Included in the user's price-priority decision is his cost of delay.

Briefly discussed for the general case, and shown in detail for the two-queue case, is the principle that the total cost to users of the computing service is minimized in a computing installation that has a price-priority service policy.

Computing center optimization by a pricing-priority policy

by S. B. Ghanem

Computer installation management typically faces the problem of distributing available computing resources among competing users or projects.[1] Whenever the current demand for computer service exceeds the current capacity to provide that service, queues of unsatisfied demand form. Excessive queuing time is costly, whether it is a social cost or the cost of idle employees. It has been difficult to formulate optimal policies to reduce the losses that occur in waiting lines.

This paper proposes the modeling and studying of an optimal operating policy for a service facility that is subject to queuing. The proposed scheme would levy admission tolls at several different priority classes. Higher tolls are charged for priorities with lower average waiting time and vice versa. The reason for suggesting a pricing system for priority allocation is that the demand for service varies in its urgency from one user to another. Some users experience high cost or great inconvenience if their computing job is not done promptly. Others can wait longer at a little cost. Users who place a high value on their time are more likely to choose priorities that result in shorter average delay, and they expect to pay higher tolls to join these classes. By setting different charging rates and by providing the necessary information, users are encouraged to weigh the relative values of services to them before picking priorities for their jobs. Hence, the pricing scheme is used as a control mechanism to guide users toward the correct decisions.

At present, user charges in some computer centers are either heuristic or based on average-cost principles.[2] Average-cost pricing means that such computing facilities yield zero profit, but produce a misallocation of computer time[3] because high tolls are charged during low-use periods, and low tolls are charged during high-use periods. It is also a common practice to assign priorities administratively to requests in a way that is often arbitrary and, thereby, ignores the effect of the priority system on computer efficiency and other factors. The pricing schemes in common use give the user little or no choice in the priority to which his job is assigned. Accordingly, the user has no means to signal urgency in having his job completed, even if he is willing to pay for his urgency. Other computation centers have instituted a differential pricing scheme, but their approach is heuristic. In any case, it is common practice to assign service prices either on average-cost principles or on administrative decisions, neither of which includes the cost of delay.

One wonders whether an optimal pricing system that includes the cost of delay can be used to achieve an optimal priority allocation among competing users. Mechanisms have been proposed for optimizing service facilities subject to queuing limitations. Cox and Smith[4] and others have considered the case in which users are divided into groups, each group having the same cost rate. The authors minimize the total time-averaged cost and develop the "c/t rule." This rule indicates that if c_i is the *cost rate for the* ith *unit waiting for service* and t_i is its *expected service time*, then the unit with highest c_i/t_i should be served next.

Although users do not intentionally degrade the service of others, they are often accidentally inconsiderate. Hence, when using the c/t rule, users are better off by overstating their cost rate. Kleinrock[5] has proposed a model in which the relative position in queue is decided according to the customer's bribe size. Naor[6] has proved that for a single queue, if the newly arriving customer observes the queue size, then an optimal allocation of resources cannot be achieved unless tolls are imposed. In his model, he assigns a constant value to time for all users, which is a gross simplification. Wirt[7] has concentrated on the demand model, which is sensitive to both price and quality of service. He uses a simulation approach to find the optimal prices. Merchand[8] has formulated a general equilibrium model that enables him to state the first-order conditions that the price and capacity should satisfy to be optimal. An alternative method, which is discussed in this paper, considers the cost of delay for the optimal scheduling of jobs using a pricing scheme. General principles are developed for this model with an arbitrary number of priority queues, and the case of two priority queues is discussed in detail.

Concepts of the general model

We begin by analyzing concepts of optimal allocation of priorities through a pricing scheme. Real-time pricing systems are ruled out in this model because of the impracticality of continually fluctuating prices. We are concerned rather with determining the constant price to charge over some time interval. At the end of that interval, prices can be adjusted. In this model, the cost per unit time delay for the group of users is a random variable c, with probability density function $f(c)$, which can vary from one installation to another according to the types of users, and the importance of the service to them.

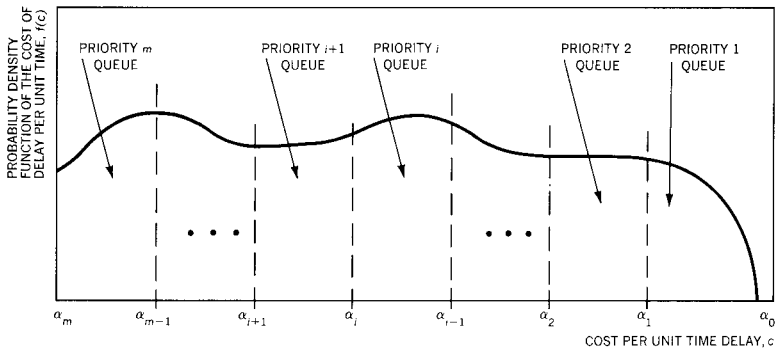
assigning priorities

A brief discussion is given later in this paper of possible techniques for estimating $f(c)$. At this point, consider the following problem: With a fixed number of priority classes, how should one assign priorities to customers so as to minimize the total expected cost of delay of all customers who use the service system? The objectives of our model are as follows:

- Develop optimal priority purchasing policies for the arriving customers, each of whom is free to decide the priority assigned to his job, so as to minimize his total cost function. That is, he compares the toll charged to join each priority queue plus the cost of delay and selects the minimum.
- Develop the optimal pricing policy for a batch computing installation for which the installation manager tries to achieve an overall optimality for the group of users. His objective is to minimize the total expected cost of delay, given the capacity limitations of his installation.
- Develop a pricing scheme in which self-optimization can lead to overall optimization.
- Investigate the adjustment of admission tolls during periods with high traffic intensity (high demand periods).

