Heavy-traffic analysis of the discriminatory random-order-of-service discipline[*]

U. Ayesta[1,2], A. Izagirre[1,3], I.M. Verloop[1,4]

[1]BCAM – Basque Center for Applied Mathematics, Derio, Spain
[2]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3]UPV/EHU, University of the Basque Country, Bilbao, Spain
[4]Université de Toulouse, IRIT-CNRS, Toulouse, France

ABSTRACT

We study the steady-state queue-length vector in a multi-class single-server queue with relative priorities. Upon service completion, the probability that the next customer to be served is from class k is controlled by class-dependent weights. Once a customer has started service, it is served without interruption until completion. This is a generalization of the random-order-of-service discipline.

We investigate the system in a heavy-traffic regime. We first establish a state-space collapse for the scaled queue length vector, that is, the scaled queue length vector is in the limit the product of an exponentially distributed random variable and a deterministic vector. As a direct consequence, we obtain that the scaled number of customers in the system reduces as classes with smaller mean service requirement obtain relatively larger weights. In addition, we present the distribution of the scaled sojourn time of a customer given its class, in heavy traffic.

1. INTRODUCTION

In this paper we consider an $M/G/1$ queue with K different classes of customers. Service is non-preemptive and upon service completion, a class-k customer, $k = 1, \dots, K$, is selected to be served with probability

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

where p_k, $k = 1, \dots, K$, are given class-dependent weights, and n_k are the number of class-k customers present in the system at the decision epoch. This model was first introduced in [6]. Expressions for the mean waiting time of a customer given its class have been obtained in [7]. In [8, 9] the authors derive differential equations that the transform of the joint queue lengths and the waiting time in steady-state must satisfy. In particular, they obtain systems of linear equations from which the moments of the queue lengths can be obtained.

When all the weights p_k, $k = 1, \dots, K$, are equal, the model reduces to the random-order-of-service (ROS) discipline. Classical papers on ROS are for example [12, 13, 14]. The Laplace transform for the waiting time distribution was obtained in [12]. In [12, 13, 16], ROS is studied in a heavy-traffic setting and for service requirements having finite variance it was shown that the scaled queue length converges to an exponential distribution and that the scaled waiting time is equal in distribution to the product of two independent exponential random variables. More recently, the authors of [4] obtained the waiting time distribution in heavy traffic for certain service requirements having infinite variance. In addition, waiting time tail asymptotics have been obtained for heavy-tailed service time distributions. In [2] the authors derive the relationship between the distribution of the waiting time under ROS and the sojourn time under the processor-sharing discipline.

Both ROS and its multi-class generalization are fundamental models with application in various domains, and in particular in telecommunication networks [3].

In the present study, we are interested in the distribution of both the queue length vector and the waiting time for the multi-class queue with relative priorities, and we will study these in the heavy-traffic regime. In particular, we establish a state-space collapse for the queue length vector. The result shows that in the limit, the scaled queue length vector is the product of an exponentially distributed random variable and a deterministic vector. Making use of the state-space collapse result, we derive interesting properties. First of all, we show that the scaled holding cost reduces as classes with higher value of $c_k/\mathbb{E}(B_k)$ obtain a relatively larger weight, where c_k is the cost associated to class k, and $\mathbb{E}(B_k)$ is the mean service requirement of a class-k customer. This can be seen as an extension of the optimality result of the $c\mu$-rule [5], the strict priority discipline that gives priority in decreasing order of $c_k/\mathbb{E}(B_k)$. Second, we study the distribution of the waiting time for a customer of a given class in heavy traffic and obtain that it is distributed as the product of two exponentially distributed random variables.

In this short paper we provide sketches of the proofs. Full proofs are available in the technical report [1].

2. MODEL DESCRIPTION

We consider a multi-class single-server queue with K classes of customers. Class-k customers, $k = 1, \dots, K$, arrive according to independent Poisson processes with rate λ_k. We denote the overall arrival rate by $\lambda = \sum_{k=1}^K \lambda_k$. We assume that class-$k$ customers have i.i.d. generally distributed service requirements B_k, with distribution function $B_k(t)$ and Laplace-Stieltjes transform $B_k^*(s) = \int_0^\infty \mathrm{e}^{-st} \mathrm{d}B_k(t)$. We assume $\mathbb{E}(B_k^2) < \infty$, for all k. The traffic intensity for

[*]Research partially supported by the grants MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain) and PI2010-2 (Department of Education and Research, Basque Government).

class-k customers is $\rho_k = \lambda_k\mathbb{E}(B_k)$ and $\rho = \sum_{k=1}^K \rho_k = \sum_{k=1}^K \lambda_k\mathbb{E}(B_k) = \lambda\sum_{k=1}^K \alpha_k\mathbb{E}(B_k)$, denotes the total traffic intensity, where we denote by $\alpha_k = \lambda_k/\lambda$ the probability an arrival is of class k. Service is non-preemptive and upon service completion a class-k customer is selected to be served with probability as given in (1).

