Discussion Paper No. 571

THE SLOW SERVER PROBLEM

by

Michael Rubinovitch

Northwestern University
Evanston, Illinois 60201*

August 1983

---

*Department of Managerial Economics and Decision Sciences, 2001 Sheridan Road, Evanston, Illinois 60201. The author is currently on leave from the Technion--Israel Institute of Technology.

Abstract

The problem is what to do with a slow server in a service facility which has fast and slow servers. Should the slow server be used to render service or is it better not to use it at all? Simple models for answering this question are formulated and studied.

1. Introduction

The objective of this communication is to introduce a problem which, although fundamental in its nature, has never been considered in the literature on queues. In the most general terms, the slow server problem is what to do with the slow servers in a multi-server service facility. What we have in mind is a system with several servers, some of which are slow while others are fast in the sense of mean service time. Should one use the slow servers for rendering service or is it better not to use them at all?

At first glance, one may be tempted to discard this question since it is obvious that, in most conceivable situations, a system which uses additional servers (even if they are slow) can handle a heavier traffic load than a system which does not use them. Furthermore, it is also true that a system which can handle a heavier traffic load will, in most cases, provide better service to individual customers. Indeed, this seems to have been the traditional point of view in the literature on queues. However, as we shall see, the latter is by no means of general truth and is in particular false in the case of the slow server problem.

Consider for example a system with two servers, one of which is fast while the other is slow. Let us focus attention on the customer who is first in line when the slow server becomes free and the fast server remains busy. This customer will join the slow server and is likely to stay in service for a long time. In fact, it may well happen that a short time after he enters service, the fast server becomes free and could have finished serving him long before termination of his service at the slow server. Our customer could have been better off had the slow server been removed from the system.

This simple example reveals several features of systems with slow servers. First, in any such system there is an inherent inefficiency which

is manifest at times when fast servers are idle while slow servers are busy. Second, it is clear that some customers may experience shorter delays in systems which do not employ the slow servers. It is in fact not difficult to think of situations when the mean delay becomes shorter when the slow server is removed. In short, while systems which employ slow servers can handle a heavier traffic load, it may be that systems which do not employ them can provide better service if traffic is below the maximum load which the system can carry. The central issue in the slow server problem is the trade off between an increasing mean waiting time and a decreased mean service time when the slow server is removed. The question is whether or not the net result is a decrease in the mean delay.

In this paper we present some simple queuing models for the slow server problem. The objective is to obtain qualitative and explicit quantitative results in the most simple set-ups for a better understanding of the general phenomenon. We shall look at Markovian queues with two servers, one of which is fast while the other is slow. We shall show that depending on the traffic load, and on how slow the slow server is, one can determine by a simple formula whether the slow server should be retained. More specifically, to each possible value of the ratio of the fast server's service rate to the service rate of the slow server there corresponds a single critical number. If the traffic load is larger than this number, the slow server should be retained; if it is smaller the slow server should be removed. In deriving this result we use mean delay (waiting plus service time) as our measure for system performance. It is of course possible to introduce other measures of system performance, or more complicated cost structures, but this will unavoidably be at the expense of simplicity and universality of the results. It is also possible to consider, under the same measure of performance, more

complicated models in which customers are allowed to stall, i.e., wait for the busy fast server in times when the slow server is free. Such systems are discussed in a forthcoming paper.

Models for the slow server problem and their results are of interest in many real life problems. Obviously, most any system with human servers will have slow and fast servers and no further elaboration is needed. However, the more interesting applications are in situations when new equipment is introduced as a replacement to old, or absolute service facilities or when older equipment is kept as a backup to a new facility. The new facility may be a faster medium of communication, a new computing device, or a new piece of machinery. In all these cases one usually has the option of using the old equipment at minimum cost, or at no cost at all. This is precisely the slow server problem as we perceive it.

In this paper we discuss only simple Markovian queues and the explicit results we obtain are applicable to situations when these assumptions are valid. However, the qualitative results are true in much more general setups. In fact, it seems that the simple formuli, e.g., (19), which specify the optimal action as it depends on the ratio of service times will be true at least in M/G/2 type situations.

2. The Simplest Slow Server Problem

Consider a system with two servers--server 1, the fast server, and server 2, the slow server. We assume that service times follow an exponential distribution with mean μ_1^{-1} for server 1 and μ_2^{-1} for server 2 ($\mu_1 > \mu_2$). The arrival process is Poisson with rate $\lambda > 0$ and the system has one waiting line. When there are customers in the waiting line and a server becomes free the customer who is first in line joins it. On the other hand, a customer who arrives when the system is empty (both servers are free) chooses his server at

random.