In our model, the arrival process is assumed to be a homogeneous Poisson process[9] with average arrival rate λ. The single server facility has an arbitrary service time distribution with mean service time $E(s)$ seconds. The computing system consists of m separate queues, the ith queue ($i = 1, 2, \cdots, m$) has priority over the jth queue, if and only if $i < j$. We use the nonpreemptive priority discipline developed by Cobham.[10] We also assume that the arrival time, the service time, and the cost per unit time delay are all independent random variables for each customer, and are independent of the values chosen for all other customers. The higher the priority of the queue, the higher the admission toll charged as an entrance fee for it, but the shorter the average waiting time before accessing the server. On joining the ith queue, a customer pays an admission toll x_i monetary units, where $x_i > x_j$ if $i < j$. We shall deal only with unsaturated queues where

Figure 1 Cost separation points for m priority queues



the traffic intensity is less than one ($\rho < 1$). This is a necessary and sufficient condition for the system to reach steady state. In our model, the waiting cost of a request is assumed to be a linear function of the waiting time. In Theorem 1 (in Appendix 1) we prove that the priority given to a user's job increases with the increase of his cost per unit waiting time. The *traffic intensity* ρ is the ratio of the mean service time and the mean interarrival time.

price per unit delay boundaries

Since priority increases with the cost of unit waiting time, for two priority queues there exists a price per unit time delay α_1 such that a pricing system should be established in a way to guarantee that all arriving customers with $c \geq \alpha_1$ join queue 1, and those with $c < \alpha_1$ join queue 2, where queue 1 has higher priority than queue 2.

Figure 1 illustrates these conditions in a computing installation in which there are m priority-level queues, such that $\alpha = \{\alpha_i : \alpha_1 > \alpha_2 \cdots > \alpha_{m-1}\}$, where α_i is the separation point between queue i and queue $(i + 1)$. In such an installation, the pricing system should motivate the arriving customers with $\alpha_i \leq c < \alpha_{i-1}$ to join the queue with priority i, and those customers with $\alpha_{i+1} \leq c < \alpha_i$ to join the queue with priority $(i + 1)$. Notice that priority i is higher than priority $(i + 1)$.

If we can determine the optimal values of α_i, where $(i = 1, 2, \cdots m - 1)$, then the optimal ratio of customers who should join the different priority queues will be known. A set of admission tolls $x = \{x_i : x_1 > x_2 > \cdots > x_m\}$, where x_i is the admission toll at priority i and $x_i > x_j$ if and only if $i < j$, is calculated and announced, together with the announcement of the expected waiting time at each priority queue. The higher the admission toll charged, the higher the priority and the lower the expected waiting at this priority. That is, $E(W_i) < E(W_j)$ if and only if priority i is higher than priority j, where $E(W_i)$ and $E(W_j)$ are the expected waiting time at priority i and j respectively.

optimal priority queue assignment criterion

A newly arrived customer makes an irrevocable decision as to the queue to which he assigns his job. That decision is made so as to minimize his total expected cost function, which is the toll charged at a certain priority and the cost of waiting at that priority. Thus the arriving customer bases his decision on self-optimization. Our pricing system has been designed so that self-optimization leads to an overall optimization for the whole group of users.

Appendix 2 and Figure 1 give the sum over all priorities of the mean waiting-time costs at each priority H. To establish an optimal priority allocation, the total expected cost of delay H should be minimized with respect to the cost separation points α_i. As a result of the minimization process, the α_i values can be obtained, as well as the proportion of customers who should join each priority queue and the mean waiting time at each queue. Note that Appendix 2 does not show the method for minimizing H, but merely shows the computation of H and asserts that its minimization is the desired criterion or goal for optimal priority queue assignment.

optimal admission tolls

Assume that users assign their jobs to priority queues so as to minimize their total cost function (i.e., tolls to join a certain priority queue plus the cost of delay at that queue). We now seek a pricing system that motivates the users to minimize their total mean cost of waiting. Since α_i is the optimal separation point between queue i and queue $i+1$, the optimal proportion of users who should join the different queues is already known. To encourage the rational customer to behave according to this optimal policy, each user who arrives at the facility with $c = \alpha_i$ must have his total mean cost function for joining queue i equal to that for joining queue $i+1$. That is, he should be indifferent to the choice between queue i and queue $i+1$. Let

x_i = admission toll charged at priority i

and

x_{i+1} = admission toll charged at priority $i+1$;

then

$$x_i + \alpha_i E(W_i) = x_{i+1} + \alpha_i E(W_{i+1}) \text{ for } i = 1, 2, \cdots, m-1. \quad (1)$$

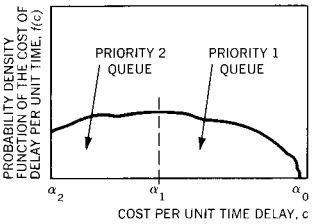
$E(W_i)$ and $E(W_{i+1})$ are the expected waiting times at priorities i and $i+1$, respectively. As soon as the optimal values of α_i for $i = 1, 2, \cdots, m-1$ are known $E(W_i)$ and $E(W_{i+1})$ can be calculated. From Equation 1 we can write

$$(x_i - x_{i+1}) = \alpha_i[E(W_{i+1}) - E(W_i)] \text{ for } i = 1, 2, \cdots, m-1. \quad (2)$$

Equation (2) specifies the optimal set of admission tolls. Clearly, if toll revenue is used for socially useful purposes, then the proposed imposition of tolls is an optimal procedure.

Optimal policy for a batch computing installation with two priority queues

Figure 2 Two priority queues



We now consider the case of two priority queues in detail. (The analysis of the m-priority queues case is discussed in Reference 11.) Given that the probability density function of the cost of delay per unit time $f(c)$ is any general continuous function, and under the general assumptions of our model, Figure 2 and Equation (6A) of Appendix 2, imply that H is total mean cost of delay for arriving users who join the two priority queues and may be expressed as follows:

$$H = \lambda K \left[\frac{\int_{\alpha_1}^{\alpha_0} cf(c)\,dc}{(1-\rho_1)} + \frac{\int_{\alpha_2}^{\alpha_1} cf(c)\,dc}{(1-\rho_1)(1-\rho)} \right].$$

optimal solution characteristics

For two priority queues, the cost separation point α_1 is chosen such that customers with $c \geq \alpha_1$ join queue 1 and customers with $c < \alpha_1$ join queue 2. The value of the cost separation point α_1 that minimizes the total expected cost of delay satisfies the following relation:

$$\alpha_1 = \frac{M - q_1\rho}{1-\rho_1}, \tag{3}$$

where

M = mean of $f(c)$, Mean of the probability density function of the cost of delay per unit time

$$q_1 = \int_{\alpha_1}^{\alpha_0} cf(c)\,dc,$$

and

$$\rho_1 = \rho \int_{\alpha_1}^{\alpha_0} f(c)\,dc.$$ Traffiic intensity of priority 1 queue

The cost function introduced by using two priority queues is less than the cost function that results from using only one priority. Hence, we assume that the optimal solution is an interior point of the feasible set. A necessary condition for α_1 to be an optimal solution is that $\partial H/\partial\alpha_1 = 0$ at the minimum point. Accordingly, Equation (3) follows.

effect of traffic intensity on cost separation point

Usually, demand for computer service is subject to periodic changes. We propose that the demand cycle be divided into time intervals, each of which has an average traffic intensity ρ. During each time interval, constant charges are levied on arriving users. At the end of each time interval, prices can be adjusted. In this section, we discuss the adjustment of the optimal cost separation point α_1, with a change of traffic intensity ρ.

