Discussion Paper No. 901

OPTIMAL SERVICE SPEEDS IN A COMPETITIVE ENVIRONMENT

by

Ehud Kalai*
Morton I. Kamien*

and

Michael Rubinovitch**

September 1990***

---

*Department of Managerial Economics and Decisions Sciences, J.L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road Evanston, IL 60208, U.S.A.

**Faculty of Industrial Engineering and Management Technion - Israel Institute of Technology, Haifa 32000, Israel.

***Kalai's research was partialy supported by NSF Economics Grant No. SES7911790; Kamien's research was partially supported by Heizer Entrepreneurship Center; Rubinovitch's research was supported by The Fund for Promotion of Research at the Technion.

Abstract

This is a study of the economic behavior of vendors of service in competition. A simple model with two competing exponential servers and Poisson arrivals is considered. Each server is free to choose his own service rate at a cost (per time unit) that is strictly convex and increasing. There is a fixed reward to a server for each customer that he serves. The model is designed to study one specific aspect of competition. Namely, competition in speed of service as a means for capturing a larger market share in order to maximize long run expected profit per time unit. A two person strategic game is formulated and its solutions are characterized. Depending on the revenue per customer served and on the cost of maintaining service rates, the following three situations may arise. (i) a unique symmetric strategic (Nash) equilibrium in which expected waiting time is infinite; (ii) a unique symmetric strategic equilibrium in which expected waiting time is finite; and (iii) several, non symmetric strategic equilibria with infinite expected waiting time. An explicit expression for the market share of each server as a function of the service rates of the two servers is also given.

1 Introduction

Our objective is to study the consequences of competition on service speed. We consider a population of potential customers each of which may, at any time, need service. There are two firms (servers) in the market able to provide it. Service times are random and each firm is free to choose independently its mean service time. The optimal choice of mean service time, in the presence of a competitor, is the subject addressed.

Problems of finding optimal (static) policies for choosing service rates are not new. They fall within the general area of optimal design and control of queueing systems and are classified as optimal design problems (Crabill, Gross and Magazine [2]). For a more recent review of the subject see Teghem [6]. In these problems there is usually a choice among several options regarding the structure of the system, given data on the service environment. The choice could be between different service rates, number of servers, queue disciplines, different rules for allocating customers to servers etc. Under a specified cost structure the designer has to choose, in an optimal manner, among the available options.

However, in studying the economics of service facilities, issues of incentives and competition also enter (customer's incentives to join a queue and competition among servers). Customer's incentives and their implications on social welfare were studied by Naor and his followers (see Bell and Stidham [1]). Here we study the role of competitive incentives on serves. We do so by constructing a simple queuing and cost model in which the competitive element is analyzed in the case of two servers [1].

Specifically, the system we posit behaves as follows. When a new request for service (a new job) arrives and both servers are free, the job is dispatched to one of them at random. If only one server is free the job is assigned to him and if no server is free the job enters a pool (queue) of jobs awaiting service.

Under these assumptions it is clear that the server with the shorter mean service time realizes a larger market share. The mean number of jobs (per time unit) that he serves exceeds that of a server with a longer mean service time. Consequently, if each job makes the same payment to his server, the faster server realizes larger revenues.

On the other hand, faster service rates (short mean service times) are usually associated with higher operating costs. Therefore, each server, seeking to maximize his profits, faces a tradeoff between revenues and operating costs.

[1] Lode Li [3] was the first to address the issue of competition in the (not unrelated) context of Inventory Theory.

For a single firm operating alone in the market, and under suitable assumptions on costs and revenues, this trade off leads to a simple optimization problem. In fact, it will be shown that its solution is a mean service time that is never below the mean inter-arrival time. This is intuitively obvious since a monopolist whose service rate is the same as the arrival rate is sure to capture the entire available market. Any service rate exceeding the arrival rate will not increase the number of customers that he captures but will incur additional costs. Unfortunately, at the monopolist's service rate, the expected waiting time is infinite, but this is of no concern to him since sooner or later he serves as many customers as he desires. Since we are assuming no discounting of future profits, the server will be able to maximize profits with a service rate that does not exceed the arrival rate.

The situation is quite different in the presence of a competitor since then, a job left waiting may be lost to the competitor. The determination of an optimal service rate is more complicated since a server's profit depends now not only on his own rate but also on the competitor's rate. Thus, we can view this situation as a two-person (noncooperative) strategic game between the servers. Each seeks to maximize expected profits by strategically choosing his service rate. Our objective is to solve this game, i.e., to characterize its strategic (Nash) equilibria. A strategic equilibrium consists of a pair of service rates having the property that each is optimal for the server who chooses it, given that the opponent chooses the other rate (see for example Owen [4]).

