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*
[b] *Department of Statistics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem*
[c] *The Sauder School of Business, The University of British Columbia*

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.

Keywords:
Relative Priorities, Selfish Routing, Queueing Network

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 $\overline{x}_i$ 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 $C_i/\overline{x}_i$. This means that upon service completion, the next customer to enter service

*Corresponding author.
Email addresses: binyamin.oz@gmail.com (Binyamin Oz), moshe.haviv@gmail.com (Moshe Haviv), martin.puterman@sauder.ubc.ca (Martin L. Puterman)

should be one whose parameter $C_i/\overline{x}_i$ 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., [5], 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 [3] 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 affects 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 Figure 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 Figure 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 pos-

sible 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 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 [4] 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 Figure 1.

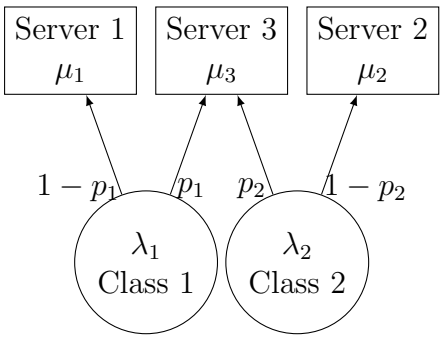

Figure 1: The W model

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]); 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 probability $p_2(q)$. Using the above notation, this condition means that

$$W_1^{(3)}(q, \lambda_1 p_1(q), \lambda_2 p_2(q)) = W^{(1)}(p_1(q)), \quad 0 < p_1(q) < 1. \tag{3}$$

Similarly for type-2 customers:

$$W_2^{(3)}(q, \lambda_1 p_1(q), \lambda_2 p_2(q)) = W^{(2)}(p_2(q)), \quad 0 < p_2(q) < 1. \tag{4}$$

More formally, define $SR_2(p_1)$ as the stable response of type-2 customers against p_1 used by all type-1 customers and define $SR_1(p_2)$ as the stable response of type-1 customers for any p_2 by

$$SR_1(p_2) = \begin{cases} 1 & W_1^{(3)}(q, \lambda_1, \lambda_2 p_2) < W^{(1)}(1) \\ 0 & W_1^{(3)}(q, 0, \lambda_2 p_2) > W^{(1)}(0) \\ p_1^*(p_2) & \text{otherwise,} \end{cases}$$

where $p_1^*(p_2)$ is p_1 as an implicit function of p_2 satisfying (3). $SR_2(p_1)$ is defined symmetrically. Equilibria correspond to the cases where the two stable response curves intersect each other; i.e., $(p_1(q), p_2(q))$ is an equilibrium point if

$$SR_2(p_1(q)) = p_2(q) \quad \text{and} \quad SR_1(p_2(q)) = p_1(q). \tag{5}$$

In the case where the equilibrium point $(p_1(q), p_2(q))$ satisfies $0 < p_1(q), p_2(q) < 1$, we call it an *interior* equilibrium point.

It is easy to see that $SR_i(p_j)$ and $p_i^*(p_j)$ are monotone non-increasing and continuous functions in p_j, $1 \le i \ne j \le 2$. This is due to the fact that the more type-j customers join server 3, the less appealing is this server to type-i customers, as well as to the fact that the mean waiting time in server 3 is a continuous function of p_j.

Figure 2 shows an example of the two curves representing $SR_1(p_2)$ and $SR_2(p_1)$. As can be seen, the curve representing $SR_1(p_2)$ is steeper than the curve representing $SR_2(p_1)$ at the equilibrium point where the two curves

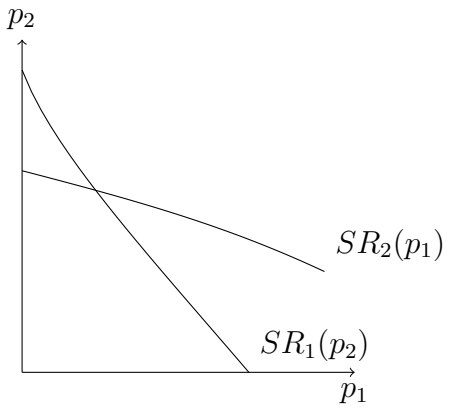

Figure 2: Stable response curves in the case where $\lambda_1 = 2,\ \lambda_2 = 3,\ \mu_1 = 2,\ \mu_2 = 2,\ \mu_3 = 3,\ q = 0.25$.

intersect each other. This phenomenon holds in general and it leads to the uniqueness of the equilibrium profile, as shown below.

