Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Operations Research Letters 34 (2006) 191–198
Operations Research Letters
www.elsevier.com/locate/orl

Who should be given priority in a queue?

Refael Hassin[a], Moshe Haviv[b,*]

[a]*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69987, Israel*
[b]*Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel*

Received 12 November 2004; accepted 11 April 2005
Available online 13 June 2005

Abstract

We consider a memoryless single server queue with two classes of customers, each having its fixed entry fee. We show that profit and social welfare may benefit from a service discipline based on relative priorities.

Keywords: Single server queues; Equilibrium behavior; Discriminatory processor sharing (DPS); Relative priorities

1. Introduction

We consider a single server queueing model where customers require service whose length follows an exponential distribution with mean which we assume without loss of generality to be 1. Each customer belongs to one out of two classes of customers. The potential arrival process of customers of class i (called i-*customers*) is Poisson with rate λ_i, $i = 1, 2$. We assume that these rates are large so that if all potential customers from some class join, the queue will explode. Class-i customers suffer a cost of C_i per unit of time while waiting in the queue (including service time), and are rewarded by R_i due to service completion. An i-customer who decides to enter pays a fee of T_i, $T_i < R_i$. Thus, the expected utility of an i-customer who joins the queue is $R_i - T_i - C_i W_i$, where W_i is his expected time in the system, $i = 1, 2$. Lastly, we assume that the customer cannot observe the queue length before joining, and that the server can distinguish between the customer types and set priorities and fees that discriminate between the classes.

Suppose that $R_1 > R_2$ and also $C_1 > C_2$. For the customers who already joined the queue, the value of R is of no relevance, and their total cost of waiting is minimized if class 1 obtains priority over class 2. However, if such priority is announced, it will affect the arrival process, encouraging more 1-customers and less 2-customers to join. Overall, the effect may be negative. In this paper we will investigate the question of which class should be given priority.

Typically, there are three criteria to consider when dealing with decision making in queueing systems. The first is that of *self-optimization*. Here an individual

* Corresponding author.
E-mail addresses: hassin@post.tau.ac.il (R. Hassin), haviv@mscc.huji.ac.il (M. Haviv).

0167-6377/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2005.04.007

wants to maximize his personal reward. In particular, an i-customer joins if his expected utility from joining is positive, he does not join if it is negative, and he is indifferent between the two options if it is zero. The second criterion is that of *social optimization*. Here, entry fees are considered as transfer payments and hence not included in the utility function. The goal is to find joining rates, q_1 and q_2 that maximize the rate of society gain, which is $q_1(R_1 - C_1W_1) + q_2(R_2 - C_2W_2)$. To administrate these arrival rates, the social planner may determine entry fees such that when customers consider their self-optimization, they end up with the socially optimal joining rates. The third criterion is *profit maximization*. Here, the service provider is interested in maximizing $q_1T_1 + q_2T_2$. Customers join if it is in their self-interest to do so. So a fee which is too high, may deter many of them from joining and hence charging too much will not be in the benefit of the profit maximizer.

A standard argument (see [5] or [8, p. 49]) can be used to show that given the unobservable nature of the queue, the server can extract all of the consumer surplus. Therefore, the objective of the profit maximizer and the social objective turn out to be identical, and the profit maximizing entry fees also maximize social welfare.

It is also well known that for *any given joining rates* social welfare is maximized by the $C\mu$-rule, which says that customer classes with a high value of $C\mu$ should get priority of those with lower values. Mendelson and Whang [12] have shown that the socially optimal arrival rates can also be obtained by setting appropriate entry fees *under the $C\mu$-rule*. It follows from the previous paragraph that the $C\mu$-rule will also be applied by a profit maximizing server.

In this paper we concentrate on two second-best optimization problems, assuming that the server is not free to select any prices of his choice. We consider two models. In the first, the fees are fixed and the server's only way to control the queue is by choosing a priority scheme. We are interested in whether giving priority to the class with higher waiting cost, which is optimal under the optimal entry fees, is still optimal when these fees are fixed. In the second, the server can choose a single price that will apply to both classes.

