STABLE PRIORITY PURCHASING IN QUEUES

Amihai GLAZER
School of Social Sciences, University of California, Irvine, CA 92717, USA

Refael HASSIN
Statistics Department, Tel-Aviv University, Tel-Aviv 69978, Israel

Received May 1985
Revised October 1985

We consider an $M/G/1$ queueing system where each customer can purchase the priority with which he will be served. Customers may differ in their time valuation. They know the statistical distrubution of the queue length and amounts paid by others, but not their actual values. We determine the payment policy from which no customer will deviate as long as the others use it, and compare it to the first-come first-served discipline.

priority queues * priority purchasing

1. Introduction

Two types of queueing models with priority models have been extensively studied. In the first type, a customer's relative priority is completely out of his control. In the second type, to which the present study belongs, customers purchase the right to be accepted to a priority group: this allows customers with high time valuations to be served first. There are here two ways to elicit information about customers' waiting costs. The queue organizer can set prices for different priority classes, so that a customer can choose which class to join. Dolan [3], Ghanem [4], and Marchand [6] show, under various assumptions, that there exist sets of prices that will cause each customer to join the same priority class that would have been assigned him by a socially optimal rule.

The other approach, which is the one we use, allows each customer to choose the price he will pay, and assigns him a priority class on this basis; the bigger the payment, the higher the priority. Each customer is assumed to minimize his private costs, consisting of his payment and of his expected waiting cost.

Balachandran [1] (see also Balachandran and Lukens [2]) investigates this setting. He assumes that all customers are identical, and that each arriving customer knows the amounts paid by all others in the queue. He then proves that the *globally optimal* policy, the one which minimizes the average cost per customer, requires that no payments be made, so that the order of service is first-come first-served (FCFS). However, Balachandran also shows that this policy is *not stable* in the sense that it pays an individual customer to deviate from it if everyone else uses it.

In this paper we assume, more realistically, that different customers have different valuations of time, and that a customer knows only the statistical distributions of the queue length and of the amounts paid by others. For this case, Kleinrock [5] shows that every policy for which a customer's payment is a strictly increasing function of his waiting cost minimizes the customer's average waiting costs. Thus, if one considers the amounts paid by customers as transfer payments that do not affect aggregate social welfare, any such policy is globally optimal. This result holds both for the preemptive and the non-preemptive queueing model.

We consider, in contrast, a customer's total costs, including his payments and his waiting cost. We find that the stable policy obtained under our

0167-6377/86/$3.50

assumptions is characterized by such a strictly increasing payment function. The system therefore leads customers to be served in the efficient order. We then compute the unique stable policy in the case of $M/G/1$ queues and determine which customers will benefit from the institution of this payment system, relative to the FCFS discipline.

2. Stable priority purchasing

Each customer is characterized by his *time valuation*, α, that defines the cost per unit time of waiting in queue (it is immaterial to the discussion whether this cost is also incurred during the service time). We denote $P(\alpha)$ = probability that an entering customer has a time valuation $\alpha' \le \alpha$.

A *policy* π is an assignment of a value x to each entering customer, denoting the prices he pays for priority. The policy is *stable* if no individual customer will wish to deviate from it if everyone else uses it.

$B_\pi(x)$ = probability that an entering customer pays a price $x' \le x$ under policy π.

The entering customer is placed in the queue according to the price he paid; the higher the price the closer to the queue's head. Customers who paid identical prices are served according to an FCFS policy. The discussion will apply both to preemptive and non-preemptive systems. A customer's goal is to minimize his expected cost, $x + \alpha w_\pi(x)$, where $w_\pi(x)$ is the expected wait of a customer who paid x under policy π. When making the payment, the customer knows neither the queue length nor the prices paid by others.

Lemma 1. *If π is stable, then b_π is continuous and strictly increasing in the payment, x.*

Proof. Suppose that B_π has a jump at x_0. Then there exists a probability $p > 0$ that a randomly chosen costumer will pay exactly x_0. If such a customer pays $x_0 + \mathrm{d}x$ instead, he will reduce his expected wait by a non-infinitesimal length, since he obtains service before a customer who paid exactly x_0. Thus π would not be stable.

