Operations Research Letters 48 (2020) 136–141

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Heavy-traffic limits for Discriminatory Processor Sharing models with joint batch arrivals

P. Vis [a], R. Bekker [a], R.D. van der Mei [b,*], R. Núñez-Queija [c]

[a] VU Amsterdam. Department of Mathematics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[b] CWI, Stochastics group, Science Park 123, 1098 XG Amsterdam, The Netherlands
[c] Korteweg–de Vries Institute for Mathematics, Postbus 94248, 1090 GE Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 24 June 2019
Received in revised form 22 January 2020
Accepted 22 January 2020
Available online 31 January 2020

Keywords:
Discriminatory Processor Sharing
Batch arrivals
Joint queue-length distribution
Heavy traffic

ABSTRACT

We study the performance of Discriminatory Processor Sharing (DPS) systems, with exponential service times and in which batches of customers of different types may arrive simultaneously according to a Poisson process. We show that the stationary joint queue-length distribution exhibits state-space collapse in heavy traffic: as the load ρ tends to 1, the scaled joint queue-length vector $(1-\rho)\mathbf{Q}$ converges in distribution to the product of a deterministic vector and an exponentially distributed random variable, with known parameters. The result provides new insights into the behavior of DPS systems. It shows how the queue-length distribution depends on the system parameters, and in particular, on the simultaneity of the arrivals. The result also suggests simple and fast approximations for the tail probabilities and the moments of the queue lengths in stable DPS systems, capturing the impact of the correlation structure in the arrival processes. Numerical experiments indicate that the approximations are accurate for medium and heavily loaded systems.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

A Discriminatory Processing Sharing (DPS) model is a multi-class queueing model where all customers are served simultaneously. The customer classes have different service rates and are assigned different weights indicating their priority. Customers of classes with relatively high weights get more of the server's capacity than customers of classes with relatively low weights. This type of systems was introduced by Kleinrock [12]. Applications of DPS models can be found in, for instance, communication networks (e.g., bandwidth sharing) and in computers (e.g., multiple processes on a shared processor).

The majority of papers on DPS consider single arrivals, i.e., whenever an arrival occurs, only one new customer joins the system. In some applications it is possible that multiple customers arrive at the same time, and these customers could possibly belong to different classes. Such an arrival pattern is captured by allowing multi-class batch arrivals. In this paper we investigate the impact of the simultaneity of arrivals on the joint queue-length distribution in heavy traffic.

The possibility of simultaneous arrivals of batches of different types strongly enhances the modeling capabilities of DPS models. Examples are found in communication networks. For instance, consider a Web server that needs to respond to numerous page-retrieval requests initiated by the end users. A Web page generally consists of a number of objects (e.g., pieces of text, in-line images, audio or video files), all of which generate separate object-retrieval requests to the Web server. Each of these requests initiates one or more data flows to be transferred via a multitude of connections (typically TCP-based connections with typically different characteristics) that compete for access on a shared medium. DPS being a convenient modeling paradigm for the sharing mechanism of TCP, see e.g. [1,2], a page request can thus be seen as a batch of flows that arrive simultaneously to a DPS node. Other examples can be found in computer systems where threads compete for access to shared processors in a processor sharing fashion. Efficient thread-spawning algorithms create batches of additional threads to reduce congestion during temporary overload situations, and vice versa, can terminate threads when no longer needed. At the operating system level, different thread types may have different priorities. This way, the creation of threads can be seen as a batch of jobs arriving to a DPS node.

DPS models have received much attention in the literature; we refer to Altman et al. [1] for a survey on DPS queues. The (conditional) moments of the response times and number of customers in a DPS queue and their finiteness are studied in [3,6,8]. Analysis of overloaded regimes in PS and DPS models can be found in [2,4,5,10]. DPS models in the heavy-traffic (HT) regime have been studied in [13,15]. These papers assume single arrivals

[*] Corresponding author.
E-mail address: mei@cwi.nl (R.D. van der Mei).

https://doi.org/10.1016/j.orl.2020.01.004
0167-6377/© 2020 Elsevier B.V. All rights reserved.

and the approach they follow differs from our approach. The analysis we follow builds on the study of Verloop et al. [16], who analyze a DPS queue with phase-type service time distributions by considering a Markovian framework. Their main result is the joint distribution of the scaled number of customers in the system in the HT regime. Grishechkin [7] allows for batch arrivals and explored the relationship between Processor Sharing models and Crump–Mode–Jagers branching processes. Recently, Izagirre et al. [9] proposed an approximation for the mean sojourn time in a DPS queue with Poisson arrivals and general service times by interpolating between heavy traffic and light traffic.