The question is whether it is better to discard the slow server and use an M/M/1 system with the slow server only instead of the two server system. In making this decision we shall use the mean delay at steady state as the decision criterion. So, the problem is to determine which of the two systems has a smaller mean delay. We shall next compute the mean number of customers in the two server system at steady state and compare this with the mean number of customers in an M/M/1 queue with service rate μ_1. This is of course equivalent to comparing mean delays on account of Little's formula.

Consider now the two server system and suppose that $\lambda < \mu_1 + \mu_2$. Then the queuing process has a unique stationary distribution and we let P_n $(n \geq 1)$ be the (stationary) probability that there are n customers in the system. We also let P_{10} and P_{01} be the probability that there is one customer in the system and he is served by server 1 and server 2 respectively. Clearly, $P_{10} + P_{01} = P_1$. With this notation we can write the following steady state equation:

$$\lambda P_0 = \mu_1 P_{10} + \mu_2 P_{01} \tag{1}$$

$$\lambda P_n = \mu P_{n+1}, \quad n \geq 1 \tag{2}$$

where

$$\mu = \mu_1 + \mu_2.$$

Thus the generating function, $G(z) = \sum z^n P_n$ $(|z| < 1)$, is

$$G(z) = \frac{\mu P_0 + (\mu_2 P_{10} + \mu_1 P_{01})z}{\mu - \lambda z}$$

Letting now $\rho = \lambda/\mu$ and using the conditions $G(1) = 1$, this yields

$$(3) \qquad G(z) = \frac{P_0 + (1 - P_0 - \rho)z}{1 - \rho z},$$

and hence the mean number of customers in the system L is

$$(4) \qquad L = G'(1) = \frac{1 - P_0}{1 - \rho}.$$

To determine P_0 we first note that from (3):

$$(5) \qquad P_n = \rho^{n-1}(1 - \rho)(1 - P_0), \qquad n \geq 1$$

and so

$$(6) \qquad P_n = \rho^{n-1}P_1, \qquad n \geq 1.$$

Then we use the two additional steady state equations:

$$(7) \qquad (\lambda + \mu_1)P_{10} = \lambda P_0/2 + \mu_2 P_2$$

$$(8) \qquad (\lambda + \mu_2)P_{01} = \lambda P_0/2 + \mu_1 P_2$$

which together with

$$(9) \qquad \rho(P_{10} + P_{01}) = P_2$$

(see (6)) can be solved to give

$$(10) \qquad P_{10} = \frac{\lambda P_0}{2\mu_1}, \qquad P_{01} = \frac{\lambda P_0}{2\mu_2}$$

and

$$(11) \qquad P_1 = \frac{\lambda P_0}{a},$$

where

$$(12) \qquad a = \frac{2\mu_1\mu_2}{\mu}.$$

Now we use $\sum P_n = 1$ together with (5) and (6) and obtain

$$(13) \qquad P_0 = \frac{1 - \rho}{1 - \rho + \lambda/a}.$$

The constant a can be thought of as an effective or average service rate when there is one customer in the system.

With this, and from (4), we can write now the expression for the mean number of customers in the system:

$$(14) \qquad L = \frac{\lambda}{(1 - \rho)(\lambda + a(1 - \rho))}.$$

Thus the slow server should be removed whenever this is larger than the mean number of customers in the M/M/1 system obtained when the slow server is removed--i.e., when

$$L > \frac{\lambda}{\mu_1 - \lambda},$$

and this is the same as

$$\rho^2(\mu - a) - 2\rho(\mu - a) - (a - \mu_1) > 0. \tag{15}$$

Now, it is not hard to check that the condition under which the slow server should be removed is

$$\rho < \rho_c \tag{16}$$

where ρ_c is the smaller root of the quadratic equation obtained from (15). (The other root is larger than 1 and is of no interest here). Thus

$$\rho_c = 1 - \sqrt{\frac{\mu_2}{\mu - a}}\,. \tag{17}$$

We can also introduce the dimension free parameter

$$r = \frac{\mu_2}{\mu_1} \tag{18}$$

$(0 \leqslant r \leqslant 1)$ and then

$$\rho_c = 1 - \sqrt{\frac{r(1 + r)}{1 + r^2}}\,. \tag{19}$$

We conclude as follows. When there is a Markovian queuing system with two servers, one slow and one fast, it is better to remove the slow server whenever the traffic intensity ρ is smaller than the single critical number ρ_c given in (19). The critical number depends on the ratio, r, of the slow

server's mean service time to the fast server's mean service time. It is equal to 1 when $r = 0$ and is 0 when $r = 1$. The single server system obtained when the slow server is removed (under this condition) is always stable.