Adjustment of the pricing system is discussed in the following section.

For two priority queues, the optimal value of α_1 that minimizes the total expected cost of delay is given by Equation (3). To study the behavior of α_1 with the change of ρ, we should examine the following function: $d\alpha_1/d\rho$.

Proof is given in Reference 11 that α_1 is a decreasing function of ρ, that is,

$$\frac{d\alpha_1}{d\rho} \leq 0. \tag{4}$$

Equation (4) is plausible since, for high traffic intensity—high-demand periods—more users tend to join the higher priority queue. This is the case because the average waiting time at the lower priority is quite high. On the other hand, as we discuss in the next section, the pricing system should be adjusted to discourage users from choosing priority 1 for jobs of lesser urgency. Otherwise, most users would choose priority 1 and, thereby, degrade the effect of the priority system.

effect of traffic intensity on the pricing system

For two priority queues, the optimal separation point between priority 1 and priority 2 is given by Equation (3), and by particularizing Equation 2, the optimal admission toll for two priority queues is given by the following equation:

$$x_1 - x_2 = \alpha_1[E(W_2) - E(W_1)]. \tag{5}$$

By using expressions for the mean waiting time at the first priority queue and at some kth priority queue [Equations (4A) and (5A),], Equation (5) reduces to

$$x_1 - x_2 = \frac{\alpha_1 \rho K}{(1-\rho_1)(1-\rho)}. \tag{6}$$

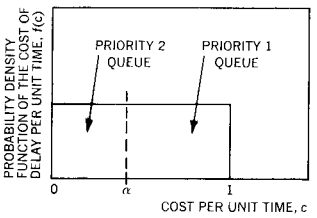
To study the adjustment of the pricing system at periods of high demand, we analyze the behavior of the difference between the tolls charged at priority 1 and 2 queues with changing traffic intensity. With reference to Equation (6), it is proved in Reference 11 that the following equation is valid:

$$\frac{d(x_1 - x_2)}{d\rho} \geq 0.$$

This implies that, for heavy traffic periods, the difference between the tolls charged at priority 1 and 2 queues should be increased. And it agrees with the reasonable policy that at periods with high demand, the toll charged at the higher priority should be increased to discourage nonurgent users from joining it.

We now use two examples to illustrate the functioning of the model developed in this paper; i.e., when the probability density function of the cost per unit time delay $f(c)$, takes a uniform distribution and when it takes an exponential distribution.

Figure 3 Uniform distribution



Uniform distribution. Consider the case illustrated in Figure 3 in which the cost per unit time delay for arriving population is a random variable c with a uniformly distributed probability density function $f(c)$. In this case, the event $c = \alpha$ is the cost per unit delay such that arriving users with $c \geq \alpha$ join priority 1 queue and those with $c < \alpha$ join priority 2 queue. From Appendix 2, the total mean cost of waiting can be written as follows:

$$H = \lambda K \left[\frac{\int_0^{\alpha} c\,dc}{(1-\rho_1)(1-\rho)} + \frac{\int_{\alpha}^{1} c\,dc}{(1-\rho_1)} \right]. \tag{7}$$

Substituting for ρ_1, Equation (7) reduces to the following expression for the total mean cost of waiting for a two-priority queuing system:

$$H_2 = \frac{\lambda K}{2(1-\rho)} \left[1 - \frac{\alpha\rho - \alpha^2\rho}{1-\rho+\alpha\rho} \right]. \tag{8}$$

But since the total mean cost of delay using one priority queue only is $H_1 = \lambda K/2(1-\rho)$, then using two priority queues is the better solution.

This conclusion is true if the gain achieved by introducing the second priority is less than the overhead cost of introducing it. The fractional saving G, obtained by using two priority queues is given by

$$G = \frac{\alpha\rho(1-\alpha)}{1-\rho+\alpha\rho}$$

which may also be expressed as

$$G = \frac{H_1 - H_2}{H_1}.$$

To get the optimal value of cost separation point α, which separates priority 1 queue and priority 2 queue, we should maximize G. This implies that the optimal value of α is

$$\alpha = \frac{(1-\rho)}{\rho} \left[\frac{1}{\sqrt{1-\rho}} - 1 \right]. \tag{9}$$

Also, Equation (6) implies that the optimal difference of admission toll for two priority queues is

$$x_1 - x_2 = \frac{K\,(1-\sqrt{1-\rho})}{1-\rho}. \tag{10}$$

Figure 4 Effects of total traffic intensity (A) On the cost separation point (B) On the traffic intensity at priority 1 queue (C) On the traffic intensity at priority 2 queue