The motivation for this paper is two-fold. First, it is of a fundamental interest to explicitly quantify the impact of correlations between the arrival processes of the different customers classes on the number of each type in the system. Using a specific class of correlation structures, we take a significant step in that direction. In fact, our results *explicitly quantify the impact of the simultaneity of the arrivals* on the joint distribution of the number of jobs in DPS systems in HT. Moreover, the result also leads to sharp approximations of the impact of batched arrivals for stable systems (i.e., with load strictly less than 1), providing new insight in the performance of DPS systems, a class of models that is notoriously hard to analyze in an exact manner. Second, in several applications of DPS models the arrival processes of the different job types are correlated (see the examples above). In view of those applications it is important to be able to predict the queueing behavior accurately, in particular when the system load is significant. The effectiveness of the existing numerical techniques (like simulations) tends to degrade strongly when the system is heavily loaded. This raises the need for the development of simple and fast approximations for the delay incurred at each of the queues, explicitly capturing the impact of correlated arrivals.

We study a DPS queue with batch arrivals that occur according to a Poisson process and exponential service times. Each arriving batch may contain customers of multiple types and the number of customers per type can be larger than one. The size of a batch is according to a general joint batch-size distribution. We are interested in the system in HT. To obtain this, we scale the arrival rate and let the total load of the system go to 1. We analyze the scaled joint queue-length distribution and show that a state-space collapse occurs in the HT limit. More specifically, the joint distribution of the scaled number of customers is given by a vector of constants multiplied by an exponential distribution. This result is similar to the result of Verloop et al. [16] for Poisson arrivals of single customers, where the authors find the same constant vector times an exponential distribution. The difference with [16] is in the parameter of the exponential distribution, which now contains the second moments and correlation structure of the batches. In the HT regime, the batch arrivals only affect the mean of the scaled joint queue-length distribution. For polling models a similar phenomenon is observed in, e.g., [14].

The remainder of this paper is organized as follows. In Section 2 the model is described in detail and we introduce the notation. In Section 3, we formulate the main result and include some intuition; the proof is given in Appendix. In Section 4 we discuss numerical results. Finally, Section 5 contains some concluding remarks and topic for further research.

2. Model description

We consider a system with N classes. Arrivals occur according to a Poisson process with rate λ. Each arrival consists of a batch of size $\mathbf{K} = (K_1, \ldots, K_N)$, where K_i stands for the number of class-i customers. Denote the joint batch-size distribution by $p(k_1, \ldots, k_N) = p(\mathbf{k}) := \mathbb{P}(K_1 = k_1, \ldots, K_N = k_N)$ and let the corresponding probability generating function (PGF) of $\mathbf{K}$ be $K(\mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_N)$ and $|z_i| < 1$, for $i = 1, \ldots, N$. The arrival rate of class-i customers is denoted by $\lambda_i := \lambda\mathbb{E}[K_i]$. Customers of type i have an exponentially distributed service requirement with mean $1/\mu_i$. The N customer types share a common resource of capacity 1. Associated with every class i, there is a strictly positive weight w_i, $i = 1, \ldots, N$. When there are $\mathbf{q} := (q_1, \ldots, q_N)$ customers present in the system, with q_i the number of type-i customers, each type-i customer is served at rate

$$\frac{w_i}{\sum_{j=1}^N w_j q_j}, \quad i = 1, \ldots, N.$$

We denote the random variable of the number of type-i customers in the system by Q_i and denote its joint steady-state distribution by $\pi(\mathbf{q}) := \mathbb{P}(\mathbf{Q} = \mathbf{q})$, with $\mathbf{Q} = (Q_1, \ldots, Q_N)$. The load of type-$i$ is given by

$$\rho_i := \frac{\lambda_i}{\mu_i},$$

and the load of the system is

$$\rho := \lambda \sum_{j=1}^N \frac{\mathbb{E}[K_j]}{\mu_j} = \sum_{j=1}^N \rho_j.$$

We analyze the system when it is near saturation, i.e., for $\rho \uparrow 1$. To obtain this regime we scale the arrival rate by letting

$$\lambda \uparrow \hat{\lambda} := \left(\sum_{i=1}^N \frac{\mathbb{E}[K_i]}{\mu_i} \right)^{-1}, \tag{1}$$

while keeping μ_i, $i = 1, \ldots, N$, and the batch-size distribution $p(\mathbf{k})$ fixed. In HT, the load per customer type is given by

$$\hat{\rho}_i = \frac{\hat{\lambda}_i}{\mu_i}, \quad i = 1, \ldots, N,$$

with $\hat{\lambda}_i = \hat{\lambda}\mathbb{E}[K_i]$. The total load is equal to $\hat{\rho} = \sum_{i=1}^N \hat{\rho}_i = 1$. We let $\mathbf{e}_i$ denote the ith unit vector.

3. Main result

In this section we state the main result. The proof proceeds along similar lines as the derivation in [16], and the details are given in Appendix.

