Interpolation approximations for the steady-state distribution in multi-class resource-sharing systems

A. Izagirre[b,e], U. Ayesta[b,c,d,e], I.M. Verloop[a,e]

[a]CNRS, IRIT, 2 rue C. Camichel, 31071 Toulouse, France

[b]CNRS, LAAS, 7 avenue du colonel Roche, 31400 Toulouse, France

[c]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

[d]UPV/EHU, Univ. of the Basque Country, 20018 Donostia, Spain

[e]Univ. de Toulouse, INP, INSA, LAAS, 31400 Toulouse, France

Abstract

We consider a single-server multi-class queue that implements relative priorities among customers of the various classes. The discipline might serve one customer at a time in a non-preemptive way, or serve all customers simultaneously. The analysis of the steady-state distribution of the queue-length and the waiting time in such systems is complex and closed-form results are available only in particular cases. We therefore set out to develop approximations for the steady-state distribution of these performance metrics. We first analyze the performance in light traffic. Using known results in the heavy-traffic regime, we then show how to develop an interpolation-based approximation that is valid for any load in the system. An advantage of the approach taken is that it is not model dependent and hence could potentially be applied to other complex queueing models. We numerically assess the accuracy of the interpolation approximation through the first and second moments.

Key words: light traffic, interpolation approximation, discriminatory processor sharing, random order of service

1 Introduction

In this paper we are interested in analyzing the steady-state performance of two multi-class single-server models: discriminatory processor sharing (DPS) and relative-priorities (RP). The behavior of both systems is determined by a vector of class-dependent weights, which we will denote by $(g_1, \ldots, g_K)$ for DPS and by $(p_1, \ldots, p_K)$ for RP. DPS is a time-sharing discipline in which all customers in the system get served simultaneously, being $\frac{g_k}{\sum_j n_j g_j}$, the fraction of the service that is allocated to a class-k customer, with n_j the number of class-j customers in

the system. On the other hand RP operates in a non-preemptive manner, and the probability that the next customer to be served is from class k is given by $\frac{n_k p_k}{\sum_j n_j p_j}$. The intra-class scheduling discipline under RP can be any non-anticipating policy, e.g. First Come First Served (FCFS), Last Come First Served (LCFS), or Random Order of Service (ROS).

Both DPS and RP are versatile queueing models providing a natural framework to model service differentiation in systems. DPS is a multi-class extension of the well-studied egalitarian Processor Sharing (PS) policy, where the various classes are assigned positive weight factors. The DPS queue has received lot of attention due to its application to model the performance of bandwidth sharing policies in communication networks, see for example [21, 8, 10, 18]. The RP model can have applications 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 [25].

The exact analysis of both DPS and RP is difficult, and closed-form results are scarce and exist only under limiting assumptions. For DPS with *exponential service time distributions*, in [26] the authors established that the generating function of the queue length vector satisfies a differential equation. From this equation, the authors further show that the moments can be determined numerically as the solution of a system of equations. RP is more amenable to analyze because it is non-preemptive. In [23] the authors established for *general service requirements* a set of equations for the generating function of the queue length vector *and* the Laplace-Stieltjes Transform (LST) of the waiting time. For both DPS and RP, a closed-form expression for the mean queue length is available only for the case of two classes, see [12] for DPS (with exponential service times) and [17] for RP. The heavy-traffic limits for DPS and RP have been studied in [20, 15, 32]. For both models it has been shown that a so-called "state-space collapse" appears, which describes that the queue lengths of the various classes become proportional in the heavy-traffic regime.

Motivated by the difficulty in analyzing both systems in exact form, in this paper we derive closed-form approximations for the steady-state distribution of the queue length vector and waiting time. We have chosen these metrics since they are among the most frequently considered measures in the performance evaluation literature. More precisely, we will first investigate the performance of both systems in light traffic, that is, when the arrival rate tends to 0. This approach was pioneered in a series of papers by Reiman & Simon, see for example [29], where the objective was the mean number of customers or mean sojourn time, and extended to the distribution of the sojourn time for Markovian queues in [14] and [28]. In one of our main contributions, we will derive the distribution of performance metrics under DPS in a light-traffic regime for general service times. We emphasize that in that case no analytical characterizations are available for DPS. In the case of RP, we will show that the light-traffic approximation can be obtained directly from the differential equations obtained in [23]. We will then combine our light-traffic approximations with the heavy-traffic characterization in order to develop an interpolation approximation that aims at capturing the performance for any load. We investigate the accuracy of our approximations for several service time distributions to illustrate the applicability of the approach.

We note that this paper is a generalization of [19] where we developed closed-form approximations for the mean conditional and unconditional sojourn times for the DPS policy. The main result in [19, Proposition IV.1]

is a particular case of Proposition 6.6, as described in Section 6.3.

The remainder of the paper is organized as follows. In Section 2 we provide a short overview of the related literature. In Section 3 we present the main modeling assumptions and notation used in this paper. In Section 4 we provide a detailed explanation of how to obtain the light-traffic derivatives and how to build the interpolation approximation. Section 5 and Section 6 focus on the RP model and the DPS model, respectively. We first introduce the known results from the literature (including known heavy-traffic results), and then explain how to derive the light-traffic approximation and the interpolation approximation. In Section 7 we numerically illustrate the accuracy of our approximations.

2 Related work

In this section we present a brief overview of the main results available on the models DPS and RP, and on light-traffic approximations.

The DPS model was introduced by Kleinrock in [24]. Despite the simplicity of the model description and the fact that the properties of the egalitarian Processor-Sharing queue (equal weights) are quite thoroughly understood, the analysis of DPS has proven to be extremely difficult. In a seminal paper Fayolle et al. [12] studied the mean conditional (on the service requirement) and unconditional sojourn time. For general service time distributions, the authors obtained the mean conditional sojourn time as the solution of a system of integro-differential equations. Asymptotics of the sojourn time have received considerable attention for example in [5] and [4]. Time-scale separations have been studied in [30] and [7]. The performance of DPS in overload and its application to model TCP flows is considered in [2]. The application of DPS to analyse the performance of TCP is also considered in [21] and for more applications of DPS in communication networks see [8, 10, 18]. DPS under a heavy-traffic regime (when the traffic load approaches the available capacity) was analysed in Grishechkin [15] assuming finite second moments of the service requirement distributions. Subsequently, assuming exponential service requirement distributions, a direct approach to establish a heavy-traffic limit for the joint queue length distribution was described by Rege & Sengupta [26] and extended to *phase-type* distributions in [32]. For an overview of the literature on DPS we refer to the survey [1].

A special case of RP is when the intra-class scheduling discipline is uniformly random, that is, within a class a customer is selected randomly. This model was proposed in [16] and it is referred to as discriminatory-random-order-of-service (DROS). In recent years several interesting studies have been published on DROS, [17, 22, 23, 20]. Expressions for the mean waiting time of a customer given its class have been obtained in [17]. In [22, 23] 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 [20] the authors obtain that the scaled waiting time of a customer of a given class in heavy traffic is distributed as the product of two exponentially distributed random variables, see Section 5.1 for more details.

The light-traffic regime concerns the performance of the system for small values of the arrival rate λ, i.e.,

when the system is almost empty. The approach relies on approximating the performance measure of interest by a Taylor series expansion at $\lambda = 0$. In order to obtain an approximation for any value of the arrival rate, in [27, 28, 29] Reiman and Simon propose an interpolation technique that consist of interpolating the light-traffic approximation and the heavy-traffic result. This technique has been applied with success to models like processor-sharing, fork-join, etc.; see examples in the literature in [9, 31, 19, 14, 28]. The method was extended to the distribution of the sojourn time for Markovian queues in [14] and [28]. In the case of models that permit a multidimensional quasi birth-and-death representation researchers have also developed light-traffic approximations of the mean queue lengths using the power-series algorithm, see for example [11].

3 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 $\lambda_k \geq 0$. We denote the overall arrival rate by $\lambda = \sum_{k=1}^{K} \lambda_k$ and let $\alpha_k = \lambda_k/\lambda$ be the probability that an arrival is of class k. Class-k customers have i.i.d. generally distributed service requirements denoted by B_k, $k = 1, \ldots, K$, with the distribution function $F_k(b) := \mathbb{P}(B_k \leq b)$, and Laplace Stieltjes transform (LST) $B_k^*(s)$. We assume that $\mathbb{E}\left[B_k^2\right] < \infty$, $k = 1, \ldots, K$. We further denote by B the service requirement of an arbitrary arriving customer. The traffic intensity for class-k customers is denoted by $\rho_k := \lambda_k \mathbb{E}[B_k]$ and the total traffic intensity is denoted by

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

We will use throughout the paper the notation $(x)^+ = \max\{0, x\}$.

In this paper we study the Discriminatory Processor Sharing (DPS) and the Relative Priorities (RP) policies.

DPS simultaneously shares the resources among the K classes. There are strictly positive class-dependent weights $g_1, \ldots, g_K$ associated with each of the classes. Whenever there are n_k class-k customers, $k = 1, \ldots, K$, in the system, each class-k customer is served at rate $g_k / \sum_{j=1}^{K} n_j g_j$.

The RP policy is a non-preemptive discipline and serves at each moment in time one customer. Upon service completion, the probability that the next customer to be served is of class k is given by $n_k p_k / \sum_j n_j p_j$, where, $p_j > 0$, $j = 1, \ldots, K$, are class-dependent weights, and n_j is the number of class-j customers at the decision epoch. 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 denote the steady-state number of class-k customers in the system at arbitrary epochs by N_k. We define the vector $\vec{N} = (N_1, \ldots, N_K)$ and the total number of customers is denoted by $N := \sum_{k=1}^{K} N_k$. For a given λ, let $\psi(\lambda, \vec{z}) := \mathbb{E}[z_1^{N_1} \cdots z_K^{N_K}]$ be the joint probability generating function (pgf) of $(N_1, \ldots, N_K)$, with $\vec{z} = (z_1, ..., z_K)$. In the remainder of the paper we will add a superscript $\{DPS, RP\}$ to the metrics in order to denote the dependency on the service discipline.

We will also be interested in the waiting time defined as the sojourn time in the system minus the service

requirement. In the case of RP, we make the assumption that the intra-class scheduling discipline is random, that is, the DROS discipline, since for that setting an expression for the scaled waiting time in heavy traffic is available. Under DROS, the probability that a particular class-k customer is selected for service is $\frac{p_k}{\sum_j p_j n_j}$. We denote the conditional (on the service requirement b) and unconditional waiting time of an arbitrary class-k customer by $W_k(b)$ and W_k, respectively, and let $W_k(\lambda, b, x) := \mathbb{P}[W_k(b) > x]$ be the complementary distribution function and $\widetilde{W}_k(\lambda, u) := \mathbb{E}\left[e^{-uW_k}\right]$ its Laplace-Stieltjes transform (LST).

The main results of the paper are the derivation of approximations for the (i) pgf of the queue length distribution for both DPS and RP, (ii) the LST of the waiting time in RP, and (iii) the distribution of the waiting time in DPS.

4 Interpolation approximation

In this section we denote by $G(\lambda, \vec{y})$, the performance metric we are interested in, as a function of the arrival rate λ and a vector $\vec{y}$. The interpretation of the function G and the vector $\vec{y}$ will change depending on the metric we are approximating. In this paper the metric G will represent either (i) the generating function $\psi(\lambda, \vec{z})$, hence $\vec{y} = \vec{z}$, (ii) the LST $\widetilde{W}_k^{DROS}(\lambda, u)$, hence $\vec{y} = u$, or (iii) the complementary distribution of the waiting time $W_k^{DPS}(\lambda, b, x)$, hence $\vec{y} = (b, x)$. We will characterize $G(\lambda, \vec{y})$ both as $\lambda \downarrow 0$, the light-traffic regime, and as $\lambda \uparrow 1/\mathbb{E}(B)$ (or equivalently $\rho \uparrow 1$), the heavy-traffic regime. In both cases, closed form expressions for the performance metrics can be derived, allowing us to obtain an approximation for arbitrary λ by an interpolation technique.

In Section 4.1 we describe how performance metrics can be derived for the light-traffic regime and in Section 4.2 the heavy-traffic regime is discussed. Section 4.3 presents the general setting for the light and heavy-traffic interpolation approximation, which we will simply refer to as *interpolation approximation*.

4.1 Light-traffic analysis

The light-traffic regime concerns the performance of the system when the arrival rate λ approaches zero, or in other words, when the amount of work arriving to the system per unit of time approaches zero. We will approximate $G(\lambda, \vec{y})$ by a Taylor series expansion at $\lambda = 0$. Assuming that the first n derivatives of $G(\lambda, \vec{y})$ at $\lambda = 0$ exist, we have the following approximation for $G(\lambda, \vec{y})$ when λ is close to zero:

$$G^{LT}(\lambda, \vec{y}) := G^{(0)}(0, \vec{y}) + \lambda G^{(1)}(0, \vec{y}) + \cdots + \frac{\lambda^n}{n!} G^{(n)}(0, \vec{y}), \tag{1}$$

where $G^{(0)}(0, \vec{y}) := G(0, \vec{y})$, to which we refer to as the *zeroth* light-traffic derivative, and, $G^{(m)}(0, \vec{y}), m = 1, 2, \ldots,$ denotes the m-th derivative at $\lambda = 0$, i.e., $G^{(m)}(0, \vec{y}) := \frac{\partial^m G(\lambda, \vec{y})}{\partial \lambda^m}\Big|_{\lambda=0}$. We will refer to Equation (1) as the light-traffic approximation of order n. The choice of the value of n will depend on the compromise between tractability and accuracy that is aimed at. In general, a characterization for $G(\lambda, \vec{y})$ might not exist and hence $G^{(m)}(0, \vec{y})$ cannot be obtained in a direct manner. In that case we use the results in [29, Section 3.] and [33, Chapter 6.3.] where it is shown how to derive the light-traffic derivatives of arbitrary order m under a general

admissibility condition. Following the discussion in [29, Appendix A] we make the next assumption on the service requirements B_k:

$$\mathbb{E}[\mathrm{e}^{\eta B_k}] = \sum_{n=0}^{\infty} \frac{\eta^n}{n!} \mathbb{E}[B_k^n] < \infty, \tag{2}$$

for some $\eta > 0, \forall k$, which entails admissibility. This finite exponential moment condition requires that all moments of the service requirement B_k to be finite. Equation (2) is likely stronger than needed but its purpose here is to provide a convenient framework where calculations can be justified. In this paper we will make use of the expressions as obtained in [29, 33] for the zeroth, first and second light-traffic derivatives. The expressions are given in the proposition below. For the sake of self-completeness a proof is provided in Appendix A.