The last part is true since the slow server is removed when $\lambda < \mu\rho_c$ and then $\lambda/\mu_1 < \rho_c\mu/\mu_1 = (1 + r)\rho_c$ which can be shown to be smaller than 1.

3. The Case of Informed Customers

The decision problem studied in the previous section provides a model for the case when customers do not have information on the two servers. Then it is reasonable to asume that a customer who arrives to an empty system chooses his server at random. In this section we propose a model for the case when customers do have information on service rates and know which server is slow and which one is fast. Then a customer who arrives to an empty system always joins the fastest server.

This different customer behavior leads to a system which makes a lesser use of the slow server. It reduces the fraction of the time when the slow server is busy and the fast server is idle and improves service. Thus, one would expect that the critical number, which determines when the slow server should be removed, will be smaller than the number ρ_c given in (19).

In studying the new two server system we shall use the same notation as before except that a_0, L_0 and ρ_0 will replace a, L, and ρ_c respectively.

The state space of the new system is the same as before and transition probabilities are also the same expect that no transitions are possible from the empty state (state 0) to the state (0,1) when server 2 is busy and server 1 is idle. In any case, the steady state equations (1) and (2) are valid here and consequently (3), (4), (5), and (6) are also true. So all we need now is to compute the probability of emptiness P_0. To this end we use the two additional steady state equations

$$(20) \qquad (\lambda + \mu_1)P_{10} = \lambda P_0 + \mu_2 P_2$$

$$(21) \qquad (\lambda + \mu_2)P_{01} = \mu_1 P_2$$

instead of (7) and (8) which, together with (9), can be solved and lead (after some lengthy algebraic work) to

$$(22) \qquad P_{10} = \frac{\lambda(\lambda + \mu)P_0}{\mu_1(2\lambda + \mu)},$$

$$(23) \qquad P_{01} = \frac{\lambda^2 P_0}{\mu_2(2\lambda + \mu)}$$

and

$$(24) \qquad P_1 = \frac{\lambda P_0}{a_0}$$

where

$$(25) \qquad a_0 = \frac{(2\lambda + \mu_1)\mu_1\mu_2}{\mu(\lambda + \mu_2)}.$$

From here, in exactly the same way as before, we obtain:

$$(26) \qquad P_0 = \frac{1 - \rho}{1 - \rho + \lambda/a_0},$$

and the mean number of customers in the system at steady state

$$(27) \qquad L_0 = \frac{\lambda}{(1 - \rho)(\lambda + a_0(1 - \rho))}.$$

Again, the condition under which the slow server should be removed is

$$L_0 < \frac{\lambda}{\mu_1 - \lambda} \tag{28}$$

which is the same as

$$\rho^2(\mu - a_0) - 2\rho(\mu - a_0) - (a_0 - \mu_1) > 0. \tag{29}$$

We note that while this looks very much like the condition of (15) there is a notable difference since here a_0 depends on λ and is hence a function of ρ. So we substitute for a_0 according to (25), rearrange terms and, using the ratio r defined in (18) obtain

$$\rho^2(1 + r^2) - \rho(2 + r^2) - (2r - 1)(1 + r) > 0 \tag{30}$$

as the condition under which the slow server should be removed.

Let ρ_0 be the smaller root and ρ_0' be the larger root of the left hand side of (30). When $r = 0$ then $\rho_0 = \rho_0' = 1$, (30) holds true for all $0 < \rho < 1$ and the slow server should always be removed. When $0 < r \leq 0.5$ then $0 \leq \rho_0 < 1$, $\rho_0' > 1$ and (30) is satisfied when $\rho < \rho_0$ or when $\rho > \rho_0'$. The latter is of no interest here so we conclude that the slow server should be removed when $\rho < \rho_0$. Finally, when $0.5 < r \leq 1$ then $\rho_0 < 0$, $\rho_0' > 1$ and the slow server should be retained for all values of $0 < \rho < 1$.

We can sum up this as follows. In the case when customers are well informed about the two servers one should never remove the slow server if $r \geq 0.5$. When $0 \leq r < 0.5$ the slow server should be removed whenever $\rho < \rho_0$,

where ρ_0 is given in (31) below. The single server system obtained when the slow server is removed (under this condition) is always stable.