The environment described above, and the mathematical problem it poses, are designed to focus on one aspect of competition between servers, namely, speed of service as a means to capturing a larger market share. This aspect is present in many real competitive service environments. The areas of telecommunication and data transmission provide examples of situations where speed of service is an important competitive factor. For instance, long line telephone service in the United States after deregulation involves competition among several carriers. Large customers often subscribe to more than one carrier and use automatic controllers to dispatch telephone calls (or transmit data) to the server who can provide switching first. Thus the faster server realizes a larger market share.

The paper is organized as follows. In Section 2 we describe the queuing model, cost and reward functions and define the servers' strategic game. We also derive expressions for the market share that each server captures as a function of the service rates of the two competitors. Section 3 contains a complete characterization of the solutions of the service game. The main result is summarized below.

If the cost of providing service is high then there is a unique symmetric equilibrium. In it each server behaves as if he was a monopolist, competition has no effect and waiting times grow beyond any bound. Here by "high cost" we mean that the marginal

cost of providing service, at half the arrival rate, exceeds the reward from serving a customer.

There is also a unique equilibrium, which is symmetric, when the cost of providing service is low. Then, the sum of the equilibrium service rates is greater than the arrival rate and the resulting queuing system is stable. Expected waiting times, in steady state, are finite. Here by "low cost" we mean that the marginal cost of providing service, at half the arrival rate, is smaller than half the reward from serving a customer.

The remaining itermediary case is when the marginal cost of providing service, at a rate equal to half the arrival rate, is between half the reward from serving a customer and the reward itself. Here we have a multiplicity of solutions. Equilibrium service rates need not be the same but their sum must equal the arrival rate. Consequently the resulting queuing system is not stable and expected waiting time is again infinite. Note that only in this case may the market shares of the two competitors not be the same.

In Section 4 we describe the effect of changes in λ on the equilibrium service rates and discuss some possible generalizations of the present model.

2 The Mathematical Model

We consider a queuing system with two servers and unlimited waiting space. New jobs arrive according to a Poisson process of rate λ and service rates are μ_1 and μ_2, for Servers 1 and Server 2, respectively. When there are jobs waiting and a server becomes free, one of the jobs proceeds to that server. A job that arrives when both servers are idle is assigned randomly to a server. Finally, if a job arrives when one server is idle and the other is busy, it goes to the idle server. The following results about this system can be found in Rubinovitch [5].

When $\mu_1 + \mu_2 > \lambda$ the system has a steady state distribution. Define the following steady state probabilities.

P_n : the probability that there are n customers in the system;

P_{10} : the probability that server 1 is busy and server 2 is idle;

P_{01} : the probability that server 2 is busy and server 1 is idle.

Then,

$$P_0 = \frac{1-\rho}{1-\rho+\frac{\lambda(\mu_1+\mu_2)}{2\mu_1\mu_2}},$$

$$P_{10} = \frac{\lambda P_0}{2\mu_1}, \quad P_{01} = \frac{\lambda P_0}{2\mu_2},$$

where $\rho = \lambda/(\mu_1 + \mu_2)$ is the system's load. Also, $P_1 = P_{10} + P_{01}$ and for n = 2,3,...

$$P_n = \rho^{n-1} P_1.$$

Now we want to compute the expected value, in steady state, of the fraction of jobs served by server i. Let $\alpha_i(\mu_1, \mu_2)$ be this number. For this, consider first the mean number of jobs per time unit that are served by server i. This is the same as the mean number of jobs per time unit that enter service with this server which equals

$$\lambda\frac{P_0}{2} + \lambda P_{01} + \mu_1(P_3 + P_4 + \ldots).$$

Dividing by the expected number of arrivals per time unit (λ) and substituting according to the formulae above we obtain

$$\alpha_i(\mu_1, \mu_2) = \frac{\lambda\mu_i^2 + \mu_1\mu_2(\mu_1+\mu_2)}{\lambda(\mu_1+\mu_2)^2 + 2\mu_1\mu_2(\mu_1+\mu_2-\lambda)}.$$

This is the expression for the market share of each server as it depends on his service rate and his opponent's service rate.

Our cost and revenue assumptions are as follows. We assume that the cost of achieving service rate μ (regardless of actual use) is the same for both servers and equals $c(\mu)$ dollars per time unit. The function $c(\mu)$ is assumed to be nonnegative, increasing, strictly convex and continuously differentiable. Also, each server receives R dollars per job processed.