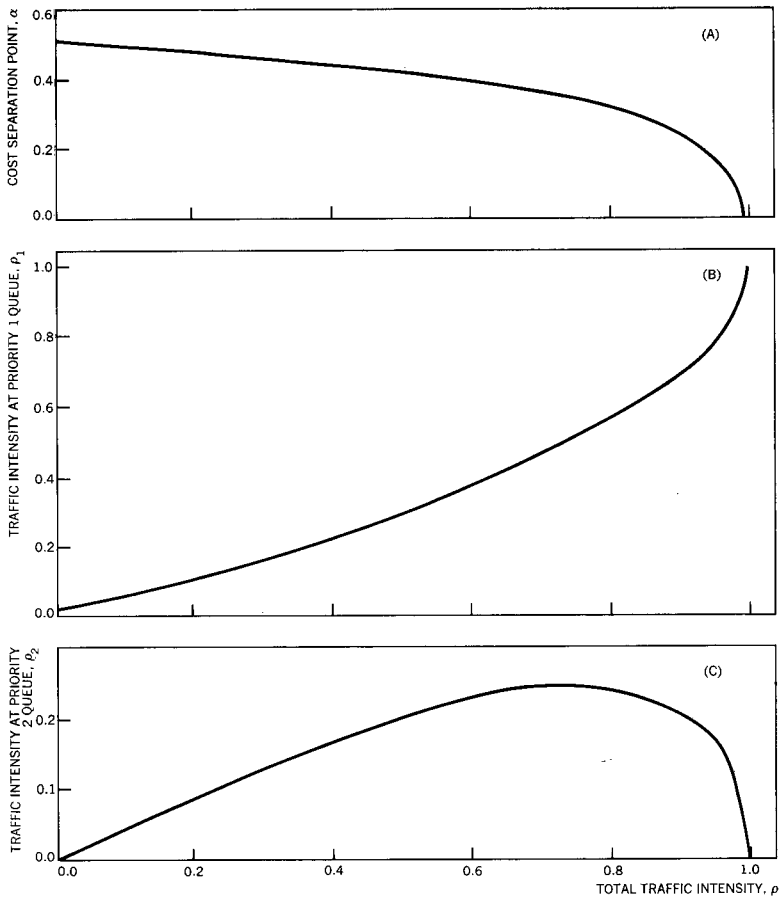
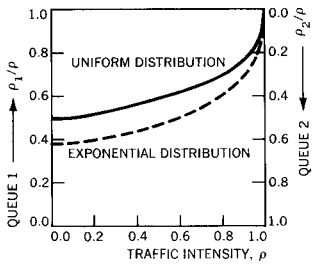


Figure 5 Ratios of queue traffic intensities to the total traffic intensity



Notice that ρ_1 and ρ_2 can be written in the following form:

$$\rho_1 = 1 - \sqrt{1-\rho}; \tag{11}$$

$$\rho_2 = \left[\sqrt{1-\rho} - (1-\rho)\right]. \tag{12}$$

Equations (8) – (12) together with Table 1 describe the relations among the model parameters G, α, $(x_1 - x_2)$, ρ_1, ρ_2, and the total traffic intensity ρ.

Some results are summarized in Figure 4. Figure 4A shows that, as the traffic intensity increases, the cost separation point decreases until the higher priority queue is saturated and there is no cost separation. Figure 4B indicates that traffic intensity ρ_1 of the higher cost queue is an increasing function of that total traffic intensity ρ. Figure 4C shows that the traffic intensity ρ_2 of the lower cost queue increases until it reaches its maximum value at $\rho = \frac{3}{4}$ then it decreases. Figure 5 describes the behavior

Table 1 Model parameters as a function of traffic intensity for the uniform distribution case

ρ	α	ρ_1	ρ_2	ρ_1/ρ	ρ_2/ρ	$(x_1 - x_2)/K$	G
0.1	0.487	0.05	0.05	0.5	0.5	0.0570	0.024
0.2	0.472	0.1056	0.0944	0.528	0.472	0.13197	0.056
0.3	0.456	0.1633	0.1367	0.544	0.456	0.2333	0.089
0.4	0.436	0.2254	0.1746	0.5635	0.4365	0.3757	0.127
0.5	0.414	0.2928	0.2072	0.5856	0.4144	0.5858	0.172
0.6	0.387	0.368	0.232	0.6133	0.3867	0.9189	0.225
0.7	0.354	0.452	0.248	0.6457	0.3543	1.5076	0.292
0.75	0.333	0.50	0.25				
0.8	0.309	0.5527	0.2473	0.691	0.309	2.7639	0.382
0.9	0.240	0.683	0.217	0.7589	0.2411	6.8377	0.519
0.95	0.183	0.776	0.174	0.817	0.183		

Figure 6 Fractional saving achieved by using a second priority queue as a function of traffic intensity

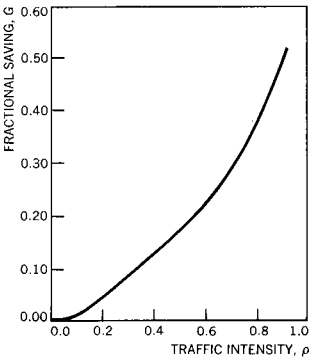


Table 2 Effect of changing the mean of $f(c)$ on the model parameters

Parameter	*Mean*		
	$\frac{1}{2}$	1	$\frac{3}{2}$
α	$\frac{1}{2}y$	y	$\frac{3}{2}y$
ρ_1	$1 - \sqrt{1-\rho}$	$1 - \sqrt{1-\rho}$	$1 - \sqrt{1-\rho}$
$x_1 - x_2$	$\frac{1}{2}z$	z	$\frac{3}{2}z$

of the traffic intensity ratios ρ_1/ρ and ρ_2/ρ with the change of total traffic intensity ρ. The relative proportion of users who join priority 1 queue increases with increasing ρ. These results are reasonable because during periods of high traffic intensity the expected waiting time at priority 2 queue is quite high. Figure 6 indicates that the fractional saving that results from using two priority queues is an increasing function of the traffic intensity ρ. (In Figure 6, the distribution is uniform and the mean M is $\frac{1}{2}$.) Accordingly, using two priority classes is a better solution than using one as long as the fractional saving is higher than the overhead cost that results from introducing the second priority. Figure 7 shows that for this example, where the mean of probability density function of the cost of delay per unit time $f(c)$ is $\frac{1}{2}$, the cost differential constant $(x_1 - x_2)/K$ is an increasing function of ρ. K is a constant given by Equation (5A) that reflects the characteristics of the stream of jobs. This means that during periods of high traffic intensity ρ, the toll charged at priority 1 queue should increase to discourage nonurgent users from joining priority 1 queue. It is interesting to study the changes in the model parameters if $f(c)$ is still uniformly distributed, but the mean M takes different values. The results obtained can be summarized in Table 2,

Figure 7 Cost differences for uniform distributions with different means

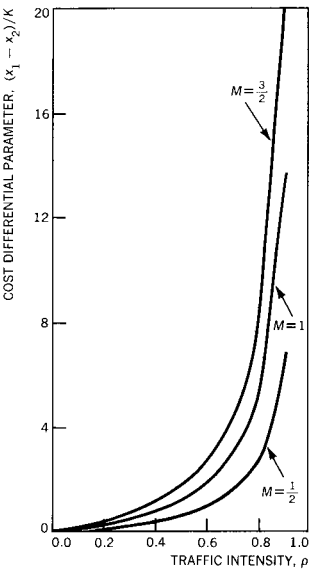
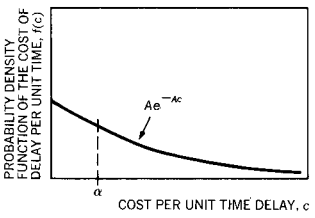


