Computer network usage control through peak load pricing

K. Ewusi-Mensah

IBM, Los Angeles Scientific Center, 9045 Lincoln Blvd, Los Angeles, Calif. 90045, USA

The problem of peak load in computer networks and computer centres is discussed. A model is presented which deals specifically with private commercial centres whose main objective is one of profit maximisation under budget and capacity constraints. Kuhn-Tucker conditions are used to determine what each centre's price structure should be under peak load. Three solution possibilities representing different stages of computer centre operation are presented. Problems of data collection and implementation of the results are addressed.

(Received February 1979; revised August 1980)

1. Introduction

A computer network is considered to consist of a network of computer centres interconnected by a transmission or communication system to enable potential users to move data and software freely from one centre to another in the network. Buyers and sellers of computer services are thereby brought together to form a market for the widespread sale and distribution of computer services (Cotton, 1975; Eric, 1975). During the past decade several commercial networks have come into existence in the business world. These include: Telenet, Tymshare (Tymnet), United Computing Services, and General Electric Information Services Division.

While technical problems remain in computer network design, some of the most important issues in the field of network development are associated with economic, legal and societal factors of the level of network management and of individual computer centres (Eric, 1975; Kriebel and Mikhail, 1975). It is in these cases that decisions relating to locations of network nodes and pricing of computer services feature prominently. In this paper we will be interested mainly in the decisions that computer network managers constantly face with respect to the pricing of the services they provide to their customers.

The peak load problem in computer networks can be attributed to the wide cyclical variations in the quantities and types of service demanded by users (Cotton, 1975). This is partly because rates uniform in time give rise to patterns of demand characterised by sharp peaks (De Salvia, 1969; Lehmah, 1972; Sharpe, 1969). The typical trend constantly exhibited is that demand for service is greater during computer prime time than at night. The demand function itself may fluctuate because of variation in need over time among similar users. Thus, a major management objective is to develop a pricing strategy to level the fluctuations in user demands. Two important reasons for this are to:

1. Increase utilisation and efficiency of the expensive system resources by spreading out the usage over a whole period, and
2. Minimise the disutility to users who cannot obtain service in the system at times of peak loading, so as to improve the overall quality of service offered to users.

Current methods in operation have consisted of prescheduling and blocking groups of users. These have been unsatisfactory and resulted in loss of flexibility. The peak load pricing formulation discussed here will provide a load levelling mechanism for reducing demand at congested computer centres in a network, and hence improve turnaround and response times during computer prime time.

2. Model formulation

The problem of pricing of services for a computer centre has received substantial attention in the computing literature. Cotton (1975) lists the following as some of the pricing objectives of different organisations: profit maximisation, market penetration, tie-in with other services and optimal use of computer resources. Nielsen (1968; 1970) indicates the consideration and procedures involved in developing a flexibly priced computer system. Lehman (1972) also provides a detailed discussion of a working flexible pricing system. Shaftel and Zmud (1974) give an analytic technique which determines a flexible pricing schedule. Finally Kriebel and Mikhail (1975) develop a 'dynamic pricing' model based on profit maximisation to allocate network resources to a relatively captive market demand through posted prices by user (or job class), network node and time period.

The model presented here deals specifically with the problem of peak load pricing as it occurs in computer networks. In this case, the centres in the network charge different prices for peak and off-peak periods for the same kind of computer service. Various approaches to the peak load pricing problem have been studied in the economic literature with special reference to the electricity and telephone industries (Steiner, 1957; Hirshleifer, 1958; Williamson, 1966; Littlechild, 1970; Pressman, 1970; Bailey, 1972; De Salvia, 1969). In what follows, we will develop a model of the problem as a constrained optimisation one based on the Williamson (1966), Littlechild (1970), Pressman (1970) and Bailey (1972) formulations. We believe the above-named papers fully capture all the important aspects of the problem especially pertinent to the management of networks of computer centres.

