i} \mathbf{1}_{\{W_i=0\}}\right] + \mathbb{E}\left[\mathrm{e}^{-\sum_{i=1}^K s_i \tilde{Q}_i} \mathbf{1}_{\{W_i>0\}}\right] \\
&= 1 - \rho + T_l(0, \mathrm{e}^{-s_1}, \ldots, \mathrm{e}^{-s_K}).
\end{aligned}$$

Hence, from Equation (17) we obtain that $T_l(0, \mathrm{e}^{-s_1}, \ldots, \mathrm{e}^{-s_K}) = \frac{\lambda p_i}{\lambda_i} \frac{\partial}{\partial z_i} r(z_1, ...z_K)|_{\vec{z}=\mathrm{e}^{-\vec{s}}}$. We have $\lim_{\rho\uparrow 1} \frac{\partial}{\partial z_i} r(z_1, ...z_K)|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} = \frac{\hat{\lambda}_i}{p_i} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} s_k}$ (see proof of Lemma 4.3), hence

$$\lim_{\rho\uparrow 1} T_l(0, \mathrm{e}^{-s_1}, \ldots, \mathrm{e}^{-s_K}) = \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} s_k}. \tag{28}$$

From (18) we obtain $\mathbb{E}(\mathrm{e}^{-\sum_{k=1}^K s_k \hat{N}_k}) = \hat{\lambda} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} s_k}$. Together with Equation (28) and Proposition 4.1 this concludes the proof. □

The following technical lemma characterizes the value that the function $W_l^1(u, z_1, \ldots, z_K)$, as defined in (27), takes in heavy traffic.

Lemma 5.4. *We consider the random intra-class scheduling discipline (i.e., DROS). Then, as* $\rho \uparrow 1$, *the limit* $W_l^1((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K})$ *exists and satisfies*

$$\lim_{\rho\uparrow 1} W_l^1((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K}) = \frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{k=1}^K s_k \frac{\hat{\lambda}_k}{p_k}},$$

with $1/\nu(\vec{p})$ *as given in (15).*

The result of Lemma 5.4 implies that in heavy traffic the function $\lim_{\rho\uparrow 1} W_l^1((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K})$ depends on s only through a linear combination of its components. The proof of Lemma 5.4 may be found in Appendix C. In the following lemma we show that the scaled waiting time of a class-l customer has a proper limit.

Lemma 5.5. *Let Assumption 1 be satisfied and consider the random intra-class scheduling discipline (i.e., DROS). Then, there exists a $\hat{W}_l$ such that $(1-\rho)W_l$ converges in distribution to $\hat{W}_l$ as $\rho \uparrow 1$.*

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

$$T_l(u, 1, \ldots, 1) = \mathbb{E}[\mathrm{e}^{-uW_l} 1_{\{W_l>0\}}], \tag{29}$$

and

$$\frac{\partial}{\partial z_i} T_l(u, z_1, \ldots, z_K)\Big|_{\vec{z}=1} = \mathbb{E}[Q_i^* \mathrm{e}^{-uW_l} 1_{\{W_l>0\}}]. \tag{30}$$

Now, considering Equation (26) with $\vec{z} = 1$ in heavy traffic we get:

$$\begin{aligned}
&\lim_{\rho\uparrow 1} W_l^1((1-\rho)u, 1, \ldots, 1) \qquad (31)\\
&= \lim_{\rho\uparrow 1} \sum_{i=1}^K \frac{p_i}{p_l}(1-\rho)\frac{\partial}{\partial z_i} T_l((1-\rho)u, z_1, \ldots, z_K)\Big|_{\vec{z}=1} \frac{1-B_i^*((1-\rho)u)}{(1-\rho)}\\
&\quad + \lim_{\rho\uparrow 1} T_l((1-\rho)u, 1, \ldots, 1)\\
&= \lim_{\rho\uparrow 1} \sum_{i=1}^K \frac{p_i}{p_l} u\mathbb{E}[B_i]\mathbb{E}\big[(1-\rho)Q_i^* \mathrm{e}^{-(1-\rho)uW_l}\mathbf{1}_{\{W_l>0\}}\big] + \lim_{\rho\uparrow 1}\mathbb{E}\big[\mathrm{e}^{-(1-\rho)uW_l}\mathbf{1}_{\{W_l>0\}}\big]\\
&= \lim_{\rho\uparrow 1}\mathbb{E}\left[\left(\sum_{i=1}^K \frac{p_i}{p_l} u\mathbb{E}[B_i](1-\rho)Q_i^* + 1\right)\mathrm{e}^{-(1-\rho)uW_l}\mathbf{1}_{\{W_l>0\}}\right]\\
&= \mathbb{E}\left[\lim_{\rho\uparrow 1}\left(\sum_{i=1}^K \frac{p_i}{p_l} u\mathbb{E}[B_i](1-\rho)Q_i^* + 1\right)\mathrm{e}^{-(1-\rho)uW_l}\mathbf{1}_{\{W_l>0\}}\right], \qquad (32)
\end{aligned}$$

where in the second step we used (29) and (30) and in the fourth step we used the hypothesis that $(1-\rho)Q_i^*$ is uniformly integrable (Assumption 1), [4, Theorem 3.5]. Note that $W_l^1((1-\rho)u, \mathrm{e}^{-(1-\rho)\vec{s}})$, which is defined in Equation (27), has a proper limit when $\rho \uparrow 1$, see Lemma 5.4. Since (31) converges, the same must hold for (32). Besides, $\sum_{i=1}^K \frac{p_i}{p_l} u\mathbb{E}[B_i](1-\rho)Q_i^*$ converges in distribution to $\sum_{i=1}^K \frac{p_i}{p_l} u\mathbb{E}[B_i]\frac{\hat{\lambda}_k}{p_k} X$ (see Lemma 5.3) and therefore, we conclude that the waiting time of an arbitrary class-l customer in heavy traffic converges in distribution to some random variable $\hat{W}_l$. □

From Lemma 5.4, we note that (32) should in fact be independent of u. It can be checked that in case $(1-\rho)(W_l, Q_1^*, \ldots, Q_K^*)$ is distributed as $X(Z_l, \frac{\hat{\lambda}_1}{p_1}, \ldots, \frac{\hat{\lambda}_K}{p_K})$, as we want to show, see Proposition 5.2, this is indeed satisfied.
We can now prove Proposition 5.2 which consists in finding $T_l(\cdot)$ by solving Equation (26) after the heavy-traffic scaling.