Theorem 3.1. *For any given q, $0 \leq q \leq 1$, $(p_1(q), p_2(q))$ is unique.*

Proof: Clearly, an equilibrium point always exists. The next step is to show that there is at most one intersection point of the two stable response curves. We show that by proving that, with respect to the p_1, p_2 plane, at any intersection point the curve representing $SR_1(p_2)$ is steeper than the curve representing $SR_2(p_1)$, i.e., $\frac{\partial SR_2}{\partial p_1} > (\frac{\partial SR_1}{\partial p_2})^{-1}$.

In the case of an equilibrium point of the form $(0, p_2)$ or $(1, p_2)$, $0 \leq p_2 \leq 1$, $(\frac{\partial SR_1}{\partial p_2})^{-1} = -\infty$, and the statement is true. In the case of an equilibrium point of the form $(p_1, 0)$ or $(p_1, 1)$, $0 \leq p_1 \leq 1$, $\frac{\partial SR_2}{\partial p_1} = 0$, and the statement is true again. In the case of an interior equilibrium point, by using the implicit function theorem on (3) and (4) we get that

$$\frac{\partial SR_2}{\partial p_1} = \frac{\partial p_2^*}{\partial p_1} = \frac{\partial W_2^{(3)}}{\partial p_1}\left[\frac{\partial W^{(2)}}{\partial p_2} - \frac{\partial W_2^{(3)}}{\partial p_2}\right]^{-1}$$

and that

$$\frac{\partial SR_1}{\partial p_2} = \frac{\partial p_1^*}{\partial p_2} = \frac{\partial W_1^{(3)}}{\partial p_2}\left[\frac{\partial W^{(1)}}{\partial p_1} - \frac{\partial W_1^{(3)}}{\partial p_1}\right]^{-1}.$$

The next step is to show that $\frac{\partial SR_2}{\partial p_1} - (\frac{\partial SR_1}{\partial p_2})^{-1}$ is positive at the intersection point. Indeed,

$$\frac{\partial SR_2}{\partial p_1} - \left(\frac{\partial SR_1}{\partial p_2}\right)^{-1} = \frac{\frac{\partial W^{(1)}}{\partial p_1}\frac{\partial W^{(2)}}{\partial p_2} - \frac{\partial W_1^{(3)}}{\partial p_1}\frac{\partial W^{(2)}}{\partial p_2} - \frac{\partial W_2^{(3)}}{\partial p_2}\frac{\partial W^{(1)}}{\partial p_1} + \left[\frac{\partial W_1^{(3)}}{\partial p_1}\frac{\partial W_2^{(3)}}{\partial p_2} - \frac{\partial W_2^{(3)}}{\partial p_1}\frac{\partial W_1^{(3)}}{\partial p_2}\right]}{\left[\frac{\partial W_2^{(3)}}{\partial p_2} - \frac{\partial W^{(2)}}{\partial p_2}\right]\frac{\partial W_1^{(3)}}{\partial p_2}}.$$

Of course, $\frac{\partial W_i^{(3)}}{\partial p_j}$ is positive and $\frac{\partial W^{(i)}}{\partial p_i}$ is negative for $i, j = 1, 2$. Therefore, the denominator and the sum of the first three terms in the numerator are positive. From item 2 of Lemma 2.1 we get that the fourth term in the numerator is positive. ∎

Remark. Finding the (unique) equilibrium pair of $(p_1(q), p_2(q))$ can be done as follows. One can first solve the two (non-linear) equations in (3) and (4). If both $p_1(q)$ and $p_2(q)$ are fractions, one is done. If not, one can check the following four options: $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. Specifically, for each pair, four inequalities need to be checked. If this fails, namely, if for each pair, at least one of the four inequalities does not hold, one needs to check the four options that are left: one value at the extreme zero or one, obeying the corresponding inequality, coupled with the other value obeying the equation for the other server, with the resulting joining probability being between zero and one.