Next we want to compute each server's expected (long run) profit per time unit. When $\mu_1 + \mu_2 > \lambda$ our simple queuing system is stable and server i's long run expected profit per time unit is $R\lambda\alpha_i(\mu_1, \mu_2) - c(\mu_i)$. If $\mu_1 + \mu_2 \leq \lambda$ the queuing system is not stable and the number of jobs awaiting service eventually exceeds any bound with probability one. However, the mean number of customers per time unit that are served by server i does converge and, in the limit, as 'time' goes to infinity, equals his service rate. Thus, in this case, server i's expected profit per time unit is $R\mu_i - c(\mu_i)$. It follows that the long run expected profit per time unit is

$$\pi_i(\mu_1, \mu_2) = \begin{cases} R\lambda\alpha_i(\mu_1, \mu_2) - c(\mu_i) & \text{if } \mu_1 + \mu_2 > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \mu_1 + \mu_2 \leq \lambda. \end{cases}$$

As already mentioned, server i's average profit depends on his choice of service rate as well as the opponent's service rate choice. We wish to characterize the solutions of the game that this profit function define, i.e., to find the strategic (Nash) equilibria of the game. Specifically, we seek the pairs of service rates $(\bar{\mu}_1, \bar{\mu}_2)$ that satisfy

$$\begin{aligned} \pi_1(\bar{\mu}_1, \bar{\mu}_2) &\geq \pi_1(\mu_1, \bar{\mu}_2) \quad \text{for all} \quad \mu_1 \geq 0, \\ \pi_2(\bar{\mu}_1, \bar{\mu}_2) &\geq \pi_2(\bar{\mu}_1, \mu_2) \quad \text{for all} \quad \mu_2 \geq 0. \end{aligned}$$

This is done in the next section.

3 Solution of the Service Game

Before proceeding to the solution of the game let us look briefly at the monopolistic case in which there is only one server in the market. Here the optimization problem is straightforward. It is obvious that the optimal $\mu \leq \lambda$ since the server will not serve any more customers if he increases his rate above λ but will have to bear the cost. Formally let μ_0 be the unique solution of $c'(\mu) = R$. If $\mu_0 < \lambda$ then the optimal service rate is μ_0 and if $\mu_0 \geq \lambda$ then the optimal service rate is λ. Notice that this outcome may be socially undesirable, as customers' expected waiting time is infinite. We now proceed to the game.

Lemma 1 *For $\mu_1 \geq 0$, $\mu_2 > 0$, $\lambda > 0$ such that $\mu_1 + \mu_2 > \lambda$ we have*

$$(i) \qquad \frac{\partial \alpha_1(\mu_1, \mu_2)}{\partial \mu_1} = \frac{\lambda \mu_2^2 [(\mu_1 + \mu_2)^2 + 2\lambda \mu_1]}{[\lambda(\mu_1 + \mu_2)^2 + 2\mu_1 \mu_2(\mu_1 + \mu_2 - \lambda)]^2};$$

$$(ii) \qquad \frac{\partial^2 \alpha_1(\mu_1, \mu_2)}{\partial \mu_1^2} < 0.$$

Proof: The proof of (i) is straightforward and is omitted. For the proof of (ii) let $A(\mu_1)$ be the numerator of the right side of (i) and $B(\mu_1)$ be the term inside the square brackets of the denominator. So

$$\alpha_1(\mu_1, \mu_2) = \frac{A(\mu_1)}{[B(\mu_1)]^2}$$

and $B(\mu_1) > 0$ when $\mu_1 + \mu_2 < \lambda$. Hence,

$$\begin{aligned} \frac{\partial^2 \alpha_1(\mu_1, \mu_2)}{\partial \mu_1^2} &= \frac{A'(\mu_1)[B(\mu_1)]^2 - 2A(\mu_1)B(\mu_1)B'(\mu_1)}{[B(\mu_1)]^4} \\ &= \frac{A'(\mu_1)B(\mu_1) - 2A(\mu_1)B'(\mu_1)}{[B(\mu_1)]^3} \end{aligned}$$

$$= -\frac{2\lambda\mu_2^2[\mu_1^3(2\mu_2+\lambda)+3\mu_1^2(2\mu_2^2+3\lambda\mu_2+\lambda^2)+\mu_2^2(2\mu_2+\lambda)(3\mu_1+\mu_2-\lambda)]}{[\lambda(\mu_1+\mu_2)^2+2\mu_1\mu_2(\mu_1+\mu_2-\lambda)]^3}.$$

For each $\lambda > 0$ and $\mu_2 > 0$ the numerator is positive at least for $\mu_1 > (\lambda - \mu_2)/3$ and the denominator is positive at least for $\mu_1 > \lambda - \mu_2$. Since the latter is a subset of the former, (ii) follows.