The critical number ρ_0 is the smaller root of the left hand side of (30) and is given by

$$\rho_0 = \frac{2 + r^2 - \sqrt{(2 + r^2) + 4(1 + r^2)(2r - 1)(1 + r)}}{2(1 + r^2)}$$

As already noted (see Figure 1) ρ_0 is always less than ρ_c.

4. A More General Model

The two queuing models presented thus far can be thought of as special cases of a more general model where a customer who arrives to an empty system joins the fast server with probability p and the slow server with probability $(1 - p)$. Our interest in this model is mainly methodological but it could serve as a model for the case of a partially informed population of customers. If, for example, the fraction of informed customers is, say, β, then a customer who arrives to an empty system will join the fast server with probability $(1 + \beta)/2$ and the slow server with probability $(1 - \beta)/2$.

We shall now outline the essential steps in solving this model and show how the results reduce to those of the two special cases discussion in Section 1 and Section 2. The derivation is lengthy (and in fact is quite messy) and we leave the algebraic details out.

Consider the two server system where customers who arrive to an empty system join the fast server with probability p and the slow server with probability $(1 - p)$. It is not difficult to check that the steady state equations (1) and (2) are valid here and hence (3)-(6) are also true. Thus, using the additional steady state equations

$$(\lambda + \mu_1)P_{10} = \lambda p P_0 + \mu_2 P_2$$

$$(\lambda + \mu_2)P_{01} = \lambda(1 - p)P_0 + \mu_1 P_0$$

and proceeding as before one can obtain

$$P_{10} = \frac{\lambda(\lambda + \mu p)P_0}{\mu_1(2\lambda + \mu)}$$

$$P_{01} = \frac{\lambda(\lambda + \mu(1 - p))P_0}{\mu_2(2\lambda + \mu)}$$

and

$$P_1 = \frac{\lambda P_0}{a(p)}$$

where

$$a(p) = \frac{(2\lambda + \mu)\mu_1\mu_2}{\mu(\lambda + (1 - p)\mu_1 + p\mu_2)}. \tag{31}$$

It follows that P_0 is the same is in (13) with a(p) replacing a. The mean number of customers in the system L(p) is the same as (14) with the same change. Consequently, the condition under which the slow server should be removed is

$$\rho^2(\mu - a(p)) - 2\rho(\mu - a(p)) - (a(p) - \mu_1) > 0, \tag{32}$$

and this can be shown to be the same as

$$(33) \qquad \rho^3(1 + r^2) - \rho^2(1 + p + r^2(2 - p)) + \rho(1 + r)(2p(1 - r) - 1) + (1 - p)(1 - r) > 0$$

where $r = \mu_2/\mu_1$ as before.

When $p = 1$, the case of fully informed customers, (33) reduces to (30) by substitution. When $p = 0.5$, the case of uninformed customers, one can show that $\rho + 0.5$ is a factor of the left hand side of (33). Clearly for $0 \leq \rho \leq 1$ this factor is positive and we can divide (33) throughout by $\rho + 0.5$. When this is done with $p = 1$ one obtains an expression which is equivalent to (15) as expected.

When $0.5 < p < 1$ and $0 < r < 1$ there exists a single critical number $\rho(p)$ such that (33) is satisfied for $0 < \rho < 1$ if and only if $\rho < \rho(p)$. To see this, let $f(\rho)$ denote the left hand side of (33). Then $f(-\rho) < 0$ for ρ sufficiently large and $f(0) > 0$. Hnece $f(\rho) = 0$ has at least one root in $(-\infty,0)$. Similarly $f(\rho) > 0$ for ρ sufficiently large and $f(1) < 0$ so there is at least one root in $(1,\infty)$. Finally $f(0) > 0$ and $f(1) < 0$ imply that there is at least one root in $(0,1)$. Since the number of real roots is at most three, we have one negative root, one root, say $\rho(p)$ in $(0,1)$, and one root in $(1,\infty)$. This together with $f(0) > 0$ implies that $f(\rho) > 0$ for $0 < \rho < \rho(p)$ and $f(\rho) < 0$ for $\rho(p) < \rho < 1$.

The root $\rho(p)$ can be computed either by a simple search method or by using the well known (and not so practical) formuli for the roots of cubic equations. We prefer, of course, the former. The results of the computation for $p = 0.5$ (i.e., the graph of ρ_c), for $p = 0.8$, and for $p = 1$ (i.e., the graph of ρ_0) are shown in Figure 1.

