Operations Research Letters 45 (2017) 498–502

Contents lists available at ScienceDirect

Operations Research Letters

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

The advantage of relative priority regimes in multi-class multi-server queueing systems with strategic customers

Binyamin Oz [a,*], Moshe Haviv [b], Martin L. Puterman [c]

[a] Department of Statistics, The University of Auckland, New Zealand
[b] Department of Statistics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel
[c] The Sauder School of Business, The University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 18 March 2017
Received in revised form 20 July 2017
Accepted 21 July 2017
Available online 29 July 2017

Keywords:
Relative priorities
Selfish routing
Queueing network

ABSTRACT

We show that relative priorities can reduce queueing costs in systems that are multi-server and multi-class as long as customers choose their routing policy strategically. This is demonstrated in two models with multi-class Poisson arrivals and parallel memoryless servers with linear cost functions of class mean waiting times. For each model we investigate the Nash equilibria under a given relative priority rule. The central planner's optimal policy is characterized and shown to be of strictly relative priorities in some cases.

1. Introduction

It is well known that the $C\mu$-rule is optimal in centrally planned queueing systems. Specifically, suppose that there is a single-server queue that is fed by various types of customers who arrive in accordance with a Poisson process. Let $\bar{x}_i$ be the mean service time of a class i customer and C_i be his waiting cost per unit of time. The $C\mu$-rule says that of all non-preemptive queue regimes, the one that minimizes the mean total (undiscounted) waiting cost is the one that gives absolute priority based on a decreasing order of $C_i/\bar{x}_i$. This means that upon service completion, the next customer to enter service should be one whose parameter $C_i/\bar{x}_i$ among all other customers present is maximal. It is of course not important which customer among those of this maximal class will be the one to enter. This observation holds also regardless of how many customers of each class are present. See, e.g., [6], p. 125. Suppose there are n classes of customers. Then there exists $n!$ absolute priority orderings of the classes and, as said above, the optimal one among them is the one based on the $C\mu$-rule.

In the above framework the introduction of the option of using lotteries does not change anything. Specifically, suppose that one is not limited to the $n!$ policies which prioritize the classes, but one is allowed to perform a lottery regarding who should enter next. Yet, under this extended set of policies the optimal policy is still the one based on the $C\mu$ rule. This may lead one to conjecture that there is no need for strategies involving lotteries when looking for optimization in queues. However, this conjecture is false. In [4] an example is shown that if customers behave strategically, then a central planner might have such strategy that yields a profit that is strictly higher than any strategy without lotteries. In this example each customer has the option of whether to join the queue or not. Joining is rewarding but it comes with a class-dependent entry fee (in addition to the waiting cost). A profit maximizer who collects the entry fees can assign *relative priority* parameters to the customers that are based on their class. The next to enter service is selected by a lottery that gives customers entrance probabilities that are proportional to their parameters. As it turns out, the priority parameters affect the joining rate at various classes, which then leads to the corresponding profits. Fine tuning of these parameters leads to an increase in the profits.

This paper shows once again how useful such a relative priority scheme can be in queueing models that are multi-server in addition to being multi-class. To this end, we consider two models. The first is a W-shaped model; see Fig. 1 below. In this model there are two classes of customers and three servers. Each class of customers has its dedicated server and one of the servers can serve both classes.

The second model is an M-shaped model; see Fig. 4 below. In this model there are three classes of customers and two servers. Each server has its exclusive class of customers and one of the classes can be served by both servers. Details are given below.

In a system with centralized planning, the decision makers choose the rules for allocating servers to demand classes so as to achieve the best possible system objectives. In this paper we use these settings to motivate the two models we consider, but in our approach customers select the server to balance their waiting

* Corresponding author.
E-mail addresses: b.oz@auckland.ac.nz (B. Oz), haviv@huji.ac.il (M. Haviv), martin.puterman@sauder.ubc.ca (M.L. Puterman).

http://dx.doi.org/10.1016/j.orl.2017.07.005
0167-6377/

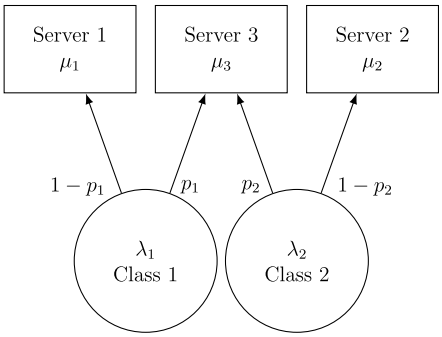

Fig. 1. The W model.

times (if possible) between the servers that are available to them. One may think that giving priority to the class with the highest cost to mean service time ratio (the $C\mu$ rule) is optimal, but as we show, this is not necessarily the case. It might be better to give each class only relative priority (in the sense to be defined later). This seeming paradox can be attributed to the fact that customers behave strategically.

2. Relative priorities, notations, and auxiliary results

In the two models studied in this paper, some of the memoryless servers serve two independent streams of Poisson arrivals of customers. The entrance regime for these servers is that of relative priority. By that we mean that whenever the server is ready to commence servicing the next customer, and there are n_i customers of type i in the queue in front of him, $i = 1, 2$, then the next to enter service is the one at the head of the line of type-i customers with probability $n_i q_i/(n_1 q_1 + n_2 q_2)$, where $q_i \geq 0$ is the relative priority parameter of type-i customers, $i = 1, 2$. We scale these parameters so that $q_1 + q_2 = 1$.