Lemma 2 *For each fixed $\lambda > 0$ and $\mu_2 > 0$ the function $\pi_1(\mu_1, \mu_2)$ is continuous and strictly concave in μ_1.*

Proof: Let $\lambda > 0$ be given. When $\mu_2 > \lambda$ then $\pi_1(\mu_1, \mu_2)$ has only its upper branch and continuity follows since $c(\mu)$ and $\alpha_1(\mu_1, \mu_2)$ are both continuous. Concavity follows from the convexity of $c(\mu)$ and from (ii) of Lemma 1. Next, suppose that $0 < \mu_2 < \lambda$. Here it is enough to show that gross profit

$$\bar{\pi}_1(\mu_1, \mu_2) = \begin{cases} R\lambda\alpha_1(\mu_1, \mu_2) & \text{if } \mu_1 > \lambda - \mu_2 \\ R\mu_1 & \text{if } \mu_1 \leq \lambda - \mu_2 \end{cases}$$

is continuous and concave in μ_1. Continuity is verifiable by checking that

$$\lambda\alpha_1(\lambda - \mu_2, \mu_2) = \lambda - \mu_2.$$

For concavity it is enough to show that the derivative from the right,

$$\left.\frac{\lambda\partial\alpha_1(\mu_1+, \mu_2)}{\partial\mu_1}\right|_{\mu_1=\lambda-\mu_2} = \lim_{\mu_1 \downarrow \lambda-\mu_2} \frac{\lambda\partial\alpha_1(\mu_1, \mu_2)}{\partial\mu_1} \leq 1.$$

But from (i) of Lemma 1,

$$\left.\frac{\lambda\partial_1(\mu_1+, \mu_2)}{\partial\mu_1}\right|_{\mu_1=\lambda-\mu_2} = \frac{\mu_2^2(3\lambda - 2\mu_2)}{\lambda^3} \equiv f(\mu_2)$$

and by differentiation it can be checked that $f(\mu_2)$ is increasing in $[0, \lambda]$ and attains its maximal value of 1 when $\mu_2 = \lambda$. Hence, the desired result.

Theorem: The following characterize the strategic equilibria of the service game.

Case a: When $c'(\lambda/2) > R$ (the marginal cost of serving one-half the customers exceeds the revenue per customer) let μ be the unique service rate satisfying $c'(\mu) = R$. Then (μ, μ) is the unique strategic equilibrium of the game. As $\mu < \lambda/2$, the long run expected waiting time is infinite and the pool of waiting customers increases beyond any bound with probability one.

Case b: When $c'(\lambda/2) < R/2$ (the marginal cost of serving one-half the customers is less than one-half the revenue per customer) let μ be the unique service rate satisfying

$$c'(\mu) = \frac{R\lambda^2}{2\mu(2\mu+\lambda)}.$$

Then (μ, μ) is the unique strategic equilibrium of the game. Here $\mu > \lambda/2$, the resulting queueing system is a simple M/M/2 queue and the long run expected waiting time is finite. It equals $[\mu(1+\rho)(1-\rho)]^{-1}$ where $\rho = \lambda/2\mu$.

Case c: When $R/2 \leq c'(\lambda/2) \leq R$, we have asymmetric solutions and (μ_1, μ_2) is a strategic equilibrium if and only if the following three conditions hold.

$$\begin{array}{ll} (i) & \mu_1 + \mu_2 = \lambda; \\ (ii) & c'(s) \geq R(\lambda - s)^2(\lambda + 2s)/\lambda^3; \\ (iii) & c'(\ell) \leq R. \end{array}$$

Here $s = \min(\mu_1, \mu_2)$ and $\ell = \max(\mu_1, \mu_2)$. In this borderline case, the long run expected waiting time is infinite again but the pool of waiting customers will be repeatedly empty with probability one.

Proof:

Case a: Let μ be the solution of $c'(\mu) = R$. Then $\mu < \lambda/2$ and we are in the noncompetitive case. Thus, each server maximizes unilaterally and (μ, μ) is an equilibrium. Conversely, if (μ_1, μ_2) is an equilibrium, then $c'(\mu_i) \leq R$ for $i = 1, 2$. Therefore, under Case a, $\mu_i < \lambda/2$, we are in the noncompetitive case, each player maximizes unilaterally and thus both μ_i's must satisfy $c'(\mu) = R$.

Case b: If $c'(\mu) = R\lambda^2/2\mu(2\mu+\lambda)$, then under Case b, $\mu > \lambda/2$. Thus, with (μ, μ) we are in the competitive case and

$$\begin{aligned} \left.\frac{\partial \pi_1(\mu_1+, \mu)}{\partial \mu_1}\right|_{\mu_1=\mu} &= \frac{R\lambda^2\mu^2(4\mu^2+2\lambda\mu)}{[4\mu^3+2\lambda\mu^2]^2} - c'(\mu) \\ &= \frac{R\lambda^2}{2\mu(2\mu+\lambda)} - c'(\mu) = 0. \end{aligned}$$

A similar condition holds for π_2 and, thus, (μ, μ) is an equilibrium.

