Optimal Priority-Purchasing and Pricing Decisions in Nonmonopoly and Monopoly Queues
Author(s): I. Adiri and U. Yechiali
Source: *Operations Research*, Vol. 22, No. 5 (Sep. - Oct., 1974), pp. 1051-1066
Published by: INFORMS
Stable URL: http://www.jstor.org/stable/169658
Accessed: 28/06/2009 05:02

Optimal Priority-Purchasing and Pricing Decisions in Nonmonopoly and Monopoly Queues

I. Adiri

Technion—Israel Institute of Technology, Haifa, Israel

and

U. Yechiali

Tel Aviv University, Tel Aviv, Israel

(Received October 22, 1971)

An $M/M/1$ service station (computer center) consists of M separate queues. The ith ($i=1, 2, \cdots, M$) queue has priority over the jth iff $i<j$. Upon arrival, a customer receives all the information regarding the state of the system and accordingly makes an irrevocable decision as to which queue to join, or rather to balk (leave) and go to a competitor. The higher the priority of the queue, the higher the toll fee to join it but the shorter the time spent in the system. This paper considers nonmonopoly and monopoly cases, and optimal priority-purchasing or balking rules for the newly arrived customer, as well as optimal pricing policies for the service station for both preemptive-resume and nonpreemptive-priority disciplines.

IN MOST OF the priority-queuing literature, the priority class that a customer belongs to is assumed to be an inherent quality over which the customer has no control. However, in real-life situations, customers who are motivated by the urgency of their service would like to influence the determination of their priority degree. In such cases, depending on the state of the system at the moment of arrival, customers may even decide not to join the system at all, i.e., balk, and look for service elsewhere.

The arrival stream to the service station arises because of the customers' potential benefit from being served. We assume that, upon completion of service, a customer is endowed, on the average, with a reward of u monetary units; in other words, u is the average value of the service. On the other hand, to reflect the customer's alternative value for his time, we assume that the waiting-time cost is c monetary units per unit time (the same for all customers).

Moreover, a customer who joins the system is charged a service fee (toll) of θ_i monetary units whenever he chooses to join the ith queue. Obviously, there is a strictly higher toll charge to join a higher-priority queue, and we will follow the convention of letting $j<i$ whenever queue j has priority over queue i (and thus $\theta_j>\theta_i$).

The customer's decision as to which priority class to purchase or, rather, to balk is based on economic considerations: the toll fee plus the expected cost of the time spent in the system (termed in the sequel the *expected service cost*) is required not to exceed u. If the minimum expected service cost is greater than u, the customer balks without being served. (We refer to this assumption as the *cost constraint*).

The service station is assumed to be a profit-making organization that is interested in maximizing its average income per unit time. A service facility operating under such a procedure collects money through the tolls that joining customers pay to purchase their priority classes. This is true, for example, in a commercial computer center where a newly arrived customer is allowed either to purchase his priority class by paying a predetermined service charge, or to balk and go to a competitor.

THE MODELS

A SINGLE SERVER dispenses service to an infinite number of potential customers. The arrival process is a homogeneous Poisson process with parameter λ. (This stream includes customers who leave the competitors and arrive at our station.) Service times are assumed to be independent identically distributed exponential random variables with mean $1/\mu$. We restrict our analysis to the steady state. A newly-arrived customer, knowing all the information regarding the state of the system, is allowed to purchase, out of M available priority classes, his class, or to balk. The set of toll fees $\boldsymbol{\theta}=\{\theta_i\colon i=1, 2, \cdots, M;\ \theta_i>\theta_j \text{ iff } i<j\}$ is determined by the service station so as to maximize its average income per unit time. An arrival who decides to pay the amount of θ_k $(k=1, 2, \cdots, M)$, is assigned to the kth priority class. Within each priority class the FIFO rule is practiced. In the sequel we consider the preemptive-resume and head-of-the-line (nonpreemptive) priority regimes where no losses are involved.

Two cases are analyzed:

I. A nonmonopoly service station where a customer has an option to leave and get service elsewhere. We reflect this by assuming u to be a positive finite number.

II. A monopoly service station where a customer does not have the option to leave, no matter how high the service cost is; i.e., u is infinite.

Our purposes in this paper are: (i) optimal decision policies for newly arrived customers, and (ii) an optimal set of toll fees for the service station. Obviously, (i) and (ii) are highly correlated.

Recently some papers dealing with related models have been published. Kleinrock[4] has studied the case where a newly arrived customer decides which priority to join without any specific prior knowledge of the current state of the system. Naor[5] and Yechiali[7,8] have studied various cases where customers can either join a single queue or balk. Balachandran[3] determined the best (and stable) prices to be paid by customers on arrival for a system with an infinite number of classes where only one or two customers are allowed in each class.

CASE I: NONMONOPOLY

Optimal Priority-Purchasing and Balking Policies

Two Priority Classes, $M=2$: Preemptive-Resume Regime

Let $s=(X_1, X_2)$ denote the state of the system observed by an arrival. X_i $(i=1, 2)$ is the number of customers present in the ith queue. Let η_m be the state observed by the mth ($m=1, 2, 3, \cdots$) arrival. We say that the system is in state s at time m if $\eta_m=s$. If the mth arrival chooses to pay θ_k and to join the kth ($k=1, 2$) queue, we say that he takes an action $\Delta_m=k$. If he decides to balk, we let $\Delta_m=0$ ($k=0$). Regarding the system as making its transitions at instants of arrival, we obtain a Markovian decision process (MDP) $\{\eta_m, \Delta_m, m=0, 1, 2, \cdots\}$. A policy R is a set of functions $\{D_k^R(h_{m-1}, \eta_m)\}$, ($m=0, 1, 2, \cdots$), where $h_{m-1}=\{\eta_0, \Delta_0, \cdots, \eta_{m-1}, \Delta_{m-1}\}$ represents the history of the MDP up to and including the $(m-1)$th step. For any history h_{m-1} and any state $\eta_m=s$, the functions $D_k^R(\cdot)$ comprise a probability distribution; i.e., $D_k^R(\cdot)\geqq 0$ for all k and $\sum_{k=0}^{k=M} D_k^R(\cdot)=1$. $D_k^R(h_{m-1}, \eta_m)$ specifies the probability of the mth arrival taking on action $\Delta_m=k$ ($k=0, 1, 2$).

An arrival who observes $s=(X_1, X_2)$ and selects to join queue 1 ($k=1$) stays, on the average, in the system $(X_1+1)/\mu$ time units. If he decides to join queue 2 ($k=2$), his average time in the system $W_u(s)$ is composed of the sum of service times of all customers (of both classes) present in the system at the moment of his arrival plus his service time plus the service times of all future type 1 customers who arrive before he leaves the system. Hence, his expected time in the system is a function of the priority-purchasing policies of future arrivals.

We assume that all customers behave 'rationally' and their optimal policy is calculated by taking into consideration that every arrival follows the same reasoning. Let b_{sk} be the expected service cost of a newly arrived customer who observes state $s=(X_1, X_2)$ and decides to join the kth ($k=1, 2$) queue. Hence,

$$b_{s1}=c(X_1+1)/\mu+\theta_1, \tag{1}$$

and

$$b_{s2}=cW_u(s)+\theta_2. \tag{2}$$

In the case where $\min(b_{s1}, b_{s2})>u$, the customer balks.

$$b_{s0}=\begin{cases} y, & \text{if } \min(b_{s1}, b_{s2})\leqq u, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where y is a large positive number.