The model studied is one of profit maximisation under budget and capacity constraints for each computer centre in the network. Ewusi-Mensah (1978) discusses other models which reflect the behaviour of different computer centre managers in the delivery of computer services. The following notation is used in the model:

$p_{jt}(p'_{jt})$	The unit price centre j charges in period t for services to local (non-local) users.
$q_{jt}(q'_{jt})$	The amount of services centre j provides to local (non-local) users.
K_j	The maximum capacity of centre j's service output in any period.
$C_j(K_j)$	Capacity cost of centre j.
T_{jt}	The total cost of centre j in period t for services to both local and non-local users, i.e. $T_{jt} = T_{jt}(q_{jt}, q'_{jt}, K_j)$

CCC-0010-4620/81/0024-0071 $03.50

λ_j The Lagrange multiplier for the budget constraint of centre j.

γ_{jt} The Lagrange multiplier for the capacity constraint of centre j in period t.

$\varepsilon_{jt}(\varepsilon'_{jt})$ The own-price elasticity of demand for centre j's services in period t to local (non-local) users.

$\varepsilon_{jkt}(\varepsilon'_{jkt})$ The cross-price elasticity of demand for centre j's services between the periods k and $t(k \neq t)$ for local (non-local) users.

U_j The maximum allowable cost for centre j.

In developing the model, the following assumptions are made:

1. The product of any computer centre is very inhomogeneous due to the many different languages, programs and uses that can be made of a large computer (Eric, 1975). Thus, to simplify the analysis, we assume all the centres in the network sell the same kind of computer service (e.g. general time sharing) which may include cpu cycles, memory storage, input/output operations and others.
2. We assume that the cost of delivering services to remote locations in the network is independent of the distance involved. Also the centres in the network may charge different prices for service to local and non-local or remote users.
3. Each user in the network has access to all the computer centres in the network. In other words, we do not assume a captive user market which management can monopolise to meet its objective.
4. The demand function at each centre in the network is assumed to be dependent on the price of that centre alone, other prices being held constant, that is,
$$q_{jt} = q_{jt}(p_{jk}) \text{ or } p_{jt} = p_{jt}(q_{jk}); \; k = 1, 2 .$$
And the price functions $p_{jt}(q_{jk})$ and $p'_{jt}(q'_{jk})$ are both differentiable and strictly decreasing in q_{jk} and q'_{jk} respectively.
5. The total cost function $T_{jt}(q_{jt}, q'_{jt}, K_j)$ for each centre j in period t is twice differentiable (except at $K_j = 0$), convex and increasing in q_{jt}, q'_{jt} and K_j with $T_{jt}(0, 0, K_j) = C_j(K_j)$ the fixed cost of capacity and $T_{jt}(0, 0, 0) = 0$. We also assume the total revenue function to be concave.
6. The centres in the network act independently of each other to maximise their objective functions. That is, though the centres recognise the mutual interdependence of their decisions and those of their rivals, they are not allowed to act in collusion.

In the analysis, we restrict ourselves to the case where $t = 1, 2$ to denote peak and off-peak periods. As Kriebel and Mikhail have shown, it is theoretically simple to deal with the case where $t = 1, 2, \ldots, n$ and thus allow the price of the services of the centres to vary continuously over the hours of the day. Two main concerns for this approach are:

1. Continuous price fluctuations can be expensive for the centres to administer and the users to keep track of and respond intelligently.
2. With short pricing intervals and low demand the centres are likely to produce highly erratic prices and this, as Smidt (1968) suggests, will add to the difficulty already faced by users in forecasting what quality of service they can obtain at a given price.

Using the notation given above, the mathematical representation of the model for each centre j in the network is thus;

$$\underset{q_{jt}, q'_{jt}}{\text{Max}} \sum_{t=1}^{2} \{p_{jt}q_{jt} + p'_{jt}q'_{jt} - T_{jt}(q_{jt}, q'_{jt}, K_j)\}$$

subject to

$$\sum_{t=1}^{2} T_{jt}(q_{jt}, q'_{jt}, K_j) \leqslant U_j \tag{1}$$

$$q_{jt} + q'_{jt} \leqslant K_j; \quad t = 1, 2$$

where $q_{jt} \geqslant 0$; $q'_{jt} \geqslant 0$; $K_j \geqslant 0$; $U_j \geqslant 0$; $p_{jt} \geqslant 0$ and $p'_{jt} \geqslant 0$. And the budget constraint is justified on the grounds that each centre operates on a budget allocated in advance with little opportunity for variation on account of price changes within a budget period.