Proposition 4.1. ([29, Section 3], [33, Chapter 6.3]) *Let $A(s,t)$ denote the number of arrivals in the interval $[s,t)$ in addition to a tagged customer who is assumed to arrive at time 0. Let $G(\lambda, \vec{y}|A)$ denote the performance metric $G(\lambda, \vec{y})$ conditioned on event A. Then the zeroth, first and second light-traffic derivative can be written as*

$$G^{(0)}(0,\vec{y}) = G\left(0,\vec{y}\middle|A(-\infty,\infty)=0\right),$$

$$G^{(1)}(0,\vec{y}) = \int_{-\infty}^{\infty} \left(G\left(0,\vec{y}\middle|A(-\infty,\infty)=1,\tau_1=t\right) - G\left(0,\vec{y}\middle|A(-\infty,\infty)=0\right)\right)\mathrm{d}t$$

and

$$G^{(2)}(0,\vec{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big(G\left(0,\vec{y}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right) - G\left(0,\vec{y}\middle|A(-\infty,\infty)=1,\tau_1=t'\right)$$
$$-G\left(0,\vec{y}\middle|A(-\infty,\infty)=1,\tau_1=t''\right) + G\left(0,\vec{y}\middle|A(-\infty,\infty)=0\right)\Big)\mathrm{d}t'\mathrm{d}t'',$$

where $\tau_i, i=1,2$, is the arrival time of the i-th customer.

4.2 Heavy-traffic regime

The heavy-traffic regime consists in investigating the queue when it is near saturation, i.e., $\rho \uparrow 1$. This regime can be obtained by letting

$$\lambda \uparrow \hat{\lambda} := \frac{1}{\mathbb{E}[B]},$$

since then $\rho = \lambda\mathbb{E}[B] \uparrow 1$. When passing to the heavy-traffic regime we keep the fraction of class-k arrivals, α_k, fixed and we define

$$\hat{\lambda}_k := \alpha_k\hat{\lambda} = \frac{\alpha_k}{\mathbb{E}[B]} \quad \text{and} \quad \hat{\rho}_k := \alpha_k\hat{\lambda}\mathbb{E}[B_k] = \alpha_k\frac{\mathbb{E}[B_k]}{\mathbb{E}[B]}. \tag{3}$$

In Sections 5 and 6 we provide a brief overview of the heavy-traffic results known for RP and DPS. The basic principle is to establish that the scaled performance metrics $(1-\lambda\mathbb{E}[B])\vec{N}$ and $(1-\lambda\mathbb{E}[B])W_k$, have a proper limit as $\lambda \uparrow \frac{1}{\mathbb{E}[B]}$. Hence, in the heavy-traffic regime we have expressions for the following scaled performance metrics:

(i) $\psi(\lambda, \vec{z}^{1-\lambda\mathbb{E}[B]}) = \mathbb{E}[\vec{z}^{(1-\lambda\mathbb{E}[B])\vec{N}}]$,

(ii) $\widetilde{W}_k^{DROS}(\lambda, u(1-\lambda\mathbb{E}[B])) = \mathbb{E}[e^{-u(1-\lambda\mathbb{E}[B])W_k}]$,

(iii) $W_k^{DPS}(\lambda, b, x/(1-\lambda\mathbb{E}[B])) = \mathbb{P}[(1-\lambda\mathbb{E}[B])W_k^{DPS}(b) > x]$,

where we used the notation $\vec{z}^{\gamma\vec{N}} := (z_1^{\gamma N_1}, \ldots, z_K^{\gamma N_K})$. We are hence interested in the scaled performance metric $G(\lambda, f_\lambda(\vec{y}))$ as $\lambda \uparrow \frac{1}{\mathbb{E}[B]}$, where $f_\lambda(\vec{y})$ is the scaling used. Depending on the three metrics described above, this function is given by

(i) $f_\lambda(\vec{z}) = \vec{z}^{1-\lambda\mathbb{E}[B]}$,

(ii) $f_\lambda(u) = u(1-\lambda\mathbb{E}[B])$,

(iii) $f_\lambda(b, x) = (b, x/(1-\lambda\mathbb{E}[B]))$.

Let $G^{HT}(\vec{y})$ be the heavy-traffic term defined as

$$G^{HT}(\vec{y}) := \lim_{\lambda\uparrow 1/\mathbb{E}[B]} G\big(\lambda, f_\lambda(\vec{y})\big). \tag{4}$$

4.3 Light-traffic and heavy-traffic interpolation

In the case expressions for a performance metric are known both for light traffic and heavy traffic, an approximation for an arbitrary λ can be derived following the light and heavy-traffic interpolation technique. This technique was popularized by Reiman and Simon [27, 28, 29] and consists in approximating the scaled performance metric, $G\left(\lambda, f_\lambda(\vec{y})\right)$, by a polynomial $\hat{G}\left(\lambda, \vec{y}\right)$ of order $n+1$:

$$\hat{G}\left(\lambda, \vec{y}\right) := h_0(\vec{y}) + h_1(\vec{y})\lambda + h_2(\vec{y})\lambda^2 + \ldots + h_{n+1}(\vec{y})\lambda^{n+1}. \tag{5}$$

Unnormalizing we then obtain the light and heavy-traffic interpolation approximation for the performance metric $G(\lambda, \vec{y})$, that is,

$$G^{INT}(\lambda, \vec{y}) := \hat{G}\left(\lambda, f_\lambda^{-1}(\vec{y})\right), \tag{6}$$

where $G^{HT}(\vec{y})$ denotes the heavy-traffic result as defined in (4).

To determine the coefficients $h_0(\vec{y}), \ldots, h_n(\vec{y})$ we take the m-th derivative to λ, $m = 0, \ldots, n$, in (6) at $\lambda = 0$ and set this equal to the m-th derivative of the performance metric to be approximated. Hence, we obtain the following light-traffic conditions:

$$\left.\frac{\partial^m G^{INT}(\lambda, \vec{y})}{\partial\lambda^m}\right|_{\lambda=0} = G^{(m)}(0, \vec{y}), \text{ for } m = 0, \ldots, n. \tag{7}$$

Note that expressions for $G^{(m)}(0, \vec{y})$ are given in Proposition 4.1. To determine $h_{n+1}(\vec{y})$, we use the heavy-traffic condition:

$$\lim_{\lambda\uparrow 1/\mathbb{E}[B]} \hat{G}\big(\lambda, \vec{y}\big) = G^{HT}(\vec{y}), \tag{8}$$

where $G\big(1/\mathbb{E}[B], \vec{y}\big)$ is the heavy-traffic result as described in Section 4.2. In the proof of Proposition 5.6 we explain how to determine the coefficients $h_0(\vec{y}), \ldots, h_{n+1}(\vec{y})$ in practice. We will refer to the approximation (6) as the light and heavy-traffic interpolation, or simply as *interpolation approximation*, of order $n+1$.

Proposition 4.2. *The interpolation approximation of order $n+1$ can equivalently be written as*

$$G^{INT}(\lambda,\vec{y}) = \sum_{i=0}^{n} \lambda^i \left(1-(\lambda\mathbb{E}[B])^{n+1-i}\right) h_i\left(f_\lambda^{-1}(\vec{y})\right) + (\lambda\mathbb{E}[B])^{n+1} G^{HT}\left(f_\lambda^{-1}(\vec{y})\right). \tag{9}$$

Proof. From the heavy-traffic condition (8) we obtain

$$h_{n+1}(\vec{y}) \quad = \quad \mathbb{E}[B]^{n+1}\left(G^{HT}(\vec{y}) - \sum_{i=0}^{n} \frac{h_i(\vec{y})}{\mathbb{E}[B]^i}\right).$$

Equation (9) follows after substituting this expression in (5) and then undoing the normalisation as in Equation (6). □

We note that in the case $G(\lambda,\vec{y})$ denotes the sojourn time distribution, (9) reduces to Equation (1) in [14].

An important observation is that the interpolation approximation obtained for the LST and pgf of the performance metrics might not correspond themselves to a random variable, that is, they might not be completely monotone functions as defined in [13, Section XIII.4]. However, we will show that they can still provide accurate approximations for the moments.

5 Relative-priorities queue

This section is devoted to the RP model. In Section 5.1 we will describe the heavy-traffic results on RP. These allow us to determine the interpolation approximation for the distribution of the joint queue length and waiting time in Section 5.2 and Section 5.3, respectively.

We recall from Section 3 that the steady-state number of class-k customers in the system is denoted by N_k^{RP}. We also recall that $\psi^{RP}(\lambda,\vec{z})$, with $\vec{z}=(z_1,...,z_K)$, denotes the joint pgf of $(N_1^{RP},\ldots,N_K^{RP})$.

5.1 Preliminaries

In [23] the distribution of the joint 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 [23].

Theorem 5.1. [23, Theorem 3 and Theorem 4] *The joint pgf $\psi^{RP}(\lambda,\vec{z})$ of the joint stationary queue lengths at arbitrary time epochs is given by*

$$\psi^{RP}(\lambda,\vec{z}) = 1-\rho+\sum_{i=1}^{K}\alpha_i z_i\left(1-\rho+\frac{p_i}{\alpha_i}\frac{\partial}{\partial z_i}r(\lambda,\vec{z})\right)\frac{1-B_i^*(\lambda-\lambda\sum_{k=1}^{K}\alpha_k z_k)}{1-\sum_{k=1}^{K}\alpha_k z_k}, \tag{10}$$

where $r(\lambda,\vec{z})$ *is defined as* $r(\lambda,\vec{z}) := \mathbb{E}\left[\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{\sum_{k=1}^K Q_k p_k}\cdot \mathbf{1}_{(\sum_{k=1}^K Q_k>0)}\right]$, *with* $Q_k, k=1,\ldots,K$, *the steady-state number of class-k customers in the system at departure epochs, and satisfies the equation*

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

Unfortunately (10) cannot be solved analytically for arbitrary λ. However, in [23, Section 3.2] the authors present a numerical scheme to obtain the moments of the total queue length. We will use this scheme in Section 7 in order to numerically estimate the accuracy of our approximation for the first and second moments of the queue length for arbitrary λ.

Under RP, once a customer enters service it is served until it has received its full service requirement. Hence, we will be interested in the waiting time. In the case of the waiting time we focus on the random intra-class scheduling discipline, that is, we consider the specific model DROS. In this case the probability that a particular class-k customer is selected for service is $\frac{p_k}{\sum_j p_j n_j}$. We denote the waiting time of an arbitrary class-k customer by W_k^{DROS}. We refer to this customer as the tagged class-k 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., $W_k^{DROS}>0$.

We now define the following joint transform:

$$T_l^{DROS}(u,z_1,\ldots,z_K) := \mathbb{E}[\mathrm{e}^{-uW_l^{DROS}} z_1^{Q_1^*}\cdots z_K^{Q_K^*} 1_{\{W_l^{DROS}>0\}}]. \tag{11}$$

Note that the transform of the waiting time $\widetilde{W}_k^{DROS}$ of the tagged class-k customer is given by

$$\widetilde{W}_k^{DROS}(\lambda,u) = \mathbb{E}[\mathrm{e}^{-uW_k^{DROS}}] = \mathbb{E}[\mathrm{e}^{-u\cdot 0}\mathbf{1}_{\{W_k^{DROS}=0\}} + \mathrm{e}^{-u\cdot W_k^{DROS}}\mathbf{1}_{\{W_k^{DROS}>0\}}] = 1-\rho+T_k^{DROS}(u,\vec{1}), \tag{12}$$

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 [23] the following result for the transform $T_k^{DROS}(u,\vec{z})$.

Theorem 5.2. [23, Theorem 8] *For the random intra-class scheduling discipline, the joint transform* $T_l^{DROS}(u,\vec{z})$ *satisfies*

$$\begin{aligned}&\sum_{i=1}^K \frac{p_i}{p_l}\left(\frac{\partial}{\partial z_i}T_l^{DROS}(u,\vec{z})\right)(z_i - B_i^*(u+\lambda-\lambda\sum_{k=1}^K\alpha_k z_k)) + T_l^{DROS}(u,\vec{z})\\ &= \sum_{i=1}^K((1-\rho)\lambda\alpha_i + \lambda p_i\frac{\partial}{\partial z_i}r(\lambda,\vec{z}))\frac{B_i^*(\lambda-\lambda\sum_{k=1}^K\alpha_k z_k) - B_i^*(u+\lambda-\lambda\sum_{k=1}^K\alpha_k z_k)}{u},\end{aligned} \tag{13}$$

with $r(\vec{z})$ *as defined in Theorem 5.1.*

The integro-differential equations as given in Theorems 5.1 and 5.2 cannot be solved in general, however they are very valuable in obtaining insights into the performance of the system. In particular, they were key in carrying out a heavy-traffic analysis of RP (see below), and they will be key in obtaining the light-traffic approximation required to derive the interpolation results in Sections 5.2 and 5.3.