Theorem 1. *As $\rho \uparrow 1$, the joint distribution of the scaled queue lengths is given by*

$$\begin{aligned}(1-\rho)(Q_1, Q_2, \ldots, Q_N) &\to_d (\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_N) \\ &=_d \left(\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_N}{w_N} \right) X,\end{aligned} \tag{2}$$

where X is exponentially distributed with mean

$$\mathbb{E}[X] = \frac{\sum_{j=1}^N \hat{\rho}_j \frac{1}{\mu_j} + \hat{\lambda} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[K_i K_j] \frac{1}{\mu_i} \frac{1}{\mu_j}}{2 \sum_{j=1}^N (\hat{\rho}_j / w_j) \frac{1}{\mu_j}}. \tag{3}$$

The intuition behind the result is as follows. Observe that the total amount of work is the same for any work-conserving M/G/1 queue and is exponentially distributed in HT. Hence, the total amount of work in a DPS queue in HT is also an exponential random variable X. From the theorem above, we see that there is balance in how the total amount of work is distributed among the N classes. This is reflected by the fact that the exponential random variable is multiplied by a constant vector; the different customer types all have their own portion of the exponential random variable equal to $\hat{\rho}_i / w_i$, $i = 1, \ldots, N$. The number of type-i customers grows with rate $\hat{\lambda}_i$ and is depleted with rate

$w_i q_i \mu_i \left(\sum_{j=1}^N w_j q_j\right)^{-1}$. Due to the balance of type-i customers for a certain realization of X, these two rates should be equal. We can solve this equation for q_i to get: $q_i = (\hat{\rho}_i/w_i)\sum_{j=1}^N w_j q_j$. We see that $\sum_{j=1}^N w_j q_j$ is a constant common to all q_i, $i = 1, \dots, N$, giving relative portions according to the vector $(\hat{\rho}_1/w_1, \dots, \hat{\rho}_N/w_N)$.

Observe that the joint batch arrivals only influence the deterministic vector in Theorem 1 through the load. Also, the random variable X remains exponential; the effect of the joint batch arrivals only appears in the mean of X. In case of single-class arrival processes, it holds that $\mathbb{E}[K_i K_j] = 0$ if $i \neq j$. Rewriting $\mathbb{E}[X]$ for single and single-class arrivals, respectively, yields

$$\mathbb{E}[X] = \frac{\sum_{j=1}^N \hat{\rho}_j \frac{1}{\mu_j}}{\sum_{j=1}^N (\hat{\rho}_j/w_j)\frac{1}{\mu_j}} \qquad \text{(for single arrivals)}$$

$$\mathbb{E}[X] = \frac{\sum_{j=1}^N \hat{\rho}_j \frac{1}{\mu_j}\left(1 + \mathbb{E}[K_j^2]/\mathbb{E}[K_j]\right)}{2\sum_{j=1}^N (\hat{\rho}_j/w_j)\frac{1}{\mu_j}} \qquad \text{(for single-type batch arrivals).}$$

Remark 1. Analogous to the derivation of Verloop et al. [16], our main result may be extended to phase type service-time distributions and even to a more general Markovian framework. In this framework, after service completion a customer of type i becomes a customer of type j with probability $\hat{p}_{ij}$, or the customer leaves the system with probability $\hat{p}_{i0}$. The service duration of a type-i customer is still exponential with rate μ_i, but now a customer has to complete multiple services with different service rates (because the customer changes type after a service completion). We conjecture that the joint distribution of the scaled queue length $(\hat{Q}_1, \dots, \hat{Q}_N)$ is still given as in Theorem 1, where (3) is replaced by

$$\mathbb{E}[X] = \frac{\sum_{j=1}^N \hat{\rho}_j \mathbb{E}[R_j] + \hat{\lambda}\sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[K_i K_j]\mathbb{E}[R_i]\mathbb{E}[R_j]}{2\sum_{j=1}^N (\hat{\rho}_j/w_j)\mathbb{E}[R_j]},$$

with R_i the remaining service duration of customers of type i, $i = 1, \dots, N$.

4. Numerical results

In this section we perform some numerical experiments and compare simulation results with the closed-form expressions from the HT limit. In Section 4.1, we plot the queue-length distribution obtained from simulation, to demonstrate the state-space collapse. In Section 4.2, we present the scaled mean queue lengths for different loads and show that the mean queue lengths indeed converge to their HT limit. We will use the heavy-traffic result as an approximation for smaller loads and show the errors in a table. Finally, in Section 4.3, we compare the mean queue lengths in the system with joint batch arrivals to the mean queue lengths in a system with batch arrivals of one customer class and a system with single arrivals.

4.1. State-space collapse

The basic DPS queue that we use for our experiments is a system with two customer classes and batches of at most 2 arrivals per class. We use the batch-size distribution $p(0, 1) = p(1, 0) = p(1, 1) = p(1, 2) = p(0, 2) = 1/5$, i.e., there are five possible batches that have the same probability of occurrence. We take $w_1 = 2$, $w_2 = 1$, $\mu_1 = 0.75$ and $\mu_2 = 1$. The arrival rate λ is varied to allow for different loads. In Fig. 1, we plot the joint queue-length distribution obtained by simulation for three different loads: $\rho = 0.8$ 1(a), $\rho = 0.9$ 1(b) and $\rho = 0.99$ 1(c). For every point (Q_1, Q_2), the color of the point represents the density. We see that for higher loads, the density is more concentrated on a single line, demonstrating the state-space collapse.