Figure 8 Exponential distribution



where

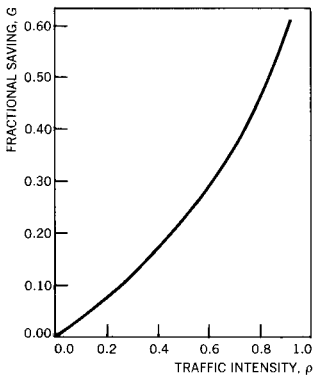
$$y = \frac{2(1-\rho)}{\rho}\left[\frac{1}{\sqrt{1-\rho}}\right]$$

and

$$z = \frac{2K(1-\sqrt{1-\rho})}{(1-\rho)}.$$

We notice that the cost differential constant $(x_1 - x_2)/K$ is directly proportional to the mean of the probability density function of the cost of unit delay $f(c)$. Figure 7 shows that $(x_1 - x_2)/K$ increases as M increases, which is plausibly based on the expectation that, with $f(c)$ more widely distributed, more users will have higher urgency. Hence, the cost differential $(x_1 - x_2)$ should be increased but the traffic intensity ratios ρ_1/ρ and ρ_2/ρ are constant and independent of the mean M. The solid curve of Figure 5 is valid for any uniformly distributed $f(c)$.

Figure 9 Fractional saving as a function of traffic intensity for the exponential case



Exponential distribution. When the probability density function for the cost per unit time delay for arriving users is exponentially distributed, as is illustrated in Figure 8, the total mean cost of delay, H, is given by

$$H = \lambda K\left[\frac{\int_0^{\alpha} cAe^{-Ac}dc}{(1-\rho_1)(1-\rho)} + \frac{\int_{\alpha}^{\infty} cAe^{-Ac}dc}{(1-\rho_1)}\right],$$

which reduces to

$$H = \frac{\lambda K}{A(1-\rho)}\left[1 - \frac{A\rho\alpha e^{-A\alpha}}{(1-\rho)e^{-A\alpha}}\right].$$

Since, however, the mean cost of delay using only one priority queue is $\lambda K/A(1-\rho)$, then the fractional saving G that results from using two priority queues as compared with using only one queue is

$$G = \frac{\rho A\alpha e^{-A\alpha}}{1-\rho e^{-A\alpha}}. \tag{13}$$

Again, we can conclude that using two priority queues is better than using only one queue, if the cost of introducing the second queue is less than the gain. The optimal value of the cost separation point α that maximizes G for the exponential case is given by

$$\alpha^* = \frac{1}{A} - \frac{\rho}{A}e^{-A\alpha^*}. \tag{14}$$

Optimal cost separation points for the exponential case

The traffic intensities for the two queues ρ_1 and ρ_2 can be written as

Table 5 Model parameters as a function of traffic intensity for the exponential distribution case

ρ	α	ρ_1	ρ_2	ρ_1/ρ	ρ_2/ρ	$(x_1 - x_2)/K$
0.1	0.480	0.0383	0.0617	0.383	0.617	0.05546
0.2	0.460	0.0797	0.1203	0.3985	0.6015	0.125
0.3	0.4375	0.1251	0.1749	0.417	0.583	0.2143
0.4	0.412	0.1753	0.2247	0.438	0.562	0.3331
0.5	0.385	0.2315	0.2685	0.463	0.537	0.501
0.6	0.3515	0.2974	0.3026	0.496	0.504	0.7504
0.7	0.3135	0.3739	0.3261	0.534	0.466	1.1683
0.8	0.264	0.4718	0.3282	0.590	0.410	1.199
0.9	0.198	0.6057	0.2943	0.673	0.327	4.519

$$\rho_1 = \rho e^{-A\alpha}; \tag{15}$$

$$\rho_2 = \rho(1 - e^{-A\alpha}). \tag{16}$$

From Equation (6), the difference between the admission toll at priority 1 queue and priority 2 queue is the following:

$$x_1 - x_2 = \frac{\alpha\rho\ K}{(1 - \rho e^{-A\alpha})(1 - \rho)}. \tag{17}$$

From Equations (13) and (14),

$$G = \rho e^{-A\alpha^*} = 1 - \alpha^* A. \tag{18}$$

In Equation (18), the fractional saving G does not depend on A because $\alpha^* A$ is uniquely determined for each ρ, and, since G is equal to $(1 - \alpha^* A)$, it is also uniquely determined for each ρ. The solution of Equation (14) can be obtained from the following relation:

$$\alpha^* A = 1 - \rho e^{-\alpha^* A}.$$

Cost separation constants $\alpha^* A$ are given in Table 3, for ρ having values from 0.1 to 0.9. The fractional saving G that is achieved by using two priority queues beyond that achieved when using one queue only as a function of traffic intensity ρ is summarized in Table 4. From Table 4 and Figure 9, we can conclude that with an increase of the traffic intensity the fractional saving G increases. This indicates that the benefit of the two-priority solution is greater at periods of high traffic intensity than the one-priority solution.

The behavior of α^*, ρ_1, ρ_2, and $(x_1 - x_2)$ with the change of ρ in Equations (14) – (18) is given in Table 5 for $A = 2$. Figures 10 and 11 show that the cost separation point α is decreasing with increasing traffic intensity ρ, and the cost differential constant $(x_1 - x_2)/K$ increases with ρ. Also the traffic intensity ρ_1 at the higher priority queue is increasing with the increase of ρ, and the intensity ρ_2 at the lower priority queue increases until a certain