Proof of Proposition 5.2: We know by Lemma 5.5 that there is a random variable $\hat{W}_l$ such that $(1-\rho)W_l$ converges in distribution to $\hat{W}_l$. Hence, we can define the function $\hat{T}_l(u, \vec{s})$ as follows:

$$\begin{aligned}
\hat{T}_l(u, \vec{s}) \quad &:= \mathbb{E}[\mathrm{e}^{-u\hat{W}_l}\mathrm{e}^{-\sum_{i=1}^K s_i\hat{Q}_i^*}]\\
&= \lim_{\rho\uparrow 1}\mathbb{E}[\mathrm{e}^{-(1-\rho)uW_l}\mathrm{e}^{-(1-\rho)s_1Q_1^*}\ldots\mathrm{e}^{-(1-\rho)s_KQ_K^*}]\\
&= \lim_{\rho\uparrow 1} T_l((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K}).
\end{aligned}$$

We will evaluate Equation (26) in the point $(u, \vec{z}) = (u(1-\rho), \mathrm{e}^{-(1-\rho)\vec{s}})$ as $\rho \uparrow 1$. We first focus on the first term. We have

$$\begin{aligned}
&\lim_{\rho\uparrow 1}(1-\rho)\frac{\partial}{\partial z_i}T_l(u,\vec{z})\Big|_{u=(1-\rho)u,\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} \\
&= \lim_{\rho\uparrow 1}(1-\rho)\mathbb{E}(\frac{Q_i^*\mathrm{e}^{-(1-\rho)uW_l}\mathrm{e}^{-(1-\rho)s_1Q_1^*}\cdots\mathrm{e}^{-(1-\rho)s_i(Q_i^*-1)}\cdots\mathrm{e}^{-(1-\rho)s_KQ_K^*}}{\mathrm{e}^{-(1-\rho)s_i}}) \\
&= \mathbb{E}[\hat{Q}_i^*\mathrm{e}^{-u\hat{W}_l}\mathrm{e}^{-s_1\hat{Q}_1^*}\cdots\mathrm{e}^{-s_K\hat{Q}_K^*}] = -\frac{\partial}{\partial s_i}\hat{T}_l(u,\vec{s}),
\end{aligned} \tag{33}$$

where in the second step we used Assumption 1 and [4, Theorem 3.5].
Moreover, realize that $\hat{T}_l(u,\vec{s}) = \mathbb{E}[\mathrm{e}^{-u\hat{W}_l}\mathrm{e}^{-\sum_{i=1}^K s_i\hat{Q}_i^*}]$ depends on $\vec{s} = (s_1,\ldots,s_K)$ only through $y = \sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k}s_k$ (see Lemma 5.3). Thus, we will write $\hat{T}_l(u,\vec{s}) = \hat{T}_l(u,y)$ and by the chain rule:

$$\frac{\partial}{\partial s_i}\hat{T}_l(u,\vec{s}) = \frac{\partial}{\partial s_i}\hat{T}_l(u,y) = \frac{\partial}{\partial y}\hat{T}_l(u,y)\frac{\partial y}{\partial s_i} = \frac{\hat{\lambda}_i}{p_i}\frac{\partial}{\partial y}\hat{T}_l(u,y)\Big|_{y=\sum_{k=1}^K s_k\frac{\hat{\lambda}_k}{p_k}}. \tag{34}$$

Then, taking the heavy-traffic limit in Equation (26) and using Equations (33), (34), Lemma 5.4 and the relation $\lim_{\rho\uparrow 1}\dfrac{(\mathrm{e}^{-(1-\rho)s_i} - B_i^*((1-\rho)u+\lambda-\sum_{k=1}^K\lambda_k\mathrm{e}^{-(1-\rho)s_k}))}{1-\rho} = -s_i + \mathbb{E}[B_i]\left(u+\sum_{k=1}^K\hat{\lambda}_k s_k\right)$, which follows from l'Hopital's rule, we arrive to the following ordinary differential equation (ODE):

$$-\sum_{i=1}^K\frac{p_i}{p_l}\frac{\hat{\lambda}_i}{p_i}\frac{\partial\hat{T}_l(u,y)}{\partial y}\Big|_{y=\sum_{i=1}^K s_i\frac{\hat{\lambda}_i}{p_i}}(-s_i+\mathbb{E}[B_i](u+\sum_{k=1}^K\hat{\lambda}_k s_k)) + \hat{T}_l(u,y) = \frac{\nu(\vec{p})}{\nu(\vec{p})+y}.$$

Since $\sum_{i=1}^K\hat{\lambda}_i\left(-s_i+\mathbb{E}[B_i]\left(u+\sum_{k=1}^K\hat{\lambda}_k s_k\right)\right) = u$, the latter can be written as

$$-\frac{u}{p_l}\frac{\partial\hat{T}_l(u,y)}{\partial y}\Big|_{y=\sum_{k=1}^K s_k\frac{\hat{\lambda}_k}{p_k}} + \hat{T}_l(u,y) = \frac{\nu(\vec{p})}{\nu(\vec{p})+y}. \tag{35}$$

The solution of the ODE (35) is

$$\hat{T}_l(u,y) = \frac{p_l}{u}\mathrm{e}^{\frac{p_l}{u}y}\int_y^\infty \mathrm{e}^{-\frac{p_l}{u}x}\frac{\nu(\vec{p})}{\nu(\vec{p})+x}\mathrm{d}x, \tag{36}$$