Our analysis indicates that with fixed prices the server may be able to do better by assigning strict relative priorities. We prove this by analyzing the model under a generalized priority scheme called *discriminatory processor sharing* (DPS). Under this model there exist nonnegative parameters p_1 and $p_2 = 1 - p_1$ representing *relative priority* of customers of the two classes. If n_i customers are present in the system, $i = 1, 2$, an i-customer obtains $p_i/(n_1p_1 + n_2p_2)$ of the service capacity of the server. In particular, the total capacity dedicated to class-i is $n_ip_i/(n_1p_1 + n_2p_2)$. Of course, $p_1 \in \{0, 1\}$ means that one of the classes obtains absolute priority.

Remark 1.1. Since service times are exponential with a common mean, the DPS model is equivalent to the following scheme: There is no physical entrance into service; the server produces some goods whose recipient will be determined when production ends; the actual recipient is decided by a lottery conducted upon production completion among all customers presented in the queue. The odds that a tagged i-customer obtains the good are proportional to p_i. Thus, the recipient is an i-customer with probability $n_ip_i/(n_1p_1 + n_2p_2)$.

The DPS discipline is well established in the queueing, computer science and communication literature. See, for example, [9,7,11,13]. In particular, delays in the Internet due to the bandwidth allocation are often measured by a DPS model, see [1,4].

Hayel and Tuffin [10] defined a similar DPS model but assume that the utility of a i-customer who joins is $W_i^{-\alpha_i} - T_i$ for some parameter $\alpha_i > 0$, $i = 1, 2$. They report that in all their numerical experiments the priority schemes for the optimal entry fees were to give priority to one of the classes. As the $C\mu$ rule does not suit their model, a different proof or an explanation for this phenomenon is still an open problem. In the case of fixed prices, their findings are qualitatively similar to ours.

In the next section we give some preliminary results. In Section 3, we characterize the equilibrium solution. We first characterize the *boundary equilibria* where only one class joins, and then the *internal equilibria* where customers from the two classes join. We derive necessary and sufficient conditions for each type of these equilibria. In Section 4, we consider profit and social welfare maximization. We give a numerical example where the optimal priority parameters are strictly between zero and one and the equilibrium arrival pattern is internal. In Section 5, we solve the model with a single price common to both classes. We show that under the common optimal entry fee, giving absolute priority to one of the classes is optimal. Yet, this class is not always the one selected by the $C\mu$-rule.

2. Preliminary results

We assume the following two-stage decision making process. First, a *central planner* determines the parameters p_1, p_2, and possibly also T_1, and T_2. These values are announced and become a common knowledge among the customers. They, in their turn, decide whether or not to enter and pay their class entry fee. Note that the queue size is not observable to them and no refund or reneging is possible once a customer joins. Not joining (called *balking*) comes with no cost or reward. Note that associating zero value with balking is without loss of generality.

As said, customers have two possible actions: To join or not to join the queue. We now introduce an additional option, that of mixing between these two options with some probability. Thus, any strategy can be characterized by a joining probability, say p. A strategy profile is now a set of joining probabilities, one for each customer. A strategy profile is called *symmetric* if according to it all use the same strategy. A symmetric *Nash equilibrium* strategy profile, is a joining probability p such that if used by all, then under the resulting steady-state conditions, no individual has any incentive to deviate to some other strategy.

Suppose that i-customers join with probability π_i, then their *effective arrival rate* or *joining rate* is $q_i = \lambda_i \pi_i$. In equilibrium, if $q_i > 0$ then joining is a best response for i-customers, and if $q_i < \lambda_i$ then balking is a best response. Thus, $0 < q_i < \lambda_i$ means that i-customers are indifferent between joining and balking (and they might as well mix between the two with any probabilities, in particular, the equilibrium probabilities).

Denote by $W_i(q_1, q_2)$, $i = 1, 2$, the expected time in the system of an i-customer who joins, when the joining rates are q_1 and q_2. We refer to this time as the *waiting time*. Note that $W_1(0, q_2)$ is the expected waiting time for an 1-customer who joins when all his peers balk. To simplify the exposition, we suppress q_1 and q_2 from $W_i(q_1, q_2)$ and use instead W_i.