Table 1
Comparison between simulation and heavy-traffic approximation for type-1 customers.

ρ	$\mathbb{E}[Q_1]$			$\mathbb{E}[Q_1^2]$		
	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.70	1.06	1.33	25.79	3.30	3.56	7.80
0.80	1.76	2.00	13.90	7.88	8.00	1.58
0.90	3.79	4.00	5.60	32.27	32.00	0.85
0.95	7.82	8.00	2.35	129.48	128.00	1.15
0.99	40.19	40.00	0.48	3212.30	3200.00	0.38

4.2. Approximation of moments

The HT result provides the following for the scaled marginal queue-length distribution: $(1-\rho)Q_i \to_d (\hat{\rho}_i/w_i)X$, with X an exponential random variable. This suggests the following approximation for the number of type-i customers in a system with $\rho < 1$. Specifically, Q_i is then approximately exponentially distributed with mean

$$\mathbb{E}[Q_i] = \frac{(\hat{\rho}_i/w_i)\mathbb{E}[X]}{1-\rho}$$

and with second moment

$$\mathbb{E}[Q_i^2] = \frac{2((\hat{\rho}_i/w_i)\mathbb{E}[X])^2}{(1-\rho)^2}, \quad i = i, \dots, N.$$

We compare the approximations above with simulation results for different values of ρ (by changing λ) using the absolute percentage error, given by

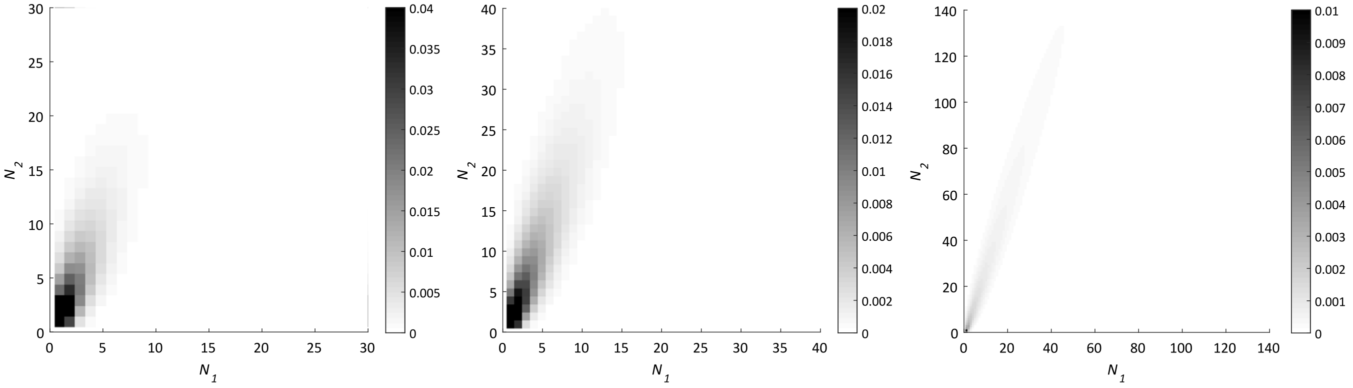
$$\Delta\% = 100\% \times \frac{|\text{App} - \text{Sim}|}{\text{Sim}}.$$

From Table 1 we see that the approximation works better for higher loads. This is to be expected, since the approximation is exact in HT. For loads around 0.9, the approximation is reasonable, for lower loads the error increases substantially. Note that we only studied one specific setting, but we expect similar results in other settings.

4.3. The impact of batch arrivals

Finally, we experiment with the impact of batch arrivals on the (scaled) mean queue length. To do so, we consider a system with joint batch arrivals to similar systems with single-class batch arrivals and systems with single arrivals only. The arrivals process is modified such that the systems have as many features in common as possible, like the load per class. In the system with joint batch arrivals we again take: $p(0, 1) = p(1, 0) = p(1, 1) = p(1, 2) = p(0, 2) = 1/5$. In the system with batch arrivals of a single type we have: $p(1, 0) = 3/7$, $p(0, 1) = 2/7$ and $p(0, 2) = 2/7$, and in the system with single arrivals: $p(1, 0) = 1/3$ and $p(0, 1) = 2/3$. We simulate the mean queue lengths of type-1 customers for different loads. The results are scaled by a factor $(1-\rho)$ and plotted in Fig. 2 (dashed lines). We see that the scaled mean queue length is smaller if the batches are of a single type and smallest if there are only single arrivals. This can be explained by the variability in the arrival process, where an arrival process consisting of only single customers has the smallest variation. This also explains why the convergence to the HT limits (solid lines) is faster in case of single arrivals.

We conclude that the influence on the queue-length distributions of the joint batch arrivals is significant and that the



(a) $\rho = 0.8$.

(b) $\rho = 0.9$.

(c) $\rho = 0.99$.

Fig. 1. Joint queue-length distribution for different values of ρ.