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

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{-uZ_l\cdot X - s_1\frac{\lambda_1}{p_1}X - \ldots - s_K\frac{\lambda_K}{p_K}X}] = \mathbb{E}[\mathrm{e}^{-uZ_l\cdot X-\sum_{k=1}^K s_k\frac{\lambda_k}{p_k}X}] = \mathbb{E}[\mathrm{e}^{-uZ_l\cdot X - yX}] \\
&= \mathbb{E}[\mathbb{E}[\mathrm{e}^{-uz_lX-yX}|z_l=Z_l]] = \mathbb{E}[\frac{\nu(\vec{p})}{\nu(\vec{p})+uZ_l+y}] = \int_0^\infty\eta\mathrm{e}^{-\eta z_l}\frac{\nu(\vec{p})}{\nu(\vec{p})+uz_l+y}\mathrm{d}z_l \\
&= \frac{1}{u}\int_{\nu(\vec{p})+y}^\infty\eta\mathrm{e}^{-\eta\frac{x-\nu(\vec{p})-y}{u}}\frac{\nu(\vec{p})}{x}\mathrm{d}x = \frac{\nu(\vec{p})\eta}{u}\mathrm{e}^{\eta\frac{\nu(\vec{p})+y}{u}}\int_{\nu(\vec{p})+y}^\infty\mathrm{e}^{-\eta\frac{x}{u}}\frac{1}{x}\mathrm{d}x \\
&= \frac{\nu(\vec{p})\eta}{u}\mathrm{e}^{\eta\frac{\nu(\vec{p})+y}{u}}\int_{\eta\frac{\nu(\vec{p})+y}{u}}^\infty\frac{\mathrm{e}^{-l}}{l}\mathrm{d}l.
\end{aligned} \tag{37}$$

Making the change of variable $z = p_l\frac{\nu(\vec{p})+x}{u}$ in Equation (36) we obtain

$$\hat{T}_l(u,y) = \nu(\vec{p})\frac{p_l}{u}\mathrm{e}^{p_l\frac{y}{u}}\int_y^\infty\mathrm{e}^{-p_l\frac{x}{u}}\frac{1}{\nu(\vec{p})+x}\mathrm{d}x = p_l\frac{\nu(\vec{p})}{u}\mathrm{e}^{p_l\frac{\nu(\vec{p})+y}{u}}\int_{p_l\frac{\nu(\vec{p})+y}{u}}^\infty\mathrm{e}^{-z}\frac{1}{z}\mathrm{d}z.$$

Hence, it coincides with the Laplace Transform of $(Z_l \cdot X, \frac{\lambda_1}{p_1}X, \ldots, \frac{\lambda_K}{p_K}X)$ obtained in (37) with $p_l = \eta$. Since the Laplace transform of a probability distribution is unique, (uniqueness theorem, [9]), we conclude that $(1-\rho)(W_l, Q_1^*, \ldots, Q_K^*)$ converges in distribution to $(Z_l \cdot X, \frac{\lambda_1}{p_1}X, \ldots, \frac{\lambda_K}{p_K}X)$, where, as we have previously mentioned, Z_l and X are exponentially distributed independent random variables with $\mathbb{E}(Z_l) = 1/p_l$ and $\mathbb{E}(X) = 1/\nu(\vec{p})$. □

6 Optimal selection of the weights

In this section we show how the results of Proposition 4.1 and Proposition 5.2 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 minimize the moments of the waiting time of an arbitrary customer.

6.1 Holding Cost

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

$$\sum_{k=1}^K c_k \mathbb{E}[\hat{N}_k(\vec{p})] = \frac{\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} c_k}{2\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k}\mathbb{E}[B_k]} \sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k^2], \tag{38}$$

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 $\frac{c_i}{\mathbb{E}[B_i]}$. This can be seen as an extension of the $c\mu$-rule for the heavy-traffic setting [10].

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

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^K c_k N_k(\vec{p}) \geq_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^K c_k N_k(\vec{q}),$$

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 $(1-\rho)\sum_{k=1}^K c_k N_k(\vec{p})$ converges in distribution to an exponentially distributed random variable with mean as stated in (38). Since exponentially distributed random variables are stochastically ordered according to their means, it only remains to check that

$$\frac{\sum_{k=1}^K \frac{c_k\hat{\lambda}_k}{p_k}}{\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k}\mathbb{E}[B_k]} \geq \frac{\sum_{k=1}^K \frac{c_k\hat{\lambda}_k}{q_k}}{\sum_{k=1}^K \frac{\hat{\lambda}_k}{q_k}\mathbb{E}[B_k]}.$$

This holds since

$$(\sum_{k=1}^K \frac{c_k\hat{\lambda}_k}{p_k})(\sum_{k=1}^K \frac{\hat{\lambda}_k}{q_k}\mathbb{E}[B_k]) = \sum_{k,i:k\neq i} \hat{\lambda}_k\hat{\lambda}_i(\frac{1}{p_k q_i}c_k\mathbb{E}[B_i] + \frac{1}{p_i q_k}c_i\mathbb{E}[B_k]) + \sum_{k=1}^K \hat{\lambda}_k^2 \frac{1}{p_k q_k}c_k\mathbb{E}[B_k]$$

$$\geq \sum_{k,i:k\neq i} \hat{\lambda}_k\hat{\lambda}_i(\frac{1}{p_i q_k}c_k\mathbb{E}[B_i] + \frac{1}{p_k q_i}c_i\mathbb{E}[B_k]) + \sum_{k=1}^K \hat{\lambda}_k^2 \frac{1}{p_k q_k}c_k\mathbb{E}[B_k] = (\sum_{k=1}^K \frac{c_k\hat{\lambda}_k}{q_k})(\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k}\mathbb{E}[B_k]).$$

Here we used that $c_i\mathbb{E}[B_k](\frac{1}{p_i q_k} - \frac{1}{p_k q_i}) \geq c_k\mathbb{E}[B_i](\frac{1}{p_i q_k} - \frac{1}{p_k q_i})$, which follows from the fact that $\frac{p_i}{p_k} \leq \frac{q_i}{q_k}$ and $\frac{c_i}{\mathbb{E}[B_i]} \geq \frac{c_k}{\mathbb{E}[B_k]}$, for $i \leq k$. □

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 limit of the scaled waiting time of a tagged class-k customer, $\hat{W}_k$. From Proposition 5.2 we know that

$$\hat{W}_k \stackrel{d}{=} X \cdot Z_k, \tag{39}$$

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 (39) 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

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

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

$$\mathbb{E}[\hat{W}^m] = \sum_{k=1}^K \frac{\hat{\lambda}_k}{\hat{\lambda}}\mathbb{E}[\hat{W}_k^m] = (m!)^2 \left(\sum_{k=1}^K \frac{\hat{\lambda}_k}{\hat{\lambda} p_k^m}\right)\left(\frac{\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k^2]}{2\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k}\mathbb{E}[B_k]}\right)^m. \tag{40}$$

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 $\mathbb{E}[\hat{W}(\vec{p})] = \frac{1}{\lambda}\sum_{k=1}^K \mathbb{E}[\hat{N}_k(\vec{p})]$. 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 $\mathbb{E}[B_i], i = 1, \ldots, K$.