We denote by (q_1^*, q_2^*) a pair of equilibrium arrival rates. There are three options regarding this pair: (1) $C_i W_i \geqslant R_i - T_i$ if $q_i^* = 0$, (2) $C_i W_i = R_i - T_i$ if $0 < q_i^* < \lambda_i$, and (3) $C_i W_i \leqslant R_i - T_i$ if $q_i^* = \lambda_i$. The third option, is ruled out here since by assumption $q_i = \lambda_i$ implies $W_i = \infty$. In particular, the condition $q_i^* < \lambda_i$ automatically holds and hence the second option reduces in fact to $q_i^* > 0$.

Denote $\gamma_k = (R_k - T_k)/C_k$, then in equilibrium

$$q_k^*(\gamma_k - W_k) = 0, \quad k = 1, 2. \tag{1}$$

Without loss of generality we assume that $R_i \geqslant T_i + C_i$ (recall that the service rate is 1), or

$$\gamma_i \geqslant 1, \quad i = 1, 2. \tag{2}$$

Suppose that 1-customers obtain absolute priority over 2-customers (i.e. $p_1 = 1$ and $p_2 = 0$) and their arrival rates are $q_1 > 0$ and $q_2 \geqslant 0$, respectively, with $q_1 + q_2 < 1$.

Without loss of generality we make the standard assumption that $R_1 > C_1/\mu$, and this implies that $q_1 > 0$ when $p_1 = 1$. Then, their mean waiting times are

$$W_1 = \frac{1}{1 - q_1}$$

and

$$W_2 = \frac{1}{(1 - q_1)(1 - q_1 - q_2)} \tag{3}$$

(see Lemma 3.1 below or [11, p. 122]).

By (1) and (3)

$$q_1^* = 1 - \frac{1}{\gamma_1} \tag{4}$$

and

$$q_2^* = \max\left\{0, \frac{1}{\gamma_1} - \frac{\gamma_1}{\gamma_2}\right\}.$$

3. The equilibrium

We now solve the general case. Let $\rho = q_1 + q_2$ denote the total arrival rate. The following lemma is due to Fayolle et al. [6].

Lemma 3.1.

$$W_1 = \frac{1 - p_1\rho}{(1 - \rho)(1 - q_1 p_1 - q_2 p_2)}, \tag{5}$$

$$W_2 = \frac{1 - p_2\rho}{(1-\rho)(1 - q_1 p_1 - q_2 p_2)}. \tag{6}$$

Remark 3.2. Fayolle et al. [6] derived the expected waiting time formulas also when the service rates μ_i of i-customers, $i = 1, 2$, differ. Let $\rho_i = q_i/\mu_i$ then for $i \neq j$

$$W_i = \frac{1}{\mu_i(1-\rho)}\left[1 + \frac{\mu_i \rho_j (p_j - p_i)}{D}\right],$$

where $\rho = \rho_1 + \rho_2$ and $D = \mu_1 p_1(1-\rho_1) + \mu_2 p_2(1-\rho_2)$. This relation can be used to solve our model also for this generalized case. However, the resulting formulas are considerably more complicated and we chose to restrict our derivation to the special case where $\mu_1 = \mu_2$.

3.1. Boundary equilibria

The arrival rates q_1 and q_2 with $q_1 > 0$ and $q_2 = 0$ define an equilibrium if and only if $C_1 W_1 = R_1 - T_1$, or equivalently

$$W_1 = \gamma_1 \tag{7}$$

and $C_2 W_2 \geqslant R_2 - T_2$, or equivalently,

$$W_2 \geqslant \gamma_2. \tag{8}$$

In this case, 1-customers face a single class FCFS $M/M/1$ queue and $W_1 = 1/(1 - q_1)$. Substituting in (7) we conclude (as in (4)) that

$$q_1^* = 1 - \frac{1}{\gamma_1}. \tag{9}$$

Considering the ratio W_2/W_1 from (5) and (6), and then substituting $\rho = q_1$, (7) and (9), gives

$$W_2 = W_1 \frac{1 - p_2\rho}{1 - p_1\rho} = \gamma_1 \frac{1 - p_2(\gamma_1 - 1/\gamma_1)}{1 - p_1(\gamma_1 - 1/\gamma_1)} = \gamma_1 \frac{\gamma_1 p_1 + p_2}{\gamma_1 p_2 + p_1}. \tag{10}$$

Note that W_2 would be the expected queueing time for a 2-customer if he joined. This of course does not contradict the fact that under the equilibrium solution none of them actually joins.