The problem a newly arrived customer is faced with may be formulated in the following way: Find an optimal policy $R=D_k^R(\cdot)$ that, for every history h_{m-1} and every state s, minimizes

$$\sum_{k=0}^{k=2} D_k^R(h_{m-1}, s)b_{sk}, \tag{4}$$

subject to the constraints

$$D_k^R(\cdot)\geqq 0 \qquad (k=0, 1, 2) \tag{5}$$

and

$$\sum_{k=0}^{k=2} D_k^R(\cdot)=1. \tag{6}$$

This is a linear programming problem in three variables. Clearly its solution is

$$\begin{aligned} &D_0^R(\cdot)=1, \quad D_1^R(\cdot)=D_2^R(\cdot)=0, \quad \text{if} \quad \min(b_{s1}, b_{s2})>u; \\ &D_1^R(\cdot)=1, \quad D_0^R(\cdot)=D_2^R(\cdot)=0, \quad \text{if} \quad \min(b_{s1}, b_{s2})\leqq u \quad \text{and} \quad b_{s1}<b_{s2}; \\ &D_2^R(\cdot)=1, \quad D_0^R(\cdot)=D_1^R(\cdot)=0, \quad \text{otherwise.} \end{aligned}$$

Hence, for all histories and for all states, the optimal policy is a stationary nonrandomized one.

We now show that the optimal policy is not only a nonrandomized one but is of the 'control-limit' type; i.e., there exists an integer n_2^* such that the optimal joining strategy is a control-limit rule of the form: whenever $s=(X_1, X_2)$, join queue 2 (i.e., purchase priority 2) if and only if $X_2<n_2^*$.

To see this, consider an arrival who observes the state $s=(r-1, j)$ and elects to purchase priority 1 by paying θ_1. This action is taken because this relation holds:

$$\theta_1+rc/\mu<\theta_2+cW_u(r-1, j). \qquad (7)$$

Let us assume that a customer observing $s=(r, j)$ decides to pay θ_2 and to join queue 2; his decision is made because

$$\theta_2+cW_u(r, j)\leqq\theta_1+(r+1)c/\mu. \qquad (8)$$

Relations (7) and (8) yield:

$$W_u(r, j)<W_u(r-1,j)+1/\mu. \qquad (9)$$

But the underlying model and the definition of $W_u(s)$ imply that $W_u(r,j)$ is greater than $W_u(r-1, j)$ by at least $1/\mu$ time units. Thus, (9) cannot hold. Hence, an optimal policy has the property that, if it pays to purchase priority 1 when observing $s=(r-1, j)$, $(r\geqq 1, j\geqq 0)$,then it also pays to purchase the highest priority when observing $s=(r, j)$.

We distinguish among the following three possible cases:

(i) $u<\theta_2+c/\mu$; customers never join the system since their cost constraint is not satisfied.

(ii) $\theta_2+c/\mu\leqq u<\theta_1+c/\mu$; queue 1 is always empty. The cost constraint eliminates the possibility of a customer paying a toll fee of θ_1 and joining the first queue. Our model becomes the usual limited-room $M/M/1$ queue, where the maximum number of customers in the system is $[(u-\theta_2)\mu/c]$, where $[x]$ denotes the largest integer smaller than or equal to x.

(iii) $u\geqq\theta_1+c/\mu$; this is the interesting case where both queues are active. Let us consider the service station at its idle period, i.e., in state $s=(0, 0)$. If the customer who initiates a busy period decides to pay a toll fee of θ_1 and to join queue 1, then, in view of our previous discussion, all the future customers will follow his decision and no customer will join queue 2, i.e., the optimal control limit n_2^* is zero. Otherwise, customers join queue 2 until, for some state $s=(0, j)$, $j>0$, the best decision is to join queue 1. This j is the optimal control limit and we set $n_2^*=j$. An arrival who meets the system in state $s=(r, n_2^*)$ joins queue 1, until, for some r_1 $(r=r_1)$, the expected service cost is greater than u. We denote by m_1 the largest r that still satisfies the cost constraint. A newly arrived customer who observes (m_1, n_2^*) balks. Hence, if $X_1=1, 2, \cdots, m_1$, then $X_2=n_2^*$ and, if $X_1=0$, then $X_2\leqq n_2^*$. We now turn to calculating m_1 and n_2^*.

Calculating the Optimal Control Limits. Owing to the cost constraint, the number of customers in the system is limited. For the case where $u\geqq\theta_1+c/\mu$, in view of our previous discussion, a newly arrived customer who initiates a busy period either joins queue 1 or queue 2. If he elects to join queue 1, then $n_2^*=0$; otherwise, he pays θ_2 and joins queue 2. All subsequent arrivals follow him as long as $X_2<n_2^*$. An arrival who observes a system with n_2^* customers in queue 2 elects either to join queue 1 or to balk, depending on the number of customers in queue 1. The maximum number of customers in queue 1 is

$$m_1=[(u-\theta_1)\mu/c]. \qquad (10)$$

An arrival who finds the system in state $s=(m_1, n_2^*)$ balks. Hence, the number of customers in the system is distributed as the number of customers in a limited-waiting-room $M/M/1$ queue where the maximum number of customers in the system is $N=m_1+n_2^*$ (designated in the sequel by $M/M/1/N$). Note that, no matter how high the load on the system is, the cost constraint eliminates the possibility of saturation.

Consider a service station in state $s=(0, j)$. For a control limit n_2 and maximum of m_1 customers in queue 1, let $H_{n_2}^{m_1}(q, j)$, $(q<j\leqq n_2)$ be the expected time elapsed from the moment a customer becomes the $(q+1)$th customer in queue 2 until the moment when he departs. Being an optimal control limit, n_2 must satisfy

$$\theta_2+cH_{n_2}^{m_1}(n_2-1,\ n_2)\leqq\theta_1+c/\mu<\theta_2+cH_{n_2+1}^{m_1}(n_2,\ n_2+1), \qquad (n_2=1, 2, \cdots) \quad (11)$$

and for $n_2=0$ we have,

$$\theta_1+c/\mu<\theta_2+cH_1^{m_1}(0, 1). \qquad (12)$$

Regarding the underlying process, it is easily seen that the function $H_{n_2}^{m_1}(n_2-1, n_2)$ increases with n_2 and m_1. Since m_1 is a nonincreasing step function of θ_1 [equation (10)], then $H_{n_2}^{m_1}(n_2-1, n_2)$ is also a nonincreasing step function of θ_1. To find n_2^* we calculate $H_{\mathrm{n}_2}^{\mathrm{m}_1}(n_2-1, n_2)$ for increasing values of $n_2 (n_2=0, 1, 2, \cdots)$ until (12) or (11) is satisfied. We calculate $H_{n_2}^{m_1}(q, j)$ for $0<q<j\leqq n_2$ recursively; i.e., we find $H_{n_2}^{m_1}(q, j)$ as a function of $H_{n_2}^{m_1}(q-1, e)$, $e=j-1, j, \cdots, n_2-1$, with $H_{n_2}^{m_1}(0, e)$, $e=1, 2, \cdots, n_2-1$, as initial values.