Heavy-traffic results for the RP model were obtained in [20]. As stated in the following proposition, a state-space collapse for the scaled queue length vector in the heavy-traffic regime was established, 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.

Proposition 5.3. [20, Proposition 3.1] *The scaled joint pgf of the stationary queue lengths, $\psi^{RP}(\lambda, \vec{z}^{(1-\lambda\mathbb{E}[B])})$, satisfies*

$$\lim_{\lambda\uparrow 1/\mathbb{E}[B]} \psi^{RP}(\lambda, \vec{z}^{(1-\lambda\mathbb{E}[B])}) = \lim_{\lambda\uparrow 1/\mathbb{E}[B]} \mathbb{E}[z_1^{(1-\lambda\mathbb{E}[B])N_1^{RP}} \cdots z_K^{(1-\lambda\mathbb{E}[B])N_K^{RP}}] = \frac{\mathbb{E}[B]\nu(\vec{p})}{\mathbb{E}[B]\nu(\vec{p}) - \sum_{i=1}^K \frac{\alpha_i}{p_i}\ln(z_i)} \quad (14)$$

where

$$\nu(\vec{p}) := \frac{2\sum_{k=1}^K \alpha_k \mathbb{E}[B_k]/p_k}{\mathbb{E}[B^2]}. \quad (15)$$

Or in other words, as $\lambda \uparrow 1/\mathbb{E}[B]$, $(1-\lambda\mathbb{E}[B])(N_1^{RP}, \ldots, N_K^{RP}) \xrightarrow{d} X \cdot (\frac{\alpha_1}{p_1}, \frac{\alpha_2}{p_2}, ..., \frac{\alpha_K}{p_K})$, where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean $1/(\mathbb{E}[B]\nu(\vec{p}))$.

The next proposition states that under the heavy-traffic regime the waiting time of a tagged class-l customer, W_l^{DROS}, is the product of two exponentially distributed independent random variables:

Proposition 5.4. [20, Proposition 5.1] *The Laplace Transform of the scaled waiting time of a class-k customer under the heavy-traffic regime satisfies*

$$\lim_{\lambda\uparrow 1/\mathbb{E}[B]} \widetilde{W}_k^{DROS}(\lambda, (1-\lambda\mathbb{E}[B])u) = \lim_{\lambda\uparrow 1/\mathbb{E}[B]} \mathbb{E}[\mathrm{e}^{-u(1-\lambda\mathbb{E}[B])W_k^{DROS}}] = \frac{\nu(\vec{p})p_k}{u}\mathrm{e}^{p_k\frac{\nu(\vec{p})}{u}} \int_{p_k\frac{\nu(\vec{p})}{u}}^{\infty} \frac{\mathrm{e}^{-l}}{l}\mathrm{d}l. \quad (16)$$

Or in other words, as $\lambda \uparrow 1/\mathbb{E}[B]$, $(1-\lambda\mathbb{E}[B])W_k^{DROS} \xrightarrow{d} Z_k \cdot X$, where $\xrightarrow{d}$ denotes convergence in distribution and X and Z_k are exponentially distributed independent random variables with $\mathbb{E}[Z_k] = 1/p_k$, $\mathbb{E}[X] = 1/\nu(\vec{p})$ and $\nu(\vec{p})$ as given in (15).

5.2 Approximation for the joint queue-length distribution

In this section we set $\vec{y} = \vec{z}$ and let $G(\lambda, \vec{z}) = \psi^{RP}(\lambda, \vec{z})$ be the pgf of the joint queue lengths under RP. Then, using Theorem 5.1 (when $\lambda = 0$) we obtain the following light-traffic approximation.

Lemma 5.5. *The light-traffic approximation (of order 2) of the joint pgf of $(N_1^{RP}, \ldots, N_K^{RP})$ is given by*

$$\psi^{RP,LT}(\lambda, \vec{z}) = 1 - \rho + \lambda\sum_{i=1}^K \alpha_i \mathbb{E}[B_i] z_i + \frac{\lambda^2}{2}\sum_{i=1}^K \alpha_i z_i \mathbb{E}[B_i^2]\left(\sum_{k=1}^K \alpha_k z_k - 1\right).$$

Proof. See Appendix B for the proof. □

We now present the interpolation approximation for the queue length.

Proposition 5.6. *The light and heavy-traffic interpolation (of order 3) of the joint pgf of $(N_1^{RP}, \ldots, N_K^{RP})$ is*

given by

$$\begin{aligned}
&\psi^{RP,INT}(\lambda,\vec{z})\\
&=\left(1-\rho^3\right)+\lambda\left(1-\rho^2\right)\left(-\mathbb{E}[B]+\sum_{i=1}^{K}\alpha_i\mathbb{E}[B_i]z_i^{(1-\rho)^{-1}}\right)+\frac{\lambda^2(1-\rho)}{2}\Bigg(-2\mathbb{E}[B]\sum_{i=1}^{K}\alpha_i z_i^{(1-\rho)^{-1}}\mathbb{E}[B_i]\frac{\ln(z_i)}{1-\rho}\\
&\quad+\sum_{i=1}^{K}\alpha_i z_i^{(1-\rho)^{-1}}\mathbb{E}[B_i^2]\left(\sum_{k=1}^{K}\alpha_k z_k^{(1-\rho)^{-1}}-1\right)\Bigg)+\rho^3\frac{\mathbb{E}[B]\nu(\vec{p})}{\mathbb{E}[B]\nu(\vec{p})-\sum_{i=1}^{K}\frac{\alpha_i}{p_i}\ln\left(z_i^{(1-\rho)^{-1}}\right)},
\end{aligned}\tag{17}$$

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

Proof. The result follows using the heavy-traffic term $G^{HT}(\vec{z})$ as given in (14), together with Lemma 5.5 and Proposition 4.2. A detailed proof is provided in Appendix C. □

Equation (17) can be readily used to derive our approximation for the first and second moments of the total number of customers in the system. The first moment is given by

$$\begin{aligned}
\mathbb{E}[N^{RP,INT}]&=\mathbb{E}[N_1^{RP,INT}+\ldots+N_K^{RP,INT}]=\left.\frac{\partial\left(\psi^{RP,INT}(\lambda,\vec{z})\Big|_{z_i=z_j=z}\right)}{\partial z}\right|_{z=1}\\
&=\rho+\frac{\lambda^2\mathbb{E}[B^2]}{2}+\frac{\rho^3}{(1-\rho)}\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]\sum_{k=1}^{K}\frac{\alpha_k}{p_k}\mathbb{E}[B_k]}\cdot\sum_{i=1}^{K}\frac{\alpha_i}{p_i}.
\end{aligned}\tag{18}$$

Under the assumption that there is one class in the system, that is, $\alpha_i=0$, $\forall i\neq k$ and $\alpha_k=1$, Equation (18) is exact. It gives $\mathbb{E}[N^{RP,INT}]=\rho+\frac{\lambda^2\mathbb{E}[B^2]}{2}\left(1+\frac{\rho}{1-\rho}\right)=\rho+\frac{\lambda^2\mathbb{E}[B^2]}{2(1-\rho)}$, that is, it coincides with the well known Pollaczek-Khinchine formula for the M/G/1 queue.

The second derivative of $\left(\psi^{RP}\right)^{INT}(\lambda,\vec{z})$ with respect to z, evaluated at $z=1$, is given by

$$\begin{aligned}
&\left.\frac{\partial^2\left(\psi^{RP,INT}(\lambda,\vec{z}))\Big|_{z_i=z_j=z}\right)}{\partial z^2}\right|_{z=1}=\mathbb{E}\left[\left(N^{RP,INT}\right)^2\right]-\mathbb{E}[N^{RP,INT}]=\frac{\lambda^2\mathbb{E}[B^2]}{1-\rho}\frac{2+\rho}{2}\\
&\quad+\frac{\rho^3}{1-\rho}\frac{2\mathbb{E}[B]}{\mathbb{E}[B^2]}\sum_{k=1}^{K}\frac{\alpha_k}{p_k}\mathbb{E}[B_k]\left(\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]\sum_{k=1}^{K}\frac{\alpha_k}{p_k}\mathbb{E}[B_k]}\right)^2\sum_{i=1}^{K}\frac{\alpha_i}{p_i}\left(\frac{2}{(1-\rho)}\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]\sum_{k=1}^{K}\frac{\alpha_k}{p_k}\mathbb{E}[B_k]}\sum_{i=1}^{K}\frac{\alpha_i}{p_i}-1\right).
\end{aligned}\tag{19}$$

Therefore, the approximation for the second moment of the total number of customers is given by the sum of Equations (18) and (19):

$$\mathbb{E}\left[\left(N^{RP,INT}\right)^2\right]=\left.\frac{\partial^2\left(\psi^{RP,INT}(\lambda,\vec{z}))\Big|_{z_i=z_j=z}\right)}{\partial z^2}\right|_{z=1}+\mathbb{E}[N^{RP,INT}].\tag{20}$$

In Section 7 we use the expression for the first and second moments, Equations (18) and (20), to numerically assess the accuracy of our interpolation approximation.

5.3 Approximation for the waiting time distribution

We recall that the waiting time in RP depends on the intra-class scheduling discipline being implemented. We will consider the particular case in which the intra-class scheduling discipline is random, that is, DROS. In this section, we set $\vec{y} = u$ and let $G(\lambda, u) = \widetilde{W}_k^{DROS}(\lambda, u)$ be the LST of a class-k customer's waiting time under DROS.

Taking the derivatives of Equation (13) with respect to λ we obtain the following light-traffic approximation.

Lemma 5.7. *The light-traffic approximation (of order 1) of the Laplace Transform of the waiting time under DROS is given by*

$$\widetilde{W}_k^{DROS,LT}(\lambda, u) = 1 - \rho + \lambda \left(\sum_{i=1}^{K} \alpha_i \frac{1 - B_i^*(u)}{u} \right).$$

Proof. See Appendix D for the proof. □

We note that the light-traffic approximation is independent of the class. Indeed, from Proposition 4.1 we know that the 1st order approximation is calculated when there is only one additional arrival to the system (apart from the tagged customer), and thus, the non-preemptive scheduling policy does not play any role. The 2nd order approximation can be calculated, however the final expression is much more cumbersome, and yet the numerical accuracy does not significantly improve. In the next proposition we present the interpolation approximation which does depend on the class due to the heavy-traffic term:

Proposition 5.8. *The light and heavy-traffic interpolation (of order 2) of the LST of the waiting time under DROS is given by*

$$\begin{aligned}
&\widetilde{W}_k^{DROS,INT}(\lambda, u) \\
&= (1-\rho)^2 + \lambda(1-\rho) \left(-\mathbb{E}[B] + \sum_{i=1}^{K} \alpha_i \frac{1 - B_i^*((1-\rho)^{-1}u)}{(1-\rho)^{-1}u} \right) + \rho^2 \frac{\nu(\vec{p}) p_k}{(1-\rho)^{-1}u} \mathrm{e}^{p_k \frac{\nu(\vec{p})}{(1-\rho)^{-1}u}} \int_{p_k \frac{\nu(\vec{p})}{(1-\rho)^{-1}u}}^{\infty} \frac{\mathrm{e}^{-l}}{l} \mathrm{d}l,
\end{aligned}$$

with $\nu(\vec{p})$ given as in Equation (15).

Proof. The result follows after using the heavy-traffic term $G^{HT}(u)$ as given in (16), together with Lemma 5.7 and Proposition 4.2. A detailed proof is omitted, but it follows similarly to that of Proposition 5.6. □

6 Discriminatory-Processor-Sharing queue

We now focus on the DPS model. In Section 6.1 we will describe the main results on DPS that are used later on. In Section 6.2 we obtain the interpolation approximation for the distribution of the queue-length vector, and in Section 6.3 for the waiting time. We recall from Section 3 that the steady-state number of class-k customers in the system at arbitrary epochs is denoted by N_k^{DPS}. We also recall that $\psi^{DPS}(\lambda, \vec{z})$, with $\vec{z} = (z_1, ..., z_K)$, denotes the joint pgf of $(N_1^{DPS}, \ldots, N_K^{DPS})$. The conditional (on the service requirement b) and unconditional waiting time of an arbitrary class-k customer is denoted by $W_k^{DPS}(b)$ and W_k^{DPS}, respectively, and $W_k^{DPS}(\lambda, b, x) = \mathbb{P}\left[W_k^{DPS}(b) > x\right]$.

6.1 Preliminaries

As mentioned in Section 1 the analysis of DPS is difficult, and therefore there is no exact analysis available for the queue-length distribution under general service time distributions. However, there are several results on DPS in heavy traffic that are available in the literature and that we will use in order to obtain our interpolation approximation.

As stated in the following proposition, in heavy traffic a state-space collapse for the scaled queue length vector appears, 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.

Proposition 6.1. [32, Proposition 2.1.] *The scaled joint pgf of the stationary queue lengths, $\psi^{DPS}(\lambda, \vec{z}^{(1-\lambda\mathbb{E}[B])})$, satisfies*

$$\lim_{\lambda\to 1/\mathbb{E}[B]} \psi^{DPS}(\lambda, \vec{z}^{(1-\lambda\mathbb{E}[B])}) = \lim_{\rho\to 1} \mathbb{E}[z_1^{(1-\lambda\mathbb{E}[B])N_1^{DPS}} \cdots z_K^{(1-\lambda\mathbb{E}[B])N_K^{DPS}}] = \frac{\mathbb{E}[B]/\mathbb{E}[Y]}{\mathbb{E}[B]/\mathbb{E}[Y] - \sum_{i=1}^K \frac{\alpha_i \mathbb{E}[B_i]}{g_i} \ln(z_i)}, \quad (21)$$

where

$$\mathbb{E}[Y] = \frac{\mathbb{E}\left[B^2\right]}{\mathbb{E}[B] \sum_{k=1}^K \alpha_k \mathbb{E}\left[B_k^2\right]/g_k}. \quad (22)$$

Or in other words, as $\lambda \uparrow 1/\mathbb{E}[B]$, $(1-\lambda\mathbb{E}[B])(N_1^{DPS}, \ldots, N_K^{DPS}) \xrightarrow{d} Y \cdot \left(\frac{\alpha_1\mathbb{E}[B_1]}{g_1}, \frac{\alpha_2\mathbb{E}[B_2]}{g_2}, \ldots, \frac{\alpha_K\mathbb{E}[B_K]}{g_K}\right)$, where $\xrightarrow{d}$ denotes convergence in distribution and Y is an exponentially distributed random variable with mean $\mathbb{E}[Y]$ as given in (22).