A necessary and sufficient condition for q_1^* as given in (9) and $q_2^* = 0$ to define a pair of equilibrium arrival rates is that $W_2 \geqslant \gamma_2$ where W_2 is given in (10). Substituting $p_2 = 1 - p_1$ in (10), this condition amounts to

$$p_1 \geqslant \frac{\gamma_1(\gamma_2 - 1)}{(\gamma_1 + \gamma_2)(\gamma_1 - 1)} \equiv B. \tag{11}$$

Remark 3.3.

1. By (2), $B \geqslant 0$.
2. If $\gamma_2 > \gamma_1^2$ then $B > 1$ and hence a boundary equilibrium with $q_1^* > 0$ and $q_2^* = 0$ is not possible regardless of the priority parameters.

By interchanging the roles of the indices 1 and 2, a solution with $q_1^* = 0$ and $q_2^* = 1 - 1/\gamma_2$ defines an equilibrium if and only if

$$p_2 \geqslant \frac{\gamma_2(\gamma_1 - 1)}{(\gamma_1 + \gamma_2)(\gamma_2 - 1)}$$

or, since $p_1 + p_2 = 1$

$$p_1 \leqslant \frac{\gamma_2^2 - \gamma_1}{(\gamma_1 + \gamma_2)(\gamma_2 - 1)} \equiv A. \tag{12}$$

Note that by (12), (11), and (2)

$$B - A = \frac{(\gamma_1 - \gamma_2)^2}{(\gamma_1 + \gamma_2)(\gamma_1 - 1)(\gamma_2 - 1)} \geqslant 0. \tag{13}$$

The following remark is the counterpart of Remark 3.3.

Remark 3.4.

1. Since w.l.o.g. $\gamma_1 \geqslant 1$, it follows that $A \leqslant 1$.
2. If $\gamma_1 > \gamma_2^2$ then $A < 0$ and hence a boundary equilibrium with $q_1^* = 0$ and $q_2^* > 0$ is not possible regardless of the priority parameters.

3.2. Internal equilibria

We now focus our attention on equilibria with arrivals from both classes, i.e. internal equilibria. In an internal equilibrium

$$W_i = \gamma_i, \quad i = 1, 2. \tag{14}$$

With Lemma 3.1, this condition leads to

$$\frac{W_1}{W_2} = \frac{1 - \rho p_1}{1 - \rho p_2} = \frac{\gamma_1}{\gamma_2}. \tag{15}$$

Remark 3.5. Assume that $\gamma_1 \neq \gamma_2$. Then, from (15), a necessary condition for the existence of an internal equilibrium is that the class with higher net benefit to cost ratio obtains a lower priority parameter.

Theorem 3.6. *Suppose that* $\gamma_1 = \gamma_2 = \gamma$. *If* $p_1 = p_2 = \frac{1}{2}$ *then any pair* $q_1, q_2 \geqslant 0$ *such that* $q_1 + q_2 = 1 - 1/\gamma$ *defines an equilibrium. If* $p_1 \neq p_2$ *then there is a unique equilibrium which is a boundary one. In particular, if* $p_i > p_j$, *then* $q_i = 1 - 1/\gamma_i$ *and* $q_j = 0$.

Proof. By (15), when $\gamma_1 = \gamma_2$ an internal equilibrium exists only if $p_1 = p_2 = \frac{1}{2}$. With these priority parameters, the model, in terms of mean waiting time, is reduced to an $M/M/1$ queue without priority. Hence, $W_i = 1/(1-\rho)$, $i = 1, 2$. With the equilibrium condition (14), $\gamma = 1/(1-\rho)$, or $\rho = 1 - 1/\gamma$. □

Remark 3.5 says that if $\gamma_1 > \gamma_2$ then a necessary condition for an internal equilibrium is that $p_1 < \frac{1}{2}$. In fact, a sharper bound exists and is derived from (11), namely, a necessary condition for an internal equilibrium is that

$$p_1 < \frac{\gamma_1(\gamma_2 - 1)}{(\gamma_1 + \gamma_2)(\gamma_1 - 1)}. \tag{16}$$