The Lagrange function for each centre j yields the following unconstrained maximisation problem.

$$\begin{aligned} \underset{q_{jt}, q'_{jt}}{\text{Max}} & \sum_{t=1}^{2} \{p_{jt}q_{jt} + p'_{jt}q'_{jt} - T_{jt}(q_{jt}, q'_{jt}, K_j)\} \\ & + \lambda_j\{U_j - \sum_{t=1}^{2} T_{jt}(q_{jt}, q'_{jt}, K_j)\} \\ & + \sum_{t=1}^{2} \gamma_{jt}[K_j - (q_{jt} + q'_{jt})] . \end{aligned} \tag{2}$$

On applying the Kuhn–Tucker conditions for maximum on (2) we get the following:

$$q_{jt} \geqslant 0; \quad p_{jt} + \sum_{k=1}^{2} q_{jk} \frac{\partial p_{jk}}{\partial q_{jt}} - (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q_{jt}} - \gamma_{jt} \leqslant 0; \quad t = 1, 2 \tag{3a}$$

$$q'_{jt} \geqslant 0; \quad p'_{jt} + \sum_{k=1}^{2} q'_{jk} \frac{\partial p'_{jk}}{\partial q'_{jt}} - (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q'_{jt}} - \gamma_{jt} \leqslant 0; \quad t = 1, 2 \tag{3b}$$

$$K_j \geqslant 0; \quad -(1 + \lambda_j) \frac{\partial T_j}{\partial K_j} + \sum_{t=1}^{2} \gamma_{jt} \leqslant 0 \tag{3c}$$

$$\lambda_j \geqslant 0; \quad U_j - \sum_{t=1}^{2} T_{jt}(q_{jt}, q'_{jt}, K_j) \geqslant 0 \tag{3d}$$

$$\gamma_{jt} \geqslant 0; \quad K_j - (q_{jt} + q'_{jt}) \geqslant 0; \quad t = 1, 2 . \tag{3e}$$

We now discuss various solution possibilities.

Case 1 *Independent demands between periods*

We consider the case of independent demands between the periods, with cross-effects between the periods assumed to be zero. Also unless otherwise stated, we assume the capacity constraint is not binding.

That is,

$$\gamma_{jt} = 0; \quad t = 1, 2 \text{ and } \frac{\partial p_{jk}}{\partial q_{jt}} = \frac{\partial p'_{jk}}{\partial q'_{jt}} = 0; \quad t \neq k; k = 1, 2$$

1. If both the budget and capacity constraints are inactive, i.e. $\lambda_j = 0$ and $\gamma_{jt} = 0$; $t = 1, 2$, then (1) reduces to an unconstrained maximisation problem with the following solution obtain from (3a)
$$p_{jt} + q_{jt} \frac{\partial p_{jt}}{\partial q_{jt}} - \frac{\partial T_{jt}}{\partial q_{jt}} = 0; \quad t = 1, 2 \tag{4}$$
which simplifies to
$$MR_{jt} = p_{jt}(1 - 1/\varepsilon_{jt}) = \frac{\partial T'_{jt}}{\partial q_{jt}} \tag{5}$$
or
$$p_{jt} = \frac{\varepsilon_{jt}}{(\varepsilon_{jt} - 1)} \frac{\partial T_{jt}}{\partial q_{jt}}; \quad t = 1, 2$$
where MR_{jt} is the marginal revenue of centre j in period t and $\varepsilon_{jt} = -\frac{q_{jt}}{p_{jt}} \frac{\partial p_{jt}}{\partial q_{jt}}$ is the own-price elasticity of demand for centre

j's services in period t.

The analogous results for non-local services obtained from (3b) are:

$$MR'_{jt} = p'_{jt}(1 - 1/\varepsilon'_{jt}) = \frac{\partial T_{jt}}{\partial q'_{jt}} \qquad (5')$$

or

$$p'_{jt} = \frac{\varepsilon'_{jt}}{(\varepsilon'_{jt} - 1)} \frac{\partial T_{jt}}{\partial q'_{jt}}; \quad t = 1, 2.$$

Thus as expected, for a profit maximising centre, the marginal revenue it receives for services to either local or non-local customers must equal its marginal cost of operation for those customers.