Suppose now that B_π is constant in $[x_1, x_2]$. A customer who pays x_2 can reduce his payment to x_1 without affecting his expected waiting time, so that again π would not be stable. □

Corollary 1. *If π is stable, then the functions $-w_\pi(x)$ and $c_\pi(\alpha) \equiv \min_x \{x + \alpha w_\pi(x)\}$ are continuous and strictly increasing.*

Corollary 2. *Let x_1 be the amount paid by a customer with time valuation α_1, and x_2 the amount paid by a customer with time valuation α_1. If $\alpha_1 > \alpha_2$ and the policy is stable, then $x_1 > x_2$.*

Corollary 3. *Suppose that $P(\alpha)$ is continuous on (a, b). If π is stable, then there exists a continuous function $x_\pi(\alpha)$ defining the payment made by customers with time valuations $\alpha \in (a, b)$.*

Proof. Were there a discontinuity in the payment level at α_0, a customer with $\alpha = \alpha_0 + \mathrm{d}\alpha$ could reduce his payment by a non-infinitesimal amount while increasing his expected wait by an infinitesimal amount, so that π would not be a stable policy.

We now turn to characterizing a stable policy in $M/G/1$ queues. We consider a system characterized by Poisson arrivals with mean λ, and a general service process with a cummulative distribution $F(\ )$ and mean $1/\mu$. For this system without preemption Kleinrock proves that

$$w(x) = w_0/[1 - \rho + \rho B(x)]^2, \tag{1}$$

where $\rho = \lambda/\mu$, and $w_0 = (\lambda/2)\int_0^\infty t^2 \mathrm{d}F(t)$ is the expected duration of the residual service time. In a preemptive system, w_0 is replaced in (1) by $1/\mu$ so that the following discussion can be modified to such a system by exchanging these terms.

To simplify the exposition, we at times consider functions which may not be differentiable everywhere; that is, opposite inequalities hold for the left and right derivatives.

Theorem 1. *Suppose $P(\ )$ is continuous on (a, b). If π is stable, then*

$$x(\alpha) = \int_0^\alpha \frac{2\rho w_0 y}{[1 - \rho + \rho P(y)]^3} \mathrm{d}P(y), \quad \alpha \in (a, b).$$

Proof. A customer with time valuation α wishes to minimize $x + \alpha w_\pi(x)$. The first order condition is

$$-1/\alpha = w_\pi'(x(\alpha)). \tag{2}$$

From Corollaries 2 and 3, $B_\pi(x(\alpha)) = P(\alpha)$ for $\alpha \in (a,b)$; together with (1) and (2) it follows that

$$\frac{1}{\alpha} = \frac{2w_0\rho B_\pi'(x(\alpha))}{[1 - \rho + \rho B_\pi(x(\alpha))]^3} = \frac{2w_0\rho P'(\alpha)/x'(\alpha)}{[1 - \rho + \rho P(\alpha)]^3},$$

or that

$$x^{1}(\alpha) = \frac{2\rho w_0 \alpha P'(\alpha)}{[1-\rho+\rho P(\alpha)]^3}.$$

The theorem follows from this expression. □

Suppose now that, contrary to the assumption of Theorem 1, $P(\ )$ has a discontinuity point at α. Now there is a positive probability, $p = P(\alpha) - P(\alpha-)$, that a randomly chosen customer has time valuation α. Stability requires that these customers spread their payments over an interval $(x_\pi(\alpha-),\ x_\pi(\alpha+))$: if instead they all paid the same amount x, a customer could reduce his expected wait by increasing his payment to $x + \mathrm{d}x$. However, for any value of x in this interval, the total expected cost, $x + \alpha w_\pi(x)$, must be constant, since otherwise those who chose an x which entails higher expected cost have an incentive to change their choice. Thus we have the following theorem.