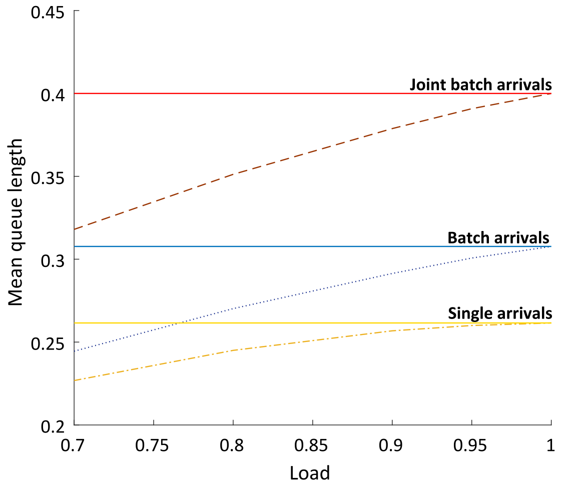


Fig. 2. Scaled mean queue lengths of type-1 customers in systems with different arrival types.

HT approximation works reasonably well for systems with high loads.

5. Conclusion and topics for further research

This paper provides new fundamental insights in the performance of DPS systems, a class of models for which hardly any exact results are known. We show explicitly how the queue-length distributions depend on the system parameters when the system is heavily loaded, and proof that the scaled queue-length distribution converges to the product of a known vector and an exponentially distributed random variable, when the load goes to 1. In particular, the asymptotic results explicitly quantify the impact of the simultaneity of the job arrivals on the queue-length distributions. The results also lead to new and simple closed-form approximations for stable systems, and numerical results illustrate the accuracy of the approximations.

The results can be extended in several directions. First, the exponentiality assumption can be relaxed to phase-type distributions, as suggested in Remark 1. Obtaining a rigorous proof is a topic for further research. Second, one may suspect that the Poisson assumption can be relaxed to renewal arrivals, using the time-scale decomposition results obtained for polling models (e.g., [14]). Finally, the asymptotic results form open up possibilities for the development of approximations for the queue-length and sojourn-time distributions for arbitrary values of the load, for the models batched arrivals. To this end, the elegant interpolation technique proposed in [9] forms an excellent basis.

Appendix. Proof of Theorem 1

In this appendix we derive the limiting distribution of the number of customers in the queue in HT (i.e., when $\rho \uparrow 1$). To this end, we start by formulating the balance equations for the limiting distribution $\pi(\mathbf{q})$; these balance equations will be used to derive the functional equation, see Appendix A.1. When we have the functional equations, we scale it with a factor $(1-\rho)$ and take the limit $\rho \uparrow 1$. This leads to a partial differential equation as presented in Appendix A.2. The solution to this equation gives the desired distribution up to a single random variable. The final step is finding this random variable, see Appendix A.3.

A.1. Balance equations and functional equation

We start by deriving the functional equation. To this end, we introduce a transformation that leads to more convenient expressions. Define

$$r(\mathbf{0}) = 0, \quad \text{and} \quad r(\mathbf{q}) = \frac{\pi(\mathbf{q})}{\sum_{j=1}^{N} w_j q_j}, \quad \text{for } \mathbf{q} \neq \mathbf{0}. \tag{A.1}$$

Let $P(\mathbf{z})$ and $R(\mathbf{z})$ denote the generating functions of $\pi(\mathbf{q})$ and $r(\mathbf{q})$, respectively. That is

$$P(\mathbf{z}) = \mathbb{E}\left[z_1^{Q_1} \cdots z_N^{Q_N}\right] = \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \cdots \cdot z_N^{q_N} \pi(\mathbf{q}),$$

and

$$R(\mathbf{z}) = \mathbb{E}\left[\frac{z_1^{Q_1} \cdot \cdots \cdot z_N^{Q_N}}{\sum_{i=1}^{N} w_i Q_i} \mathbb{1}_{\{\sum_{j=1}^{N} Q_j > 0\}}\right] = \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \cdots \cdot z_N^{q_N} r(\mathbf{q}).$$

The following lemma formulates a functional equation for $R(\mathbf{z})$:

Lemma 1. *For $\rho < 1$, a functional equation for $R(\mathbf{z})$ is given by*

$$\lambda(1-\rho)(1-K(\mathbf{z})) = \sum_{i=1}^{N} (\lambda z_i (K(\mathbf{z}) - 1) + \mu_i (1 - z_i))\, w_i \frac{\partial}{\partial z_i} R(\mathbf{z}). \tag{A.2}$$

Proof. Assuming $\rho < 1$, the equilibrium distribution $\pi(\mathbf{q})$ satisfies the following balance equations

$$\lambda \pi(\mathbf{0}) = \sum_{i=1}^{N} \mu_i \pi(\mathbf{e}_i), \tag{A.3}$$

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

$$\left(\lambda + \frac{\sum_{i=1}^N w_i \mu_i q_i}{\sum_{i=1}^N w_i q_i}\right)\pi(\mathbf{q}) = \lambda \sum_{k_1=0}^{q_1} \cdots \sum_{k_N=0}^{q_N} p(\mathbf{k})\pi(\mathbf{q}-\mathbf{k}) + \sum_{i=1}^N \mu_i \frac{w_i(q_i+1)}{\sum_{j=1}^N w_j q_j + w_i}\pi(\mathbf{q}+\mathbf{e}_i).$$