We investigate the queue when it is near saturation, i.e., $\rho \uparrow 1$, which is commonly referred to as the heavy-traffic regime. This regime can be obtained by letting $\lambda \uparrow \hat{\lambda} := (\sum_{k=1}^K \alpha_k\mathbb{E}(B_k))^{-1}$, since then $\rho = \lambda\sum_{k=1}^K \alpha_k\mathbb{E}(B_k) \uparrow 1$. When passing to the heavy-traffic regime we keep the fraction of class-k arrivals, α_k, fixed and we define $\hat{\lambda}_k := \alpha_k\hat{\lambda}$.

We denote the steady-state number of class-k customers in the system at departure epochs by Q_k and at arbitrary epochs by N_k.

3. STATE-SPACE COLLAPSE

In this section we present the state-space collapse result for the steady-state queue length distribution at departure and arbitrary epochs. The result shows that in the limit, the queue length vector is the product of an exponentially distributed random variable and a deterministic vector.

PROPOSITION 3.1. *When scaled by $1-\rho$, the queue length vector at departure and arbitrary time epochs has a proper limiting distribution as $(\lambda_1,\dots,\lambda_K) \to (\hat{\lambda}_1,\dots,\hat{\lambda}_K)$:*

$$(1-\rho)(Q_1,\dots,Q_K) \xrightarrow{d} (\hat{Q}_1,\dots,\hat{Q}_K) \overset{d}{=} Y(\frac{\hat{\lambda}_1}{p_1},\frac{\hat{\lambda}_2}{p_2},\dots,\frac{\hat{\lambda}_K}{p_K})$$

and

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

where $\xrightarrow{d}$ denotes convergence in distribution and X and Y are exponentially distributed random variables with mean

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

REMARK 1 (RANDOM-ORDER-OF-SERVICE). *In the case of one class, i.e., $K=1$, the system reduces to the ROS discipline. Proposition 3.1 implies that the queue length is exponentially distributed with mean $\frac{\hat{\lambda}\mathbb{E}(B^2)}{2}$. This has been obtained previously by Kingman [12]. Note that [12] states the result for normalized arrival rate and service times, and thus obtains that the mean number of customers is $(1+\sigma^2)/2$, with σ^2 the variance of the service time distribution.*

Proof of Proposition 3.1. We sketch the proof for the departure epochs. We note that the proof technique is similar to that of the state-space collapse result in [15]. For the proof for the arbitrary epochs we refer to the technical report [1].

Let $\pi(\vec{q})$ be the stationary distribution of $(Q_1,\dots,Q_K)$ and let

$$p(\vec{z}) = \mathbb{E}(z_1^{Q_1}\cdots z_K^{Q_K}) = \sum_{q_1=0}^{\infty}\cdots\sum_{q_K=0}^{\infty} z_1^{q_1}\dots z_K^{q_K}\pi(\vec{q}) \qquad (3)$$

be its joint probability generating function. We define

$$r(\vec{z}) = \sum_{(q_1,\dots,q_K)\neq(0,\dots,0)} \frac{\pi(\vec{q})}{q_1p_1+\dots+q_Kp_K} z_1^{q_1}\dots z_K^{q_K}.$$

In [9] it is shown that $p(\cdot)$ satisfies

$$p(z_1,\dots,z_K) = 1-\rho+\sum_{k=1}^K p_kz_k\frac{\partial}{\partial z_k}r(z_1,\dots,z_K), \qquad (4)$$

and that $r(\cdot)$ satisfies

$$\sum_{k=1}^K p_k(z_k - B_k^*(\lambda-\sum_{j=1}^K\lambda_jz_j))\frac{\partial}{\partial z_k}r(z_1,\dots,z_K)$$
$$= (\rho-1)(1-\sum_{k=1}^K\frac{\lambda_k}{\lambda}B_k^*(\lambda-\sum_{j=1}^K\lambda_jz_j)). \qquad (5)$$

Denote $\mathrm{e}^{-(1-\rho)\vec{s}} = (\mathrm{e}^{-(1-\rho)s_1},\dots,\mathrm{e}^{-(1-\rho)s_K})$. We study the Laplace transform of $(1-\rho)(Q_1,\dots,Q_K)$ as $\rho\uparrow 1$, i.e.,

$$\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \lim_{\rho\uparrow 1}\mathbb{E}(\mathrm{e}^{-(1-\rho)s_1Q_1}\cdots\mathrm{e}^{-(1-\rho)s_KQ_K}).$$