Now we define the criterion used by the central planner. Recall that the social cost associated with type-i customers is $c_i > 0$ per customer per unit of time, $i = 1, 2$. Given q, $(p_1(q), p_2(q))$ are the routing proportions resulting from the above-described equilibrium behavior from the customers' side. Note that they are not functions of c_i, $i = 1, 2$. The central planner's objective is to minimize the total cost per unit of time in the entire system. This is given by the social cost function

$$SC(q) = C^{(3)}(c_1, c_2, q, \lambda_1 p_1(q), \lambda_2 p_2(q)) + \sum_{i=1,2} c_i \lambda_i (1 - p_i(q)) W^{(i)}(p_i(q)),$$

where $C^{(3)}$ is the cost associated with server 3 as defined in (2); i.e, he looks for q^* where

$$q^* = \arg \min_{0 \le q \le 1} SC(q).$$

An alternative way to write the social cost function is

$$SC(q) = \sum_{i=1,2} c_i \lambda_i W_i^e(q) = \sum_{i=1,2} c_i L_i^e(q),$$

where $W_i^e(q)$, $i = 1, 2$, is the equilibrium mean waiting time of type-i customers in the system and it equals $\min\{W_i^{(3)}(q, \lambda_1 p_1(q), \lambda_2 p_2(q)), W^{(i)}(p_i(q))\}$ and $L_i^e(q) = \lambda_i W_i^e(q)$, $i = 1, 2$, is the equilibrium sum of mean queue lengths of type-i customers, namely, the queue length in server i plus the queue length of type i customers in server 3, $i = 1, 2$.

Note that the customers' profile needs to specify $(p_1(q), p_2(q))$ for *all* q, $0 \le q \le 1$, while only q^* is required by the part of the equilibrium profile for the central planner. Also, $(q^*, p_1(q^*), p_2(q^*))$ is known as the *equilibrium path*, which in fact is the actual behavior predicted by the model.

Figure 3 shows an example of the pairs L_1^e and L_2^e achievable by the relative priority parameter values $0 \le q \le 1$. The two extreme points are $(L_1^e(0), L_2^e(0))$ and $(L_1^e(1), L_2^e(1))$ and all other points are $(L_1^e(q), L_2^e(q))$ for some $0 < q < 1$. Since $SC(q)$ is a linear combination of L_1^e and L_2^e, the optimal point $(L_1^e(q^*), L_2^e(q^*))$ is its point of tangency with a line with slope $-c_1/c_2$. In this example, where $c_1 = c_2$, the optimal point is the point of tangency with a line with slope -1. This indicates that the optimal relative priority parameter is at some $0 < q < 1$. In fact, the optimal policy in server 3 is of relative priorities with priority parameter strictly between zero and one for all c_1 and c_2 such that $-c_1/c_2$ is between the two slopes in the two extreme points.

4. The M model

Consider the following two-server memoryless queueing system. Server i works at a service rate of μ_i, $i = 1, 2$, and is fed by a Poisson stream of customers, $i = 1, 2$. There is a third Poisson stream of customers with a rate of λ_3. A fraction p of the third stream goes to server 1 while the rest goes to server 2. See Figure 4. We denote p by p_1, and $1 - p$ by p_2.

The queueing policy at each queue is that of relative priority. The priority parameter of class i customers in their queue is q_i, $0 \le q_i \le 1$, while for the customers of the third stream, it is $1 - q_i$, $i = 1, 2$.

The scenario here is as follows. First, the central planner announces q_i, $i = 1, 2$ and, given that, the customers of the third class decide on the

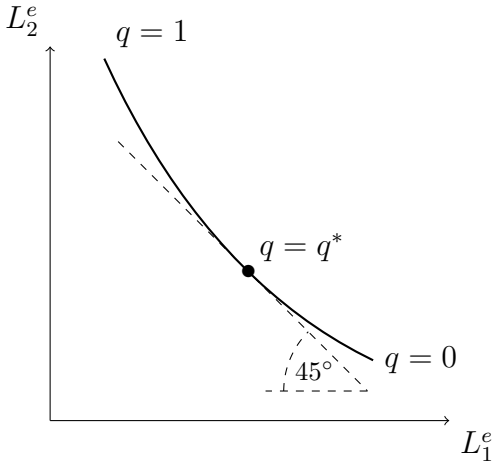

Figure 3: The achievable pairs L_1^e and L_2^e and the optimal point $(L_1^e(q^*), L_2^e(q^*))$ in the case where $\lambda_1 = 2,\ \lambda_2 = 3,\ \mu_1 = 5,\ \mu_2 = 5,\ \mu_3 = 6,\ c_1 = c_2$.