In [15] it was obtained that under the heavy-traffic regime the conditional sojourn time of a tagged class-k customer is the product of an exponentially distributed random variable and a deterministic factor. Since under the heavy-traffic scaling the sojourn time is equal to the waiting time we have the following result:

Proposition 6.2. [15, Theorem 4.2] *The Laplace Transform of the scaled conditional waiting time of a class-k customer satisfies*

$$\lim_{\lambda\to 1/\mathbb{E}[B]} \mathbb{P}[(1-\rho)W_k^{DPS}(b) \le x] = \mathrm{e}^{-x\frac{g_k}{b\mathbb{E}[V]}}, \quad (23)$$

where

$$\mathbb{E}[V] = \frac{\mathbb{E}[B^2]}{\sum_{i=1}^K \alpha_i \mathbb{E}[B_i^2]/g_i}. \quad (24)$$

Or in other words, as $\lambda \uparrow 1/\mathbb{E}[B]$, $(1-\lambda\mathbb{E}[B])W_k^{DPS}(b) \xrightarrow{d} \frac{b}{g_k}V$, where $\xrightarrow{d}$ denotes convergence in distribution and V is exponentially distributed with mean $\mathbb{E}[V]$ as given in (24).

6.2 Approximation for the joint queue-length distribution

In this section we set $\vec{y} = \vec{z}$ and let $G(\lambda, \vec{z}) = \psi^{DPS}(\lambda, \vec{z})$ be the joint pgf of the joint queue lengths under DPS. Since there is no characterization available for the queue length distribution, we derive in the next lemma the light-traffic derivatives using the result given in Proposition 4.1. The proof method is constructive, and it can readily be applied to other queueing systems for which no analytical results are available. We thus believe that this represents in itself one of the main contributions of the paper.

Lemma 6.3. *The light-traffic approximation (of order 2) of the pgf of* $(N_1^{DPS},\dots,N_K^{DPS})$ *is given by*

$$\begin{aligned}
&\psi^{DPS,LT}(\lambda,\vec{z}) = \left(\psi^{DPS}\right)^{(0)}(\lambda,\vec{z})\Big|_{\lambda=0} + \lambda\left(\psi^{DPS}\right)^{(1)}(\lambda,\vec{z})\Big|_{\lambda=0} + \frac{\lambda^2}{2}\left(\psi^{DPS}\right)^{(2)}(\lambda,\vec{z})\Big|_{\lambda=0}\\
&= 1-\rho+\lambda\sum_{i=1}^{K}\alpha_i z_i\mathbb{E}\left[B_i\right] + \frac{\lambda^2}{2}\cdot 2\Bigg(\sum_{i,j=1}^{K}\alpha_i\alpha_j\left(z_i-1\right)\mathbb{E}\Bigg[\left(B_i - B_j\frac{g_i}{g_j}\right)\left(B_i-\min\{B_i,B_j\frac{g_i}{g_j}\}\right)\\
&\qquad -\frac{1}{2}\left(B_i-\min\{B_i,B_j\frac{g_i}{g_j}\}\right)^2 - \frac{B_i^2}{2}\Bigg]\\
&\quad + \sum_{i,j=1}^{K}\alpha_i\alpha_j\left(z_i\cdot z_j-1\right)\mathbb{E}\Bigg[B_j\left(1+\frac{g_i}{g_j}\right)\left(B_i-\min\{B_i,\frac{g_i}{g_j}B_j\}\right) + \frac{1}{2}\left(1+\frac{g_j}{g_i}\right)\min\{B_i,\frac{g_i}{g_j}B_j\}^2\Bigg]\\
&\quad + \sum_{i,j=1}^{K}\alpha_i\alpha_j\left(z_j-1\right)\mathbb{E}\left[\frac{g_i}{2g_j}\min\{\frac{g_j}{g_i}B_i,B_j\}^2 - B_i\min\{\frac{g_j}{g_i}B_i,B_j\}\right]\Bigg).
\end{aligned}$$

Proof. To calculate the zeroth, first and second light-traffic derivatives of the joint pgf of the queue length we measure how many customers are in the system when the tagged customer arrives (at time 0), given that $0,1$ or 2 customers might arrive to the system at most, respectively. For instance, for the zeroth derivative we need to consider the system with no other arrivals, hence $\vec{N}=\vec{0}$. For the first derivative we consider one additional arrival. Hence, different cases might happen: the customer arriving at time t might come before the tagged customer and leave before or after its arrival. Then the tagged customer observes either $\vec{N}=\vec{0}$ or $\vec{N}=e_k$, with k the class of the arrival, respectively. Or the one customer arrives after the tagged customer, in which case the tagged customer observes $\vec{N}=\vec{0}$. To obtain the second light-traffic derivative we analyse all different cases in a similar way. See Appendix E for the detailed proof. □

Proposition 6.4. *The light and heavy-traffic interpolation (of order 3) of the joint pgf of* $(N_1^{DPS},\dots,N_K^{DPS})$ *is given by*

$$\begin{aligned}
&\psi^{DPS,INT}(\lambda,\vec{z})\\
&= \left(1-\rho^3\right) + \lambda\left(1-\rho^2\right)\left(\sum_{i=1}^{K}\alpha_i z_i^{(1-\rho)^{-1}}\mathbb{E}\left[B_i\right] - \mathbb{E}\left[B\right]\right) + \lambda^2\left(1-\rho\right)\sum_{i,j=1}^{K}\alpha_i\alpha_j\Bigg(\\
&\quad \mathbb{E}\left[\left(z_i^{(1-\rho)^{-1}}-1\right)\left(\frac{g_i}{g_j}B_j\left(\min\{B_i,\frac{g_i}{g_j}B_j\}-B_i\right) - \frac{1}{2}\min\{B_i,\frac{g_i}{g_j}B_j\}^2\right)\right]\\
&\quad +\mathbb{E}\Bigg[\left(z_i^{(1-\rho)^{-1}}\cdot z_j^{(1-\rho)^{-1}}-1\right)\left(B_j\left(1+\frac{g_i}{g_j}\right)\left(B_i-\min\{B_i,\frac{g_i}{g_j}B_j\}\right)\right.\\
&\qquad \left.+\frac{1}{2}\left(1+\frac{g_j}{g_i}\right)\min\{B_i,\frac{g_i}{g_j}B_j\}^2\right)\Bigg]\\
&\quad +\mathbb{E}\left[\left(z_j^{(1-\rho)^{-1}}-1\right)\left(\frac{g_j}{2g_i}\min\{B_i,\frac{g_i}{g_j}B_j\}^2 - \frac{g_j}{g_i}B_i\min\{B_i,\frac{g_i}{g_j}B_j\}\right)\right]\Bigg)\\
&\quad -\mathbb{E}[B]\sum_{i=1}^{K}\alpha_i\mathbb{E}[B_i]z_i^{(1-\rho)^{-1}}\ln\left(z_i^{(1-\rho)^{-1}}\right) + \rho^3\frac{\mathbb{E}[B]/\mathbb{E}[Y]}{\mathbb{E}[B]/\mathbb{E}[Y]-\sum_{i=1}^{K}\frac{\alpha_i\mathbb{E}[B_i]}{g_i}\ln\left(z_i^{(1-\rho)^{-1}}\right)},
\end{aligned}$$

with $\mathbb{E}[Y]$ *as given in Equation* (22).

Proof. The result follows after using the heavy-traffic term $G^{HT}(\vec{z})$ as given in (21), together with Lemma 6.3 and Proposition 4.2. We omit the details since the proof follows the same steps as the proof of Proposition 5.6 $\square$

We now derive the first and second moment of our approximation for the total number of customers in the system. The approximation for the first moment of the total number of customers is given by

$$\begin{aligned}
&\mathbb{E}[N^{DPS,INT}] = \mathbb{E}[N_1^{DPS,INT} + \ldots + N_K^{DPS,INT}] = \frac{\partial\left(\psi^{DPS,INT}(\lambda,\vec{z})\Big|_{z_i=z_j=z}\right)}{\partial z}\Bigg|_{z=1} \\
&= \rho + \lambda^2 \sum_{i,j=1}^{K} \alpha_i\alpha_j \mathbb{E}\Bigg[\left(2+\frac{g_i}{g_j}\right)B_iB_j - \left(2+\frac{g_i}{g_j}\right)B_j \min\{B_i, B_j\frac{g_i}{g_j}\} \\
&\quad + \left(\frac{1}{2}+\frac{3g_j}{2g_i}\right)\min\{B_i, B_j\frac{g_i}{g_j}\}^2 - \frac{g_j}{g_i}B_i\min\{B_i, \frac{g_i}{g_j}B_j\}\Bigg] + \frac{\rho^3}{1-\rho}\frac{\mathbb{E}[Y]}{\mathbb{E}[B]}\sum_{i=1}^{K}\frac{\alpha_i\mathbb{E}[B_i]}{g_i}. \qquad (25)
\end{aligned}$$

The second derivative of $\psi^{DPS,INT}(\lambda,\vec{z})$ with respect to z, evaluated at $z=1$, is

$$\begin{aligned}
&\frac{\partial^2\left(\psi^{DPS,INT}(\lambda,\vec{z})\Big|_{z_i=z_j=z}\right)}{\partial z^2}\Bigg|_{z=1} = \mathbb{E}\left[\left(N^{DPS,INT}\right)^2\right] - \mathbb{E}\left[N^{DPS,INT}\right] \\
&= \lambda^2\sum_{i,j=1}^{K}\alpha_i\alpha_j\Bigg(\frac{\rho}{(1-\rho)}\mathbb{E}\Bigg[-B_j\frac{g_i}{g_j}B_i + B_j\frac{g_i}{g_j}\min\{B_i, B_j\frac{g_i}{g_j}\} - \frac{1}{2}\min\{B_i, B_j\frac{g_i}{g_j}\}^2\Bigg] \\
&\quad + \frac{2(1+\rho)}{(1-\rho)}\mathbb{E}\Bigg[\left(B_j\left(1+\frac{g_i}{g_j}\right)\left(B_i - \min\{B_i, \frac{g_i}{g_j}B_j\}\right) + \frac{1}{2}\left(1+\frac{g_j}{g_i}\right)\min\{B_i, \frac{g_i}{g_j}B_j\}^2\right)\Bigg] \\
&\quad + \frac{\rho}{(1-\rho)}\mathbb{E}\Bigg[\left(\frac{g_j}{2g_i}\min\{B_i, \frac{g_i}{g_j}B_j\}^2 - \frac{g_j}{g_i}B_i\min\{B_i, \frac{g_i}{g_j}B_{U_{t''}}\}\right)\Bigg]\Bigg) \\
&\quad + \rho^3\frac{\mathbb{E}[Y]}{\mathbb{E}[B]}\left(\frac{2\mathbb{E}[Y]}{\mathbb{E}[B]}\left(\sum_{i=1}^{K}\frac{\alpha_i\mathbb{E}[B_i]}{g_i}(1-\rho)^{-1}\right)^2 - \left(\sum_{i=1}^{K}\frac{\alpha_i\mathbb{E}[B_i]}{g_i}(1-\rho)^{-1}\right)\right). \qquad (26)
\end{aligned}$$

Therefore, the second moment of the total number of customers is obtained from Equations (25) and (26):

$$\mathbb{E}\left[\left(N^{DPS,INT}\right)^2\right] = \frac{\partial^2\left(\psi^{DPS,INT}(\lambda,\vec{z}))\Big|_{z_i=z_j=z}\right)}{\partial z^2}\Bigg|_{z=1} + \mathbb{E}\left[N^{DPS,INT}\right]. \qquad (27)$$

We observe that under the assumption that the service time distributions are exponential with the same mean $1/\mu$, the DPS queue behaves as an $M/M/1$ queue. The first and second moment of our approximation are $\mathbb{E}\left[N^{DPS,INT}\right] = \frac{\rho}{1-\rho}$ and $\mathbb{E}\left[\left(N^{DPS,INT}\right)^2\right] = \frac{2\rho^2}{(1-\rho)^2} + \frac{\rho}{1-\rho}$, hence, they are exact. The approximation is also exact with general service time distributions in the case that there is only one class in the system, that is, $\alpha_i = 0$, $\forall i \neq k$ and $\alpha_k = 1$, since then $\mathbb{E}[N^{DPS,INT}] = \rho + \rho^2 + \frac{\rho^3}{1-\rho} = \frac{\rho}{1-\rho}$.

In Section 7 we use the expression for the first and second moment, Equations (25) and (27) to numerically test the accuracy of the interpolation approximation.

6.3 Approximation for the waiting time distribution

In this section we set $\vec{y} = (b, x)$ and let $G(\lambda, b, x) = W_k^{DPS}(\lambda, b, x) = \mathbb{P}\left[W_k^{DPS}(b) > x\right]$ be the complementary distribution function of the conditional waiting time. We note that in the case of DPS, the waiting time has an

atom at the point $x = 0$ of size $1-\mathbb{P}(W_k^{DPS}(b) > 0)$. In the ensuing we develop the interpolation approximations for $\mathbb{P}(W_k^{DPS}(b) > x), x \geq 0$.