Corollary 6.2. *Without loss of generality we assume that the classes are ordered such that* $\mathbb{E}[B_1] \leq \ldots \leq \mathbb{E}[B_K]$. *If* $\frac{p_j}{p_{j+1}} \leq \frac{q_j}{q_{j+1}}$, *for all* $j = 1, \ldots, K-1$, *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* ρ, *not necessarily close to one. This can be seen as follows. The expression for the mean waiting time for K=2 is the following:*

$$\begin{aligned}\mathbb{E}[W(\vec{p})] &= \sum_{i=1}^2 \frac{\lambda_i}{\lambda}\mathbb{E}[W_i] \\ &= \frac{\lambda_1\mathbb{E}[B_1^2] + \lambda_2\mathbb{E}[B_2^2]}{2\lambda}\frac{\lambda_1(1-\rho p_1) + \lambda_2(1-\rho p_2)}{(1-\rho_1-p_2\rho_2)(1-\rho_2-p_1\rho_1) - p_1p_2\rho_1\rho_2},\end{aligned} \tag{41}$$

where the expression of $\mathbb{E}[W_i], i = 1, 2$, *was obtained in [14], Equation (38). Without loss of generality we assume that* $p_1 + p_2 = 1$. *Then, taking the derivative of (41) 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 [14] for any value of K, i.e., for any number of classes of customers. We choose the weights such that $\frac{p_j}{p_{j+1}} = \frac{1}{r}, \forall j$. *In the figures we chose exponentially distributed service requirements, however, the monotonicity observed 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 for these examples that 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 limit of the scaled waiting time,* $\mathbb{E}[\hat{W}^m(\vec{p})]$, *is minimized in* $\vec{p^*} = (p_1^*, \ldots, p_K^*)$, *with*

$$p_k^* := \frac{1/\mathbb{E}[B_k]^{1/m-1}}{\sum_{i=1}^K 1/\mathbb{E}[B_i]^{1/m-1}}, \tag{42}$$

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

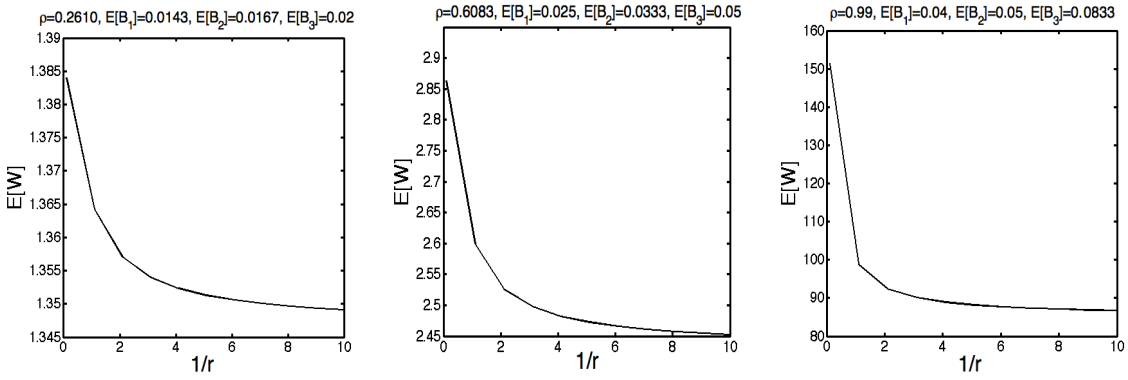

Figure 1: The mean waiting time for three classes of customers in the system, K=3, under DROS for the loads $\rho = \ 0.2610, \rho = \ 0.6083$ and $\rho = \ 0.99$, respectively. The horizontal axis corresponds to $\frac{1}{r} = \frac{p_j}{p_{j+1}}, j = \ 1, \ldots, K-1$.

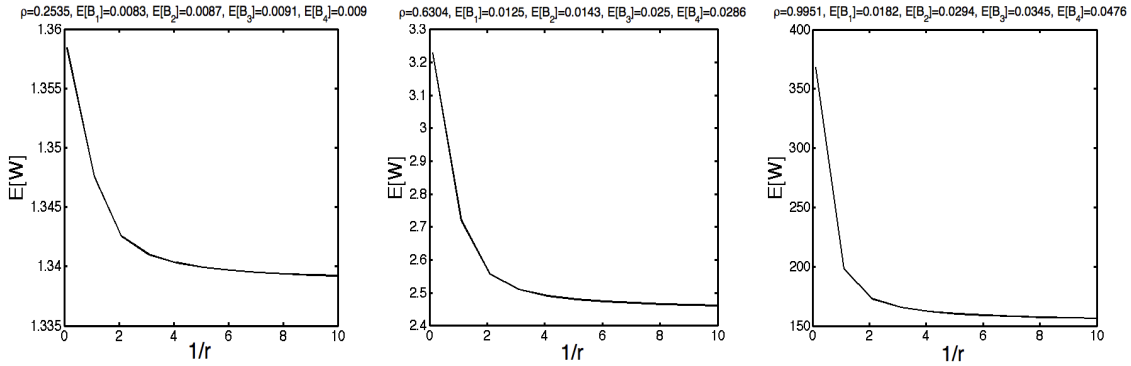

Figure 2: The mean waiting time for four classes of customers in the system, K=4, under DROS for the loads $\rho = \ 0.2535, \rho = \ 0.6304$ and $\rho = \ 0.9951$, respectively. The horizontal axis corresponds to $\frac{1}{r} = \frac{p_j}{p_{j+1}}, j = \ 1, \ldots, K-1$.