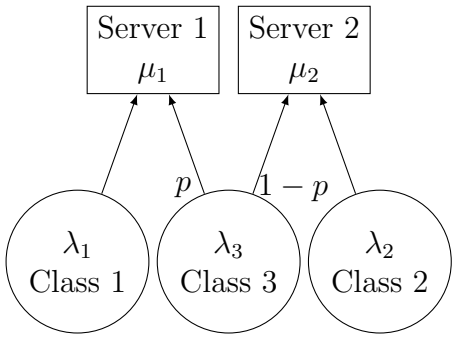

Figure 4: The M model

corresponding p in an equilibrium fashion. The social cost associated with type-i customers is $c_i > 0$ per customer per unit of time, $i = 1, 2, 3$. The central planner's objective is to minimize the total cost per unit of time in the system; i.e, the decision maker selects q_i, $i = 1, 2$, so as to minimize

$$\sum_{i=1,2} C^{(i)}(c_i, c_3, q_i, \lambda_i, p_i\lambda_3),$$

where $C^{(i)}$, $i = 1, 2$ is the cost associated with server i as defined in (2).

Equilibrium conditions: In equilibrium, at least one of the following three conditions hold:

1. $W_3^{(1)}(q_1, \lambda_1, 0) - W_3^{(2)}(q_2, \lambda_2, \lambda_3) \geq 0$ and $p = 0$.
2. $W_3^{(1)}(q_1, \lambda_1, \lambda_3) - W_3^{(2)}(q_2, \lambda_2, 0) \leq 0$ and $p = 1$.
3. $W_3^{(1)}(q_1, \lambda_1, p\lambda_3) - W_3^{(2)}(q_2, \lambda_2, (1-p)\lambda_3) = 0$ and $0 < p < 1$.

Theorem 4.1. *For any given set of parameters λ_i, $i = 1, 2, 3$, μ_i, $i = 1, 2$ and q_i, $i = 1, 2$, there exists a unique $p^e(q_1, q_2)$ that obeys exactly one of the above equilibrium conditions and is given by, respectively,*

$$p^e(q_1, q_2) = \begin{cases} 0 & W_3^{(1)}(q_1, \lambda_1, 0) > W_3^{(2)}(q_2, \lambda_2, \lambda_3) \\ 1 & W_3^{(1)}(q_1, \lambda_1, \lambda_3) < W_3^{(2)}(q_2, \lambda_2, 0) \\ p^*(q_1, q_2) & \textit{otherwise,} \end{cases}$$

where $p^*(q_1, q_2)$ is p as an implicit function of q_1 and q_2 satisfying

$$W_3^{(1)}(q_1, \lambda_1, p\lambda_3) - W_3^{(2)}(q_2, \lambda_2, (1-p)\lambda_3) = 0.$$

Proof: Since both $W_3^{(1)}$ and $W_3^{(2)}$ are continuous functions in their third argument, from the intermediate value theorem we get that if conditions 1 and 3 do not hold, then condition 2 holds, which proves existence. Recalling Lemma 2.1, $W_3^{(1)}$ and $W_3^{(2)}$ are monotone increasing in their third argument, and hence $W_3^{(1)}(q_1, \lambda_1, p\lambda_3) - W_3^{(2)}(q_2, \lambda_2, (1-p)\lambda_3)$ is monotone increasing in p. From this monotonicity we have that if one condition holds, then the other two do not. ∎

For simplicity, we use the following short notations below: $p_1^e = p^e(q_1, q_2)$, $p_2^e = 1 - p^e(q_1, q_2)$, and $W_3^e = \min\{W_3^{(1)}(q_1, \lambda_1, p_1^e\lambda_3), W_3^{(2)}(q_2, \lambda_2, p_2^e\lambda_3)\}$. Note that W_3^e is the equilibrium mean waiting time for type-3 customers in the system.