As was the case for the queue-length distribution, under DPS there is no characterisation available for the waiting time distribution with general service time distributions. Thus, in the next lemma we obtain the light-traffic derivatives using the result given in Proposition 4.1. Again, the proof method is constructive and it could be applied to other queueing systems for which no analytical results are available. The proof can be found in Appendix F.

Lemma 6.5. *The light-traffic approximation (of order 1) of the complementary distribution function of the conditional waiting time of a tagged class-k customer with a given service requirement b is given by*

$$\begin{aligned} &W_k^{DPS,LT}(\lambda, b, x) \\ &= \lambda \sum_{j=1}^{K} \alpha_j \mathbb{E}\Bigg[\left(1 + \frac{g_k}{g_j}\right) \left(-x + \min\{B_j, \frac{g_j}{g_k} b\}\right)^+ \\ &\quad + \mathbf{1}\left[\frac{g_j}{g_k} b > x\right] \left(B_j - \min\{B_j, \frac{g_j}{g_k} b\}\right) + \mathbf{1}\left[B_j > x\right] \left(b - \frac{g_k}{g_j} \min\{B_j, \frac{g_j}{g_k} b\}\right) \Bigg] \end{aligned} \tag{28}$$

Proof. To calculate the first light-traffic derivative of the complementary distribution function of the waiting time we measure which is the waiting time of the tagged customer, that arrives at time 0, given that 1 customer might arrive to the system at most. For the first derivative six different cases might happen. See Appendix F for the detailed proof. □

We note that the first order light-traffic approximation is class and weight dependent, unlike the RP model. This happens due to the time-sharing property of the DPS policy.

We can now present the interpolation approximation for the complementary distribution function of the conditional waiting time in DPS.

Proposition 6.6. *The light and heavy-traffic interpolation (of order 2) of the complementary distribution of the conditional waiting time of a tagged class-k customer with a given service requirement b is given by*

$$\begin{aligned} &W_k^{DPS,INT}(\lambda, b, x) \\ &= \lambda(1-\rho) \sum_{j=1}^{K} \alpha_j \mathbb{E}\Bigg[\left(1 + \frac{g_k}{g_j}\right) \left(-(1-\rho)x + \min\{B_j, \frac{g_j}{g_k} b\}\right)^+ + \mathbf{1}\left[\frac{g_j}{g_k} b > (1-\rho)x\right] \left(B_j - \min\{B_j, \frac{g_j}{g_k} b\}\right) \\ &\quad + \mathbf{1}\left[B_j > (1-\rho)x\right] \left(b - \frac{g_k}{g_j} \min\{B_j, \frac{g_j}{g_k} b\}\right) \Bigg] + \rho^2 \mathrm{e}^{-(1-\rho)x \frac{g_k}{b\mathbb{E}[V]}}, \end{aligned} \tag{29}$$

with $\mathbb{E}[V]$ as given in (24).

Proof. The result is obtained by using the heavy-traffic term $G^{HT}(b, x)$ as given in (23), together with Lemma 6.5 and Proposition 4.2. The detailed proof is omitted since it is similar to that of Proposition 5.6. □

From Equation (29) we obtain that the mean conditional waiting time, $\mathbb{E}\left[W_k^{DPS,INT}(b)\right]$, of a class-$k$ cus-

tomer satisfies the equation:

$$\mathbb{E}\left[W_k^{DPS,INT}(b)\right] = b\rho + \lambda\sum_{j=1}^{K}\alpha_j\mathbb{E}\left[\frac{1}{2}\big(1+\frac{g_k}{g_j}\big)\min\{B_j, b\frac{g_j}{g_k}\}^2 - \big(b\frac{g_j}{g_k}+\frac{g_k}{g_j}B_j\big)\min\{B_j, b\frac{g_j}{g_k}\} + b\frac{g_j}{g_k}B_j\right] + \frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{b}{g_k}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K}\alpha_j\mathbb{E}[B_j^2]/g_j}.$$

We observe that this coincides with that obtained in [19, Proposition IV.1]. In particular, in [19] the authors showed that the mean conditional sojourn time of a customer was decreasing as its relative priority increased, that it was uniformly bounded in the second moments of the service requirements and that the approximation was exact in various scenarios: one class $K = 1$, multi class with equal weights, total mean sojourn time for exponentially distributed service requirements.

From Proposition 6.6 we get as a corollary the interpolation approximation for the unconditional waiting time:

Corollary 6.7. *The interpolation approximation (of order 2) of the complementary distribution function of the unconditional waiting time of a tagged class-k customer is given by*

$$\begin{aligned}
&W_k^{DPS,INT}(\lambda,x) := \int_0^\infty W_k^{DPS,INT}(\lambda,b,x)\mathrm{d}F_k(b)\\
&= \lambda(1-\rho)\sum_{j=1}^{K}\alpha_j\Bigg(\left(1+\frac{g_k}{g_j}\right)\int_{(1-\rho)x\frac{g_k}{g_j}}^{\infty}\left(\int_{(1-\rho)x}^{\frac{g_j}{g_k}b}(1-F_j(b_j))\,\mathrm{d}b_j\right)\mathrm{d}F_k(b)\\
&\quad+\left(\int_{(1-\rho)x\frac{g_k}{g_j}}^{\infty}\left(\int_{\frac{g_j}{g_k}b}^{\infty}(1-F_j(b_j))\,\mathrm{d}b_j\right)\mathrm{d}F_k(b)\right)\\
&\quad+\int_{(1-\rho)xg_k/g_j}^{\infty}\left(\left(b-\frac{g_k}{g_j}(1-\rho)x\right)(1-F_j((1-\rho)x)) - \frac{g_k}{g_j}\int_{(1-\rho)x}^{b\frac{g_j}{g_k}}(1-F_j(b_j))\mathrm{d}b_j\right)\mathrm{d}F_k(b)\Bigg)\\
&\quad+\rho^2\mathrm{e}^{-(1-\rho)x\frac{g_k}{\mathbb{E}[B_k]\mathbb{E}[V]}}.
\end{aligned} \tag{30}$$

From Equation (30) it can easily be obtained the approximation derived in [14, Section 2.2.] for the unconditional sojourn time distribution under Processor Sharing.

7 Numerical results

In this section we numerically investigate the accuracy of the approximations obtained in Proposition 5.6 and Proposition 6.4.

To measure the accuracy, for the RP model we use as reference the algorithm proposed by Kim et al. in [23, Section 3.2], that allows us to obtain the moments numerically for any service-time distribution, and we denote their results by $(*)^{KIM}$, where $*$ refers to the metric studied. For the DPS model we use the algorithm proposed by Rege et al. in [26, Section 1] which is only valid for exponential service times and we denote their results by $(*)^{REGE}$ where, again, $*$ refers to the metric under consideration.

We review now the service time distributions we will use. We recall that a random variable B_i is exponentially

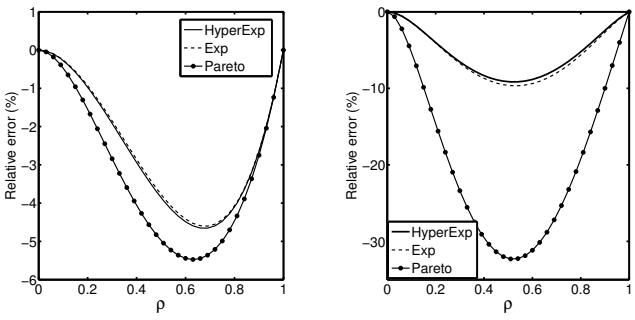

Figure I: Scenario 1. Relative error under RP of the first (left) and second moments (right) of the total number of customers in the system for hyper-exponential, exponential and Pareto service-time distributions.

distributed if $F_i(b_i) = 1 - \mathrm{e}^{-b_i/\mathbb{E}[B_i]}$. We say that B_i has Pareto distribution with scale parameter c_i and shape parameter γ_i if $F_i(b_i) = 1 - \left(\frac{1}{1+c_i b_i}\right)^{\gamma_i}$. We will further consider hyper-exponential distributions. We say that B_i has a hyper-exponential distribution with m_i phases if

$$F_i(b_i) = 1 - \sum_{k=1}^{m_i} \beta_{ik} \mathrm{e}^{(-b_i/\mathbb{E}[B_{ik}])}, \tag{31}$$

where β_{ik} is the probability that a class-i customer is exponentially distributed with mean $\mathbb{E}[B_{ik}]$. A particular case of the hyper-exponential distribution is the so-called *degenerate hyper-exponential* distribution. In this case one of the phases has mean 0. For instance, let us consider the case of 2 phases, $m_i = 2$, and let $\beta_{i1} = w, \beta_{i2} = 1 - w, w \in [0,1], \mathbb{E}[B_{i1}] = 1/(\mu_i w)$ and $\mathbb{E}[B_{i2}] = 0$. It then follows that $\mathbb{E}[B_i] = 1/\mu_i$ and $\mathbb{E}[B_i^2] = \frac{2w}{(w\mu_i)^2} = \frac{2}{w\mu_i^2}$. Hence, the coefficient of variation is $C_{B_i}^2 = 2/w - 1$, so that it ranges from 1 to ∞ as w changes from 1 till 0.

We make the observation that if classes $k = 1, \ldots, m_i$ are exponentially distributed (where class k has arrival rate λ_k and mean service requirement $\mathbb{E}[B_k]$) and have the same DPS weight, $g_1 = \ldots = g_{m_i}$, then they can be seen as a single (merged) class i with a hyperexponential distribution with parameters $\beta_{ik} = \lambda_k / \sum_{l=1}^{m_i} \lambda_l$ and $\mathbb{E}[B_{ik}] = \mathbb{E}[B_k]$, for each phase $k = 1, \ldots, m_i$. This allows us to calculate the moments in DPS with hyperexponential distribution using the algorithm of [26].

We note that the exponential distribution has a constant hazard rate, while the hyper-exponential and Pareto distributions have a decreasing hazard rate, and their second moment can be made arbitrarily large. Finally we remark that the hyperexponential distribution satisfies the sufficient condition (2) in order for the admissibility condition to hold, whereas Pareto does not satisfy it; moments of an order higher than γ_i are unbounded.

Throughout this section the performance criteria will be the relative error. For the first and second moments of the number of customers, we will hence calculate $100\% \times \frac{\mathbb{E}[N] - \mathbb{E}\left[N^{INT}\right]}{\mathbb{E}[N]}$ and $100\% \times \frac{\mathbb{E}[N^2] - \mathbb{E}\left[\left(N^{INT}\right)^2\right]}{\mathbb{E}[N^2]}$, respectively.

RP model

We measure the accuracy of the approximation obtained in Proposition 5.6 by considering the first and second moments that are given in Equations (18) and (20).

Scenario 1. In Figure I we plot the relative error of the first and second moments of the total number of customers in the system with respect to the load for exponential, hyper-exponential and Pareto service time

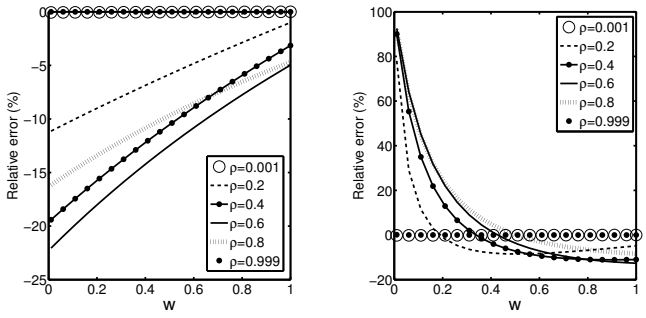

Figure II: Scenario 2. Relative error under RP of the first (left) and second moments (right) of the total number of customers in the system with respect to w for different values of the load.

distributions. We consider two classes and set $\mathbb{E}[B_1] = 11/3$ and $\mathbb{E}[B_2] = 44/3$. We assume that an arriving customer is of class 1 (class 2) with probability $\alpha_1 = 8/12$ ($\alpha_2 = 4/12$). The weights are set equal to $p_1 = 2$ and $p_2 = 5$. We observe in Figure I that the first moment remains accurate for any choice of the service time distribution. The absolute relative error of the second moment is small for the exponential and hyper-exponential distribution, but reaches the value of 30% for Pareto distributions. The fact that Pareto does not satisfy the admissibility condition (2) might explain the large relative error.

Scenario 2. In Figure II we consider 2 classes of customers. Class-1 customers' service requirements follow an exponential distribution of rate μ_1, while class-2 customers' service requirements follow a degenerate hyper-exponential distribution as defined in Equation (31) with parameters $m_2 = 2, \beta_{21} = \ w, \beta_{22} = 1 - w, \mathbb{E}[B_{21}] = 1/(\mu_2 w)$ and $\mathbb{E}[B_{22}] = 0$. We consider $p_1 = 2$, $p_2 = 5$, $\alpha_1 = 7/12$, $\alpha_2 = 5/12$, $\mathbb{E}[B_1] = 11/3$, $\mathbb{E}[B_2] = 1/mu_2 = 44/3$. In Figure II we plot the relative error of the first and second moments of the total number of customers in the system with respect to w for different values of the load. Observe that, as expected, for $\rho \approx 0$ and $\rho \approx 1$ our interpolation approximation is exact. The absolute largest error occurs for intermediate values of the load, as w approaches 0, that is, as the coefficient of variation of B_2 goes to ∞.

DPS model

We first measure the accuracy of the approximation obtained for the queue length in Proposition 6.4 by considering the first and second moments that are given in Equations (25) and (27). We do this both for exponentially distributed service times and for degenerate hyper-exponential distributed service times.