From (15), in an internal equilibrium

$$\rho = \frac{\gamma_2 - \gamma_1}{\gamma_2 p_1 - \gamma_1 p_2}. \tag{17}$$

The equilibrium rates, q_1^* and q_2^* are determined by $W_1 = \gamma_1$, where W_1 is given in (5), and $q_1 + q_2 = \rho$, where ρ satisfies (17). Solving for q_1^* and q_2^* the result is

$$q_1^* = \frac{\gamma_2}{\gamma_2 p_1 - \gamma_1 p_2} - \frac{1}{\gamma_1 p_1 - \gamma_2 p_2} \tag{18}$$

and

$$q_2^* = \frac{1}{\gamma_1 p_1 - \gamma_2 p_2} - \frac{\gamma_1}{\gamma_2 p_1 - \gamma_1 p_2}. \tag{19}$$

The following theorem summarizes the results.

Theorem 3.7. *If* $\gamma_1 \neq \gamma_2$, *then for any* $p_1 \in [0, 1]$ *there exists a unique equilibrium. The corresponding joining rates are*:

1. *If* $0 \leqslant p_1 \leqslant A$ *then* $q_1^* = 0$ *and* $q_2^* = 1 - 1/\gamma_2$.
2. *If* $A < p_1 < B$ *then* q_1^* *and* q_2^* *are given by* (18) *and* (19).
3. *If* $1 \geqslant p_1 \geqslant B$ *then* $q_1^* = 1 - 1/\gamma_1$ *and* $q_2^* = 0$.

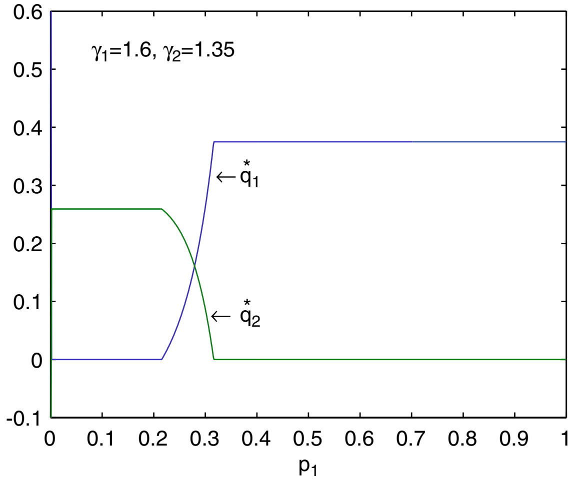

Fig. 1. Equilibrium arrival rates.

Remark 3.8. Theorem 3.7 partitions the unit interval, representing possible values for p_1, into three (not necessarily non-empty) subintervals according to the type of equilibrium. Depending on the parameters γ_i, $i = 1, 2$, one or even two of these subintervals may be empty.

Fig. 1 illustrates the equilibrium arrival rates when $\gamma_1 = 1.6$ and $\gamma_2 = 1.35$. Note that with these values $A = 0.2155$ and $B = 0.3164$.

Remark 3.9. The assumption of large potential arrival rates can be somewhat relaxed now. We only need to assume that the potential arrival rates are larger than the values of the equilibrium arrival rates specified in Theorem 3.7.

Remark 3.10. It is well known, see [2], [3] or [8, pp. 56–57], that in a FCFS system without class discrimination, a *class dominance* phenomenon prevails. Specifically, in case of high potential arrival rates, there exists a unique equilibrium in which customers only from a single class join the queue. As we have seen, the introduction of relative priorities may lead to an equilibrium in which customers from more than one class join the queue.

Remark 3.11. The analysis in [10] is quite similar. In particular, a distinction between boundary and internal equilibria is made. Yet, an explicit derivation for the optimal p_1 (given T_i, $i=1,2$) seems harder.

4. Profit maximizing priority parameters

We now consider the design of profit maximizing priority parameters when the other parameters, in particular entry fees, are given. As shown in Section 3 the relevant input parameters required are γ_1 and γ_2. The problem is to select $p_1 \in [0,1]$ such that $q_1^* T_1 + q_2^* T_2$ is maximized.

We maximize profits over the non-empty interval $\max\{0, A\} \leqslant p_1 \leqslant \min\{B, 1\}$. At the ends of the interval, one class actually has absolute priority over the other. For example, suppose that $p_1 = \max\{A, 0\}$. If $0 < A < 1$ then we are in the first case of Theorem 3.7, where $q_1^* = 0$ and $q_2^* = 1 - 1/\gamma_2$. If $A < 0$ then the 2-customers enjoy absolute priority, but some 1-customers join too. A similar situation occurs with the two classes swapping roles when $p_1 = \min\{B, 1\}$.