As mentioned, the central planner looks for a pair $(q_1, q_2) \in [0, 1] \times [0, 1]$ that minimizes

$$SC(q_1, q_2) = \sum_{i=1,2} C^{(i)}(c_i, c_3, q_i, \lambda_i, p_i^e\lambda_3),$$

for some given $c_i \geq 0$, $i = 1, 2, 3$.

Figure 5 shows an example for the achievable social cost as a function of the resulting p^e from the corresponding (undisplayed) relative priority

parameters q_1, q_2. Each point (SC, p^e) is an outcome of some point $(q_1, q_2) \in [0,1] \times [0,1]$. The outcome points of the boundary points of $[0,1] \times [0,1]$ are indicated by solid and dashed lines and the corner points, $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$, are indicated by bullets. As can be seen, the minimum of SC is achieved in a boundary point between $(0,0)$ and $(0,1)$ (indicated by a star); i.e., the optimal priorities are absolute priority to type-3 customers in server 1 and some relative priority parameter $0 < q_2 \approx 0.1 < 1$ in server 2. The fact that the optimal relative priority parameter at one of the two servers is in the boundary is not specific for this example but is true in general, as Theorem 4.2 below claims.

Another observation is that the boundary points can be divided into two sections. The first is the boundary points connecting $(1,0)$, $(0,0)$, and $(0,1)$ (their outcome points are indicated by dashed lines). The second section is the points connecting $(1,0)$, $(1,1)$ and $(0,1)$ (their outcome points are indicated by solid lines). In both sections, p^e monotonically increases from the first point, $(1,0)$, to the last point, $(0,1)$. As a result, each achievable p^e of a point in the first section is achievable with a point in the second section with larger relative priority parameters. These two points do not share the same SC except for exactly one case ($p^e = 0.5$). The order of the costs depends on the value of p^e. If $p^e c_1 - (1-p^e)c_2 < c_3$ (small values of p^e), then the cost of the first section is less than the cost of the second, and vice versa. In fact, the above discussion is not limited to this example but holds in general.

Theorem 4.2. *For any given set of parameters λ_i, $i = 1,2,3$, μ_i, $i = 1,2$, there exists an optimal policy that uses absolute priority at least in one server (but not necessarily in both). In other words, there exists a point $(q_1^*, q_2^*) \in Bd([0,1] \times [0,1])$ such that*

$$(q_1^*, q_2^*) \in \arg \min_{(q_1,q_2)\in[0,1]\times[0,1]} SC(q_1, q_2),$$

where $Bd([0,1] \times [0,1])$ is the boundary of the set $[0,1] \times [0,1]$.

Proof: Since $SC(q_1, q_2)$ is a continuous function in both q_1 and q_2, the minima always exists over the compact set $[0,1] \times [0,1]$. We will show that for every point $(q_1, q_2) \in [0,1]\times[0,1]$, there exists a point $(q_1^*, q_2^*) \in \mathrm{Bd}([0,1]\times[0,1])$ such that $SC(q_1^*, q_2^*) \leq SC(q_1, q_2)$. We do so by showing that increasing

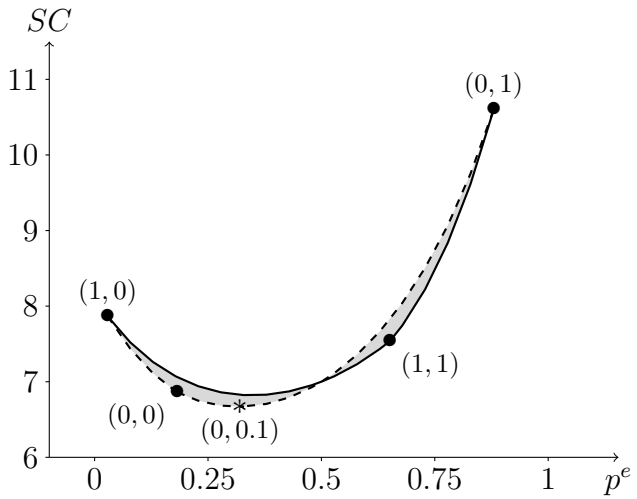

Figure 5: The achievable social cost as a function of the resulting p^e from the corresponding relative priority parameters q_1, q_2 in the case where $\lambda_1 = 1,\ \lambda_2 = 3,\ \lambda_3 = 2,\ \mu_1 = 4,\ \mu_2 = 6,\ c_1 = 5,\ c_2 = 1,\ c_3 = 3$.

both q_1 and q_2 while keeping $p^e(q_1, q_2)$ constant, monotonically increases or decreases the social cost. From this it follows that for any interior point, there exists a boundary point (with larger or smaller relative priority parameters) with no greater social cost.