Using (4), we show the existence of a function $\hat{r}(\vec{s})$ such that

$$\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{k=1}^K p_k\frac{\partial\hat{r}(\vec{s})}{\partial s_k}. \qquad (6)$$

From (5) it follows that $\hat{r}$ satisfies:

$$0 = \sum_{k=1}^K p_k(-s_k+\mathbb{E}(B_k)\sum_{j=1}^K\hat{\lambda}_js_j)\frac{\partial\hat{r}(\vec{s})}{\partial s_k}.$$

From the above partial differential equation it can be obtained that the function $\hat{r}(\vec{s})$ is constant on the $(K-1)$-dimensional hyperplane $H_c := \{\vec{s}\geq\vec{0}\colon \sum_{j=1}^K\frac{\hat{\lambda}_j}{p_j}s_j = c\}$, $c>0$, see [1] for the full proof. As $\hat{r}(\vec{s})$ is constant on H_c, it depends on $\vec{s}$ only through $\sum_{j=1}^K\hat{\lambda}_js_j/p_j$. Then, from (6) it follows that $\mathbb{E}(\mathrm{e}^{-\sum_{k=1}^K s_k\hat{Q}_k}) = \lim_{\rho\to 1}p(\mathrm{e}^{-(1-\rho)\vec{s}})$ depends on $\vec{s}$ only through $\sum_{j=1}^K\hat{\lambda}_js_j/p_j$. This implies that $\frac{p_i}{\hat{\lambda}_i}\hat{Q}_i = \frac{p_j}{\hat{\lambda}_j}\hat{Q}_j$ almost surely for all i,j, and we obtain $(\hat{Q}_1,...,\hat{Q}_K)\overset{d}{=}(\frac{\hat{\lambda}_1}{p_1},\frac{\hat{\lambda}_2}{p_2},...,\frac{\hat{\lambda}_K}{p_K})Y$, with Y distributed as $\frac{p_1}{\hat{\lambda}_1}\hat{Q}_1$.

To conclude that Y is exponentially distributed we use the fact that the scaled workload in the $M/G/1$ queue in heavy-traffic is exponentially distributed [10, 11]. □

4. SIZE-BASED SCHEDULING

In this section we investigate how the choice of the weights influences the performance of the system. With each class of customers we associate a cost $c_j\geq 0, j=1,\dots,K$, and we are interested in the holding cost $\sum_{j=1}^K c_jN_j$. In the heavy-traffic regime we obtain that the scaled holding cost decreases "stochastically" as classes with lower value for $\frac{c_k}{\mathbb{E}(B_k)}$ have larger weights. We will write $N_j^{(p)}$ ($\hat{N}_j^{(p)}$) instead of N_j ($\hat{N}_j$) to emphasize the dependence on the weights $p_1,\dots,p_K$.

PROPOSITION 4.1. *Let $c_j\geq 0, j=1,\dots,K$. Without loss of generality we assume that $\frac{c_1}{\mathbb{E}(B_1)}\geq\frac{c_2}{\mathbb{E}(B_2)}\geq\dots\geq\frac{c_K}{\mathbb{E}(B_K)}$. If $\frac{p_j}{p_{j+1}}\leq\frac{\tilde{p}_j}{\tilde{p}_{j+1}}$, for all $j=1,\dots,K-1$, then*

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^K c_jN_j^{(p)} \geq_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^K c_jN_j^{(\tilde{p})},$$

where $\geq_{st}$ denotes the usual stochastic ordering, i.e., $X\geq_{st} Y$ if and only if $\mathbb{P}(X\geq z)\geq\mathbb{P}(Y\geq z)$ for all z.

Sketch of proof. From Proposition 3.1 we obtain that, as $\rho \uparrow 1$, the scaled holding cost, $(1-\rho)\sum_{j=1}^K c_j N_j^{(p)}$, converges in distribution to an exponentially distributed random variable with mean

$$\sum_{j=1}^{K} c_j \mathbb{E}(\hat{N}_j^{(p)}) = \frac{\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j} c_j}{2\sum_{j=1}^K \frac{\hat{\lambda}_j}{p_j}\mathbb{E}(B_j)} \sum_{j=1}^{K} \hat{\lambda}_j \mathbb{E}(B_j^2). \qquad (7)$$

In order to prove the stochastic ordering result it is therefore sufficient to check that the ordering result holds for the *mean* holding cost (7). □

It is well-known that the so-called $c\mu$-rule (the non-preemptive version) minimizes the mean number of customers in a non-preemptive M/G/1 queue [5]. Under this rule, when the server gets idle, the next customer to be served is the one having the highest value for $c_i/\mathbb{E}(B_i)$. The $c\mu$-rule is a particular case of the relative-priority policy which can be retrieved by letting the ratios $\tilde{p}_j/\tilde{p}_{j+1}$ go to ∞, $j = 1, \ldots, K$ (assuming that the classes are ordered such that $\frac{c_1}{\mathbb{E}(B_1)} \geq \frac{c_2}{\mathbb{E}(B_2)} \geq \cdots \geq \frac{c_K}{\mathbb{E}(B_K)}$). Hence, Proposition 4.1 can be seen as an extension of the optimality of the $c\mu$-rule in the heavy-traffic regime: the performance improves as larger weights are assigned according to the values of $c_k/\mathbb{E}(B_k)$.