By (18) and (19), the expected rate of profit is

$$\begin{aligned}\Pi &= \left[\frac{\gamma_2}{\gamma_2 p_1 - \gamma_1 p_2} - \frac{1}{\gamma_1 p_1 - \gamma_2 p_2}\right] T_1 \\ &\quad + \left[\frac{1}{\gamma_1 p_1 - \gamma_2 p_2} - \frac{\gamma_1}{\gamma_2 p_1 - \gamma_1 p_2}\right] T_2 \\ &= \left[\frac{\gamma_2}{p_1(\gamma_1+\gamma_2) - \gamma_1} - \frac{1}{p_1(\gamma_1+\gamma_2) - \gamma_2}\right] T_1 \\ &\quad + \left[\frac{1}{p_1(\gamma_1+\gamma_2) - \gamma_2} - \frac{\gamma_1}{p_1(\gamma_1+\gamma_2) - \gamma_1}\right] T_2 \\ &= \left[\frac{\gamma_2}{x-\gamma_1} - \frac{1}{x-\gamma_2}\right] T_1 + \left[\frac{1}{x-\gamma_2} - \frac{\gamma_1}{x-\gamma_1}\right] T_2 \\ &= \frac{D}{x-\gamma_1} - \frac{\Delta}{x-\gamma_2}, \qquad (20)\end{aligned}$$

where $x = p_1(\gamma_1+\gamma_2)$, $D = \gamma_2 T_1 - \gamma_1 T_2$, and $\Delta = T_1 - T_2$. This function needs to be optimized over $p_1 \in [\max\{0, A\}, \min\{B, 1\}]$.

The first-order optimality condition in the case of an internal solution is

$$\frac{d\Pi}{dx} = -\frac{D}{(x-\gamma_1)^2} + \frac{\Delta}{(x-\gamma_2)^2} = 0$$

or

$$x^2(D-\Delta) - 2x(D\gamma_2 - \Delta\gamma_1) + (D\gamma_2^2 - \Delta\gamma_1^2) = 0.$$

The roots x^+ and x^- of this quadratic equation, which exist if $D\Delta \geqslant 0$, are candidates for an optimal value of x (and hence $x/(\gamma_1+\gamma_2)$ is candidate for an optimal value of p_1)

$$x^{\pm} = \frac{D\gamma_2 - \Delta\gamma_1 \pm \sqrt{D\Delta}(\gamma_2-\gamma_1)}{(D-\Delta)}. \qquad (21)$$

We now check which of these solutions corresponds to a local maximum of Π

$$\frac{d^2\Pi}{dx^2} = \frac{2D}{(x-\gamma_1)^3} - \frac{2\Delta}{(x-\gamma_2)^3}.$$

Note that $x^{\pm} - \gamma_1 = [(\gamma_2-\gamma_1)/(D-\Delta)](D \pm \sqrt{D\Delta})$ and $x^{\pm} - \gamma_2 = (\gamma_2-\gamma_1)/(D-\Delta)(\Delta \pm \sqrt{D\Delta})$. Also, $D(x^{\pm}-\gamma_2)^3 - \Delta(x^{\pm}-\gamma_1)^3 = ((\gamma_1-\gamma_2)/(D-\Delta))^3 D\Delta(D-\Delta)(\Delta^{1/2} \pm D^{1/2})^2$. Substituting these equations, we obtain that the sign of $d^2\Pi/dx^2$ is as that of $(\gamma_1-\gamma_2)(\Delta \pm \sqrt{D\Delta})(D \pm \sqrt{D\Delta})$. Note that $(\Delta - \sqrt{D\Delta})(D - \sqrt{D\Delta}) = 2\sqrt{D\Delta}(\sqrt{D\Delta} - ((D+\Delta)/2)) < 0$. We conclude that: If $\gamma_1 > \gamma_2$ then only x^- is a local maximizer of Π (and x^+ gives a local minimum). If $\gamma_1 < \gamma_2$ then only x^+ gives a localmaximum.

If the candidate solution is internal to $[\max\{0, A\}, \min\{B, 1\}]$ then it is optimal. Otherwise, the optimal value is the appropriate edge of this interval. The following summarizes it all.