2. If the budget constraint is active ($\lambda_j > 0$) and the capacity constraint inactive ($\gamma_{jt} = 0$; $t = 1, 2$) then (3a) gives

$$p_{jt} + \sum_{k=1}^{2} q_{jk} \frac{\partial p_{jk}}{\partial q_{jt}} = (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q_{jt}}; \quad t = 1, 2 \qquad (6)$$

which simplifies to

$$MR_{jt} = (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q_{jt}} > \frac{\partial T_{jt}}{\partial q_{jt}} \text{ (since } \lambda_j > 0) \qquad (7)$$

or

$$p_{jt} = \frac{\varepsilon_{jt}(1 + \lambda_j)}{(\varepsilon_{jt} - 1)} \frac{\partial T_{jt}}{\partial q_{jt}} > \frac{\varepsilon_{jt}}{(\varepsilon_{jt} - 1)} \frac{\partial T_{jt}}{\partial q_{jt}}; \quad t = 1, 2.$$

For non-local services, the results derived from (3b) are;

$$MR'_{jt} = (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q'_{jt}} > \frac{\partial T_{jt}}{\partial q'_{jt}} \qquad (7')$$

or

$$p'_{jt} = \frac{\varepsilon'_{jt}(1 + \lambda_j)}{(\varepsilon'_{jt} - 1)} \frac{\partial T_{jt}}{\partial q'_{jt}} > \frac{\varepsilon'_{jt}}{(\varepsilon'_{jt} - 1)} \frac{\partial T_{jt}}{\partial q'_{jt}}; \quad t = 1, 2.$$

3. We now assume period $t = 1$ to be the peak period, thus making the capacity constraint binding or active in that period.

Therefore $\gamma_{jt} > 0$ for $t = 1$
$= 0 \quad t = 2.$

If the budget constraint is either active or inactive ($\lambda_j \geqslant 0$) then by substitution in (3a) we get

$$p_{jt} + \sum_{k=1}^{2} q_{jk} \frac{\partial p_{jk}}{\partial q_{jt}} = (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q_{jt}} + \gamma_{jt}; \quad t = 1, 2$$

that is,

$$MR_{jt} = p_{jt}(1 - 1/\varepsilon_{jt}) = (1 + \lambda_j) \frac{\partial T_{jt}}{\partial q_{jt}} + \gamma_{jt}; \quad t = 1, 2 \qquad (8)$$

from (3c) we get by substitution

$$\sum_{t=1}^{2} \gamma_{jt} = (1 + \lambda_j) \frac{\partial T_j}{\partial K_j}$$

$$\gamma_{jt} = \delta_{tl}(1 + \lambda_j) \frac{\partial T_j}{\partial K_j}; \quad t = 1, 2 \qquad (9)$$

where $\delta_{tl} = \begin{cases} 1 \text{ for } t = l = 1 \\ 0 \text{ otherwise.} \end{cases}$

Substituting (9) in (8) gives

$$MR_{jt} = p_{jt}(1 - 1/\varepsilon_{jt}) = (1 + \lambda_j) \left\{ \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\} \qquad (10)$$

or

$$p_{jt} = \frac{\varepsilon_{jt}(1 + \lambda_j)}{(\varepsilon_{jt} - 1)} \left\{ \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\}; \quad t = 1, 2.$$

Now when the budget constraint is not binding ($\lambda_j = 0$) then we get from (10)

$$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \qquad (11)$$

and

$$p_{jt} = \frac{\varepsilon_{jt}}{(\varepsilon_{jt} - 1)} \left\{ \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\}; \quad t = 1, 2.$$

Hence when the budget constraint of a profit maximising centre is not binding, its marginal revenue in its peak period must equal the sum of its marginal cost of operation and its marginal cost of capacity. However, when the budget constraint is active ($\lambda_j > 0$), then from (10) we get

$$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \qquad (12)$$

and

$$p_{jt} > \frac{\varepsilon_{jt}}{(\varepsilon_{jt} - 1)} \left\{ \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\}; \quad t = 1, 2.$$