Note that despite the fact that $W_3^{(i)}(q_i, \lambda_i, p_i^e\lambda_3)$ and W_3^e are not necessarily equal, $p_i^e W_3^{(i)}(q_i, \lambda_i, p_i^e\lambda_3) = p_i^e W_3^e$. Therefore, the objective function can be rewritten as follows:

$$\begin{aligned}
&\textstyle\sum_{i=1,2} C^{(i)}(c_i, c_3, q_i, \lambda_i, p_i^e\lambda_3) \\
&= \textstyle\sum_{i=1,2} C^{(i)}(c_i, c_i, q_i, \lambda_i, p_i^e\lambda_3) + \\
&\qquad\qquad \textstyle\sum_{i=1,2} (c_3 - c_i) p_i^e \lambda_3 W_3^{(i)}(q_i, \lambda_i, p_i^e\lambda_3) \\
&= \textstyle\sum_{i=1,2} C^{(i)}(c_i, c_i, q_i, \lambda_i, p_i^e\lambda_3) + \sum_{i=1,2} (c_3 - c_i) p_i^e \lambda_3 W_3^e \\
&= \textstyle\sum_{i=1,2} C^{(i)}(c_i, c_i, q_i, \lambda_i, p_i^e\lambda_3) + (c_3 - c_1 p_1^e - c_2 p_2^e)\lambda_3 W_3^e.
\end{aligned}$$

By Lemma 2.3, increasing both q_1 and q_2 while keeping $p^e(q_1, q_2)$ constant does not change the first term. For the second term, W_3^e increases and so the objective function decreases monotonically if $c_3 - c_1 p_1^e - c_2 p_2^e < 0$, increases monotonically if $c_3 - c_1 p_1^e - c_2 p_2^e > 0$, and does not change if $c_3 - c_1 p_1^e - c_2 p_2^e = 0$.

■

5. Extensions

The two models discussed here can be generalized in two directions. First, service distribution at each server can be type-dependent. Second, service times need not be exponential. In fact, we can have both these extensions simultaneously. Indeed, for servers that serve a single class of customers, we get by the famous Khintchine–Pollazcek formula that mean waiting times are monotone decreasing and concave functions of their arrival rates. Mean waiting times in servers that serve multiple classes have to be slightly modified: from [4] we can see that the multiplicative term of ρ/μ in (1) needs to replaced by the mean residual service time of the one in service. Our conjecture is that one would get the same qualitative results.

Furthermore, an additional performance measure of the system can be looked at. Consider the following alternative social optimization for both models, where it is the central planner who selects the relative priority parameters as well as the routing fractions. In such systems, all mean waiting times can be computed by the formulas given above. Then the objective is as before. Note that according to the $c\mu$ rule, under these settings, the optimal priorities in every server are absolute priorities for one type of customers. One can divide the central planner objective values under the two criteria and get the price of anarchy (PoA), i.e., the (proportional) loss due to the fact that the customers themselves select which server to join (in an equilibrium fashion) in comparison with the case where the splits between the servers is decided centrally.

Acknowledgement

The majority of this work was done while the first author was a Ph.D. student in the Department of Statistics and the Federmann Center for the Study of Rationality, the Hebrew University of Jerusalem. Research was partially funded by the Programme of Canadian Studies, Co-sponsored by the Government of Canada and Ralph and Roz Halbert of Toronto.

[1] Edelson, M. N. and D. K. Hilderbrand (1975), "Congestion tolls for Poisson queuing processes," *Econometrica*, 43, 81–92.

[2] Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer Academic Publisher, Boston. Also in http://www.tau.ac.il/ hassin/book.html

[3] Hassin, R. and M. Haviv (2006), "Who should be given priority in a queue?," *Operations Research Letters* 34, 191–198.

[4] Haviv, M. and J. van der Wal (2006), "Waiting times in queues with relative priorities," *Operations Research Letters*, 35, 591–594.

[5] Kleinrock, L. (1976) *Queueing Systems, Volume 2: Computer Applications*, Wiley, New York.