Let K be the number of arrivals during a service time, and let a_k denote the probability of exactly k arrivals. Hence,

$$a_k=P(K=k)=\int_0^\infty \{(\lambda t)^k/k!\}e^{-\lambda t}\mu e^{-\mu t}\,dt=\{\rho/(1+\rho)\}^k\{1/(1+\rho)\}, \qquad \rho=\lambda/\mu. \qquad (13)$$

The probability of k or more arrivals during a service time is denoted by α_k,

$$\alpha_k=P(K\geqq k)=\textstyle\sum_{i=k}^{i=\infty} a_i=\{\rho/(1+\rho)\}^k. \qquad (14)$$

The density function of a service time L during which k arrivals occur is

$$\begin{aligned} P(l\leqq L\leqq l+dl|K=k) &= \{(\lambda l)^k/k!\}e^{-\lambda l}e^{-\mu l}\mu dl/P(K=k) \\ &= (\lambda+\mu)\{(\lambda+\mu)l\}^k e^{-(\lambda+\mu)l}/k!; \qquad (k=0, 1, \cdots) \end{aligned} \qquad (15)$$

i.e., this is an Erlang distribution with parameter $(\lambda+\mu)$ and $(k+1)$ phases. Thus,

$$E(L|K=k)=(k+1)/(\lambda+\mu). \qquad (16)$$

Consider a system in the state $s=(0, j)$ with n_2 and m_1 as its control limits. If during the service of the currently served customer there are k arrivals, and $k\leqq n_2-j$, the current service is not preempted and $H_{n_2}^{m_1}(q, j)$ is equal to $(k+1)/(\lambda+\mu)+H_{n_2}^{m_1}(q-1, j+k-1)$. If $k>n_2-j$, then the (n_2-j+1)th arrival preempts our customer's service and initiates a busy period that is distributed as a busy period in an $M/M/1/m_1$ queue. The average length of such a busy period[6] is $(1-\rho^{m_1})/\{\mu(1-\rho)\}$. The service of the current customer is resumed when the state of the system is changed from $s=(1, n_2)$ to $s=(0, n_2)$.

This procedure is repeated with any arrival while the system is in state $s=(0, n_2)$. Thus, each of the (n_2-j+1)th, (n_2-j+2)th, $\cdots$, arrivals during the service of the current customer initiates an independent $M/M/1/m_1$ busy period of average length $(1-\rho^{m}{}_{1})/\{\mu(1-\rho)\}$. Owing to the memoryless property of the exponential service time, the number of additional busy periods during a service time of a customer A is distributed as K, equation (13). Upon completion of the service of the current customer, our customer becomes the qth in queue 2.

Hence, for $0<q<j\leqq n_2$ we have

$$\begin{aligned} H_{n_2}^{m_1}(q,j)=\textstyle\sum_{k=0}^{n_2-j}&[(k+1)/(\lambda+\mu)+H_{n_2}^{m_1}(q-1,j+k-1)]a_k\\ &+\textstyle\sum_{k=n_2-j+1}^{k=\infty}[(k+1)/(\lambda+\mu)+\{k-(n_2-j)\}(1-\rho^{m_1})/\mu(1-\rho)\\ &\qquad\qquad +H_{n_2}^{m_1}(q-1,n_2-1)]a_k. \end{aligned} \tag{17}$$

Rearranging and simplifying (17) yield

$$\begin{aligned} H_{n_2}^{m_1}(q,j)=1/\mu+\alpha_{n_2-j+1}&[(1+\rho)(1-\rho^{m_1})/\mu(1-\rho)+H_{n_2}^{m_1}(q-1,n_2-1)]\\ &+\textstyle\sum_{k=0}^{n_2-j}H_{n_2}^{m_1}(q-1,j+k-1)a_k. \end{aligned} \tag{18}$$

The initial values $H_{n_2}^{m_1}(0,j)$, $(j=1,2,\cdots,n_2)$, are calculated by using the same arguments. The result is

$$\begin{aligned} H_{n_2}^{m_1}(0,j)&=\textstyle\sum_{k=0}^{n_2-j}\{(k+1)/(\lambda+\mu)\}a_k+\sum_{k=n_2-j+1}^{k=\infty}[(k+1)/(\lambda+\mu)\\ &\qquad+\{k-(n_2-j)\}(1-\rho^{m_1})/\mu(1-\rho)]a_k\\ &=1/\mu+\alpha_{n_2-j+1}(1+\rho)(1-\rho^{m_1})/\mu(1-\rho). \qquad (1\leqq j\leqq n_2) \end{aligned} \tag{19}$$

Equation (19), as well as (18), agrees with our intuition. The unconditional expected service time is $1/\mu$. With probability $P(K>n_2-j)=\alpha_{n_2-j+1}$, the service of the current customer is preempted. Given that the service of the customer is preempted at least once, the average number of busy periods is $E(A+1)=1+\rho$ each of average length $(1-\rho^{m_1})/\{\mu(1-\rho)\}$.

To find n_2^*, one calculates $H_{n_2+1}^{m_1}(n_2,n_2+1)$ for $n_2=0,1,2,\cdots$, [equations (19) and (18)] and compares sequentially the calculated values until (11) is satisfied. The value of n_2 that satisfies (11) is the optimal control limit of the second queue n_2^*. [$n_2^*=0$ if (12) is satisfied.]

Two Priority Classes, $M=2$: Head-of-the-Line Regime

In a nonpreemptive discipline, an arrival who observes the system in state $s=(0,j)$ and decides to join the first queue is admitted for service only at the completion of the current service. By using the same arguments as in the previous case, we arrive at the conclusion that, in this case too, the optimal purchasing policy is a stationary nonrandomized one of a control-limit type. However, in this case, we may encounter states of the form $s=(r,n_2^*-1)$, $r=1,2,\cdots,m_1$. An arrival who observes $s=(r,n_2^*-1)$, $r=1,2,\cdots,m_1-1$, joins the first queue. This is true since, if it pays for a customer who observes $s=(0,n_2^*)$ to join queue 1, then it also pays to join the first queue while observing $s=(1,n_2^*-1)$. In both cases the expected service cost of joining queue 1 is θ_1+2c/μ and joining queue 2 means paying a toll fee of θ_2 and waiting for the completion of the service of n_2^* customers and all future customers that will join queue 1. Clearly, a newly arrived customer balks if he finds the system in either state $s=(m_1,n_2^*-1)$ or state $s=(m_1-1,n_2^*)$; hence, the maximum number of customers in the system is now $m_1+n_2^*-1$. Thus, for $s=(X_1, X_2)$, if $0<X_1 \leqq m_1$, then $X_2=n_2^*$ or $X_2=n_2^*-1$, and if $X_1=0$, then $X_2=0, 1, 2, \cdots, n_2^*$. In the event that $X_1 \geqq 0$ and $X_2=n_2^*$, the currently served customer belongs to queue 2, but, in the case where $X_1>0$ and $X_2=n_2^*-1$, the current customer belongs to queue 1.