Theorem 2. *Suppose $P'(\ )$ has a jump at α. If π is stable, then $x_\pi(\alpha) \in (x_\pi(\alpha-), x_\pi(\alpha+))$; the distribution of payments, $B_\pi(x)$, in this interval satisfies*

$$x + \frac{\alpha w_0}{[1-\rho+\rho B_\pi(x)]^2} = x_\pi(\alpha-) + \frac{\alpha w_0}{[1-\rho+\rho P(\alpha-)]^2} = x_\pi(\alpha+) + \frac{\alpha w_0}{[1-\rho+\rho P(\alpha)]^2}.$$

3. Customer welfare

We let a variable with an F subscript denote its value under an FCFS discipline, and a variable with an S subscript denote its value under a stable payment system. We denote by $C(\alpha)$ the expected cost for a customer with time valuation α. Theorem 2 assures that $C_S(\alpha)$ is also defined when α is a discontinuity point of $P(\ )$. For the other points we have $C_S(\alpha) = x(\alpha) + \alpha w(x(\alpha))$. Differentiate with respect to α and substitute (2) to obtain

$$C_S'(\alpha) = w(x(\alpha)). \tag{3}$$

Differentiate again to find that $C_S''(\alpha) = w'(x)x'(\alpha) < 0$, so that C_S is concave. Clearly C_F is linear. Customers with very low time valuation will pay very little; their expected wait will be larger under a payment system and $C_S(\alpha) > C_F(\alpha)$ for these customers.

Altogether we conclude that one of the following alternatives must hold:

(i) $C_S(\alpha) \geq C_F(\alpha)$ for all possible values of α.

(ii) The two curves intersect once, and $C_S(\alpha) < C_F(\alpha)$ when $\alpha > \alpha_0$ for some critical value α_0.

Thus a payment system is certain to decrease the welfare of some customers. It may increase the welfare of a section of the population with high valuation of time.

This section considers several examples to demonstrate the results.

3.1. Identical customers

Suppose all customers share a common time valuation, α. The payments in the population will be distributed on $[0,\ a]$ for some $a > 0$ that will be specified later. For each $x \in [0,\ a]$ the cost $x + \alpha w(x)$ must be constant. Substitute $x = 0$ to find

$$x + \alpha w(x) = x + \frac{\alpha w_0}{[1-\rho+\rho B(x)]^2} = \frac{\alpha w_0}{(1-\rho)^2}, \quad x \in [0,\ a], \tag{4}$$

so that

$$B(x) = \frac{1}{\rho}\left\{\left[\frac{1}{(1-\rho)^2} - \frac{x}{\alpha w_0}\right]^{-\frac{1}{2}} - (1-\rho)\right\}, \quad x \in [0,\ a].$$

From the condition that $B(a) = 1$ it follows that $a = \alpha w_0[(1-\rho)^{-2} - 1]$, and from (4) that

$$C_S = \frac{\alpha W_0}{(1-\rho)^2} > \frac{\alpha w_0}{1-\rho} = C_F.$$

All customers are worse off under the payment system. This is a parallel result to the one obtained by Balachandran [1] under his assumptions, and could be expected since for identical customers the service order is immaterial.

3.2. Two types of customers

We suppose now that a proportion p of the population has $\alpha = \alpha_1$ and that a proportion $(1-$

p) has $\alpha = \alpha_2$, with $\alpha_1 < \alpha_2$. Customers with α_1 will then choose payments in [0, a], and the others choose payments in $[a, b]$ for $0 < a < b$ that will be specified below.

For customers with $\alpha = \alpha_1$, any $x \in [0, a]$ must yield identical costs. In particular, if $x = 0$,

$$x + \alpha_1 w_1(x) = x + \frac{\alpha_1 w_0}{[1 - \rho + \rho B(x)]^2} = \frac{\alpha_1 w_0}{(1-\rho)^2} \equiv C_1,$$

$$x \in [0, a]. \tag{5}$$

Since $B(a) = p$ we have

$$a = \alpha_1 w_0 \left[\frac{1}{(1-\rho)^2} - \frac{1}{(1-\rho+\rho p)^2} \right]. \tag{6}$$