Now we take the generating function, yielding

$$\sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} \mathbb{1}_{\{\sum_{j=1}^N q_j > 0\}} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \left(\lambda + \frac{\sum_{i=1}^N w_i \mu_i q_i}{\sum_{i=1}^N w_i q_i}\right)\pi(\mathbf{q})$$
$$= \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \lambda \sum_{k_1=0}^{q_1} \cdots \sum_{k_N=0}^{q_N} p(\mathbf{k})\pi(\mathbf{q}-\mathbf{k}) + \sum_{i=1}^N \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} \mathbb{1}_{\{\sum_{j=1}^N q_j > 0\}} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \mu_i \times \frac{w_i(q_i+1)}{\sum_{j=1}^N w_j q_j + w_i}\pi(\mathbf{q}+\mathbf{e}_i).$$

To get rid of the indicator functions, we add Eq. (A.3), change the order of summation in the second line and start the corresponding summations at 0 by a change of variable. This leads to

$$\lambda\pi(\mathbf{0}) + \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} \mathbb{1}_{\{\sum_{j=1}^N q_j > 0\}} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \left(\lambda + \frac{\sum_{i=1}^N w_i \mu_i q_i}{\sum_{i=1}^N w_i q_i}\right)\pi(\mathbf{q})$$
$$= \lambda \sum_{k_1=0}^{\infty} \cdots \sum_{k_N=0}^{\infty} z_1^{k_1} \cdot \ldots \cdot z_N^{k_N} p(\mathbf{k}) \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \pi(\mathbf{q}) + \sum_{i=1}^N \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \mu_i \frac{w_i(q_i+1)}{\sum_{j=1}^N w_j q_j + w_i}\pi(\mathbf{q}+\mathbf{e}_i).$$

We now apply the transformation from (A.1). Note that for the first term on the right-hand side, we have to take into account that $r(\mathbf{0}) = 0$, but $\pi(\mathbf{0}) \neq 0$. Hence, we obtain

$$\lambda\pi(\mathbf{0}) + \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \left(\lambda \sum_{i=1}^N w_i q_i + \sum_{i=1}^N w_i \mu_i q_i\right) r(\mathbf{q})$$
$$= \lambda K(\mathbf{z}) \left(\pi(\mathbf{0}) + \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} r(\mathbf{q}) \sum_{i=1}^N w_i q_i\right) + \sum_{i=1}^N \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} \mu_i w_i (q_i+1) r(\mathbf{q}+\mathbf{e}_i).$$

Taking partial derivatives of $R(\mathbf{z})$ with respect to z_i, we get

$$\frac{\partial}{\partial z_i} R(\mathbf{z}) = z_i^{-1} \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_N^{q_N} q_i r(\mathbf{q}).$$

Using this, we can rewrite the functional equation as

$$\lambda\pi(\mathbf{0}) + \sum_{i=1}^N \left(\lambda w_i z_i \frac{\partial}{\partial z_i} R(\mathbf{z}) + \mu_i w_i z_i \frac{\partial}{\partial z_i} R(\mathbf{z})\right)$$
$$= \lambda K(\mathbf{z}) \left(\pi(\mathbf{0}) + \sum_{i=1}^N w_i z_i \frac{\partial}{\partial z_i} R(\mathbf{z})\right) + \sum_{i=1}^N \mu_i w_i \frac{\partial}{\partial z_i} R(\mathbf{z}).$$

Rearranging the terms and using $\pi(\mathbf{0}) = 1 - \rho$ completes the proof. □

A.2. Heavy-traffic limit and partial differential equation

For convenience we use the change of variables $z_i = e^{-(1-\rho_i)s_i}$, with $s_i > 0$, $i = 1, \ldots, N$. We use the notation $\mathbf{s} = (s_1, \ldots, s_N)$ and $e^{-(1-\rho)\mathbf{s}} = (e^{-(1-\rho)s_1}, \ldots, e^{-(1-\rho)s_N})$. For the heavy-traffic limit, we define

$$\hat{R}(\mathbf{s}) = \mathbb{E}\left[\frac{1 - e^{-s_1 \hat{Q}_1} \cdot \ldots \cdot e^{-s_N \hat{Q}_N}}{\sum_{j=1}^N \hat{Q}_j w_j} \mathbb{1}_{\{\sum_{j=1}^N \hat{Q}_j > 0\}}\right]. \tag{A.4}$$

Now we can formulate the following lemma.