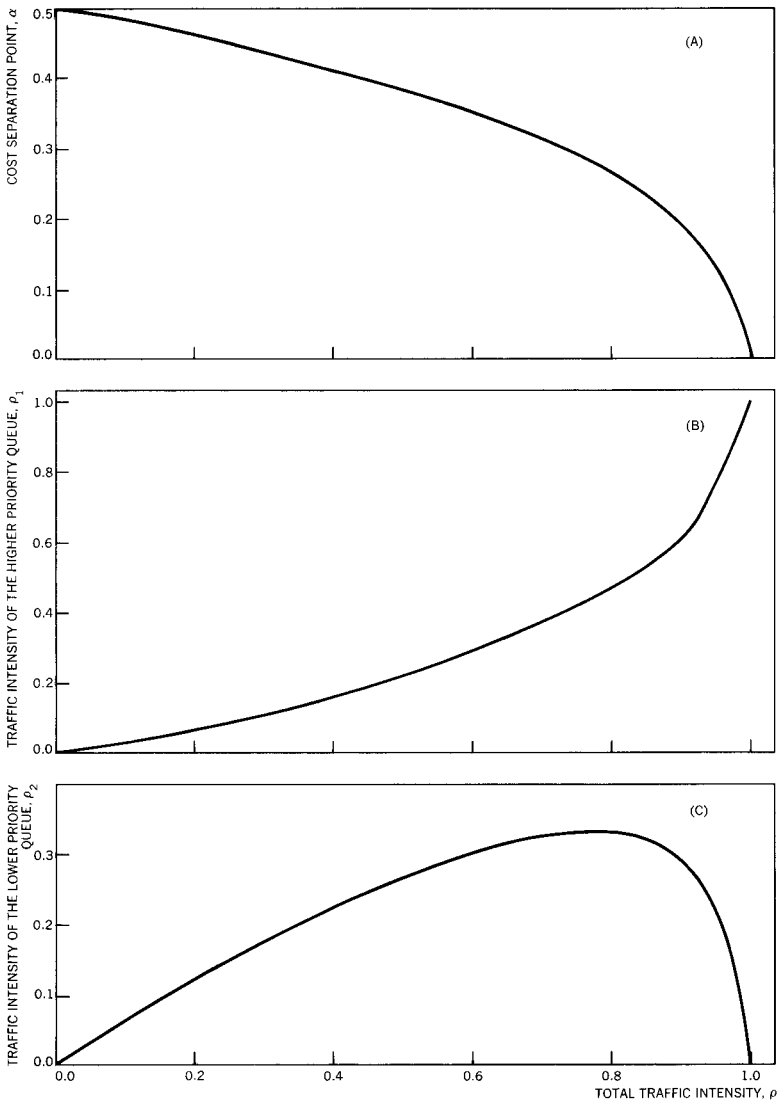
Table 3 Cost separation constants for a series of traffic intensities in the exponential case

ρ	$\alpha^* A$
0.1	0.960
0.2	0.92
0.3	0.875
0.4	0.825
0.5	0.77
0.6	0.702
0.7	0.627
0.8	0.528
0.9	0.396

Table 4 Fractional saving achieved by introducing a second priority queue in the exponential case of a function of traffic intensity

ρ	$G = 1 - \alpha^* A$
0.1	0.040
0.2	0.08
0.3	0.125
0.4	0.175
0.5	0.23
0.6	0.298
0.7	0.373
0.8	0.472
0.9	0.604

Figure 10 Model parameters as a function of traffic intensity for the exponential case with a mean of $\frac{1}{2}$ (A) Cost separation point (B) Traffic intensity of the higher priority queue (C) Traffic intensity of the lower priority queue



level is achieved, then it decreases. In Figure 5, for the exponential distribution case, the dotted curve indicates that ρ_1/ρ is an increasing function of ρ, and ρ_2/ρ is a decreasing function of ρ. The reason for this behavior is the same as that for the uniform distribution case. As in the uniform case, it is also true for exponential distributions that key parameters vary with $f(c)$ having different means, as is shown in Table 6,

where

$$W = 1 - \rho e^{-A\alpha}$$

Table 6 Effect of changing the mean of $f(c)$ on the model parameters

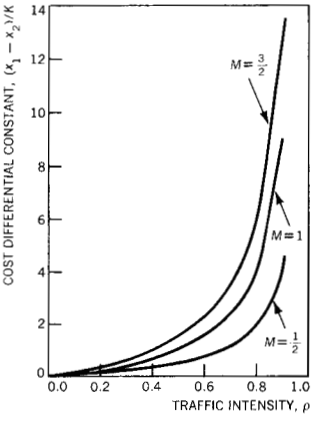
Parameter	Mean $M = \frac{1}{2}, A = 2$	$M = 1, A = 1$	$M = \frac{3}{2}, A = \frac{2}{3}$
$-\alpha$	$\frac{1}{2}W$	W	$\frac{3}{2}W$
ρ_1	$\rho e^{-A\alpha}$	$\rho e^{-A\alpha}$	$\rho e^{-A\alpha}$
$x_1 - x_2$	$\frac{1}{2}\beta$	β	$\frac{3}{2}\beta$

and

$$\beta = \frac{2\alpha\rho K}{(1 - \rho e^{-A\alpha})(1 - \rho)}.$$

Figure 11 shows that the cost differential constant $(x_1 - x_2)/K$ is higher for exponential distributions with higher values of M. Another interesting observation is the relative insensitivity of the optimal pricing policy to different distributions of $f(c)$ with the same mean, especially when $\rho \leq 0.75$.

Figure 11 Cost differential constant and traffic intensity for three values of the mean of $f(c)$