For customers with $\alpha = \alpha_2$ any $x \in [a, b]$ must yield identical costs: in particular if $x = a$, then

$$x + \frac{\alpha_2 w_0(x)}{[1 - \rho + \rho B(x)]^2} = a + \frac{\alpha_2 w_0}{(1-\rho+\rho p)^2} \equiv C_2,$$

$$x \in [0, \alpha]. \tag{7}$$

Since $B(b) = 1$ we have

$$b = a + \alpha_2 w_0 \left[\frac{1}{(1-\rho+\rho p)^2} - 1 \right]. \tag{8}$$

The distribution $B(x)$ can be extracted from (5), (6), (7) and (8). The mean cost in the population is

$$\bar{C}_S = pC_1 + (1-p)C_2 = \frac{p\alpha_1 w_0}{(1-\rho)^2} + (1-p)\left[a + \frac{\alpha_2 w_0}{(1-\rho+\rho p)^2} \right],$$

while under FCFS

$$\bar{C}_F = (p\alpha_1 + (1+p)\alpha_2) w_0/(1-\rho).$$

We demonstrate next that if the majority of the population has a very low time valuation relative to the minority, then a payment system not only optimizes social welfare, but also lowers the mean cost per customer (including his payment). Fixing p and ρ, and increasing the ratio α_2/α_1, the dominating term in $\bar{C}_F - \bar{C}_S$ becomes

$$(1-p)\alpha_2 w_0 \left[\frac{1}{1-\rho} - \frac{1}{(1-\rho+\rho p)^2} \right].$$

If $p \geq \frac{1}{2}$, then $(1 - \rho + \rho p)^2 \geq (1 - \rho/2)^2 > (1 - \rho)$. Hence in this case $\bar{C}_F > \bar{C}_S$.

3.3. Exponential distribution of α

Suppose now that α is exponentially distributed. A customer's average total cost is obtained by using (3),

$$\bar{C}_S = \int_{\alpha=0}^{\infty} \left[\int_{\beta=0}^{\alpha} w(x(\beta))\, d\beta \right] dP(\alpha) = \int_{\beta=0}^{\infty} w(x(\beta)) \left[\int_{\alpha=\beta}^{\infty} dP(\alpha) \right] d\beta = \int_{\beta=0}^{\infty} w(x(\beta))[1 - P(\beta)]\, d\beta.$$

For the exponential distribution, $1 - P(\alpha) = \bar{\alpha} p(\alpha)$, hence

$$\bar{C}_S = \bar{\alpha} \int_0^{\infty} w(x(\beta))\, dP(\beta) = \bar{\alpha}\bar{w} = \bar{C}_F,$$

where the right equality holds since in FCFS the waiting time is independent of α. Thus, in this case the mean value of customers' costs is not affected by the institution of a payment system.

3.4. Uniform distribution of α

Suppose that α is uniformly distributed on [0, 1]. Then from (1) and (3),

$$C_S(\alpha) = \int_0^{\alpha} \frac{w_0}{(1-\rho+\rho y)^2} dy = \frac{w_0 \alpha}{[(1-\rho+\alpha\rho)(1-\rho)]^2} > \frac{w_0 \alpha}{1-\rho} = C_F(\alpha),$$

so that each customer is worse off under the payment system.

References

[1] K.R. Balachandran, "Purchasing priorities in queues", *Management Science* **18** (1), 319–326 (1972).

[2] K.R. Balachandran and J.C. Lukens "Stable pricing policies in service systems", *Zeitschrift für Operations Research* **20**, 189–201 (1976).

[3] R.J. Dolan, "Incentive mechanisms for priority queuing problems", *Bell Journal of Economics* **9**, 421–436 (1978).

[4] S.B. Ghanem, "Computer center optimization by a pricing priority policy", *IBM Systems Journal* **14**, 272–292 (1975).

[5] L. Kleinrock, "Optimal bribing for queue position", *Operations Research* **15**, 304–318 (1967).

[6] M.G. Marchand, "Priority pricing", *Management Science* **20**, 1131–1140 (1974).