Few modifications are needed in the calculation of n_2^*. Since preemptions are not allowed and $\theta_1>\theta_2$, the optimal control limit is never zero. The necessary and sufficient condition for a positive integer n_2 to be the optimal control limit n_2^* takes the form

$$\theta_2+cH_{n_2}^{m_1}(n_2-1, n_2) \leqq \theta_1+2c/\mu<\theta_2+cH_{n_2+1}^{m_1}(n_2, n_2+1). \tag{20}$$

This equation replaces (11). Owing to the fact that a service is uninterruptable, we have

$$H_{n_2}^{m_1}(0, j)=1/\mu. \qquad (j=1, 2, \cdots, n_2) \tag{21}$$

Equation (21) replaces (19).

In addition, since a customer balks when the system is in state $s=(m_1-1, n_2^*)$ and the average time elapsed from the moment when there are i customers in queue 1 until, for the first time, there are $(i-1)$ customers in queue 1 is $(1-\rho^{m_1-i+1})/\{\mu(1-\rho)\}$, (the capacity is limited to m_1), then the analogous of equation (17) is

$$\begin{aligned} H_{n_2}^{m_1}(q, j) = & \textstyle\sum_{k=0}^{n_2-j} [(k+1)/(\lambda+\mu)+H_{n_2}^{m_1}(q-1, j+k-1)]a_k \\ & +\textstyle\sum_{k=n_2-j+1}^{n_2-j+m_1-1} [(k+1)/(\lambda+\mu)+\sum_{i=1}^{k-(n_2-j)} \{(1-\rho^{m_1-i})/\mu(1-\rho)\} \\ & \quad +H_{n_2}^{m_1}(q-1, n_2-1)]a_k+\textstyle\sum_{k=n_2-j+m_1}^{k=\infty} [(k+1)/(\lambda+\mu) \\ & \quad +\textstyle\sum_{i=m_1-1}^{i=1} \{(1-\rho^{m_1-i})/\mu(1-\rho)\}+H_{n_2}^{m_1}(q-1, n_2-1)]a_k. \end{aligned} \tag{22}$$

$M \geqq 2$ Priority Classes

This is a generalization of the case discussed previously where there were only two priority queues. Upon arrival, a customer who requires that the expected service cost will not exceed u monetary units is supplied with the following information:

(i) The set of toll fees $\boldsymbol{\theta}=\{\theta_i: i=1, 2, \cdots, M;\ \theta_i<\theta_j \text{ iff } i>j\}$.

(ii) The number of customers in each priority queue X_i, $i=1, 2, \cdots, M$.

Accordingly, the customer makes an irrevocable decision as to which queue to join or to balk. Paying a toll fee of θ_i, the customer joins the end of the ith queue

$M \geqq 2$ Priority Classes: Preemptive-Resume Discipline

Arguing in the same manner as in the previous case, we can prove that the optimal purchasing policy remains stationary nonrandomized and of the control-limits type. The optimal purchasing-priority policy is comprised of a set of M control limits, n_M^*, $\overset{*}{n}_{M-1}$, $\cdots$, n_2^*, n_1^*, such that a new arrival purchases priority i iff he observes the system in state $s=(0, 0, \cdots, X_i, \overset{*}{n}_{i+1}, \cdots, n_M^*)$, where $0 \leqq X_i<n_i^*$.

Let f be the smallest integer that satisfies

$$u \geqq \theta_f+c/\mu. \tag{23}$$

Only queues numbered $f, f+1, \cdots, M$ are active. Owing to the cost constraint, a customer balks rather than joining any of the queues $1, 2, \cdots, f-1$. Furthermore, the maximum number of customers in the fth queue is

$$m_f = [(u-\theta_f)\mu/c]. \tag{24}$$

Customers who observe the system in state $s=(0, 0, \cdots, 0, m_f, \overset{*}{n}_{f+1}, \cdots, n_M{}^*)$ balk without receiving service. Hence, we set $n_i{}^*=0$ for $i=1, 2, \cdots, f-1$. The maximum number of customers in the system is $m_f+\overset{*}{n}_{f+1}+\overset{*}{n}_{f+2}+\cdots+n_M{}^*$. It is left to calculate the optimal control limits for the $M-f$ lowest-priority queues, namely, $\overset{*}{n}_{f+1}, \overset{*}{n}_{f+2}, \cdots, n_M{}^*$.

A newly arrived customer who observes the system in state $s=(0, 0, \cdots, X_i, \overset{*}{n}_{i+1}, \overset{*}{n}_{i+2}, \cdots, n_M{}^*)$, $i>f$, has to decide whether to join queue i or $i-1$, and the one who observes $s=(0, 0, \cdots, X_f, \overset{*}{n}_{f+1}, \overset{*}{n}_{f+2}, \cdots, n_M{}^*)$ has to decide whether to join the fth queue or to balk. Since we deal with a preemptive-priority discipline, at most two consecutive priority queues are taken simultaneously into consideration by a newly arrived customer and all other queues are ignored. Hence, calculation of $n_i{}^*$, $i=f+1, f+2, \cdots, M$, is basically the same as in the case $M=2$.

Let $H^{m_f}_{n_i}(q, j)$, $i>f$, $q<j\leq n_i$ be the expected time elapsed between the moments a customer becomes the $(q+1)$th in the ith queue and the moment of his departure; there are j customers in the ith queue; the control limit in the ith queue is n_i; the maximum number of customers in the fth queue is m_f. The set of optimal control limits satisfies the relations

$$cH^{m_f}_{n_i}(n_i-1, n_i)+\theta_i \leqq cH^{m_f}_{n^*_{i-1}}(0, 1)+\theta_{i-1} < cH^{m_f}_{n_i+1}(n_i, n_i+1)+\theta_i, \qquad (i=f+2, f+3, \cdots, M) \tag{25}$$

and

$$cH^{m_f}_{n_{f+1}}(n_{f+1}-1, n_{f+1})+\theta_{f+1} \leqq c/\mu+\theta_f < cH^{m_f}_{n_{f+1}+1}(n_{f+1}, n_{f+1}+1)+\theta_{f+1}. \tag{26}$$

The calculation procedure is recursive. We begin with queues f and $f+1$ and apply the equations of the section on two priority classes with m_f and n_{f+1} replacing m_1 and n_2, respectively. The value of n_{f+1} that satisfies (26) is the optimal control limit of the $(f+1)$th queue $\overset{*}{n}_{f+1}$. To calculate $\overset{*}{n}_{f+2}$, we consider queues $f+1$ and $f+2$. We combine the fth and $(f+1)$th queues to one queue in which the maximum number of customers is $m_f+\overset{*}{n}_{f+1}$. [Note $\overset{*}{n}_{f+1}$ has been calculated in the previous step.] We apply again the equations of the section on two priority classes with $(m_f+\overset{*}{n}_{f+1})$ and n_{f+2} replacing m_1 and n_2, respectively. The value of n_{f+2} that satisfies (25) is $\overset{*}{n}_{f+2}$. [Note $H^{m_f}_{n^*_{f+1}}(0, 1)$ has been calculated in the previous step.] In general, to calculate $n_i{}^*$ $(i>f)$ we combine queues $f, f+1, \cdots, i-1$ in to one queue and apply the equations of the section on two priority classes with $(m_f+\overset{*}{n}_{f+1}+\cdots+\overset{*}{n}_{i-1})$ and n_i replacing m_1 and n_2, respectively. The value of n_i that satisfies (25) is $n_i{}^*$. [Note $H^{m_f}_{n^*_{i-1}}(0, 1)$ and $\overset{*}{n}_{i-1}, \overset{*}{n}_{i-2}, \cdots, \overset{*}{n}_{f+1}, m_f$ have been calculated previously.]