Example 1 (Two classes of customers). *We consider two classes of customers ($K = 2$) and assume that $c_1/\mathbb{E}(B_1) \geq c_2/\mathbb{E}(B_2)$. Without loss of generality, assume that $p_1 + p_2 = \tilde{p}_1 + \tilde{p}_2 = 1$. By Proposition 4.1 we obtain that the scaled holding cost in the (p)-system will be stochastically smaller than that in the $(\tilde{p})$-system if and only if $p_1 \geq \tilde{p}_1$. Hence, the performance improves as a larger weight, i.e., more preference, is given to class 1.*

5. WAITING TIME DISTRIBUTION

We denote by W_k the waiting time for a class-k customer. We have the following result:

Proposition 5.1. *As $(\lambda_1, \ldots, \lambda_K) \to (\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$,*

$$(1-\rho)(W_k, N_1, \ldots, N_K) \xrightarrow{d} X(Z_k, \frac{\hat{\lambda}_1}{p_1}, \ldots, \frac{\hat{\lambda}_K}{p_K}),$$

where X and Z_k are exponentially distributed independent random variables with $\mathbb{E}(X) = 1/\nu(p)$, $\mathbb{E}(Z_k) = 1/p_k$, and $1/\nu(p)$ given by Equation (2).

Sketch of proof. We refer to the report [1] for the full proof.

Let $T_k(u, z_1, \ldots, z_K) = \mathbb{E}(e^{-uW_k} z_1^{N_1} \cdots z_K^{N_K})$ denote the joint Laplace transform of the queue-length vector and the waiting time. Our goal is to study the Laplace transform

$$\begin{aligned}&\lim_{\rho\uparrow 1} T_k((1-\rho)u, \mathrm{e}^{-(1-\rho)\vec{s}}) \\ &= \lim_{\rho\uparrow 1} \mathbb{E}(\mathrm{e}^{-u(1-\rho)W_k} \mathrm{e}^{-\sum_{j=1}^K s_k (1-\rho) N_k}).\end{aligned}$$

In [9] it is shown that $T_k(u, z_1, \ldots, z_K)$ is the unique solution of a partial differential equation. After taking the limit $\rho \uparrow 1$ we derive that its solution is given by

$$\begin{aligned}&\lim_{\rho\uparrow 1} T_k((1-\rho)u, \mathrm{e}^{-(1-\rho)\vec{s}}) \\ &= \frac{p_k \nu(p)}{u} \mathrm{e}^{\frac{p_k(\nu(p)+y)}{u}} \int_{p_k(\nu(p)+y)/u}^{\infty} \frac{\mathrm{e}^{-x}}{x} \mathrm{d}x,\end{aligned}$$

with $y = \sum_{j=1}^{K} \frac{\hat{\lambda}_j}{p_j} s_j$. This coincides with the Laplace transform of the random vector $X(Z_k, \frac{\hat{\lambda}_1}{p_1}, \ldots, \frac{\hat{\lambda}_K}{p_K})$ and hence we obtain the result. □

Remark 2 (Random-order-of-service). *We note that in the case of one class, $K = 1$, the distribution of the waiting time has been obtained previously in [12, 16].*

6. REFERENCES

[1] U. Ayesta, A. Izagirre, and I.M. Verloop. Heavy-traffic analysis of a multi-class queue with relative priorities. *Technical report*, 2011.

[2] S.C. Borst, O.J. Boxma, J.A. Morrison, and R. Núñez-Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, (31):254–262, 2003.

[3] O.J. Boxma, D. Denteneer, and J.A.C. Resing. Some models for contention resolution in cable networks. *Lect. Notes in Comp. Sc.*, 2345:117–128, 2002.

[4] O.J. Boxma, S.G. Foss, J.-M. Lasgouttes, and R. Núñez-Queija. Waiting time asymptotics in the single server queue with service in random order. *Queueing Systems*, (46):35–73, 2004.

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

[6] 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.

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

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

[9] 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.

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

[11] J.F.C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B, Methodological*, 24:383–392, 1962.

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

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

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

[15] I.M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *To appear in Operations Research*, 2011.

[16] A.P. Zwart. Heavy-traffic asymptotics for the single-server queue with random order of service. *Operations Research Letters*, (33):511–518, 2005.