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

$$\frac{\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k]^{m/m-1}}{\hat{\lambda}(\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k]^{m/m-1})^m} \cdot \left(\frac{\sum_{j=1}^K \hat{\lambda}_j \mathbb{E}[B_j^2]}{2}\right)^m \leq \frac{\sum_{k=1}^K \frac{\hat{\lambda}_k}{\hat{\lambda} p_k^m}}{(\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k])^m} \cdot \left(\frac{\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k^2]}{2}\right)^m,$$

which follows by definition. This is equivalent to

$$\frac{1}{(\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k]^{m/m-1})^{m-1}} \leq \frac{\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k^m}}{(\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k])^m}$$

and rewriting it we obtain

$$(\sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k} \mathbb{E}[B_k])^m \leq \sum_{k=1}^K \frac{\hat{\lambda}_k}{p_k^m} (\sum_{k=1}^K \hat{\lambda}_k \mathbb{E}[B_k]^{m/m-1})^{m-1}. \tag{43}$$

The latter holds by Hölder's inequality. □

Remark 5. *By Equation (42) we get that the ratio of two optimal weights is the following:*

$$\frac{p_k^*}{p_j^*} = \left(\frac{\mathbb{E}[B_j]}{\mathbb{E}[B_k]}\right)^{1/(m-1)}. \tag{44}$$

In general, the optimal choice for the weights is non trivial. However, note that when $m \to 1$ we deduce that under the optimal weights (for the m-th moment) 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 cμ-rule states. In addition, when $m \to \infty$, from (44) we see that the ratio of the optimal weights converges to 1. This implies that as m gets larger, it becomes optimal (for the m-th moment) 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 $\frac{\hat{N}_2}{\hat{N}_1} = \frac{p_1}{\hat{\lambda}_1}\frac{\hat{\lambda}_2}{p_2} \approx 3.1$, 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 (41)) and $(1-\rho)^2\mathbb{E}[N^2]$ (obtained by simulation) for different values of the load ρ. 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

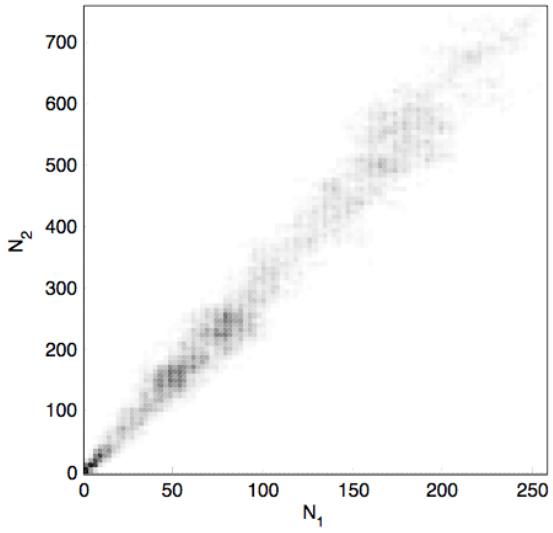

Figure 3: Joint queue length probability. The darkness of the points specifies the probability of being into a particular state. The darker the point is, the higher the probability of being in that state.

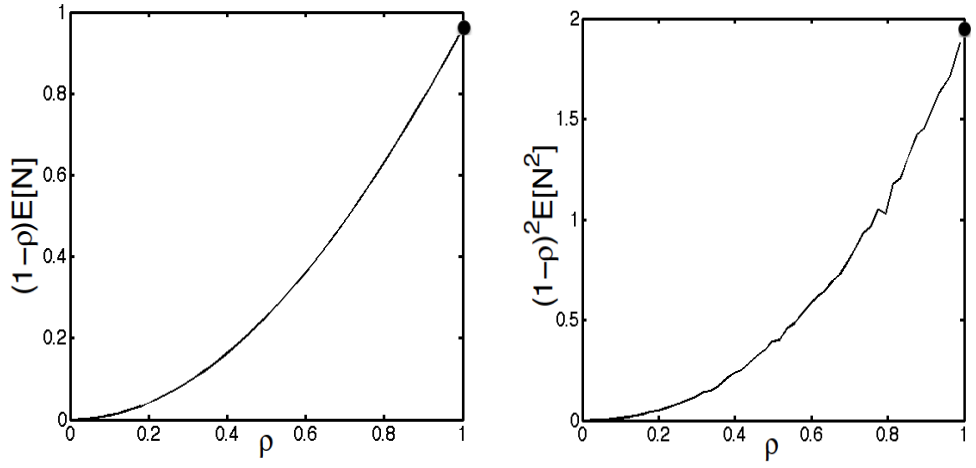

Figure 4: First and second moment of the scaled queue length obtained for different values of the load ρ. The dots in both pictures are calculated by using (14) giving as a result $\mathbb{E}[\hat{N}] = 0.9589$ and $\mathbb{E}[\hat{N}^2] = 1.9510$.

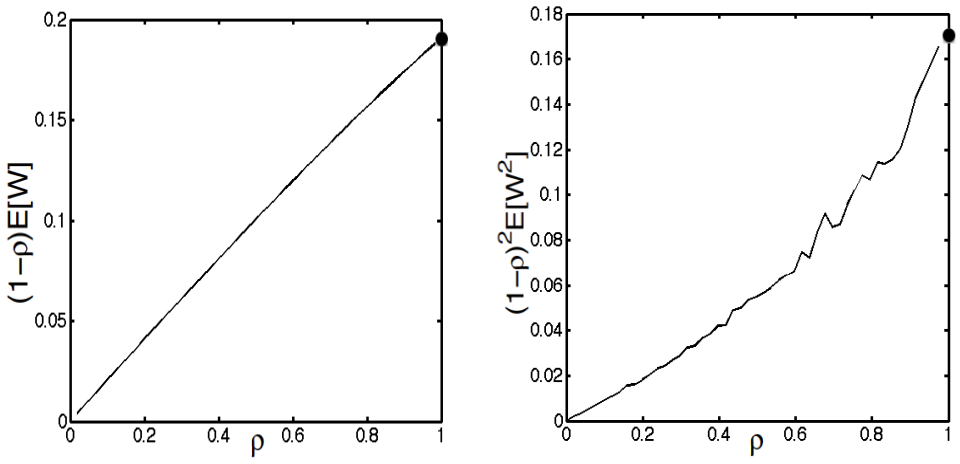

Figure 5: First and second moment of the scaled waiting time obtained for different values of the load ρ. The dots in both pictures are calculated by using Equation (40) giving as a result $\mathbb{E}[\hat{W}] = 0.1906$ and $\mathbb{E}[\hat{W}^2] = 0.1713$.

second moment of the limit of the scaled queue length, i.e., $\mathbb{E}[\lim_{\rho\uparrow 1}(1-\rho)^m N^m] = \mathbb{E}[\hat{N}^m], m = 1, 2$, using (14), 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., $\lim_{\rho\uparrow 1}\mathbb{E}[(1-\rho)^m N_k^m] = \mathbb{E}[\lim_{\rho\uparrow 1}(1-\rho)^m N_k^m], m = 1, 2$.
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 [4, Theorem 3.5]), as assumed in Assumption 1.