$M\geqq 2$ Priority Classes: Nonpreemptive Discipline

The generalization in this case is not simple. It is easy to show that the optimal policy is a stationary nonrandomized one and let us assume that it is of a control-limit type. The index of the highest-priority active queue f is determined as the smallest integer that satisfies

$$u \geqq \theta_f + 2c/\mu. \tag{27}$$

We will demonstrate the difficulties in the generalization of this case through a simple example. Let us consider a special case where there are four active queues, namely, queues numbered $f, f+1, f+2$, and $M=f+3$. An arrival who observes $s=(0, X_{f+1}, n_{f+2}^*, n_{f+3}^*)$ knows that the server attends to a customer who belongs to the $(f+3)$th queue. However, one who meets $s=(0, X_{f+1}, n_{f+2}^*, n_{f+3}^*-1)$ does not know whether the currently served customer belongs to the $(f+2)$th or $(f+1)$th queue. Moreover, the calculation of the probabilities that specify the queue that the currently served customer belongs to seem to be difficult, and is left for further research. Since the maximum effect of the existence of a customer in a lower-priority queue [in our example, the $(f+2)$th or the $(f+3)$th queues] on the expected service cost of a newly arrived customer is equivalent to adding an additional customer ahead of him in his class [in our example, the $(f+1)$th], good approximations may be obtained by using the same approach as in the previous case. To overcome this difficulty we may assume that the server (operator) informs an arrival to which queue the currently served customer belongs (an assumption that is not unrealistic).

Optimal Pricing for the Service Station

As was mentioned previously, the service station is a profit-making organization that collects money through the toll fees. Its objective is to determine a set of prices $\boldsymbol{\theta}=\{\theta_i: i=1, 2, \cdots, M, \ \theta_i<\theta_j \text{ iff } i>j\}$ so as to maximize its average income per unit time. However, any change in the set of toll fees causes an immediate change in the customer's behavior. For instance, increasing the level of the toll fees, i.e., retaining the same difference $\theta_i-\theta_{i+1}$ for $i=1, 2, \cdots, M-1$ but increasing θ_M, affects the calculation of the set of optimal control limits through the values of f and m_f. Under the assumption described previously, it is clear that, for any set of toll fees, the number of customers in the system is distributed as in an $M/M/1/N_f$ model, where

$$N_j=\textstyle\sum_{i=j}^{i=M} n_i^*, \qquad (j=f+1, f+2, \cdots, M) \tag{28}$$

and

$$N_f=N_{f+1}+m_f, \tag{29}$$

where f is given by (23). Thus, the maximum number N_f of customers in the system is a function of the u and $\boldsymbol{\theta}$. (In the case of nonpreemptive regime and $M=2$, the maximum number of customers in the system is $m_1+n_2{}^*-1$.)

A customer who arrives at an empty station joins the Mth queue (lowest priority). This policy is followed until a newly arrived customer who observes $n_M{}^*$ customers, all of whom are in the Mth queue, decides to join the $(M-1)$th queue. Now, an arrival who observes more than $n_M{}^*$ but less than $n_M{}^*+n_{M-1}^*$ customers in the system joins the $(M-1)$th queue. A newly arrived customer who observes n_{M-1}^* customers in the $(M-1)$th queue ($n_M{}^*+n_{M-1}^*$ customers in the system) decides to join the $(M-2)$th queue, and so on. An arrival who observes m_f customers in the fth queue chooses to balk and looks for service elsewhere. Each of the customers who leaves (balks) the system without being served damages the reputation of our service station. We assume that this damage amounts to ζ monetary units.

In an $M/M/1/N_f$ queue the stationary probability of there being x customers is

$$p_x=\rho^x(1-\rho)/(1-\rho^{N_f+1}), \qquad \rho=\lambda/\mu. \qquad (x=0, 1, 2, \cdots, N_f) \quad (30)$$

It should be noted that ρ may assume any positive value, and in all cases a unique nonzero distribution $\{p_x\}$ exists.

The expected number of customers who balk per unit time is

$$\lambda p_{N_f}=\lambda\rho^{N_f}(1-\rho)/(1-\rho^{N_f+1}). \qquad (31)$$

Hence, the expected loss per unit time due to the damage to reputation is $\zeta\lambda p_{N_f}$.

The service station's net income per unit time z is composed of the expected income per unit time gained through the toll fees paid by joining customers minus the expected loss per unit time due to the damage of reputation caused by balking customers. Hence, the service station is interested in maximizing

$$z=\lambda\theta_M\sum_{x=0}^{N_M-1} p_x+\lambda\theta_{M-1}\sum_{x=N_M}^{N_{M-1}-1} p_x+\cdots+\lambda 6_f\sum_{x=N_{f+1}}^{N_f-1} p_x-\zeta\lambda p_{N_f}. \quad (32)$$

The process is iterative: the service station determines an initial set of toll fees $\boldsymbol{\theta}^{(0)}=\{\theta_i^{(0)}: i=1, 2, \cdots, M, \theta_i^{(0)}<\theta_j^{(0)} \text{ iff } i>j\}$. The set of toll fees and the customers' cost constraint determine the number of active queues $M-f+1$, and the maximum number of customers in the highest-priority active queue m_f. For this set of prices the customers calculate a set of optimal control limits, $n_{f+1}^{*(0)}, n_{f+2}^{*(0)}, \cdots, n_M^{*(0)}$. Under these optimal control limits, the service station changes the set of toll fees so that (32) is maximized. The customers, in view of the new set of toll fees, calculate a new set of optimal control limits, which in turn affects (32), and so on. The process is repeated iteratively until, hopefully, a 'saddle point' is reached—both the customers and the service station cannot improve their positions.

To demonstrate the behavior of the optimal policies of both parties, the customers and the service station, we now analyze in detail a simple numerical example. The general theoretical treatment concerning the interesting questions of convergence of the procedure and of existence and uniqueness of the solutions remains open for further research.

A Numerical Example

We consider a service station with two priority queues ($M=2$) where the preemptive-resume discipline is obeyed. The cost of a unit time spent by a customer in the system is taken to be one monetary unit, i.e., $c=1$. Hence, all other parameters (u, θ_1, θ_2, ζ) are measured in these units. We assume $u\geqq\theta_1+c/\mu$, so that both queues are active. Substituting (28) through (31) in (32) yields

$$z=\lambda[\theta_2(1-\rho^{n_2^*})+\theta_1\rho^{n_2^*}(1-\rho^{m_1})-\zeta(1-\rho)\rho^{n_2+m_1}]/(1-\rho^{n_2^*+m_1}). \qquad (33)$$

The objective of the service station is to find a set of toll fees $\theta=\{\theta_1, \theta_2\}$ that maximizes z. On the other hand, the objective of the customers is to find the control limit n_2^* that minimizes their expected service cost. Note that n_2^* is a function of θ. Since we deal with a system with only two queues, we term the first and the second queues as the high- and the low-priority queues, respectively.