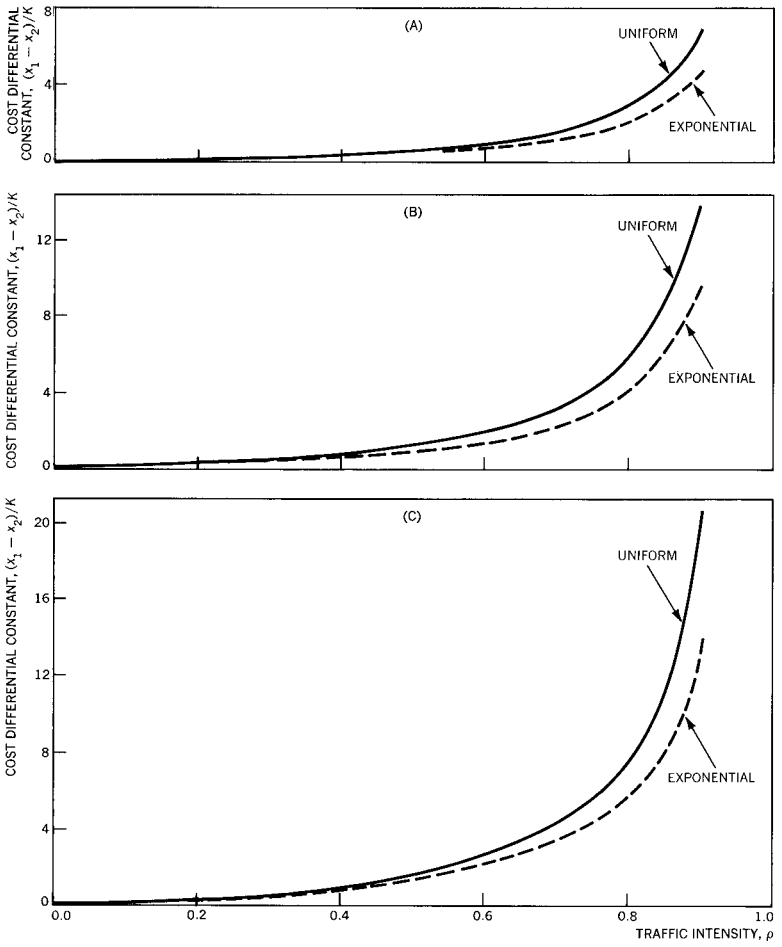


In Figures 12A – 12C the cost differential constant $(x_1 - x_2)/K$ is plotted against traffic intensity ρ for both the uniform and the exponential distributions with the same mean, where M takes the values $\frac{1}{2}$, 1, and $\frac{3}{2}$. The two curves tend to be close to each other as long as $\rho \leq 0.75$. Hence, if we can redistribute the demand to achieve this level of ρ, then it is enough to know an estimate for the mean of $f(c)$. An approximate distribution for $f(c)$ with the estimated average can lead to a near-optimal pricing scheme. An estimate of $f(c)$ can be obtained by observing the behavior of the users. We can start with any arbitrary value of $(x_1 - x_2)$ and $[E(W_1), E(W_2)]$, where $x_1 > x_2$ and $E(W_1) < E(W_2)$. By observing the behavior of the users who join the different priority queues, an estimate of $f(c)$ can be obtained. For a more detailed discussion of methods at estimating $f(c)$, see Reference 12.

Conclusions and extensions

In this paper, a general model for the optimal allocation of priorities through pricing is considered. The case of two priority queues is discussed in detail. (For the m-priority queue analysis, the reader may refer to Reference 11.) In both cases, it is shown that a set of admission tolls can be established at the different priority queues. These tolls are based on user urgency, the job arrival rate, the expected service time, and the number of priority classes. By setting a different admission toll at each priority queue and by providing the user with information and motivation, he is encouraged to weigh the relative values of the services before picking the priority for his job. According to his urgency,

Figure 12 (A) Comparison of cost differential constants for uniform and exponential distributions with a mean of $\frac{1}{2}$ (B) Comparison of cost differential constants for uniform and exponential distributions with a mean of 1 (C) Comparison of cost differential constants for uniform and exponential distributions with a mean of $\frac{3}{2}$



the user minimizes his total average cost function (cost to join a certain priority plus the cost of delay). Pricing is addressed to the allocation of priorities in a computing system to minimize the total average cost of delay of an installation's user community. Each user is free to decide on the priority assigned to his job, and the optimality of the system is maintained. As a result, the computer can process first those jobs that are urgent and then proceed with less urgent jobs. The uncertainty of the cost per unit time delay c is considered in this model. It has been found, for the case of two priority queues, that if you have an estimate of the mean of the probability density function of the cost of delay per unit time $f(c)$, the cost differential $(x_1 - x_2)$ is insensitive to the exact distribution. This result is feasible if we can redistribute the demand such that $\rho \leq 0.75$. During periods with high traffic

intensity it is shown that the admission tolls at the higher priority queues should be increased.

The model can provide answers to important questions that face the system designer. Introducing one more priority queue reduces the total average cost of delay. By comparing the overhead cost that results from introducing an additional queue with the resulting gain, the system designer can determine the optimal number of priorities that should be used. By similar argument, the system designer can also decide whether he is willing to extend the capacity of his facility.

The results of this paper have opened several areas in the study of planning and management of service facilities. Many extensions can be made that are of practical as well as theoretical interest. For example, it is more realistic to assume that the cost of delay is a nonlinear function of the delay. With this assumption, the model becomes more complicated, but it is still worthwhile to consider this extension. Another feasible extension is to combine the cost of delay with the job length in the optimal strategy for priority allocation. The study of the dynamic behavior of this model is also recommended. It is usually observed that during the course of a day, the demand for computer service varies significantly with time. Hence, by allowing the load to change with time, the dynamic behavior can be studied.

ACKNOWLEDGMENTS

The author thanks Professor W. K. Linvill of Stanford University and M. Z. Ghanem and D. N. Streeter of the IBM Thomas J. Watson Research Center for their interest and their valuable comments.

CITED REFERENCES

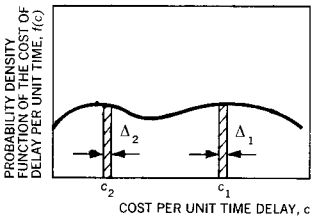
1. D. N. Streeter, *The Scientific Process and the Computer*, John Wiley and Sons, Inc., New York, 1973.
2. R. C. Rettus and R. A. Smith, "Accounting control of data processing," *IBM Systems Journal* **11**, 74–92 (1972).
3. S. Smidt, "Flexible pricing of computer services," *Management Science* **14**, No. 10 (June 1968).
4. R. D. Cox and W. L. Smith, *Queues*, John Wiley and Sons, Inc., New York, 1961.
5. L. Kleinrock, "Optimum bribing for queue position," *Operations Research* **15**, No. 2 (1967).
6. P. Naor, "The regulation of queue size by levying tolls," *Econometrica* **37**, No. 1 (1969).
7. J. Wirt, *Optimization of Price and Quality in Service Systems*, Ph. D. Thesis, Engineering Economic Systems Department, Stanford University (1971).
8. M. G. Merchand, "Priority pricing," *Management Science* **20**, No. 7 (1974).
9. A. O. Allen, "Elements of probability for system design," *IBM Systems Journal* **13**, 325–348 (1974).
10. A. Cobham, "Priority assignment in waiting line problems," *Operations Research* **2**, No. 1 (1954).