Now assume that (μ_1, μ_2) is an equilibrium. If $\mu_1 + \mu_2 < \lambda$ then $c'(\mu_i) = R$ for $i = 1, 2$. So $\mu_1 = \mu_2 = \mu < \lambda/2$. But then $c'(\lambda/2) \geq R$, which contradicts case b. If $\mu_1 + \mu_2 = \lambda$ then

$$\begin{aligned} 0 &= \geq \left.\frac{\lambda \partial \pi_1(\mu_1+, \mu_2)}{\partial \mu_1}\right|_{\mu_1=\lambda-\mu_2} = R\lambda \left.\frac{\partial \alpha_1(\mu_1+, \mu_2)}{\partial \mu_1}\right|_{\mu_1=\lambda-\mu_2} - c'(\lambda - \mu_2) \\ &= \frac{R\mu_2^2(\lambda + 2(\lambda - \mu_2))}{\lambda^3} - c'(\lambda - \mu_2). \end{aligned}$$

Therefore,

$$c'(\mu_1) \geq \frac{R(\lambda-\mu_1)^2(\lambda+2\mu_1)}{\lambda^3} = \left(\frac{R}{2} \text{ at } \mu_1 = \frac{\lambda}{2}\right).$$

Similarly,

$$c'(\mu_2) \geq \frac{R(\lambda-\mu_2)^2(\lambda+2\mu_2)}{\lambda^3}.$$

Since $R(\lambda-\mu_i)^2(\lambda+2\mu_i)/\lambda^3$ is a decreasing function in $[0,\lambda]$ assuming the value $R/2$ at $\lambda/2$, we must have $\mu_i > \lambda/2$ under Case b, which contradicts the assumption that $\mu_1+\mu_2=\lambda$.

So assuming that $\mu_1+\mu_2 > \lambda$, we first show that $\mu_1 = \mu_2$. Since we are in the steady state case

$$R\lambda\frac{\partial\alpha_1(\mu_1,\mu_2)}{\partial\mu_1} - c'(\mu_1) = 0 = R\lambda\frac{\partial\alpha_2(\mu_1,\mu_2)}{\partial\mu_2} - c'(\mu_2)$$

or

$$R\lambda\left(\frac{\partial\alpha_1(\mu_1,\mu_2)}{\partial\mu_1} - \frac{\partial\alpha_2(\mu_1,\mu_2)}{\partial\mu_2}\right) + c'(\mu_2) - c'(\mu_1) = 0.$$

Substituting and subtracting, we obtain

$$R\lambda\frac{\lambda(\mu_2^2-\mu_1^2)(\mu_1+\mu_2)^2 + 2\lambda^2\mu_1\mu_2^2 - 2\lambda^2\mu_1^2\mu_2}{[2\mu_1\mu_2^2+2\mu_1^2\mu_2+\lambda\mu_1^2+\lambda\mu_2^2]^2} + c'(\mu_2) - c'(\mu_1) = 0$$

or

$$R\lambda^2\frac{(\mu_2-\mu_1)[(\mu_1+\mu_2)^3+2\lambda\mu_1\mu_2]}{[2\mu_1\mu_2^2+2\mu_1^2\mu_2+\lambda\mu_1^2+\lambda\mu_2^2]^2} + [c'(\mu_2)-c'(\mu_1)] = 0.$$

If $\mu_1 \neq \mu_2$ then both terms in the last expression are strictly negative or both are strictly positive, depending on which of the μ_i's is larger. They cannot sum to zero unless $\mu_1=\mu_2$. It follows that $\mu_1=\mu_2=\mu>\lambda/2$. We must have now

$$R\lambda\frac{\partial\alpha_1(\mu_1,\mu)}{\partial\mu_1}\bigg|_{\mu_1=\mu} - c'(\mu) = 0,$$

or

$$\frac{R\lambda^2\mu_2(4\mu^2+2\lambda\mu)}{(4\mu^3+2\lambda\mu^2)^2} = c'(\mu),$$

or

$$c'(\mu) = \frac{R\lambda^2}{2\mu(2\mu+\lambda)}.$$

Case c: Suppose (μ_1,μ_2) satisfies conditions (i), (ii), and (iii), and assume without loss of generality that $\mu_1 \leq \mu_2$. We want to show that $\partial\pi_1(\mu_1-,\mu_2)/\partial\mu_1 \geq 0$, $\partial\pi_2(\mu_1+,\mu_2)/\partial\mu_1 \leq 0$ and similarly for π_2. Thus, for Server 1,

$$\frac{\partial\pi_1(\mu_1-,\mu_2)}{\partial\mu_1} = R - c'(\mu_1),$$

but $\mu_1 \leq \lambda/2$ so $c'(\mu_1) \leq R$ and $R - c'(\mu_1) \geq 0$. Also