Figures 1, 2, and 3 illustrate n_2^* as a function of θ_1 for different values of ρ, u, and θ_2, respectively. The graphs obtained agree with our intuition. As we increase θ_1, while all other parameters are fixed, n_2^* is increased in steps. Since $H_{n_2+1}^{m_1}(n_2, n_2+1)$ is an increasing function of ρ and u, then owing to (12) and (11) for a fixed value of θ_1, the higher ρ or u the lower the optimal control limits n_2^*. In view of (12) and (11), the influence of θ_2 and n_2^* is clear. Since we deal with the case where both queues are active, the maximum value θ_1 may assume is $\theta_1 = u - 1/\mu (c=1)$.

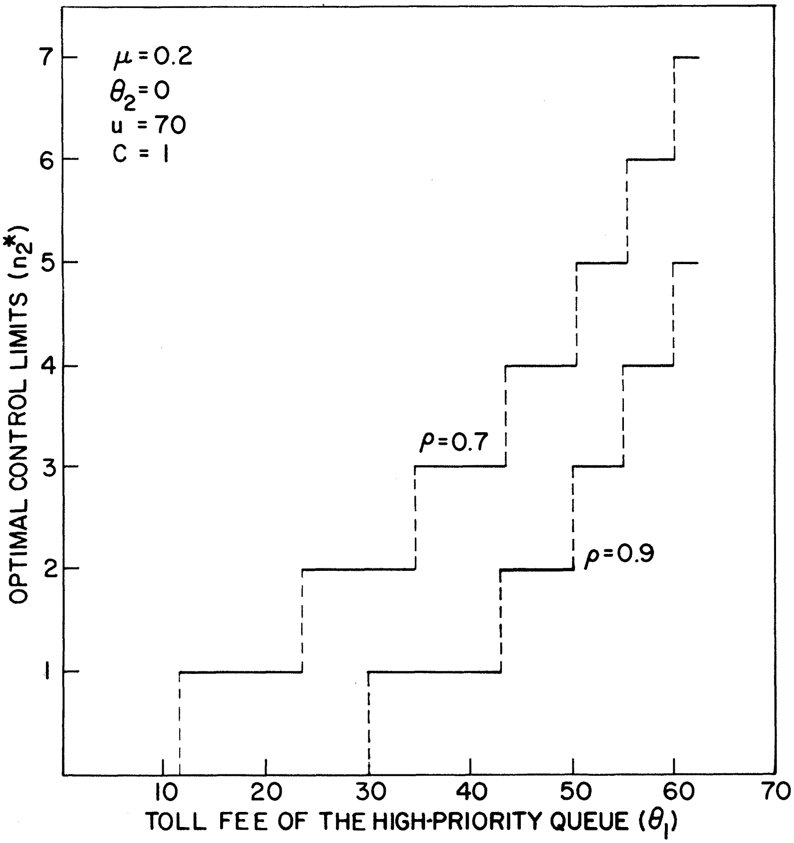

Fig. 1. Optimal control limits as a function of the toll fee of the high-priority queue (for different values of traffic intensity).

For $\zeta = 0$, the values of θ_1 and θ_2 that maximize z, the expected net income per unit time of the service station, are $\theta_1^* = 60$ and $\theta_2^* = 51.4$. The maximum value of z is $z^* = 8.06$. Figure 4 illustrates, for $\theta_2^* = 51.4$, the fluctuations of z as a function of θ_1. The sawteeth are due to changes of m_1 or n_2^*. Since $\theta_1 > \theta_2$, the minimum value of θ_1 is $51.4+\epsilon$, where ϵ is a small positive number. m_1 is changed from 3 to 2 at $\theta_1 = 55$, from 2 to 1 at $\theta_1 = 60$. On the other hand, at $\theta_1 = 59.95$, n_2^* is changed from 0 to 1. With $\theta_1 = 59.95$, the expected net income per unit time is 8.056, and,

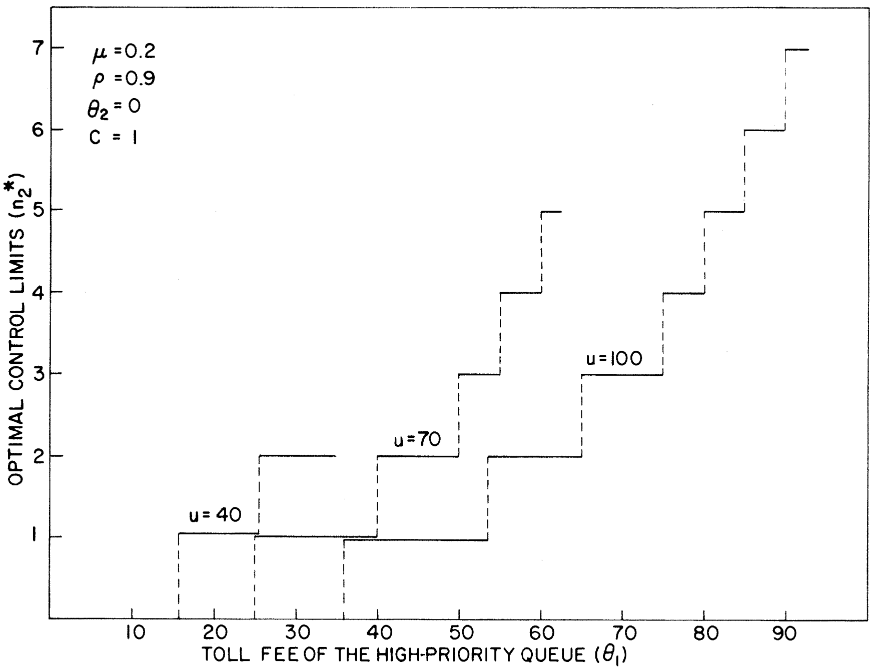

Fig. 2. Optimal control limits as a function of the toll fee of the high-priority queue (for different values of the cost constraint).

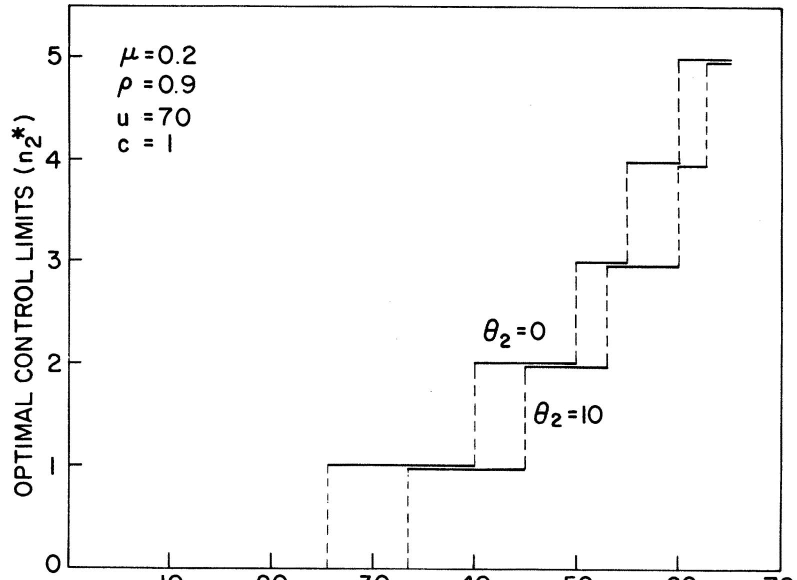

Fig. 3. Optimal control limits as a function of the toll fee of the high-priority queue (for different values of the toll fee of the low-priority queue).