In Figure 5 we plot $(1-\rho)\mathbb{E}[W]$ (using Equation (41)) and $(1-\rho)^2\mathbb{E}[W^2]$ (obtained by simulation) for different values of the load ρ. The simulation setting is the same as the one used for the queue length. We calculate the value of Equation (40) 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 (40), 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 (41) it is easy to see that it indeed converges to the heavy-traffic limit as characterized by Equation (40) 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 (42)). 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.

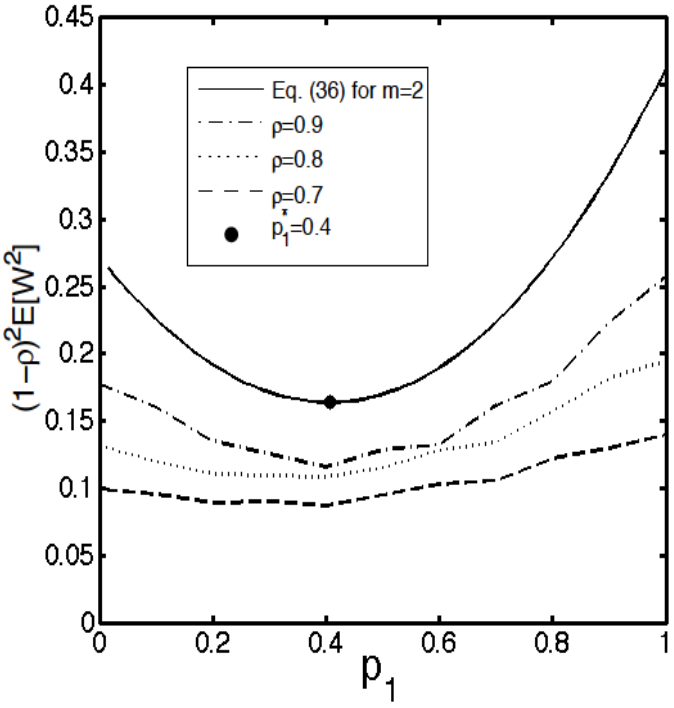

Figure 6: The second moment $\mathbb{E}[(1-\rho)^2 W^2]$ for different values of the weights $(p_1, p_2) = (p_1, 1-p_1)$.

References

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

[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] A. Banerjea and S. Keshav. Queueing delays in rate controlled ATM networks. In *Proceedings of INFOCOM 1993*, pages 547–556 vol.2, 1993.

[4] P. Billingsley. *Convergence of Probability Measures*. Wiley, 1999.

[5] 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.

[6] 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.

[7] 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.

[8] M. Bramson and J.G.Dai. Heavy traffic limits for some queueing networks. *Annals of Applied Probability*, 11:49–90, 2001.

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

[10] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. Imperial College Press, 2010.

[11] 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.

[12] M. Haviv and J. van der Wal. Waiting times in queues with relative priorities. *Operations Research Letters*, 35:591–594, 2007.

[13] J. Kim. Queue length distribution in a queue with relative priorities. *Bull. Korean Math. Soc.*, 46:107–116, 2009.

[14] 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.

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

[16] J.F.C. Kingman. On queues in which customers are served in random order. *Proc. Cambridge Philos. Soc.*, 58:79–91, 1962.

[17] J.F.C. Kingman. Queue disciplines in heavy traffic. *Mathematics of Operations Research*, 7(2):262–271, 1982.

[18] W.H. Mather, N.A. Cookson, J. Hasty, L.S. Tsimring, and R.J. Williams. Correlation resonance generated by coupled enzymatic processing. *Biophysical Journal*, 99:3172–3181, 2010.

[19] C. Palm. Waiting times with random served queue. *Tele1 (English edition; original 1938)*, 1–107, 1957.

[20] 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.

[21] 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

$$V^{\text{dep}} = \sum_{k=1}^{K} \sum_{h=1}^{Q_k} B_{k,h},$$

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

For $\epsilon' > 0$ we have

$$
\begin{aligned}
&\mathbb{P}\big((1-\rho)Q_k \geq M\big) = \mathbb{P}\bigg(Q_k \geq \frac{M}{(1-\rho)}\bigg) \\
&\leq \mathbb{P}\bigg(\sum_{h=1}^{Q_k} B_{k,h} \geq \sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h}\bigg) \\
&\leq \mathbb{P}\bigg((1-\rho)\sum_{k=1}^{K}\sum_{h=1}^{Q_k} B_{k,h} \geq M\frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h}\bigg) \\
&= \mathbb{P}\bigg(\frac{(1-\rho)V^{\mathrm{dep}}}{M} - \mathbb{E}[B_k] \geq \frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k]\bigg) \\
&= \mathbb{P}\bigg(\frac{(1-\rho)V^{\mathrm{dep}}}{M} - \mathbb{E}[B_k] \geq \frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k], \frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k] > -\epsilon\bigg) \\
&\quad + \mathbb{P}\bigg(\frac{(1-\rho)V^{\mathrm{dep}}}{M} - \mathbb{E}[B_k] \geq \frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k]\bigg|\frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}(B_k) \leq -\epsilon\bigg) \\
&\quad \cdot \mathbb{P}\bigg(\frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k] \leq -\epsilon\bigg) \\
&\leq \mathbb{P}\bigg(\frac{(1-\rho)V^{\mathrm{dep}}}{M} - \mathbb{E}[B_k] > -\epsilon\bigg) + \tilde{\epsilon} \\
&= \mathbb{P}\bigg((1-\rho)V^{\mathrm{dep}} \geq M(\mathbb{E}[B_k] - \epsilon)\bigg) + \tilde{\epsilon} \\
&< \bar{\epsilon} + \tilde{\epsilon} = \epsilon^{'}, \text{ for } \rho \text{ close enough to 1 and } M \text{ large enough.}
\end{aligned} \tag{45}
$$