Consider a server with exponentially distributed service times with rate μ and let $W_i(q, \lambda_1, \lambda_2)$ be the corresponding mean waiting time (service inclusive) of type-i customers, $i = 1, 2$, given that the priority parameter of type-1 customers is $q_1 = q$ (and that of type 2 is $q_2 = 1 - q$) and the arrival rate of type-i customers is λ_i, $i = 1, 2$. Based on [5] we learn that this mean waiting time equals

$$W_i(q, \lambda_1, \lambda_2) \equiv \frac{1 - \rho q_i}{(1-\rho)(1 - q\rho_1 - (1-q)\rho_2)} \frac{\rho}{\mu} + \frac{1}{\mu}, \tag{1}$$

where ρ is the traffic intensity, $\rho_i = \lambda_i/\mu$, $i = 1, 2$, and, clearly, $\rho = \rho_1 + \rho_2$. Moreover, the system is stable if and only if $\rho < 1$, which is assumed.

In the following lemmas we introduce some algebraic results on these mean waiting times. The proofs are technical and are available online at https://sites.google.com/site/binyaminoz/RPSM.pdf.

Lemma 2.1. *The following inequalities hold:*

1. $\frac{\partial W_i}{\partial \lambda_j} > 0, i = 1, 2, j = 1, 2$
2. $\frac{\partial W_1}{\partial \lambda_1}\frac{\partial W_2}{\partial \lambda_2} - \frac{\partial W_1}{\partial \lambda_2}\frac{\partial W_2}{\partial \lambda_1} > 0.$

Lemma 2.2. *The derivatives of the mean waiting times with respect to the relative priority parameter q satisfy*

$$\frac{\partial W_1}{\partial q} < 0 \quad \textit{and} \quad \frac{\partial W_2}{\partial q} > 0$$

if $\lambda_1, \lambda_2 > 0$.

Suppose that the social cost associated with type-i customers is $c_i \geq 0$ per customer per unit of time, $i = 1, 2$. The total cost associated with a server that uses the relative priority discipline with parameter q is hence

$$C(c_1, c_2, q, \lambda_1, \lambda_2) = \sum_{i=1,2} c_i \lambda_i W_i(q, \lambda_1, \lambda_2). \tag{2}$$

Lemma 2.3. *The derivative of the total cost with respect to the relative priority parameter q satisfies*

$$\frac{\partial C}{\partial q} = (c_1 - c_2)\lambda_1 \frac{\partial W_1}{\partial q} = (c_2 - c_1)\lambda_2 \frac{\partial W_2}{\partial q}.$$

Remark. Lemma 2.3 combined with Lemma 2.2 shows that the function C is monotone in q. Moreover, it follows that minimizing C is done by setting $q = 1$, i.e., absolute priority for type-1 customers, if $c_1 > c_2$, and $q = 0$, i.e., absolute priority for type-2 customers, if $c_1 < c_2$, and that, of course, agrees with the $c\mu$ rule.

3. The W model

Consider the following memoryless three-server model. Server-i serves at the rate of μ_i, $1 \leq i \leq 3$. There are two independent streams of Poisson arrivals, stream-i with rate λ_i, $i = 1, 2$. A fraction of $1-p_i$ of the customers of stream i go to server i, $i = 1, 2$. The others (with a total rate of $p_1\lambda_1 + p_2\lambda_2$) go to server 3. See Fig. 1.

Server i uses the first-come first-served entrance policy, $i = 1, 2$. The entrance regime for server 3 is that of relative priority, with priority parameter $q_1 = q$ for customers from stream 1, $0 \leq q \leq 1$, and priority parameter $q_2 = 1 - q$ for customers of stream 2.

It is well known that the mean waiting time (service inclusive) at server i equals

$$W^{(i)}(p_i) = \frac{1}{\mu_i - (1-p_i)\lambda_i} = \frac{1}{\mu_i(1-\rho^{(i)})}, \quad i = 1, 2,$$

where $\rho^{(i)} = (1-p_i)\lambda_i/\mu_i$ is the traffic intensity of server i, $i = 1, 2$. The corresponding mean waiting time in server 3 is given in (1) above, i.e.,

$$W_i^{(3)}(q, \lambda_1 p_1, \lambda_2 p_2) \equiv \frac{1 - \rho^{(3)} q_i}{(1-\rho^{(3)})(1 - q\rho_1^{(3)} - (1-q)\rho_2^{(3)})} \frac{\rho^{(3)}}{\mu_3} + \frac{1}{\mu_3}, \quad i = 1, 2,$$

where $\rho_i^{(3)} = \lambda_i p_i/\mu_3$, $i = 1, 2$ and $\rho^{(3)} = \rho_1^{(3)} + \rho_2^{(3)}$.

We are interested in the following decision-making model. A player, called a *central planner*, decides on the parameter q (we will define his selection criterion shortly). This value is announced and becomes known to the arrivals. The arrivals, independently of each other, have to decide which server to seek service from (but recall that customers of type i can choose only between server i and server 3, $i = 1, 2$). We assume that decision making here is as in [1] (see also [2,3]); namely, customers behave in a Nash equilibrium way. By that we mean that if all customers of type i join server 3 with probability $p_i(q)$, while all others go to server i, $i = 1, 2$, then (under the resulting steady-state conditions) an individual customer cannot do better by using a different lottery between his possible servers. This means that if $p_1(q) = 1$ (respectively, $p_1(q) = 0$), then a type-1 customer is not worse off by joining server 3 (respectively, server 1), given that all type-i customers do the same, while all type-2 customers select server 3 with probability $p_2(q)$. Using the above notation, this means that $W^{(1)}(1) > W_1^{(3)}(q, \lambda_1, \lambda_2 p_2(q))$ (respectively, $W^{(1)}(0) < W_1^{(3)}(q, 0, \lambda_2 p_2(q))$). As importantly, in the case where $0 < p_1(q) < 1$, this customer is indifferent between the two options (i.e., they come with the same mean waiting time), given that all type-1 customers use this strategy and given that all type-2 customers select server 3 with