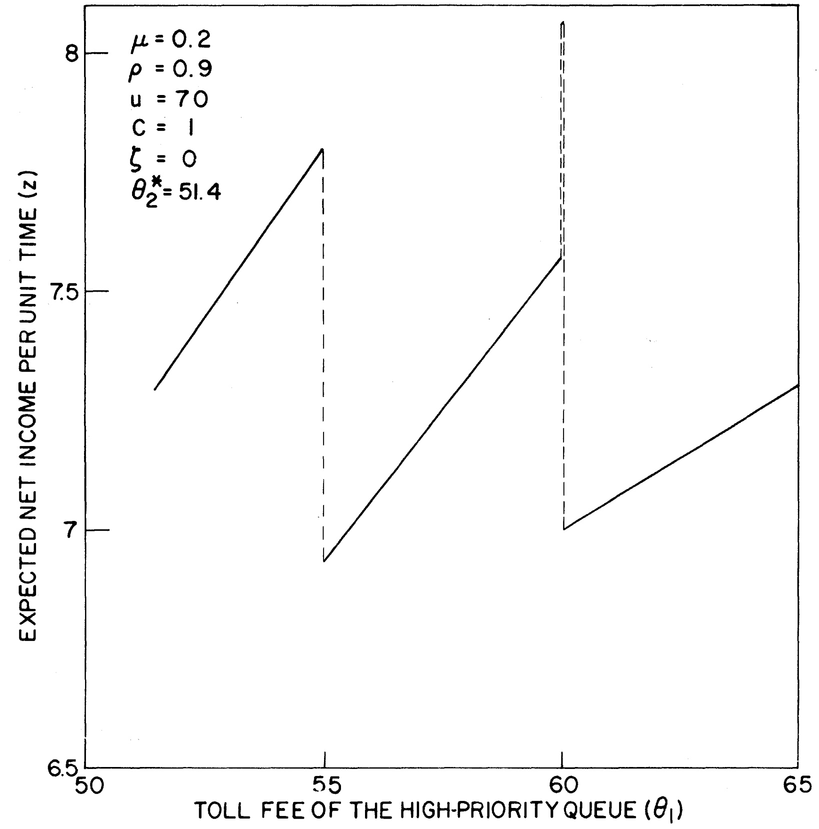

Fig. 4. The expected net income per unit time as a function of the toll fee of the high-priority queue.

just before the change of m_1 from 2 to 1 at $\theta_1=60$, we have $z=z^*=8.063$. Hence, the 'saddle point' is at $\theta_1^*=60$, $\theta_2^*=51.4$, and $n_2^*=1$ when $z^*=8.063$.

Table I shows the influence of ζ on z^*, the optimal expected net income per unit time, as well as on θ_1^*, θ_2^*, and n_2^*. The decrease of the values of z^*, θ_1^*, and θ_2^* with the increase of ζ agrees with our intuition.

TABLE I

ζ	z^*	Optimal policies		
		θ_1^*	θ_2^*	n_2^*
0	8.063	60	51.4	1
20	7.30	60	51.4	1
50	6.35	55	42.8	1
100	5.01	50	34.5	1
200	2.97	45	26.5	1
300	1.30	35	11.5	1

$u = 70$, $c = 1$, $\rho = 0.9$, $\mu = 0.2$.

CASE II: MONOPOLY

In this case a newly arrived customer does not have the option of leaving the system, no matter how high the service cost is; i.e., there is no alternative way to obtain service but in our station. Mathematically this situation is described by letting u be infinite. A necessary and sufficient condition for the system to reach steady state is $\rho=\lambda/\mu<1$. All the proofs regarding the optimality of the control limit policies remain unchanged. The only changes that need to be made are in the calculations of the control limits $\{n_i^*\}$ and m_f. These changes are easily made. First, it is clear that now $f=1$ and $m_1=\infty$. The results for the preemptive-resume discipline are obtained by letting m_1 go to infinity in (18) and (19):

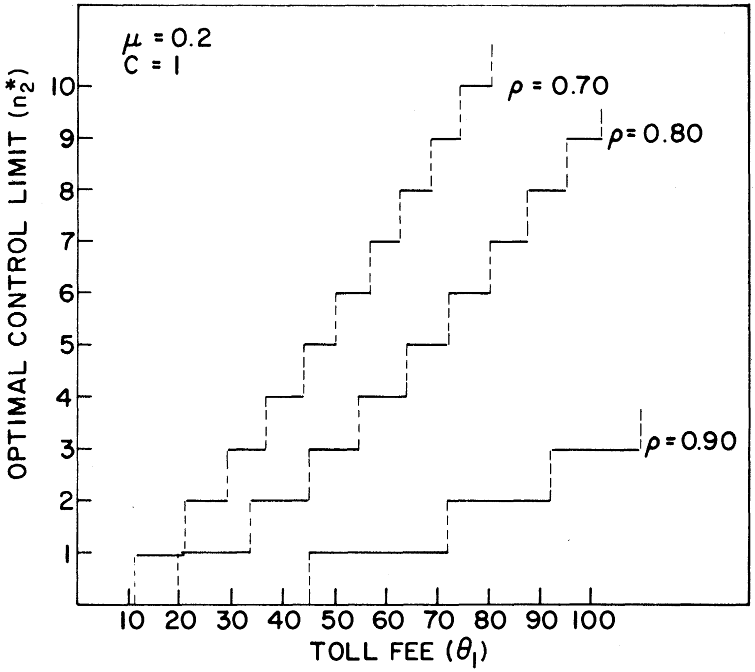

Fig. 5. Optimal control limits as a function of the toll fee (for different values of traffic intensity).

$$H_{n_2}(q,j)=1(\mu+\alpha_{n_2-j+1}[(1+\rho)/\mu(1-\rho)+H_{n_2}(q-1,n_2-1)] + \textstyle\sum_{k=0}^{n_2-j} H_{n_2}(q-1,j+k+1)a_k, \qquad (0<q<j\leqq n_2) \tag{34}$$

and

$$H_{n_2}(0,j)=1/\mu+\alpha_{n_2-j+1}(1+\rho)/\mu(1-\rho). \qquad (j=1,2,\cdots,n_2) \tag{35}$$

Since we are dealing with a preemptive-resume regime where no time losses are involved, (34) is also true for the nonpreemptive discipline, while the initial values are given by (21). [Note that the difference between the preemptive and the nonpreemptive monopoly case is represented by the difference in the initial values of $H_{n_2}(0,j)$ as given by (35) and (21), respectively.]