In Figure III we consider Scenario 1 with weights $g_1 = 2$ and $g_2 = 5$. We plot the relative error of the first and second moments of the number of customers in the system, respectively, for exponentially distributed service requirements. We observe that our approximation for the first and second moments is accurate with at most 1.2% and 3% absolute relative error, respectively.

In Figure IV we consider Scenario 2 with weights $g_1 = 2$, $g_2 = 5$. Hence, class 1 has exponential service times and class 2 has degenerate hyper-exponential service times. We plot the relative error of the first and second moments of the total number of customers in the system with respect to w for different values of the load. Observe, again, that as for the RP model, when $\rho \approx 0$ and $\rho \approx 1$ our approximation is exact. The absolute

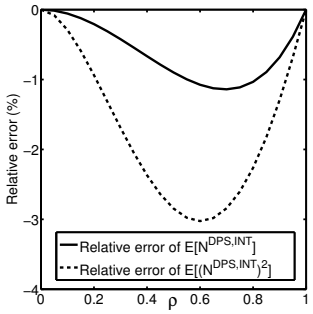

Figure III: Scenario 1. Relative error under DPS of the first and second moments of the total number of customers in the system.

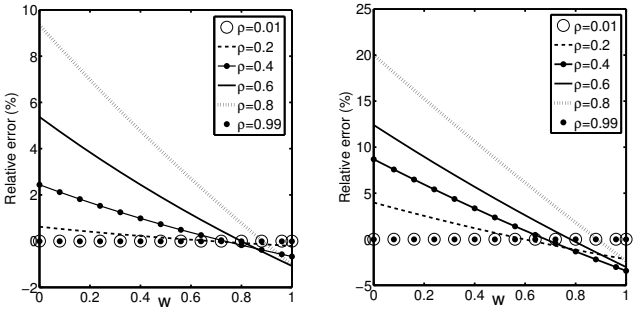

Figure IV: Scenario 2. Relative error under DPS of the first (left) and second moments (right) of the total number of customers in the system for different values of the load.

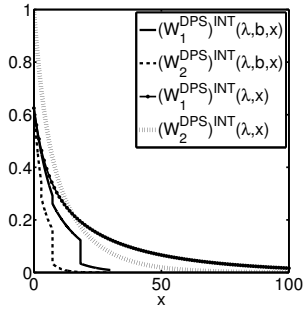

Figure V: Scenario 1. Complementary distribution of the conditional and unconditional waiting time of a class-k customer under DPS.

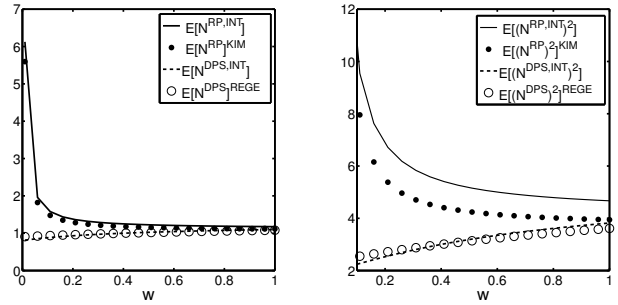

Figure VI: Scenario 3. First (left) and second moments (right) of the total number of customers in the system under DPS and RP for exponential service-time distributions.

largest relative error occurs for intermediate values of the load, as w approaches 0, that is, as the coefficient of variation of B_2 approaches ∞.

In Figure V we use Proposition 6.6 to plot the complementary distribution of the conditional and unconditional waiting time of a class-k customer for Scenario 1. For the conditional waiting time we set the service time of the tagged customer to $b = 11/3$. Class 2 gets relatively a larger weight ($g_1 = 2$, $g_2 = 5$), and as a consequence, we see in Figure V that the conditional waiting time of class 2 is stochastically smaller than that of class 1. However, the service time of class 1 customers is smaller than that of class 2. As a result, we see that the probability that the unconditional waiting time of class 1 is bigger than x is larger than that of class 2, for x small enough.

Comparing RP and DPS

Scenario 3. In Figure VI we plot the first and second moments of the total number of customers in the system for RP and DPS. We plot both our interpolation approximation as well as the exact results obtained from the literature. We consider two classes, class 1 is exponentially distributed with $\mathbb{E}[B_1] = 5$, and class 2 is degenerate

hyper- exponential with $\mathbb{E}[B_2] = 2$. We assume that an arriving customer is of class 1 (class 2) with probability $\alpha_1 = 8/12$ ($\alpha_2 = 4/12$). The weights of the DPS and RP are the same, namely, $g_1 = p_1 = 5$ and $g_2 = p_2 = 1$. We observe in Figure VI that our approximation is rather accurate. In addition, as $w \to 0$, that is, as the coefficient of variation grows large, the performance of DPS is better than that of RP, both for our approximation as for the exact results. This is something we could expect, since as $w \to 0$, the second moment of class 2 tends to ∞, and therefore the performance of RP (which is non-preemptive) is worse than DPS (which is time-sharing).

References

[1] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing systems*, 53(1-2):53–63, 2006.

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

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

[5] S.C. Borst, D.T.M.B. van Ooteghem, and A.P. Zwart. Tail asymptotics for discriminatory processor sharing queues with heavy-tailed service requirements. *Performance Evaluation*, 61(2–3):281–298, 2005.

[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, N. Hegde, and R. Núñez-Queija. Exact and approximate analysis of sojourn times in finite discriminatory processor sharing queues. *AEU International Journal on Electronic Communications*, 60:109–115, 2006.

[8] T. Bu and D. Towsley. Fixed point approximation for TCP behaviour in an AQM network. In *Proceedings of ACM SIGMETRICS/Performance*, pages 216–225, 2001.

[9] D. Burman and D. Smith. An asymptotic analysis of a queueing system with markov-modulated arrivals. *Operations Research*, 34:105–119, 1986.

[10] S.K. Cheung, J.L. van den Berg, R.J. Boucherie, R. Litjens, and F. Roijers. An analytical packet/flow-level modelling approach for wireless LANs with quality-of-service support. In *Proceedings of ITC-19*, 2005.

[11] J.L. Dorsman, R.D. van der Mei, and M. Vlasiou. Analysis of a two-layered network with correlated queues by means of the power-series algorithm. *Performance Evaluation*, 70:1072–1089, 2013.

[12] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27(3):519–532, 1980.

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

[14] P.J. Fleming and B. Simon. Interpolation approximations of sojourn time distributions. *Operations Research*, 39(2):251–260, 1991.

[15] S. Grishechkin. On a relationship between processor sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Prob.*, 24(3):653–698, 1992.

[16] 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.

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

[18] Y. Hayel and B. Tuffin. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *Proceedings of Networking*, 2005.

[19] A. Izagirre, U. Ayesta, and I.M. Verloop. Sojourn time approximations in a multi-class time-sharing system. *IEEE INFOCOM*, 2014.

[20] A. Izagirre, U. Ayesta, and I.M. Verloop. Heavy-traffic analysis of a multi-class queue with relative priorities. *Probability in Engineering and Informational Sciences*, 29(2):153–180, 2015.

[21] A.A. Kherani and R. Núñez-Queija. TCP as an implementation of age-based scheduling: fairness and performance. In *Proceedings of IEEE INFOCOM*, 2006.

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

[23] 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.

[24] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14(2):242–261, 1967.

[25] 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.

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

[27] M.I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36:454–469, 1988.

[28] M.I. Reiman and B. Simon. Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models*, 4:191–233, 1988.

[29] M.I. Reiman and B. Simon. Open queueing systems in light traffic. *Oper. Res.*, 14:26–59, 1989.

[30] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM*, 2005.

[31] S. Varma and A.M. Makowski. Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, 20:245–265, 1994.

[32] I.M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Operations Research*, 59(3):648–660, 2011.

[33] J. Walrand. An introduction to queueing networks. In *Prentice-Hall, Englewood Cliffs, NJ*, 1988.

Appendix A: Proof of Proposition 4.1

We provide the proof of how to obtain the zeroth and first light-traffic derivatives. This is based on the analysis of J. Walrand in [33, Chapter 6.3]. Higher order light-traffic derivatives can be obtained in a similar way.

Consider a system that starts at time $-Z$ and that keeps going until time T, being $Z, T > 0$ given. Let $G(\lambda, \vec{y}, -Z, T)$ denote the term we are interested in approximating and note that $\lim_{Z,T\to\infty} G(\lambda, \vec{y}, -Z, T) = G(\lambda, \vec{y})$. Let $A(s,t)$ denote the number of arrivals in the interval $[s,t)$ in addition to the tagged customer who is assumed to arrive at time 0. Throughout this section we assume that the limits (with respect to Z and T) and expectations can be interchanged. We then have

$$G(\lambda, \vec{y}, -Z, T) = \sum_{a=0}^{\infty} G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = a\right) \cdot \frac{(\lambda(T+Z))^a}{a!} \mathrm{e}^{-\lambda(T+Z)}, \tag{32}$$

where $G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = a\right)$ is conditioned on the fact that there are exactly a arrivals in the interval $[-Z, T)$. Evaluating it at $\lambda = 0$ gives

$$G(\lambda, \vec{y}, -Z, T)\Big|_{\lambda=0} = G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = 0\right), \tag{33}$$

and now taking the limit $Z, T \to \infty$ we obtain the zeroth light-traffic derivative

$$G^{(0)}(0,y) := \lim_{Z,T\to\infty} G(\lambda, \vec{y}, -Z, T)\Big|_{\lambda=0} \quad = \quad G\left(0, \vec{y}, -Z, T \middle| A(-\infty, \infty) = 0\right)$$

where the second equality follows from (33).

Next, consider the derivative with respect to λ in Equation (32) and evaluate it at $\lambda = 0$. This gives

$$\begin{aligned}
&\frac{\partial}{\partial\lambda} G(\lambda, \vec{y}, -Z, T)\Big|_{\lambda=0} \\
&= -G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = 0\right) \cdot (T+Z) + G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = 1\right) \cdot (T+Z) \\
&= \int_{-Z}^{T} \left(G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = 1, \tau_1 = t\right) - G\left(\lambda, \vec{y}, -Z, T \middle| A(-Z,T) = 0\right) \right) \mathrm{d}t,
\end{aligned} \tag{34}$$

where τ_1 is the arrival time of the first customer. The second equality holds because the arrivals follow a Poisson process. Hence given that the number of arrivals in $[-Z, T)$ is one ($A(-Z,T) = 1$), we have that τ is uniformly distributed on $[-Z, T)$.

Now taking $Z, T \to \infty$ we obtain the first light-traffic derivative

$$G^{(1)}(0, \vec{y}) := \lim_{Z,T\to\infty} \frac{\partial}{\partial\lambda} G(\lambda, \vec{y}, -Z, T)\Big|_{\lambda=0} = \int_{-\infty}^{\infty} \left(G\left(0, \vec{y} \middle| A(-\infty,\infty) = 1, \tau_1 = t\right) - G\left(0, \vec{y} \middle| A(-\infty,\infty) = 0\right) \right) \mathrm{d}t,$$

where the second equality follows from (34).

Appendix B: Proof of Lemma 5.5

As explained in Equation (1) the light-traffic approximation can be written as

$$\psi^{RP,LT}(\lambda,\vec{z}) = \left(\psi^{RP}\right)^{(0)}(0,\vec{z}) + \lambda\left(\psi^{RP}\right)^{(1)}(0,\vec{z}) + \lambda^2\left(\psi^{RP}\right)^{(2)}(0,\vec{z}). \tag{35}$$

We now obtain the zeroth, first and second light-traffic derivatives for the joint pgf $\psi^{RP}(\lambda,\vec{z})$ of the joint stationary queue lengths at arbitrary time epochs.

From (10) it follows directly that the zeroth derivative in $\lambda = 0$ satisfies

$$\left(\psi^{RP}\right)^{(0)}(\lambda,\vec{z})\Big|_{\lambda=0} = 1. \tag{36}$$

Taking the derivative in (10) we obtain that the first derivative satisfies

$$\begin{aligned}
&\left(\psi^{RP}\right)^{(1)}(0,\vec{z}) = \frac{\partial\psi^{RP}(0,\vec{z})}{\partial\lambda} \\
&= -\mathbb{E}[B] + \frac{1}{1-\sum_{k=1}^K \alpha_k z_k}\sum_{i=1}^K \alpha_i z_i\Bigg(\left(-\mathbb{E}[B] + \frac{p_i}{\alpha_i}\frac{\partial^2 r(\lambda,\vec{z})}{\partial\lambda\partial z_i}\right)\left(1 - B_i^*\left(\lambda - \lambda\sum_{k=1}^K \alpha_k z_k\right)\right) \\
&\quad + \left(1-\rho+\frac{p_i}{\alpha_i}\frac{\partial r(\lambda,\vec{z})}{\partial z_i}\right)\left(-B_i^{*\prime}\left(\lambda-\lambda\sum_{k=1}^K\alpha_k z_k\right)\left(1-\sum_{k=1}^K \alpha_k z_k\right)\right)\Bigg)\Bigg|_{\lambda=0} \\
&= -\mathbb{E}[B] + \sum_{i=1}^K \alpha_i \mathbb{E}[B_i] z_i\left(1 + \frac{p_i}{\alpha_i}\frac{\partial r(\lambda,\vec{z})}{\partial z_i}\Bigg|_{\lambda=0}\right) = -\mathbb{E}[B] + \sum_{i=1}^K \alpha_i \mathbb{E}[B_i] z_i.
\end{aligned} \tag{37}$$