Theorem 4.1. *Denote the optimal value by p_1^*. If $0 \leqslant A < p_1^* < B \leqslant 1$ then the corresponding equilibrium is internal with rates given in* (18) *and* (19). *Otherwise, if $p_1^* = \max\{0, A\}$, there are two possibilities*:

1. $A \geqslant 0$ (*and* $p_1^* \geqslant 0$): *The resulting equilibrium is a boundary one with $q_1^* = 0$ and $q_2^* = 1 - 1/\gamma_2$. Also, any $p_1 \in [0, A]$ is optimal too.*
2. $A < 0$ (*and* $p_1^* = 0$): *The resulting equilibrium is internal with $q_1^* = 1/\gamma_2 - \gamma_2/\gamma_1$ and $q_2^* = 1 - 1/\gamma_2$.*

The case where $p_1^* = \min\{B, 1\}$ is dealt similarly.

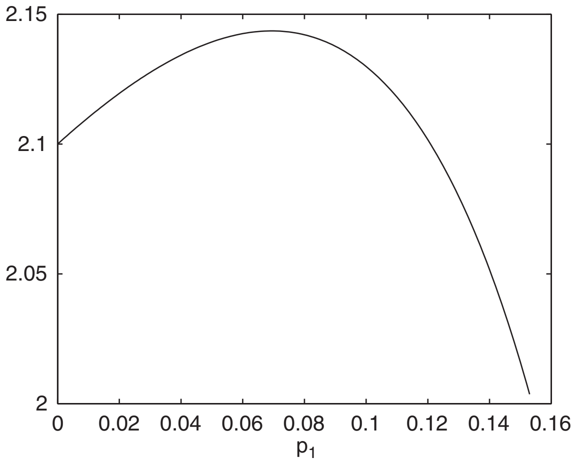

Fig. 2. Profits as a function of p_1 in the interval $[A, B]$.

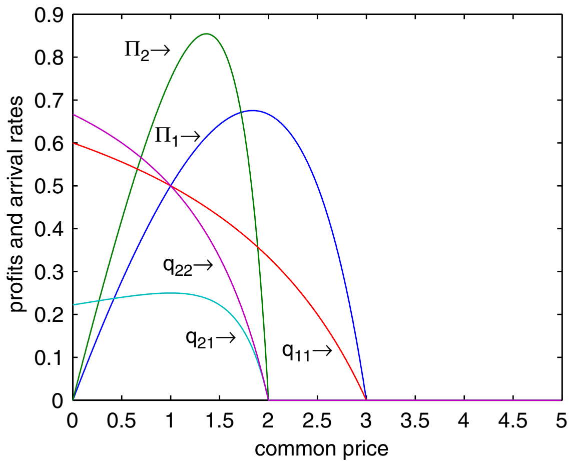

Fig. 3. Profits and joining rates as a function of T.

Fig. 2 shows an instance where for given fees ($T_1 = 4$, $T_2 = 7$) and net benefit to cost ratios ($\gamma_1 = 2$, $\gamma_2 = 1.25$), the optimal value of p_1 is 0.694, and it is strictly within the interval $A = 0$, $B = 0.1538$.

Remark 4.2. The paper [10] contains a numerical example showing that in their model it is possible that the optimal p_1 is strictly between zero and one.

5. Profit maximization with a common entry fee

We now assume that the entry fees are chosen by the server but they are restricted to be equal (an interesting example is when both must be 0). Our main finding is that giving absolute priority to one of the classes is optimal under the optimal common entry. Moreover, the class obtaining this absolute priority is not necessarily the one prescribed by the $C\mu$ rule.