In the fifth step we used that $\frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h}$ converges in distribution to $\mathbb{E}[B_k]$ as $\rho \uparrow 1$, hence $\mathbb{P}\bigg(\frac{(1-\rho)}{M}\sum_{h=1}^{\lfloor M/(1-\rho)\rfloor} B_{k,h} - \mathbb{E}[B_k] \leq -\epsilon\bigg) \leq \bar{\epsilon}$, for ρ close enough to 1. 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 [15], that is $\forall \bar{\epsilon}\ \exists M'$ such that $\mathbb{P}((1-\rho)V^{\mathrm{dep}} \geq M') < \bar{\epsilon}$. From (45) we conclude that $(1-\rho)(Q_1, Q_2, \ldots, Q_K)$ is tight.

Appendix B: Proof of Lemma 3.6

The proof of Lemma 3.6 is based on the proof of Lemma 3 in [20]. We have

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

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

$$
\vec{f}(0) = \vec{s} \quad \text{and} \quad \frac{\partial f_i(u)}{\partial u} = F_i(\vec{f}(u)), \quad \text{for all } i \text{ and } u \geq 0. \tag{46}
$$

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

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

which follows from the chain rule, Lemma 3.5, and Equation (46). 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 $c \cdot \vec{s}^* \geq 0$ as $u \to \infty$.
From (10) we get that (46) can be written as $\vec{f}(0) = \vec{s}$ and $\vec{f}'(u)^T = A\vec{f}(u)^T$ with

$$A = \begin{pmatrix} p_1(-1+\mathbb{E}[B_1]\hat{\lambda}_1) & p_1\mathbb{E}[B_1]\hat{\lambda}_2 & \cdots & p_1\mathbb{E}[B_1]\hat{\lambda}_K \\ p_2\mathbb{E}[B_2]\hat{\lambda}_1 & p_2(-1+\mathbb{E}[B_2]\hat{\lambda}_2) & \cdots & p_2\mathbb{E}[B_2]\hat{\lambda}_K \\ \vdots & \vdots & \ddots & \vdots \\ p_K\mathbb{E}[B_K]\hat{\lambda}_1 & p_K\mathbb{E}[B_K]\hat{\lambda}_2 & \cdots & p_K(-1+\mathbb{E}[B_K]\hat{\lambda}_K) \end{pmatrix}. \tag{47}$$

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 $\lim_{u\to\infty} \vec{g}(u) = \vec{0}$ 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.
The following technical lemma is used in the proof of Lemma 3.6.

Lemma 7.1. *Consider the matrix A as defined in (47). 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 $\vec{\eta} = (\eta_1, \ldots, \eta_K) \geq \vec{0}$ with $\sum_{j=1}^{K} \eta_j = 1$ such that $\vec{s}^* = (s_1^*, \ldots, s_K^*)$ with $s_j^* := \frac{p_j}{\hat{\lambda}_j}\eta_j$ is an eigenvector of A corresponding to the eigenvalue 0, and $\vec{s}^* \in H_1$.*

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

$$S = DAD^{-1} = \begin{pmatrix} p_1(-1+\mathbb{E}[B_1]\hat{\lambda}_1) & p_2\hat{\lambda}_1\mathbb{E}[B_1] & \cdots & p_K\hat{\lambda}_1\mathbb{E}[B_1] \\ p_1\hat{\lambda}_2\mathbb{E}[B_2] & p_2(-1+\mathbb{E}[B_2]\hat{\lambda}_2) & \cdots & p_K\hat{\lambda}_2\mathbb{E}[B_2] \\ \vdots & \vdots & \ddots & \vdots \\ p_1\hat{\lambda}_K\mathbb{E}[B_K] & p_2\hat{\lambda}_K\mathbb{E}[B_K] & \cdots & p_K(-1+\mathbb{E}[B_K]\hat{\lambda}_K) \end{pmatrix}. \tag{48}$$

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 $\sum_{k=1}^{K} \eta_k = 1$. 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 $A\vec{s}^{*^T} = D^{-1}DAD^{-1}\vec{\eta}^T = D^{-1}S\vec{\eta}^T = \vec{0}^T$.

□

Appendix C: Proof of Lemma 5.4

Taking $(u, z_1, \ldots, z_K) = ((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K})$ in (27) we get

$$W_l^1\big((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K}\big) =$$
$$\sum_{i=1}^{K}\big((1-\rho)\lambda_i + \lambda p_i \frac{\partial}{\partial z_i} r(z_1, \ldots, z_K)\Big|_{z_i = \mathrm{e}^{-(1-\rho)s_i}}\big) \cdot$$
$$\frac{B_i^*\big(\lambda - \sum_{k=1}^{K}\lambda_k \mathrm{e}^{-(1-\rho)s_k}\big) - B_i^*\big((1-\rho)u + \lambda - \sum_{k=1}^{K}\lambda_k \mathrm{e}^{-(1-\rho)s_k}\big)}{(1-\rho)u}.$$

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

$$\begin{aligned}
&\lim_{\rho\uparrow 1} W_l^1\big((1-\rho)u, \mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_K}\big) \\
&= -\frac{1}{u}\sum_{i=1}^{K}\left(\hat{\lambda}_i \cdot 0 + \hat{\lambda} p_i \frac{\partial}{\partial s_i}\hat{r}(s_1, \ldots, s_K)\left(B_i^{*'}(0)\left(-\sum_{k=1}^{K}\hat{\lambda}_k s_k\right) - B_i^{*'}(0)\left(-u - \sum_{k=1}^{K}\hat{\lambda}_k s_k\right)\right)\right) \\
&= \hat{\lambda}\sum_{i=1}^{K}\mathbb{E}[B_i] p_i \frac{\hat{\lambda}_i}{p_i}\frac{\mathrm{d}}{\mathrm{d}v}\hat{r}^*(v)\Big|_{v=\sum_{k=1}^{K}\frac{\hat{\lambda}_k}{p_k}s_k} = \hat{\lambda}\frac{\mathrm{d}}{\mathrm{d}v}\hat{r}^*(v)\Big|_{v=\sum_{k=1}^{K}\frac{\hat{\lambda}_k}{p_k}s_k},
\end{aligned}$$

with $\hat{r}(\vec{s})$ as defined in (6) and $\hat{r}^*(\vec{s})$ as defined in the proof of Lemma 4.3. The result now follows from Equation (18), together with the fact that the latter is equal to $\frac{\nu(\vec{p})}{\nu(\vec{p}) + \sum_{k=1}^{K} s_k \frac{\hat{\lambda}_k}{p_k}}$ (since $(\hat{N}_1, \ldots, \hat{N}_K) \overset{d}{=} X \cdot (\frac{\lambda_1}{p_1}, \ldots, \frac{\lambda_K}{p_K})$, with X exponentially distributed with mean $1/\nu(\vec{p})$).