$$\begin{aligned}\frac{\partial \pi_1(\mu_1+, \mu_2)}{\partial \mu_1} &= R\lambda \frac{\lambda \mu_2^2(\lambda^2 + 2\lambda\mu_1)}{[2\lambda\mu_1\mu_2 + \lambda\mu_1^2 + \lambda\mu_2^2]^2} - c'(\mu_1) \\ &= \frac{R\mu_2^2(\lambda + 2\mu_1)}{\lambda^3} - c'(\mu_1) \leq 0\end{aligned}$$

by Condition (ii). For Server 2,

$$\frac{\partial \pi_2(\mu_1, \mu_2-)}{\partial \mu_2} = R - c'(\mu_2) \geq 0$$

by Condition (iii) . Also

$$\frac{\partial \pi_2(\mu_1, \mu_2+)}{\partial \mu_2} = \frac{R\mu_1^2(\lambda + 2\mu_2)}{\lambda^3} - c'(\mu_2)$$

and

$$c'(\mu_2) \geq c'(\mu_1) \geq \frac{R(\lambda - \mu_1)^2(\lambda + 2\mu_1)}{\lambda^3} \geq \frac{R(\lambda - \mu_2)^2(\lambda + 2\mu_2)}{\lambda^3},$$

with the first and third inequality following from $\mu_2 \geq \mu_1$ and the second inequality from Condition (ii). So, $\partial \pi_2(\mu_1, \mu_2+)/\partial \mu_2 \leq 0$.

Now assume that (μ_1, μ_2) is an equilibrium. If $\mu_1 + \mu_2 < \lambda$ then we are in the noncompetitive case and hence $R = c'(\mu_1)$. Also, $R = c'(s)$ with $s = \min(\mu_1, \mu_2)$ so $c'(\lambda/2) > R$, which contradicts Case c. If $\mu_1 + \mu_2 > \lambda$ then we are at the competitive case. By the argument already made in the proof of Case b, $\mu_1 = \mu_2 = \mu > \lambda/2$. Hence

$$R\lambda \frac{\partial \alpha_1(\mu_1, \mu)}{\partial \mu_1}\Big|_{\mu_1 = \mu} = c'(\mu),$$

or

$$\frac{R\lambda^2}{2\mu(2\mu + \lambda)} = c'(\mu) > R/2.$$

So, $\lambda^2/\mu(2\mu + \lambda) > 1$ or $\mu < \lambda/2$, a contradiction. It follows that $\mu_1 + \mu_2 = \lambda$, i.e., condition (i) holds.

Now assume, without loss of generality, that $s = \min(\mu_1, \mu_2) = \mu_2$ and $\ell = \max(\mu_1, \mu_2) = \mu_1$. As Server 2 optimizes at μ_2 it follows that

$$\frac{\partial \pi_2(\mu_1, \mu_2+)}{\partial \mu_2}\Big|_{\mu_1 + \mu_2 = \lambda} \leq 0$$

or

$$\frac{R(\lambda - \mu_2)^2(\lambda + 2\mu_2)}{\lambda^3} \leq c'(\mu_2),$$

i.e., Condition (ii) is satisfied. Similarly, as Server 1 maximizes at μ_1,

$$\frac{\partial \pi_1(\mu_1-, \mu_2)}{\partial \mu_1}\Big|_{\mu_1 + \mu_2 = \lambda} - c'(\mu_1) \geq 0,$$

or $R - c'(\mu_1) \geq 0$, i.e., Condition (iii) is satisfied. This completes the proof.

4 Some Comments

Let us see how the equilibrium service rates change when demand for service changes, i.e. when λ changes. Suppose than c and R are fixed, let λ_0 be the solution of $c'(\lambda/2) = R/2$ and let λ_1 be the solution of $c'(\lambda/2) = R$.

When $\lambda < \lambda_0$ we are in Case b (Figure 1), we have a unique symmetric equilibrium

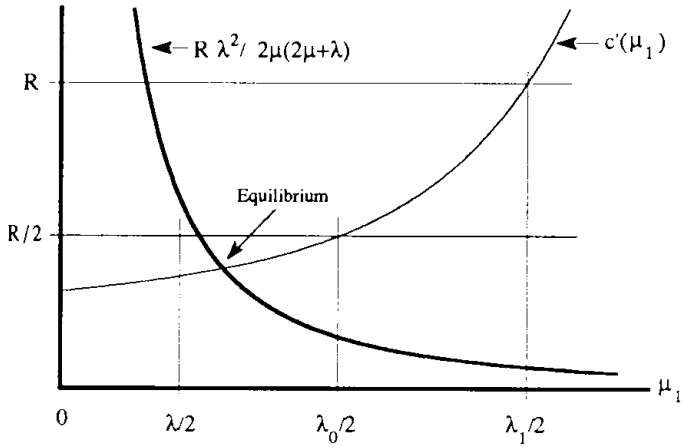

Figure 1: $\lambda < \lambda_0$ (Case b)

and a stable $M/M/2$ queuing system. In this range of λ values any increase in demand for service will cause and increase in the equilibrium service rate, since the function $R\lambda^2/2\mu(2\mu + \lambda)$ increases with λ. The equilibrium service rate may assume any value larger than $\lambda/2$ since the same function never vanishes for finite μ.