11. S. B. Ghanem, *Optimal Admission Tolls for Priority Queues*, Ph.D. Thesis, Engineering Economic Systems Department, Stanford University (1975).
12. R. A. Howard, *Dynamic Probabilitic Systems*, John Wiley and Sons, Inc., New York, 1971.

Appendix 1

Theorem 1 The priority of a given user's job increases with the increase of his cost per unit waiting time.

Figure 1A Relationship of priority level with cost per unit time of delay



Let $f(c)$ be the probability density function of the cost of delay per unit time for the arriving customers, as shown in Figure 1A. Assume that a proportion of customers of size Δ_2, who have a value per unit time delay of $c = c_2$, have joined queue i, and a proportion of customers of size Δ_1, who have $c = c_1$, have joined queue j. If $c_1 > c_2$, then queue j should have higher or at least equal priority relative to queue i for the optimality of this queuing system.

Proof. Assume the opposite, i.e., queue i has higher priority than queue j. Let

r_k = a group of customers whose cost per unit delay is c and who join queue k

and

$E(W_k)$ = mean waiting time for customers who join queue k

Then for m priority levels, the total expected cost of delay $E(c_1)$ is given as follows:

$$E(c_1) = \left[\lambda \int_{r_1} cf(c)\,dc\right] E(W_1) + \left[\lambda \int_{r_2} cf(c)\,dc\right] E(W_2) + \cdots$$
$$+ \left[\lambda \int_{r_i - \{c_2\}} cf(c)\,dc + c_2\Delta_2\right] E(W_i) + \cdots$$
$$+ \left[\lambda \int_{r_j - \{c_1\}} cf(c)\,dc + c_1\Delta_1\right] E(W_j) + \cdots$$
$$+ \left[\lambda \int_{r_m} cf(c)\,dc\right] E(W_m), \qquad (1A)$$

where

$\{c_1\}$ = proportion of customers of size Δ_1 and with $c = c_1$ who join queue j

and

$\{c_2\}$ = proportion of customers of size Δ_2 and with $c = c_2$ who join queue i.

Let

$\Delta_1 = \Delta + \epsilon_1$;

$\Delta_2 = \Delta + \epsilon_2$,

where

ϵ_1 = proportion of customers with $c = c_1$ and

ϵ_2 = proportion of customers with $c = c_2$.

Now, after switching a proportion Δ who have a value of unit time $c = c_1$ from queue j to queue i, and switching a proportion Δ with $c = c_2$ from queue i to queue j, the new total expected cost of delay $E(c_2)$ is given as follows:

$$E(c_1) = \left[\lambda \int_{r_1} cf(c)\,dc\right] E(W_1) + \left[\lambda \int_{r_2} cf(c)\,dc\right] E(W_2) + \cdots$$
$$+ \left[\lambda \int_{r_i - \{c_2\}} cf(c)\,dc + c_2\epsilon_2 + c_1\Delta\right] E(W_i) + \cdots$$
$$+ \left[\lambda \int_{r_j - \{c_1\}} cf(c)\,dc + c_1\epsilon_1 + c_{2\Delta}\right] E(W_j) + \cdots$$
$$+ \left[\lambda \int_{r_m} cf(c)\,dc\right] E(W_m). \qquad (2A)$$

Notice that by this switching procedure $W_1, W_2, \cdots, W_m$ are not affected.

Comparing Equations (1A) and (2A), we can write

$$E(c_2) = E(c_1) + (c_1 - c_2)[E(W_i) - E(W_j)]\Delta. \qquad (3A)$$

But since

$c_1 > c_2$ and $E(W_i) < E(W_j)$

by assumption, then

$E(c_2) < E(c_1)$,

which contradicts the optimality of the system. Thus, queue i should have a lower priority than queue j for the optimality of the system. Notice also that if queue i is at the same level as queue j then $E(W_i) = E(W_j)$, and from Equation (3A), we notice that, in this case, $E(c_2) = E(c_1)$. We conclude, therefore, that queue j should have higher or at least equal priority relative to queue i.

Appendix 2

In the nonpreemptive priority discipline, when a service for a customer starts, it proceeds without interruption until it has been completed. The next customer to be serviced is the one with the highest priority present in the system. Within each class, a FIFO

discipline is observed. The mean (expected or average) waiting time is given by the following equations:

$$E(W_1) = \frac{K}{(1-\rho_1)}$$

= mean waiting time at the first priority queue;

$$E(W_k) = \frac{K}{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)} \qquad k = 2, \cdots, m \tag{4A}$$

= mean waiting time for the customers who join the kth priority queue,

where

$$K = \lambda E(S^2)/2, \tag{5A}$$

ρ_i = Traffic intensity at priority queue i

$$= \rho \int_{\alpha_i}^{\alpha_{i-1}} f(c)\,dc,$$

and

$\int_{\alpha_i}^{\alpha_{i-1}} f(c)\,dc$ is the probability that $\alpha_i \leq c < \alpha_{i-1}$.

Under the assumption that the waiting costs of a request are a linear function of the waiting time, total mean cost of delay H can be written as follows:

H = the sum over all priority queues of the mean waiting costs at each priority queue (6A)

$$= \lambda K \left[\frac{\int_{\alpha_1}^{\alpha_0} cf(c)\,dc}{(1-\rho_1)} + \frac{\int_{\alpha_2}^{\alpha_1} cf(c)\,dc}{(1-\rho_1)(1-\rho_1-\rho_2)} + \cdots \right.$$

$$+ \frac{\int_{\alpha_i}^{\alpha_{i-1}} cf(c)\,dc}{(1-\rho_1-\rho_2\cdots\rho_i)(1-\rho_1-\rho_2\cdots\rho_{i-1})} + \cdots$$

$$\left. + \frac{\int_{\alpha_m}^{\alpha_{m-1}} cf(c)\,dc}{(1-\rho)(1-\rho_1-\rho_2\cdots\rho_{m-1})} \right]$$

where $\rho = \rho_1 + \rho_2 + \cdots + \rho_m$ is the total traffic intensity

and

$f(c)$ = Probability density function of the cost of delay per unit time

The α_i where $i = 1, 2, \cdots, m-1$, are the separation points between the priority queues. Our goal is to minimize H with respect to

α_i. As a result of the minimization process, the values of α_i can be obtained. Accordingly, the proportion of customers who should join the different priority queues, and the mean waiting time at each queue are known.

Reprint Form No. G321-5015