In the last step we used that

$$\begin{aligned}
\frac{\partial r(\lambda,\vec{z})}{\partial z_i}\Bigg|_{\lambda=0} &= \mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^K Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}\cdot\mathbf{1}_{(\sum_{k=1}^K Q_k>0)}\right]\Bigg|_{\lambda=0} \\
&= \mathbb{P}(\textstyle\sum Q_k>0)\Big|_{\lambda=0}\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^K Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}|\sum_{k=1}^K Q_k>0\right]\Bigg|_{\lambda=0} \\
&= \rho\Big|_{\lambda=0}\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^K Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}|\sum_{k=1}^K Q_k>0\right]\Bigg|_{\lambda=0} = 0,
\end{aligned} \tag{38}$$

since the RP model is a work-conserving policy $\mathbb{P}(\sum Q_k>0)=\rho$ is equal to the probability of the server being busy, which is independent of the scheduling policy. The other term is finite since it satisfies

$$\begin{aligned}
&\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^K Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}\Bigg|\sum_{k=1}^K Q_k>0\right]\Bigg|_{\lambda=0} \\
&= \sum_{\substack{q_1=0,\ldots,q_K=0\\ \sum_{k=1}^K q_k>0}}^{\infty}\frac{q_i}{\sum_{k=1}^K q_k p_k}\cdot\frac{z_1^{q_1}\cdot\ldots\cdot z_K^{q_K}}{z_i}\mathbb{P}\left(Q_1=q_1,\ldots,Q_K=q_K\Bigg|\sum_{k=1}^K Q_k>0\right)\Bigg|_{\lambda=0} \\
&= \sum_{\substack{q_1=0,\ldots,q_K=0\\ \sum_{k=1}^K q_k=1}}^{\infty}\frac{q_i}{\sum_{k=1}^K q_k p_k}\cdot\frac{z_1^{q_1}\cdot\ldots\cdot z_K^{q_K}}{z_i}\mathbb{P}\left(Q_1=q_1,\ldots,Q_K=q_K\Bigg|\sum_{k=1}^K Q_k>0\right)\Bigg|_{\lambda=0} + o(\lambda^2) \\
&= \frac{q_i}{q_i p_i}\cdot 1\cdot\ldots\cdot 1\cdot\mathbb{P}\left(\vec{Q}=e_i\Bigg|\sum_{k=1}^K Q_k>0\right)\Bigg|_{\lambda=0} = \frac{\alpha_i}{p_i},
\end{aligned} \tag{39}$$

due to

$$\mathbb{P}\left(\vec{Q}=e_i\middle|\sum_{k=1}^{K}Q_k>0\right)\Big|_{\lambda=0}=\frac{\mathbb{P}\left(\vec{Q}=e_i\cap\sum_{k=1}^{K}Q_k>0\right)}{\mathbb{P}\left(\sum_{k=1}^{K}Q_k>0\right)}\Big|_{\lambda=0}=\frac{\mathbb{P}\left(\vec{Q}=e_i\right)}{\mathbb{P}\left(\sum_{k=1}^{K}Q_k>0\right)}\Big|_{\lambda=0}=\frac{\alpha_i\rho(1-\rho)}{\rho}\Big|_{\lambda=0}=\alpha_i, \tag{40}$$

which follows from $A\cdot\mathbb{P}\left(\vec{Q}=\vec{0}\right)=B\cdot\mathbb{P}\left(\vec{Q}=e_i\right)+o(\lambda)$, where $A=\alpha_i\rho$ and $B=1$ from [22, Equation 1] and again the fact that the RP model is a work-conserving policy.

The second derivative satisfies

$$\begin{aligned}
&\left(\psi^{RP}\right)^{(2)}(\lambda,\vec{z})\Big|_{\lambda=0}=\frac{\partial^2\psi^{RP}(\lambda,\vec{z})}{\partial\lambda^2}\Big|_{\lambda=0}\\
&=\frac{1}{1-\sum_{k=1}^{K}\alpha_k z_k}\sum_{i=1}^{K}\alpha_i z_i\left(\frac{p_i}{\alpha_i}\frac{\partial^3 r(\lambda,\vec{z})}{\partial\lambda^2\partial z_i}\left(1-B_i^*\left(\lambda-\lambda\sum_{k=1}^{K}\alpha_k z_k\right)\right)\right.\\
&\quad+2\left(-\mathbb{E}[B]+\frac{p_i}{\alpha_i}\frac{\partial^2 r(\lambda,\vec{z})}{\partial\lambda\partial z_i}\right)\left(-B_i^{*'}\left(\lambda-\lambda\sum_{k=1}^{K}\alpha_k z_k\right)\right)\left(1-\sum_{k=1}^{K}\alpha_k z_k\right)\\
&\quad\left.+\left(1-\rho+\frac{p_i}{\alpha_i}\frac{\partial r(\lambda,\vec{z})}{\partial z_i}\right)\left(-B_i^{*''}\left(\lambda-\lambda\sum_{k=1}^{K}\alpha_k z_k\right)\right)\left(1-\sum_{k=1}^{K}\alpha_k z_k\right)^2\right)\Big|_{\lambda=0}\\
&=\sum_{i=1}^{K}\alpha_i z_i\left(2\left(-\mathbb{E}[B]+\frac{p_i}{\alpha_i}\frac{\partial^2 r(\lambda,\vec{z})}{\partial\lambda\partial z_i}\Big|_{\lambda=0}\right)\mathbb{E}[B_i]-\left(1+\alpha_i p_i\frac{\partial r(\lambda,\vec{z})}{\partial z_i}\Big|_{\lambda=0}\right)\mathbb{E}[B_i^2]\left(1-\sum_{k=1}^{K}\alpha_k z_k\right)\right)\\
&=\sum_{i=1}^{K}\alpha_i z_i\mathbb{E}[B_i^2]\left(\sum_{k=1}^{K}\alpha_k z_k-1\right),
\end{aligned} \tag{41}$$

where in the last step we used

$$\begin{aligned}
&\frac{\partial^2 r(\lambda,\vec{z})}{\partial\lambda\partial z_i}\Big|_{\lambda=0}=\frac{\partial\left(\rho\cdot\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^{K}Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}\middle|\sum_{k=1}^{K}Q_k>0\right]\right)}{\partial\lambda}\Big|_{\lambda=0}\\
&=\mathbb{E}[B]\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^{K}Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}\middle|\sum_{k=1}^{K}Q_k>0\right]\Big|_{\lambda=0}+\rho|_{\lambda=0}\cdot\frac{\partial\mathbb{E}\left[\frac{Q_i}{\sum_{k=1}^{K}Q_k p_k}\cdot\frac{z_1^{Q_1}\cdot\ldots\cdot z_K^{Q_K}}{z_i}\middle|\sum_{k=1}^{K}Q_k>0\right]}{\partial\lambda}\Big|_{\lambda=0}\\
&=\mathbb{E}[B]\frac{\alpha_i}{p_i},
\end{aligned} \tag{42}$$

which follows from Equation (39).

From Equations (36), (37) and (41), together with Equation (35), we obtain the result in Lemma 5.5 and conclude the proof.

Appendix C: Proof of Proposition 5.6

We obtain the light and heavy-traffic interpolation of the joint pgf of the queue length under the RP policy. As explained in Section 4.3 we approximate $G\left(\lambda,\vec{z}^{(1-\rho)}\right)=\psi^{RP}\left(\lambda,\vec{z}^{(1-\rho)}\right)$ by the polynomial

$$\hat{G}(\lambda,\vec{z})=h_0(\vec{z})+\lambda h_1(\vec{z})+\lambda^2 h_2(\vec{z})+\lambda^3 h_3(\vec{z}).$$

Unnormalizing, that is, for $f_\lambda^{-1}(\vec{z}) = \vec{z}^{(1-\rho)^{-1}}$, we have

$$\psi^{RP,INT}(\lambda,\vec{z}) = \hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right) = h_0\left(\vec{z}^{(1-\rho)^{-1}}\right) + \lambda h_1\left(\vec{z}^{(1-\rho)^{-1}}\right) + \lambda^2 h_2\left(\vec{z}^{(1-\rho)^{-1}}\right) + \lambda^3 h_3\left(\vec{z}^{(1-\rho)^{-1}}\right).$$

Then, from the light-traffic conditions (7) we obtain $h_0(\vec{z}), h_1(\vec{z}), h_2(\vec{z})$. First we have, $\psi^{RP,INT}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)\Big|_{\lambda=0} = \psi^{RP,INT}(0,\vec{z}) = h_0(\vec{z})$. Together with (36) we obtain $h_0(\vec{z}) = 1$.

Second,

$$\begin{aligned}
&\frac{\partial\psi^{RP,INT}(\lambda,\vec{z})}{\partial\lambda}\bigg|_{\lambda=0} = \frac{\mathrm{d}\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0} = \frac{\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\partial\lambda}\bigg|_{\lambda=0} + \sum_{i=1}^{K}\frac{\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\partial z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0}\\
&= \left(h_1\left(\vec{z}^{(1-\rho)^{-1}}\right) + 2\lambda h_2\left(\vec{z}^{(1-\rho)^{-1}}\right) + 3\lambda^2 h_3\left(\vec{z}^{(1-\rho)^{-1}}\right)\right)\Big|_{\lambda=0}\\
&+\sum_{i=1}^{K}\left(\frac{\mathrm{d}h_0\left(\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}z_i} + \lambda\frac{\mathrm{d}h_1\left(\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}z_i} + \lambda^2\frac{\mathrm{d}h_2\left(\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}z_i} + \lambda^3\frac{\mathrm{d}h_3\left(\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}z_i}\right)\Big|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\Big|_{\lambda=0}\\
&= h_1(\vec{z}) + \sum_{i=1}^{K}\frac{\mathrm{d}h_0\left(\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\Big|_{\lambda=0} = h_1(\vec{z}) + \sum_{i=1}^{K}\frac{\mathrm{d}(1)}{\mathrm{d}z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\Big|_{\lambda=0} = h_1(\vec{z}).
\end{aligned}$$

Together with (37) we obtain $h_1(\vec{z}) = -\mathbb{E}[B] + \sum_{i=1}^{K}\alpha_i\mathbb{E}[B_i]z_i$.

Third,

$$\begin{aligned}
&\frac{\partial^2\psi^{RP,INT}(\lambda,\vec{z})}{\partial\lambda^2}\bigg|_{\lambda=0} = \frac{\mathrm{d}^2\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda^2}\bigg|_{\lambda=0}\\
&= \frac{\partial^2\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\partial\lambda^2}\bigg|_{\lambda=0} + \sum_{i=1}^{K}\frac{\partial\left(\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)/\partial\lambda\right)}{\partial z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0}\\
&\quad+\sum_{i=1}^{K}\left(\left(\frac{\partial\left(\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)/\partial z_i\right)}{\partial\lambda}\bigg|_{\lambda=0} + \frac{\partial\left(\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)/\partial z_i\right)}{\partial z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0}\right)\right.\\
&\qquad\left.\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0} + \frac{\partial\hat{G}\left(\lambda,\vec{z}^{(1-\rho)^{-1}}\right)}{\partial z_i}\bigg|_{\lambda=0}\cdot\frac{\mathrm{d}^2\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda^2}\bigg|_{\lambda=0}\right)\\
&= 2h_2(\vec{z}) + 2\sum_{i=1}^{K}\frac{\mathrm{d}h_1(\vec{z})}{\mathrm{d}z_i}\cdot\frac{\mathrm{d}\left(z_i^{(1-\rho)^{-1}}\right)}{\mathrm{d}\lambda}\bigg|_{\lambda=0} = 2h_2(\vec{z}) + 2\mathbb{E}[B]\sum_{i=1}^{K}\alpha_i\mathbb{E}[B_i]z_i\ln(z_i).
\end{aligned}$$

Together with (41) we obtain $h_2(\vec{z}) = \frac{1}{2}\left(\sum_{i=1}^{K}\alpha_i z_i\mathbb{E}[B_i^2]\left(\sum_{k=1}^{K}\alpha_k z_k - 1\right) - 2\mathbb{E}[B]\sum_{i=1}^{K}\alpha_i z_i\mathbb{E}[B_i]\ln(z_i)\right)$.

Finally, from Proposition 4.2 and noting that $G^{HT}(\vec{z})$ is equal to Equation (21), we conclude the proof.

Appendix D: Proof of Lemma 5.7

As explained in Equation (1) the light-traffic approximation can be written as

$$\widetilde{W}_k^{DROS,LT}(\lambda,u) = \left(\widetilde{W}_k^{DROS}\right)^{(0)}(0,u) + \lambda\left(\widetilde{W}_k^{DROS}\right)^{(1)}(0,u).$$

We now obtain the zeroth and first light-traffic derivatives of the Laplace Transform of the waiting time of

a class-l customer under DROS using the result presented in Theorem 5.2.

The zeroth derivative satisfies $\left(\widetilde{W}_k^{DROS}\right)^{(0)}(\lambda,u)\Big|_{\lambda=0} = [1-\rho+T_l^{DROS}(u,\vec{1})]\Big|_{\lambda=0} = 1+T_l^{DROS}(u,\vec{1})\Big|_{\lambda=0} = 1$, since from Equation (13) we obtain

$$T_l^{DROS}(u,\vec{1})\Big|_{\lambda=0}\left(1+\sum_{i=1}^{K}\frac{p_i}{p_l}(1-B_i^*(u))\right) = 0 \Rightarrow T_l^{DROS}(u,\vec{1})\Big|_{\lambda=0} = 0.$$

And the first derivative satisfies

$$\left(\widetilde{W}_k^{DROS}\right)^{(1)}(\lambda,u)\Big|_{\lambda=0} = -\mathbb{E}[B] + \frac{\partial T_l^{DROS}(u,\vec{1})}{\partial\lambda}\Big|_{\lambda=0} = -\mathbb{E}[B] + \left(\sum_{i=1}^{K}\alpha_i\frac{1-B_i^*(u)}{u}\right),$$

since, again from Equation (13),

$$\begin{aligned}&\frac{\partial T_l^{DROS}(u,\vec{1})}{\partial\lambda}\Big|_{\lambda=0}\left(1+\sum_{i=1}^{K}\frac{p_i}{p_l}\mathbb{E}[Q_i^*]\Big|_{\lambda=0}(1-B_i^*(u))\right) = \sum_{i=1}^{K}\left(\alpha_i+p_i r(\lambda,\vec{1})\Big|_{\lambda=0}\right)\frac{1-B_i^*(u)}{u}\\ &\Rightarrow \frac{\partial T_l^{DROS}(u,\vec{1})}{\partial\lambda}\Big|_{\lambda=0} = \left(\sum_{i=1}^{K}\alpha_i\frac{1-B_i^*(u)}{u}\right),\end{aligned}$$

where $\mathbb{E}[Q_i^*]\Big|_{\lambda=0} = 0$ and $r(\lambda,\vec{1})\Big|_{\lambda=0} = 0$ from Equation (38).