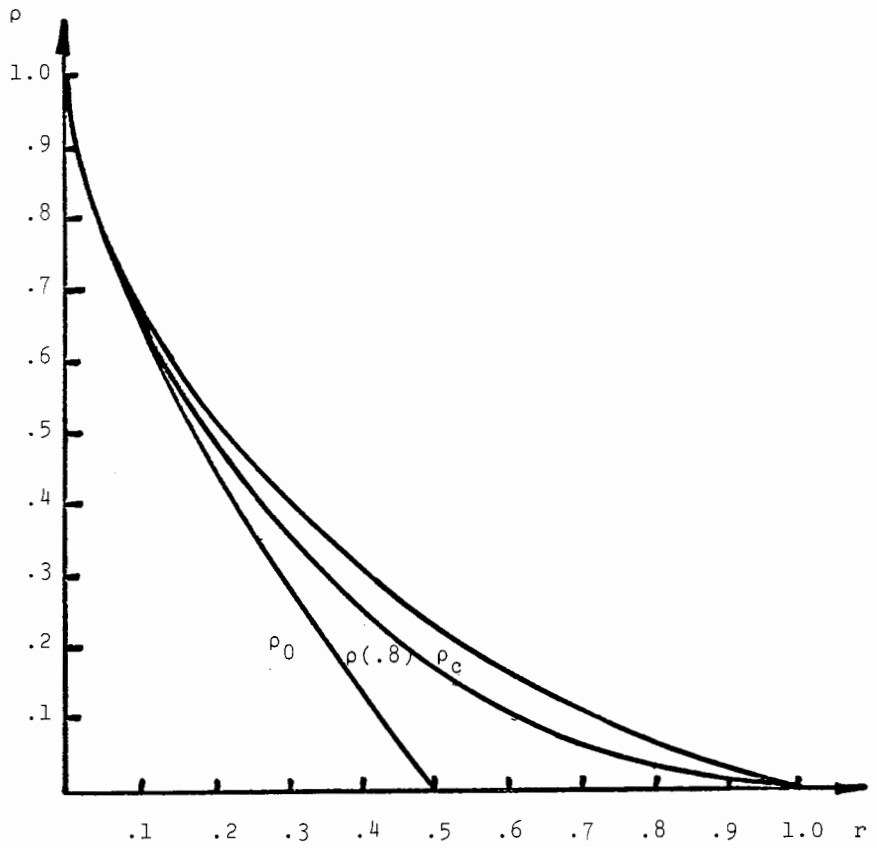

Figure 1: The Critical Numbers ρ_0, $\rho(.8)$ and ρ_c as functions of r

5. Comments

(a) Although the slow server problem as it is formulated here has not been discussed before there have been studies of systems with several different servers. The earliest reference I am aware of is Satty [5], page 290, where the system of Section 4 is discussed for the case $p = \mu_1/\mu$. It is shown that equation (4) is true but an explicit expression for p_0 is not given. The most general treatment of a queue with different servers I know of is Neuts and Takahashi [4]. Other references are Moder and Phillips [3], Singh [6], the recent paper by Bell and Stidham [7], the latest monograph by Cohen and Boxma [1], and the references they cite.

(b) It seems that the basic condition under which the slow server should be removed is in (17) which is valid in the three models studied here. The expression under the square root sign is the ratio of the slow server's service rate to the difference between the nominal service capacity of the two servers and their average service output when there is one customer in the system. I could not see the intuition behind this condition. In any case it is of interest to know how general this condition is. Is it true in the case of more than two servers? Is it true for more general service and arrival patterns?

(c) There is a heuristic argument for a quick derivation of P_0 (for all three models presented here) which may be of some interest. Let a be defined as the average service output when there is one customer in the system. Then the total service output of the sytem is $aP_1 + \mu(P_2 + P_3 + \ldots)$, which must be equal to λ since the system is at steady state. Equating these two quantities and using (5) one immediately obtains (13).