Lemma 2. *If $\lim_{\rho\uparrow 1} P(e^{-(1-\rho)\mathbf{s}})$ exists, then the function $\hat{R}(\mathbf{s})$ satisfies the following partial differential equation:*

$$0 = \sum_{i=1}^N F_i(\mathbf{s}) \frac{\partial \hat{R}(\mathbf{s})}{\partial s_i} = \mathbf{F}(\mathbf{s}) \nabla \hat{R}(\mathbf{s}), \quad \forall\ \mathbf{s} \geq \mathbf{0},$$

where $\mathbf{F}(\mathbf{s}) = (F_1(\mathbf{s}), \ldots, F_N(\mathbf{s}))$, and

$$F_i(\mathbf{s}) = w_i \left(\mu_i s_i - \hat{\lambda} \sum_{j=1}^N s_j \mathbb{E}[K_j]\right), \quad i = 1, \ldots, N,$$

with $\hat{\lambda}$ as defined in (1).

Proof. We divide both sides of (A.2) by $(1-\rho)$ and apply the change of variables. Note that $\lim_{\rho\uparrow 1}(1 - K(e^{-(1-\rho)\mathbf{s}}))/(1-\rho) = \sum_{j=1}^N s_j \mathbb{E}[K_j]$. Taking the limit $\rho \uparrow 1$ gives the partial differential equation

$$0 = \sum_{i=1}^N \left(\mu_i s_i - \hat{\lambda} \sum_{j=1}^N s_j \mathbb{E}[K_j]\right) w_i \frac{\partial}{\partial s_i} \hat{R}(\mathbf{s}). \tag{A.5}$$

This completes the proof. □

The lemma above is similar to Lemma 2 in [16]; in our case $p_{ij} = 0$ if $j > 0$ and it turns out that $p_{0j} = \mathbb{E}[K_j]$ due to batch arrivals. The next step is to establish the state-space collapse. Due to the similarity between our functional equation in Lemma 2 and the functional equation in [16, Lemma 2], we can rely on Lemma 3 of [16]. Specifically, [16, Lemma 3] gives that $\hat{R}(\mathbf{s})$ is constant on an $(N-1)$-dimensional hyperplane, see [16] for a geometric interpretation. Essentially, this provides that the N-dimensional random vector of queue lengths reduces to a deterministic vector times a single random variable in heavy traffic. Applying [16, Lemma 3] and the subsequent analysis, we get

$$(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_N) =_d \left(\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_N}{w_N}\right) \frac{w_1}{\hat{\rho}_1} \hat{Q}_1, \tag{A.6}$$

Note that [16, Lemma 3] holds in our case, as its proof does not depend on the fact that the p_{0j} add up to 1, and we take p_{0j} equal to $\mathbb{E}[K_j]$, $j = 1, \ldots, N$. Eq. (A.6) is now equivalent to (2), with X distributed as $\frac{w_1}{\hat{\rho}_1}\hat{Q}_1$. It remains to find the distribution of X.

A.3. Specifying the common distribution

The distribution of X is given in the following lemma.

Lemma 3. *X is exponentially distributed with mean*

$$\mathbb{E}[X] = \frac{\sum_{j=1}^N \hat{\rho}_j \frac{1}{\mu_j} + \hat{\lambda} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[K_i K_j] \frac{1}{\mu_i} \frac{1}{\mu_j}}{2 \sum_{j=1}^N (\hat{\rho}_j / w_j) \frac{1}{\mu_j}}.$$

Proof. Denote by B the total amount of work that an arbitrary arriving batch brings into the system. From Kingman [11] we have that the total amount of work in the system W in the GI/GI/1 queue, when scaled by $(1-\rho)$, has a proper distribution as $\rho \uparrow 1$. In particular,

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

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

$$\mathbb{E}[\hat{W}] = \frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]}.$$

For our DPS model, we can represent the total workload as

$$W = \sum_{j=1}^{N}\sum_{h=1}^{Q_j} R_{j,h},$$

where $R_{j,h}$ is the remaining service requirement of the hth type-j customer. Since we have exponential service requirements, the remaining service requirements are in distribution equal to the original service requirements: $R_{j,h} =_d B_{j,h}$, with $B_{j,h}$ exponentially distributed with mean $\mathbb{E}[B_j] = 1/\mu_j$. Using the representation of the total workload, we may write

$$(1-\rho)W = \sum_{j=1}^{N}(1-\rho)Q_j \times \frac{1}{Q_j}\sum_{h=1}^{Q_j} B_{j,h}.$$

Observe that $(1-\rho)Q_j \to X\hat{\rho}_j/w_j$ according to Eq. (A.6), and as $Q_j \to \infty$ (a.s.) for $\rho \uparrow 1$, we have that $\frac{1}{Q_j}\sum_{h=1}^{Q_j} B_{j,h} \to \mathbb{E}[B_j]$ due to the law of large numbers. This suggests that

$$\hat{W} = X\sum_{j=1}^{N}\frac{\hat{\rho}_j}{w_j}\mathbb{E}[B_j],$$

which in turn implies that X is also exponentially distributed; this equation is formally shown in [16, Equation (17)].