Theorem 5.1. *Under the optimal common entry fee one of the classes obtains absolute priority.*

Proof. By (20) with $T_1 = T_2 = T$ and $p_1 \in [\max\{0, A\}, \min\{1, B\}]$, the profit rate is

$$\Pi = -\frac{(\gamma_2 - \gamma_1)T}{p_1(\gamma_1 + \gamma_2) - \gamma_1}.$$

For any given value of T, the objective of the queue manager is to maximize Π over $p_1 \in [\max\{0, A\}, \min\{1, B\}]$. The (partial) derivative of Π with respect to p_1 is proportional to $(\gamma_2 - \gamma_1)$. Therefore, if $\gamma_2 > \gamma_1$, the optimal p_1 is in the upper end of the interval, and this is equivalent to giving absolute priority to 1-customers (though 2-customers may still arrive). Similarly, if $\gamma_2 < \gamma_1$ then $p_1 = 0$ is optimal. □

The optimal common price, with absolute priorities, is computed as follows. Let q_{ij} denote the joining rate of j-customers when class i obtains absolute priority. Then $q_{ii} = 1 - C_i/(R_i - T)$ and $q_{ij} = 1 - q_{ii} - C_j/((R_j - T)(1 - q_{ii}))$, as long as these values are in [0,1]. A value not in this interval should be set to 0. The profit Π_i which is defined as profit associated with the price T, $p_i = 1$ and $p_j = 0$ is $\Pi_i = T(q_{i1} + q_{i2})$. We compute Π_1 and Π_2 and take the maximum.

For example, suppose that $R_1 = 5$, $R_2 = 3$, $C_1 = 2$, and $C_2 = 1$. Fig. 3 shows the profits and joining rates. In this example, $q_{12} = 0$ for every value of T. The maximum profit is obtained for $T \approx 1.4$ and $p_1 = 0$. We observe that the $C\mu$ rule dictates $p_1 = 1$.

We conclude with some observations that can be deduced from Fig. 3.

With $p_1 = 1$ 1-customers join as long as $q_{11} = 1 - C_1/(R_1 - T) > 0$, or $T < 3$. The joining of 1-customers discourages all 2-customers from joining, and hence $q_{12} = 0$ for any $T \geqslant 0$. The profit is as in a single class model, and it is maximized at $T = R_1 - \sqrt{C_1 R_1} = 1.84$.

With $p_1 = 0$, 2-customers join as long as $q_{22} = 1 - C_2/(R_2 - T) > 0$ or $T < 2$. 1-customers still join, and it is interesting to observe that their joining rate initially increases with T. The reason is that 2-customers are more sensitive to changes in T, having a smaller value of R. Their joining rate q_{22} steeply decreases and the reduced expected waiting time for 1-customers caused by this more than compensates for the increase in T.

The maximum profit is obtained with $p_1 = 0$ and $T \approx 1.4$. We observe that this outcome differs from the $C\mu$ rule which dictates $p_1 = 1$.

Acknowledgements

This research was supported by The Israel Science Foundation Grant no. 237/02.

References

[1] E. Altman, T. Jimenez, D. Kofman, DPS queues with stationary ergodic service times and performance of TCD in overloads, in: Proceedings of INFOCOM, Hong-Kong, 2004.

[2] K.R. Balachandran, M.E. Schaefer, Class dominance characteristics at a service facility, Econometrica 47 (1979) 515–519.

[3] K.R. Balachandran, M.E. Schaefer, Regulation by price of arrivals to a congested facility, Cah. C.E.R.O 21 (1979) 149–158.

[4] T. Boland, L. Massoulie, Impact of fairness in Internet performance, in: Sigmetrics, Cambridge, MA, 2001.

[5] N.M. Edelson, K. Hildebrand, Congestion tolls for Poisson queueing processes, Econometrica 43 (1975) 81–92.

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

[7] F.R. Fernández, R. Hassin, J. Puerto, The use of relative priorities in optimizing the performance of a queueing system, 2004 manuscript.

[8] R. Hassin, M. Haviv, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Kluwer International Series, 2003.

[9] M. Haviv, J. van der Wal, Waiting times in queues and processor sharing systems with relative priorities, 2004 manuscript.

[10] Y. Hayel, B. Tuffin, Pricing for heterogeneous services at a discriminatory processor sharing queue, IRISA 1633 Research Report, http://www.irisa.fr/bibli/pi/pi04.html, 2004.

[11] L. Kleinrock, Queueing Systems, Volume 2: Computer Applications, Wiley Interscience, New York, 1976.

[12] H. Mendelson, S. Whang, Optimal incentive-compatible priority pricing for the $M/M/1$ queue, Oper. Res. 38 (1990) 870–883.

[13] K.M. Rege, A decomposition theorem and related results for the discriminatory processor sharing queue, Queueing Syst. 18 (1994) 333–351.