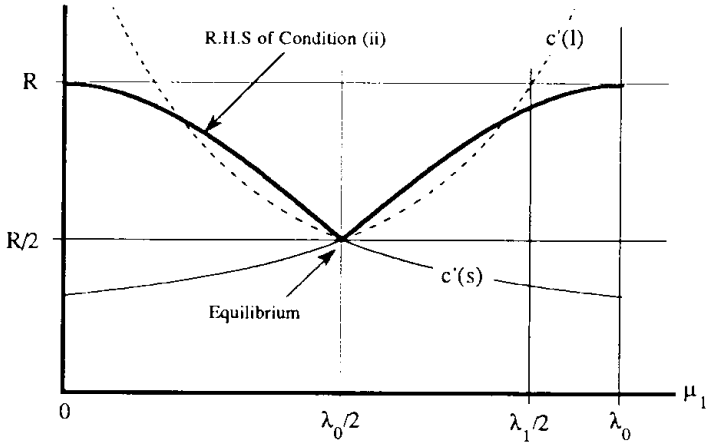

Figure 2: $\lambda = \lambda_0$ (Case c)

As λ increases and equals λ_0 we are in Case c but Condition (ii) is satisfied only when $s = \lambda_0/2$ (Figure 2). Hence the set of all service rates that satisfy conditions

(i), (ii) and (iii) shrinks to the point $(\lambda_0/2, \lambda_0/2)$, which is the unique equilibrium. For larger values of λ in Case c, i.e. for $\lambda_0 < \lambda < \lambda_1$, the set of μ values that satisfy the three conditions is of positive length (Figure 3) and its midpoint is always equal

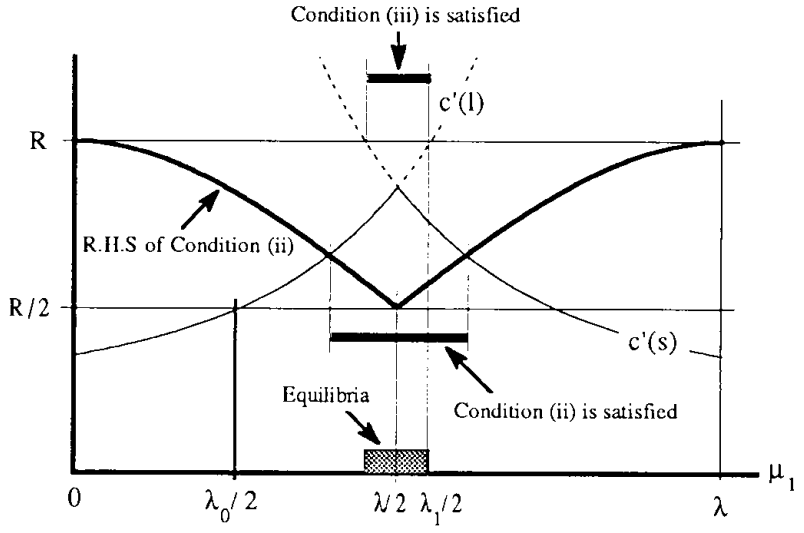

Figure 3: $\lambda_0 < \lambda < \lambda_1$ (Case c)

to $\lambda/2$. Total service capacity is equal to λ and this obviously increases with λ. When $\lambda = \lambda_1$ this set shrinks again to one point, $(\lambda_1/2, \lambda_1/2)$, since $\lambda_1/2$ is the only λ value that satisfies Condition (iii). We again have a unique symmetric equilibrium.

When $\lambda > \lambda_1$ we are in Case a (Figure 4) and the unique equilibrium is $(\lambda_1/2, \lambda_1/2)$.

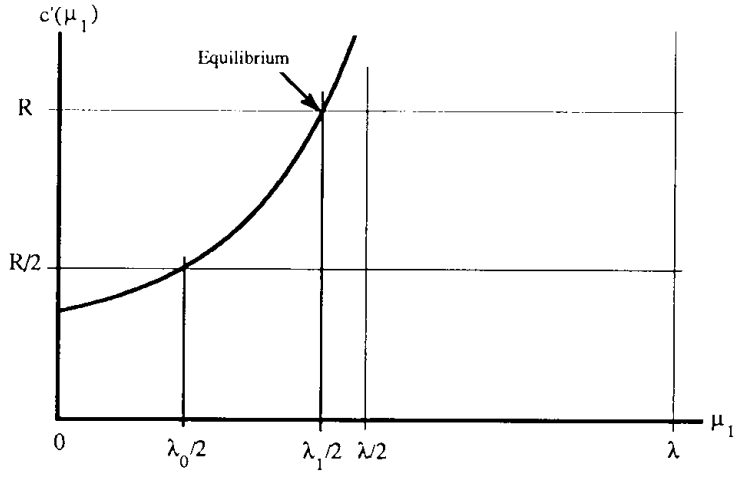

Figure 4: $\lambda > \lambda_1$ (Case a)

This remains the unique equilibrium for any value of λ exceeding λ_1.