Combining the two expressions for $\mathbb{E}[\hat{W}]$ above gives

$$\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]} = \mathbb{E}[X]\sum_{j=1}^{N}\frac{\hat{\rho}_j}{w_j}\mathbb{E}[B_j],$$

and thus

$$\mathbb{E}[X] = \frac{\mathbb{E}[B^2]/(2\mathbb{E}[B])}{\sum_{j=1}^{N}(\hat{\rho}_j/w_j)\mathbb{E}[B_j]}.$$

Note that B can be rewritten as

$$B = \sum_{j=1}^{N}\sum_{i=1}^{K_j} B_{j,i}.$$

Using the law of total expectation, we derive the moments of B:

$$\mathbb{E}[B] = \mathbb{E}[\mathbb{E}[B|\mathbf{K}]] = \mathbb{E}\left[\sum_{j=1}^{N}\sum_{i=1}^{K_j}\mathbb{E}[B_{j,i}]\right] = \sum_{j=1}^{N}\mathbb{E}[K_j]\frac{1}{\mu_j} = \frac{1}{\hat{\lambda}},$$

and

$$\begin{aligned}\mathbb{E}[B^2] &= \mathbb{E}\left[\mathbb{E}[B^2|\mathbf{K}]\right] = \mathbb{E}\left[\mathrm{Var}\,[B|\mathbf{K}] + (\mathbb{E}[B|\mathbf{K}])^2\right]\\ &= \mathbb{E}\left[\sum_{j=1}^{N}\sum_{i=1}^{K_j}\mathrm{Var}[B_{j,i}] + \sum_{i=1}^{N}\sum_{j=1}^{N}K_iK_j\mathbb{E}[B_i]\mathbb{E}[B_j]\right]\\ &= \mathbb{E}\left[\sum_{j=1}^{N}K_j\frac{1}{\mu_j^2} + \sum_{i=1}^{N}\sum_{j=1}^{N}K_iK_j\frac{1}{\mu_i}\frac{1}{\mu_j}\right]\\ &= \sum_{j=1}^{N}\mathbb{E}[K_j]\frac{1}{\mu_j^2} + \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[K_iK_j]\frac{1}{\mu_i}\frac{1}{\mu_j}.\end{aligned}$$

Substituting the above in the equation for $\mathbb{E}[X]$ completes the proof. □

Proof of Theorem 1. The proof follows directly from combining Lemma 2 and [16, Lemma 3], leading to Eq. (A.6), and Lemma 3 for the distribution of the remaining random variable. □

References

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

[2] E. Altman, T. Jimenez, D. Kofman, DPS queues with stationary ergodic service times and the performance of TCP in overload, in: Proceedings IEEE INFOCOM 2004, Vol. 2, IEEE, 2004, pp. 975–983.

[3] K. Avrachenkov, U. Ayesta, P. Brown, R. Núñez-Queija, Discriminatory processor sharing revisited, in: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 2, IEEE, 2005, pp. 784–795.

[4] A. Ben Tahar, A. Jean-Marie, The fluid limit of the multiclass processor sharing queue, Queueing Syst. 71 (4) (2012) 347–404.

[5] R. Egorova, S.C. Borst, A.P. Zwart, Bandwidth-sharing networks in overload, Perform. Eval. 64 (9) (2007) 978–993.

[6] G. Fayolle, I. Mitrani, R. Iasnogorodski, Sharing a processor among many job classes, J. ACM 27 (3) (1980) 519–532.

[7] S.A. Grishechkin, On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes, Adv. Appl. Probab. 24 (03) (1992) 653–698.

[8] M. Haviv, J. van der Wal, Mean sojourn times for phase-type discriminatory processor sharing systems, European J. Oper. Res. 189 (2) (2008) 375–386.

[9] A. Izagirre, U. Ayesta, I.M. Verloop, Sojourn time approximations for a discriminatory processor sharing queue, ACM Trans. Model. Perform. Eval. Comput. Syst. 1 (1) (2016) 5.

[10] A. Jean-Marie, P. Robert, On the transient behavior of the processor sharing queue, Queueing Syst. 17 (1) (1994) 129–136.

[11] J.F.C. Kingman, The single server queue in heavy traffic, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 57, (04) Cambridge Univ Press, 1961, pp. 902–904.

[12] L. Kleinrock, Time-shared systems: A theoretical treatment, J. ACM 14 (2) (1967) 242–261.

[13] K.M. Rege, B. Sengupta, Queue-length distribution for the discriminatory processor-sharing queue, Oper. Res. 44 (4) (1996) 653–657.

[14] R.D. van der Mei, Waiting-time distributions in polling systems with simultaneous batch arrivals, Ann. Oper. Res. 113 (1–4) (2002) 155–173.

[15] G. van Kessel, R. Núñez-Queija, S.C. Borst, Asymptotic regimes and approximations for discriminatory processor sharing, SIGMETRICS Perform. Eval. Rev. 32 (2) (2004) 44–46.

[16] I.M. Verloop, U. Ayesta, R. Núñez-Queija, Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing, Oper. Res. 59 (3) (2011) 648–660.