The procedures for obtaining the control limits are identical to the ones specified in the sections on the nonmonopoly case.

As for the optimal pricing, the service station's objective is to maximize

$$z=\lambda\theta_M\sum_{x=0}^{N_M-1} p_x+\lambda\theta_{M-1}\sum_{x=N_M}^{N_{M-1}-1} p_x+\cdots+\lambda\theta_2\sum_{x=N_3}^{N_2-1} p_x+\lambda\theta_1\sum_{x=N_2}^{x=\infty} p_x, \quad (37)$$

where, in this case, the steady-state probabilities distribution $\{p_x\}$ is the same as in

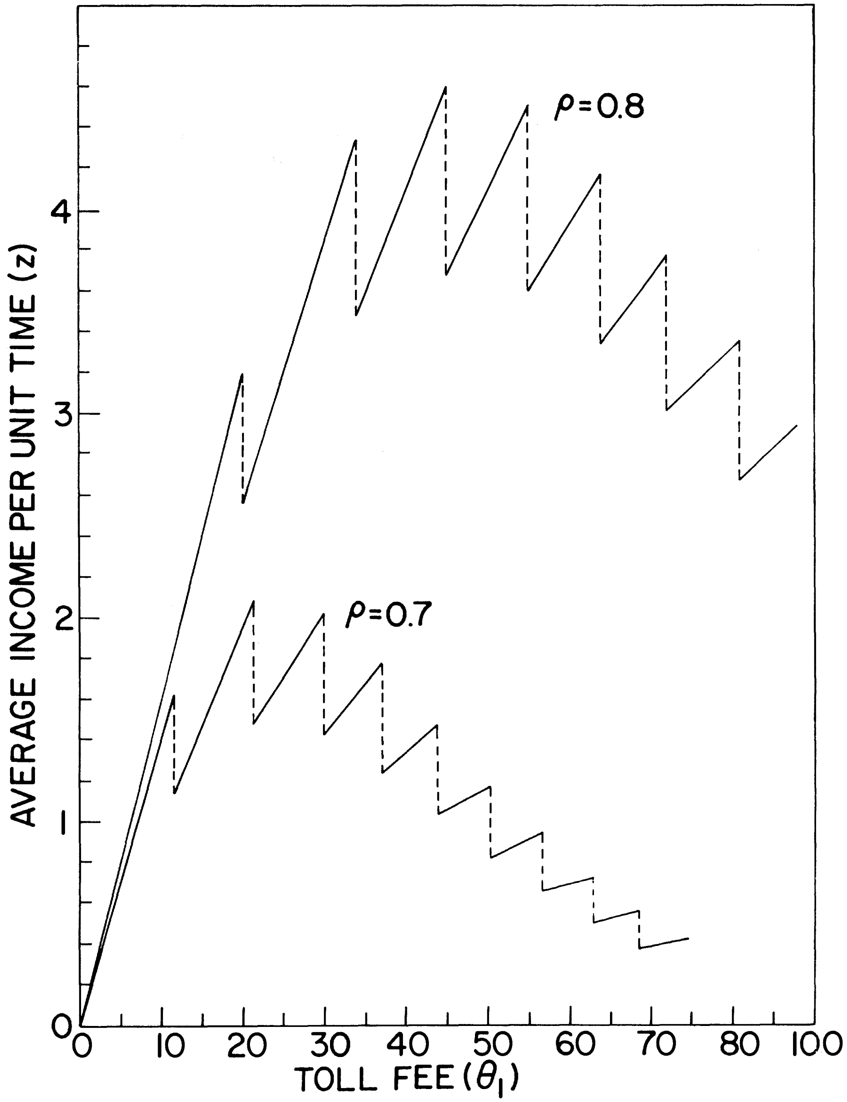

Fig. 6. Average income per unit time as a function of the toll fee (for different values of traffic intensity).

an $M/M/1$ queue,

$$p_x=(1-\rho)\rho^x. \qquad (x=0,1,2,\cdots) \quad (38)$$

For the discussion to be meaningful, we assume that θ_M is a bounded constant and that its value is determined by some other considerations and is not under the control of the service station. (Otherwise, we set $\theta_i=\infty$ for all i.)

As an example, we consider the preemptive-resume regime with $M=2$ priority classes. The station's objective is to find θ_1 so as to maximize

$$z=\lambda\theta_2+\lambda(\theta_1-\theta_2)\rho^{n_2^*}. \qquad (\theta_2 \text{ constant}) \quad (39)$$

For simplicity in obtaining numerical examples (and with no loss of generality) we let $\theta_2=0$ and $c=1$. The numerical results for $M=2$ are given in Figs. 5 and 6.

Figure 5 illustrates n_2^* as a function of θ_1 for different values of $\rho(1/\mu=5$ time units). For example, for $\theta_1=50$ we have the optimal control limits 5, 3, and 1 for traffic intensity 0.7, 0.8, and 0.9, respectively. Figure 6 illustrates the average income per unit time z as a function of the toll fee θ_1, equation (39). We have a unique saddle point. In our example, for $\rho=0.7$, $1/\mu=5$ time units and $\theta_1=21.5$, the optimum control limit is $n_2^*=1$ and $z=2.1$, and for any other value of θ_1 the optimal control limit is such that z gets a smaller value. For $\rho=0.8$, $1/\mu=5$ time units, z gets its optimal value $(z=4.6)$ when $\theta_1=45.0$ and $n_2^*=2$.

ACKNOWLEDGMENTS

THIS RESEARCH WAS partly carried out while I. ADIRI was with the IBM T. J. Watson Research Center and while U. YECHIALI was at New York University.

REFERENCES

1. I. ADIRI AND U. YECHIALI, "Optimal Pricing and Priority Purchasing Policies in Queues," IBM Research Report RC-3581. Also, Operations Research, Stat. and Econ. Mimeograph Series No. 95, Fac. of Ind. and Manag. Eng., Technion—Israel Institute of Technology, Haifa, Israel.
2. ———, "Optimal Decision Policies in a Nonmonopoly Service Station," IBM Research Report RC-3718. Also, Operations Research, Stat. and Econ. Mimeograph Series No. 96, Fac. of Ind. and Manag. Eng., Technion—Israel Institute of Technology, Haifa, Israel.
3. K. R. BALACHANDRAN, "Purchasing Priorities in Queues," *Management Sci.* **18,** 319–326 (1972).
4. L. KLEINROCK, "Optimal Bribing for Queuing Position," *Opns. Res.* **15,** 304–318 (1967).
5. P. NAOR, "On the Regulation of Queue Size by Levying Tolls," *Econometrica* **37,** 15–24 (1969).
6. J. RIORDAN, *Stochastic Service Systems*, Wiley, New York, 1962.
7. U. YECHIALI, "On Optimal Balking Rules and Toll Charges in the *GI/M/*1 Queuing Process," *Opns. Res.* **19,** 349–370 (1971).
8. ———, "Customers' Optimal Joining Rules for the *GI/M/s* Queue," *Management Sci.* **18,** 434–443 (1972).