Appendix E: Proof of Lemma 6.3

We obtain the zeroth and first light-traffic derivatives and give the main steps to obtain the second light-traffic derivative of the joint pgf of the queue length under DPS. The zeroth derivative satisfies

$$\left(\psi^{DPS}\right)^{(0)}(\lambda,\vec{z}) = \psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=0\right) = z_1^0\cdots z_K^0 = 1.$$

Let t denote the time epoch in which a customer arrives to the system, and let U_t denote its class. For the first derivative there might happen two different cases:
If $t>0$, we have $z_{U_t}^0 = 1$ and therefore $\psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=1,\tau_1=t\right) - \psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=0\right) = 0$.
If $t<0$, we have

$$\psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=1,\tau_1=t\right) = \begin{cases} z_{U_t}^0 & \text{if } t+B_{U_t}<0\\ z_{U_t} & \text{if } t+B_{U_t}>0,\end{cases}$$

such that $\psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=1,\tau_1=t\right) = \mathbb{E}\left[\mathbf{1}\left[t+B_{U_t}<0\right]+\mathbf{1}\left[t+B_{U_t}>0\right]z_{U_t}\right]$.

Therefore,

$$\begin{aligned}\left(\psi^{DPS}\right)^{(1)}(\lambda,\vec{z}) &= \int_{-\infty}^{0}\left(\psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=1,\tau_1=t\right)-\psi^{DPS}\left(0,\vec{z}\Big|A(-\infty,\infty)=0\right)\right)\mathrm{d}t\\ &= \mathbb{E}\left[\int_{-\infty}^{0}\left(\mathbf{1}\left[t+B_{U_t}<0\right]+\mathbf{1}\left[t+B_{U_t}>0\right]z_{U_t}-1\right)\mathrm{d}t\right] = \mathbb{E}\left[(z_{U_t}-1)B_{U_t}\right].\end{aligned}$$

To calculate the second derivative let us assume $t'<t''$, where t' and t'' denote the arrival epochs of two

customers. At the end, because of symmetry, we multiply the final result by 2. Then, we separate three main different cases:

If $0 < t' < t''$

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t'\right)$$
$$-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t''\right)+\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=0\right)=0.$$

If $t' < 0$ & $0 < t''$ we have two cases:

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)=\begin{cases} z^0_{U_{t'}}\cdot z^0_{U_{t''}} & \text{if } t'+B_{U_{t'}}<0 \\ z_{U_{t'}} & \text{if } t'+B_{U_{t'}}>0, \end{cases}$$

such that,

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)=\mathbb{E}\left[\mathbf{1}\left[t'+B_{U_{t'}}<0\right]+\mathbf{1}\left[t'+B_{U_{t'}}>0\right]z_{U_{t'}}\right]$$
$$=\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t'\right).$$

Therefore,

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t'\right)$$
$$-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t''\right)+\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=0\right)=0.$$

If $t' < t'' < 0$ there might happen several cases as shown below.

First, if $t'+B_{U_{t'}} < t''$ we have

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)=\begin{cases} z^0_{U_{t'}}\cdot z^0_{U_{t''}} & \text{if } t''+B_{U_{t''}}<0 \\ z_{U_{t''}} & \text{if } t''+B_{U_{t''}}>0, \end{cases}$$

such that,

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)=\mathbb{E}\left[\mathbf{1}\left[t''+B_{U_{t''}}<0\right]+\mathbf{1}\left[t''+B_{U_{t''}}>0\right]z_{U_{t''}}\right]$$
$$=\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t''\right).$$

Therefore,

$$\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=2,\tau_1=t',\tau_2=t''\right)-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t'\right)$$
$$-\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=1,\tau_1=t''\right)+\psi^{DPS}\left(0,\vec{z}\middle|A(-\infty,\infty)=0\right)=0.$$

Second, if $t' + B_{U_{t'}} > t''$ we have

$$
\begin{aligned}
&\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 2, \tau_1 = t', \tau_2 = t''\right) \\
&= \begin{cases}
z_{U_{t'}} & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} > B_{U_{t''}} \quad \& \quad t'' + B_{U_{t''}}\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t''}}} < 0 \quad \& \quad t' + B_{U_{t'}} + B_{U_{t''}} > 0 \\
1 & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} > B_{U_{t''}} \quad \& \quad t'' + B_{U_{t''}}\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t''}}} < 0 \quad \& \quad t' + B_{U_{t'}} + B_{U_{t''}} < 0 \\
z_{U_{t'}} \cdot z_{U_{t''}} & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} > B_{U_{t''}} \quad \& \quad t'' + B_{U_{t''}}\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t''}}} > 0 \\
z_{U_{t''}} & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} < B_{U_{t''}} \quad \& \quad t'' + (B_{U_{t'}} - t'' + t')\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t'}}} < 0 \quad \& \quad t' + B_{U_{t'}} + B_{U_{t''}} > 0 \\
1 & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} < B_{U_{t''}} \quad \& \quad t'' + (B_{U_{t'}} - t'' + t')\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t'}}} < 0 \quad \& \quad t' + B_{U_{t'}} + B_{U_{t''}} < 0 \\
z_{U_{t'}} \cdot z_{U_{t''}} & \text{if } (B_{U_{t'}} - t'' + t')\frac{g_{U_{t''}}}{g_{U_{t'}}} < B_{U_{t''}} \quad \& \quad t'' + (B_{U_{t'}} - t'' + t')\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t'}}} > 0
\end{cases}
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\left(\psi^{DPS}\right)^{(2)}(\lambda, \vec{z}) \\
&= 2 \cdot \int_{-\infty}^{0} \left(\int_{t'}^{0} \left(\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 2, \tau_1 = t', \tau_2 = t''\right) - \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = t'\right) \right.\right. \\
&\qquad \left.\left. -\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = t''\right) + \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 0\right) \right) \mathrm{d}t'' \right) \mathrm{d}t' \\
&= 2 \cdot \int_{0}^{\infty} \left(\int_{0}^{r} \left(\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 2, \tau_1 = -r, \tau_2 = -s\right) - \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = -r\right) \right.\right. \\
&\left.\left. -\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = -s\right) + \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 0\right) \right) \mathrm{d}s \right) \mathrm{d}r
\end{aligned}
$$

where

$$
\begin{aligned}
&\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 2, \tau_1 = -r, \tau_2 = -s\right) - \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = -r\right) \\
&\quad -\psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 1, \tau_1 = -s\right) + \psi^{DPS}\left(0, \vec{z} \middle| A(-\infty,\infty) = 0\right) \\
&= \mathbb{E}\Bigg[\mathbf{1}\left[-r + B_{U_{t'}} > -s\right] \Bigg(\\
&\quad \mathbf{1}\left[(B_{U_{t'}} + s - r)\frac{g_{U_{t''}}}{g_{U_{t'}}} > B_{U_{t''}}, -s + B_{U_{t''}}\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t''}}} < 0\right] \mathbf{1}\left[-r + B_{U_{t'}} + B_{U_{t''}} > 0\right]\left(z_{U_{t'}} - 1\right) \\
&\quad +\mathbf{1}\left[(B_{U_{t'}} + s - r)\frac{g_{U_{t''}}}{g_{U_{t'}}} > B_{U_{t''}}, -s + B_{U_{t''}}\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t''}}} > 0\right]\left(z_{U_{t'}} \cdot z_{U_{t''}} - 1\right) \\
&\quad +\mathbf{1}\left[(B_{U_{t'}} + s - r)\frac{g_{U_{t''}}}{g_{U_{t'}}} < B_{U_{t''}}, -s + (B_{U_{t'}} + s - r)\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t'}}} < 0\right] \cdot \mathbf{1}\left[-r + B_{U_{t'}} + B_{U_{t''}} > 0\right]\left(z_{U_{t''}} - 1\right) \\
&\quad +\mathbf{1}\left[(B_{U_{t'}} + s - r)\frac{g_{U_{t''}}}{g_{U_{t'}}} < B_{U_{t''}}, -s + (B_{U_{t'}} + s - r)\frac{g_{U_{t'}} + g_{U_{t''}}}{g_{U_{t'}}} > 0\right]\left(z_{U_{t'}} \cdot z_{U_{t''}} - 1\right) \\
&\quad +\mathbb{E}\left[\mathbf{1}\left[-r + B_{U_{t'}} > 0\right](1 - z_{U_{t'}})\right] \\
&\quad +\mathbb{E}\left[\mathbf{1}\left[-s + B_{U_{t''}} > 0\right](1 - z_{U_{t''}})\right] \Bigg)\Bigg].
\end{aligned}
$$

After working out the six integrals we end up with the result in Lemma 6.3.

Appendix F: Proof of Lemma 6.5

We give the main steps to obtain the first light-traffic derivative of the complementary distribution function of the conditional waiting time under DPS.

To calculate the first derivative $\left(W_k^{DPS}\right)^{(1)}(0,b,x)$ we need to calculate $\int_{-\infty}^{\infty}\mathbb{E}\Big[\mathbf{1}\left[W_k^{DPS}\left(b\Big|A(-\infty,\infty)=1,\tau_1=t\right)>x\right]\Big]\mathrm{d}t$, where $W_k^{DPS}\left(b\Big|A(-\infty,\infty)=1,\tau_1=t\right)$ denotes the conditional waiting time of the tagged class-k customer when there is exactly one arrival at time t on $\mathbb{R}$ and satisfies

$$W_k^{DPS}\left(b\Big|A(-\infty,\infty)=1,\tau_1=t\right)=\begin{cases} t+b_{u_t} & \text{if } t\le 0\le t+b_{u_t} \text{ and } \frac{b}{g_k}>\frac{t+b_{u_t}}{g_{u_t}}\\ \frac{g_{u_t}}{g_k}b & \text{if } t\le 0\le t+b_{u_t} \text{ and } \frac{b}{g_k}\le\frac{t+b_{u_t}}{g_{u_t}}\\ 0 & \text{if } t+b_{u_t}<0\\ b_{u_t} & \text{if } 0<t<b \text{ and } \frac{b-t}{g_k}>\frac{b_{u_t}}{g_{u_t}}\\ -t\frac{g_{u_t}}{g_k}+b\frac{g_k+g_{u_t}}{g_k} & \text{if } 0<t<b \text{ and } \frac{b-t}{g_k}\le\frac{b_{u_t}}{g_{u_t}}\\ 0 & \text{if } 0<b<t, \end{cases} \tag{43}$$

where u_t describes the class of the customer arriving at time t and b_{u_t} the service requirement of the customer arriving at time t.

We will focus on the calculation corresponding to the first term of (43), that is, the case when $t\le 0\le t+B_{U_t}$ and $t<\frac{g_{U_t}}{g_k}b-B_{U_t}$, (where the inequalities of the random variables hold sample-path wise). We have

$$\begin{aligned}&\int_{-\infty}^{0}\mathbb{E}\left[\mathbf{1}\left[-B_{U_t}\le t<\frac{g_{U_t}}{g_k}b-B_{U_t}\right]\mathbf{1}\left[t+B_{U_t}>x\right]\right]\mathrm{d}t\\ &=\int_{0}^{\infty}\mathbb{E}\left[\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right]\mathbf{1}\left[-t+B_{U_t}>x\right]\right]\mathrm{d}t\\ &=\mathbb{E}\left[\int_{0}^{\infty}\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right]\mathbf{1}\left[B_{U_t}-x>t\right]\mathrm{d}t\right],\end{aligned}$$

as we make use of Tonelli's Theorem. It follows that

$$\begin{aligned}&\int_{0}^{\infty}\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right]\mathbf{1}\left[B_{U_t}-x>t\right]\mathrm{d}t=\int_{\left(B_{U_t}-\frac{g_{U_t}}{g_k}b\right)^+}^{\max\{\left(B_{U_t}-\frac{g_{U_t}}{g_k}b\right)^+,\min\{B_{U_t},B_{U_t}-x\}\}}\mathrm{d}t\\ &=\max\left\{0,B_{U_t}-x-\left(B_{U_t}-\frac{g_{U_t}}{g_k}b\right)^+\right\}\\ &=\left(B_{U_t}-x-\left(B_{U_t}-\min\{B_{U_t},\frac{g_{U_t}}{g_k}b\}\right)\right)^+=\left(-x+\min\{B_{U_t},\frac{g_{U_t}}{g_k}b\}\right)^+.\end{aligned}$$

We thus obtain

$$\int_{0}^{\infty}\mathbb{E}\Big[\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right]\mathbf{1}\left[t+B_{U_t}>x\right]\Big]\mathrm{d}t=\mathbb{E}\Big[\left(-x+\min\{B_{U_t},\frac{g_{U_t}}{g_k}b\}\right)^+\Big].$$

The other five cases in (43) can be calculated in a similar way and after some simplifications, this will give us the result as stated in (28).