In summary, so long as demand does not exceed λ_1 the total equilibrium service capacity increases as demand for service increases. Past this point we are in the a noncompetitive situation and the equilibrium service rates are totally insensitive to changes in demand. Similar sensitivity analyses, e.g. for the effect of changes in R when λ and c are fixed, can be carried out with no difficulty.

There are two new concepts in this paper. One is the *the competitive game of servers* and the other is *the market share* of a server in a multiserver facility. Our model brings out the fact that a monopolist, seeking maximal profits only, will not provide fast service. Also, the fact that even when there is competition, speed of service may be poor if costs of providing service is high relative to the reward the servers receive, or if demand for service is high. The model used to present these concepts is simple and may be extended in several directions to answer a variety of interesting questions. Some of these are next listed in brief.

In reality infinite mean waiting times are not common. Vendors may, on their own initiative, increase service rates so as to keep customer's satisfaction above a minimal level, or they may be regulated by society. Alternatively, customers may simply refuse to join long waiting lines and prefer not to obtain the service at all. These issues are not incorporated in our model. It would be of interest to study models that take them into account. For instance, our basic model can be extended by assuming that customers also participate in the game. They are rewarded by a 'prize' if they obtain the service and they suffer a loss for time spent waiting. They can decide whether or not to join the queue.

Another interesting issue regards the number of servers that will operate in the market. Suppose that there is a monopolist that operates at a rate $\mu \leq \lambda$ but other vendors of service can also enter the market. The question is whether they will indeed do so, and if so, how many will enter the market if there is a cost associated with the act of entry. A related question is what is the optimal number of servers in the market from society's point of view.

We believe that the concept of market share of a server in an environment with several servers is important, irrespective of whether or not the servers actually compete. This may be of interest in the context of communication where servers are computer hardware, and of production management, where servers are work stations in a production facility. The 'market share' is then the share of work carried by each server. There are some interesting questions here. For example in the production context, what is the reward that each server is entitled to as a work incentive? This depends on the share of work that he carries, namely, on the 'market share' in the above sense.

In the present model we made the most simple assumptions regarding arrival and service times. Relaxing these assumptions (e.g. assuming that service times, inter-

arrival times, or both, are not exponential) will give rise to challenging mathematical problems in computing market shares and answering some of the questions listed above.

It is hoped that answers to some of these questions, as well as others that are not listed here, will be available soon.

References

[1] Bell, C. E. and S. Stidham Jr. (1983), "Individual Versus Social Optimization in the Allocation of Customers to Alternative Servers". *Management Sci.*, **29**, 831-839.

[2] Crabill, T. B., Gross D. and M. Magazine (1977), "A Classified Bibliography of Research on Optimal Control of Queues," *Opns. Res.*, **25**, 219-232.

[3] Li, Lode (1989), "The Role of Inventory in Delivery-Time Competition," *Yale School of Organization and Management.*

[4] Owen, Guillermo. (1982), *Game Theory*, Second Edition, Academic Press.

[5] Rubinovitch, Michael (1985), "The Slow Server Problem," *J. Applied Probability*, **22**, 205-213.

[6] Teghem J. Jr. (1986), "Control of the service process in a queueing system," *Eur. J. Opnl. Res.*, **23** , 141-158.

arrival times, or both, are not exponential) will give rise to challenging mathematical problems in computing market shares and answering some of the questions listed above.

It is hoped that answers to some of these questions, as well as others that are not listed here, will be available soon.

References

[1] Bell, C. E. and S. Stidham Jr. (1983), "Individual Versus Social Optimization in the Allocation of Customers to Alternative Servers". *Management Sci.*, **29**, 831-839.

[2] Crabill, T. B., Gross D. and M. Magazine (1977), "A Classified Bibliography of Research on Optimal Control of Queues," *Opns. Res.*, **25**, 219-232.

[3] Li, Lode (1989), "The Role of Inventory in Delivery-Time Competition," *Yale School of Organization and Management.*

[4] Owen, Guillermo. (1982), *Game Theory*, Second Edition, Academic Press.

[5] Rubinovitch, Michael (1985), "The Slow Server Problem," *J. Applied Probability*, **22**, 205-213.

[6] Teghem J. Jr. (1986), "Control of the service process in a queueing system," *Eur. J. Opnl. Res.*, **23** , 141-158.