Appendix D: Solution of the ODE (35)

The solution of (35) is given by the sum of the solution of the homogeneous case, $\hat{T}_l^H(u, y)$, and a particular solution, $\hat{T}_l^P(u, y)$. The homogeneous solution is given by:

$$\hat{T}_l^H(u, y) = C(u)\mathrm{e}^{\frac{p_l}{u}y}, \tag{49}$$

where $C(u)$ is an arbitrary function of u. In order to find the particular solution we rewrite (35) as

$$\frac{\partial \hat{T}_l^P(u, y)}{\partial y}\Big|_{y=\sum_{k=1}^{K} s_k \frac{\hat{\lambda}_k}{p_k}} - \frac{p_l}{u}\hat{T}_l^P(u, y) = -\frac{p_l}{u}\frac{\nu(\vec{p})}{\nu(\vec{p}) + y}. \tag{50}$$

Let us solve the new equation using the integrating factor technique. In order to do so, we define the function $\mu(y) = \mathrm{e}^{-\frac{p_l}{u}y}$ and multiply (50) by it. The derivative of $\mu(y)$ satisfies $\frac{\mathrm{d}\mu(y)}{\mathrm{d}y} = -\mu(y)\frac{p_l}{u}$. Then, our equation becomes

$$\begin{aligned}
-\mu(y)\frac{p_l}{u}\frac{\nu(\vec{p})}{\nu(\vec{p}) + y} &= \mu(y)\frac{\partial \hat{T}_l^P(u, y)}{\partial y}\Big|_{y=\sum_{k=1}^{K} s_k \frac{\hat{\lambda}_k}{p_k}} - \mu(y)\frac{p_l}{u}\hat{T}_l^P(u, y) \\
&= \mu(y)\frac{\partial \hat{T}_l^P(u, y)}{\partial y}\Big|_{y=\sum_{k=1}^{K} s_k \frac{\hat{\lambda}_k}{p_k}} + \hat{T}_l^P(u, y)\frac{\mathrm{d}\mu(y)}{\mathrm{d}y} \\
&= \frac{\partial}{\partial y}(\mu(y)\hat{T}_l^P(u, y)),
\end{aligned}$$

which can be solved by integration. Integrating each side with respect to y gives us a particular solution for (35), which is,

$$\hat{T}_l^P(u, y) = -\frac{p_l}{u}\mathrm{e}^{\frac{p_l}{u}y}\int_0^y \mathrm{e}^{-\frac{p_l}{u}x}\frac{\nu(\vec{p})}{\nu(\vec{p}) + x}\mathrm{d}x,$$

which is identically written as

$$\hat{T}_l^P(u,y) = \frac{p_l}{u} \mathrm{e}^{\frac{p_l}{u}y} \int_y^\infty \mathrm{e}^{-\frac{p_l}{u}x} \frac{\nu(\vec{p})}{\nu(\vec{p})+x} \mathrm{d}x. \tag{51}$$

In conclusion, the general solution of the ODE (35) is given by

$$\hat{T}_l(u,y) = \hat{T}_l^H(u,y) + \hat{T}_k^P(u,y) = C(u)\mathrm{e}^{\frac{p_l}{u}y} + \frac{p_l}{u} \mathrm{e}^{\frac{p_l}{u}y} \int_y^\infty \mathrm{e}^{-\frac{p_l}{u}x} \frac{\nu(\vec{p})}{\nu(\vec{p})+x} \mathrm{d}x. \tag{52}$$

We will now show that the constant $C(u)$ is equal to zero. First, note that $\hat{T}_l^H(u,y) \to \infty$ as $y \to \infty$. Second, since $\hat{T}_l(0,y) = \frac{\nu(\vec{p})}{\nu(\vec{p})+y}$ (from Equation (24) and Lemma 5.3) it is immediate that $\hat{T}_l(u,y)$ converges to 0 when $y = \sum_{k=1}^{K} s_k \frac{\hat{\lambda}_k}{p_k} \to \infty$. Moreover, if we take the particular solution (51), applying l'Hopital's rule for $y \to \infty$ we obtain that it also converges to zero, namely:

$$\lim_{y\to\infty} \hat{T}_l(u,y) = \lim_{y\to\infty} \frac{\nu(\vec{p})p_l}{u} \frac{\int_y^\infty \mathrm{e}^{-\frac{p_l}{u}x} \frac{1}{\nu(\vec{p})+x} \mathrm{d}x}{\mathrm{e}^{-\frac{p_l}{u}y}} = \lim_{y\to\infty} \frac{\nu(\vec{p})p_l}{u} \frac{-\mathrm{e}^{-\frac{p_l}{u}y} \frac{1}{\nu(\vec{p})+y}}{\frac{-p_l}{u}\mathrm{e}^{-\frac{p_l}{u}y}} = \lim_{y\to\infty} \frac{\nu(\vec{p})}{\nu(\vec{p})+y} = 0.$$

Then, the necessary condition for $C(u)\mathrm{e}^{\frac{p_l}{u}y}$ to converge to zero as $y \to \infty$ is $C(u) = 0$. As a consequence, we conclude that the solution of (35) is

$$\hat{T}_l(u,y) = \frac{p_l}{u} \mathrm{e}^{\frac{p_l}{u}y} \int_y^\infty \mathrm{e}^{-\frac{p_l}{u}x} \frac{\nu(\vec{p})}{\nu(\vec{p})+x} \mathrm{d}x.$$