(d) It may also be of interest to compute the gain (reduction) in the mean delay when a population of customers, who are not informed, are provided with informations on the service rates of the two servers. This can be thought of as the value of information in the present set up. Clearly, this information is of value only to customers who arrive when both servers are free since other customers enter the waiting line and must join the first server which becomes free in their turn. (The value of information to the former customers is $0.5\mu_1 - 0.5\mu_2$.) Thus, the reduction in the mean delay achieved with customers are informed is the average value of the information over all the customers. In economic terms this is the social value of the information as discussed, for example, in Hirshleifer [2]. Let $V(\lambda, \mu_1, \mu_2)$ be

this social value of information, in the present model, when the arrival and service rates are λ, μ_1 and μ_2 respectively. Let $W_1(\lambda, \mu_1)$ be the mean delay in an M/M/1 system with arrival rate λ and service rate μ_1. Also let $W_{UI}(\lambda, \mu_1, \mu_2)$ be the mean delay in a system with two servers, when customers are not informed (Section 2) and $W_I(\lambda, \mu_1, \mu_2)$ be the mean delay in the case of informed customers (Section 3). Then $W_1 = 1/(\mu_1 - \lambda)$, $W_{UI}(\lambda, \mu_1, \mu_2) = L/\lambda$ and $W_I(\lambda, \mu_1, \mu_2) = L_0/\lambda$. With this, and putting $\rho_c(r) = \rho_c$, $\rho_0(r) = \rho_0$ one can write

$$\{ V(\lambda, \mu_1, \mu_2) = \begin{cases} 0 & \rho \leq \rho_0(r) \\ W_1(\lambda, \mu_1) - W_I(\lambda, \mu_1, \mu_2) & \rho_0(r) < \rho \leq \rho_c(r) \\ W_{UI}(\lambda, \mu_1, \mu_2) - W_I(\lambda, \mu_1, \mu_2) & \rho > \rho_c(r), \end{cases}$$

since when $\rho < \rho_0(r)$ the slow server is not used at all, when $\rho_0(r) < \rho < \rho_c(r)$ the system with informed customes uses the slow server while the system with uninformed customers does not use it, and, when $\rho > \rho_c(r)$ both systems use both servers. From here one can readily compute the social value of information as a function of the system parameters. It is interesting to note that when λ and μ_1 are fixed $V(\lambda, \mu_1, \mu_2)$ is zero for $r \leq \rho_0^{-1}(\rho)$, then it increases for a while as r increases and eventually decreases to become zero when $r = 1$. Similarly, when μ_1 and μ_2 are kept fixed $V(\lambda, \mu_1, \mu_2)$ is zero for $\rho < \rho_0(r)$ then it increases with λ and eventually decreases to become zero when $\rho = 1$. This kind of behavior of the function V has not been cited before in the literature on the social value of information.

(e) It should be emphasized that in the present study customers are really not allowed to make any decisions. If they are waiting in line they are dispatached to the first server that becomes free when their turn comes

up. If they arrive when the system is empty they are dispatched to the fast server in the model of Section 3 and are sent randomly to one of the two servers in the model of Section 2. Thus issues like optimal customer decisions or social optimization versus individual customer optimization do not arise. These issues are of interest and will be discussed in a separate article which studies the case when customers are allowed to stall, i.e., to wait for a busy fast server in times when the slow server is free. It will be shown that a system with stalling can provide better service and, in fact, never discards the slow server. Specifically to each set of specified values of λ, μ_1 and μ_2 there corresponds one socially optimal system in which at most k customers are allowed to stall, where k is a function of λ, μ_1 and μ_2. The issue of social versus individual optimization will also be discussed in detail.

Acknowledgement

I am grateful to J. Michael Harrison for being so outspoken about his impatience with the inconveniences which many "little slow servers" cause us in day-to-day life.

References

[1] Cohen, J. W. and D. J. Boxma (1983). Boundary Value Problems in Queuing System Analysis. Amsterdam: North-Holland Publishing Co.

[2] Hirshleifer, J. (1971). The Private and Social Value of Information and the Reward of Inventive Activities. American Economic Rev. 61, 561-574.

[3] Moder, J. J. and C. R. Phillips (1962). Queuing with Fixed and Variable Channels. Opns. Res. 10, 218-231.

[4] Neuts, M. F. and Y. Takahashi (1981). Asymptotic Behavior of the Stationary Distribution in the GI/PH/C Queue with Heterogeneous Servers. Z. Wahrsh. Verw. Gebiete. 57, 441-452.

[5] Satty, L. S. (1961). Elements of Queueing Theory. McGraw-Hill.

[6] Singh, V. P. (1973). Queue-Dependent Servers. J. Eng. Math. 7, 123-126.

[7] Stidham, S. S. and C. E. Bell (1983). Individual Versus Social Optimization in the Allocation of Customers to Altnerative Servers. Mgmt. Sci. 29, 831-839.