Analogous results for non-local services can be derived from (3b) by going through the analysis so that for inactive budget constraint ($\lambda_j = 0$) we get

$$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \qquad (11')$$

and

$$p'_{jt} = \frac{\varepsilon'_{jt}}{(\varepsilon'_{jt} - 1)} \left\{ \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\}; \quad t = 1, 2.$$

Similarly for active budget constraint ($\lambda_j > 0$) we obtain

$$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \qquad (12')$$

and

$$p'_{jt} > \frac{\varepsilon'_{jt}}{(\varepsilon'_{jt} - 1)} \left\{ \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl} \frac{\partial T_j}{\partial K_j} \right\}; \quad t = 1, 2.$$

Case II Dependent demands between periods

In this case we have dependent demands between the periods, thus the cross-effects are non-zero. That is,

$$p_{jt} = p_{jt}(q_{jt}, q_{jk}) \text{ and } \frac{\partial p_{jk}}{\partial q_{jt}} \neq 0 \text{ for } t \neq k; \quad t, k = 1, 2$$

for local users or customers and

$$p'_{jt} = p'_{jt}(q'_{jt}, q'_{jk}) \text{ and } \frac{\partial p'_{jk}}{\partial q'_{jt}} \neq 0 \text{ for } t \neq k; \quad t, k = 1, 2$$

for non-local users or customers.

There are three solution possibilities to this case also.

1. With both the budget and capacity constraints inactive ($\lambda_j = 0$ and $\gamma_{jt} = 0$; $t = 1, 2$), then substitution in (3a) gives

$$p_{jt} + \sum_{k=1}^{2} q_{jk} \frac{\partial p_{jk}}{\partial q_{jt}} = \frac{\partial T_{jt}}{\partial q_{jt}}$$

i.e.

$$p_{jt} + p_{jt} \frac{q_{jt}}{p_{jt}} \frac{\partial p_{jt}}{\partial q_{jt}} + p_{jt} \frac{q_{jk}}{p_{jt}} \frac{\partial p_{jk}}{\partial q_{jt}} = \frac{\partial T_{jt}}{\partial q_{jt}}; \quad t \neq k; \quad t, k = 1, 2. \qquad (13)$$

Because the computer services offered in the two periods k and t can be treated as related commodities, we can use Hotelling's results showing the applicability of the integrability condition to related commodities (see Hotelling, 1935 and Pressman, 1970 for a detailed discussion), that is

$$\frac{\partial p_{jk}}{\partial q_{jt}} = \frac{\partial p_{jt}}{\partial q_{jk}} \text{ for } t \neq k; \quad t, k = 1, 2 .$$

Hence (13) becomes

$$p_{jt} + p_{jt}\frac{q_{jt}}{p_{jt}}\frac{\partial p_{jt}}{\partial q_{jt}} + p_{jt}\frac{q_{jt}}{p_{jt}}\frac{\partial p_{jt}}{\partial q_{jk}} = \frac{\partial T_{jt}}{\partial q_{jt}}$$

$$MR_{jt} - \frac{p_{jt}}{\varepsilon_{jkt}} = \frac{\partial T_{jt}}{\partial q_{jt}}$$

where

$$\varepsilon_{jkt} = -\frac{p_{jt}}{q_{jk}}\frac{\partial q_{jk}}{\partial p_{jt}} .$$

Thus,

$$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \frac{p_{jt}}{\varepsilon_{jkt}} \tag{14}$$

and

$$p_{jt} = \varepsilon_{jkt}\left(MR_{jt} - \frac{\partial T_{jt}}{\partial q_{jt}}\right) \text{ for } t \neq k; \quad t, k = 1, 2 .$$

The results for non-local services can be similarly derived from (3b) and they are:

$$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \frac{p'_{jt}}{\varepsilon'_{jkt}} \tag{14'}$$

and

$$p'_{jt} = \varepsilon'_{jkt}\left(MR'_{jt} - \frac{\partial T_{jt}}{\partial q'_{jt}}\right) \text{ for } t \neq k; \quad t, k = 1, 2 .$$

Hence under profit maximisation, the price the centre charges in each period for services to local or non-local customers depends on the marginal operating cost for that period, the marginal revenue for the period and the cross-price elasticity of demand between the two periods when capacity is under utilised and the budget constraint inactive.

2. If the budget constraint alone is active ($\lambda_j > 0$) with capacity constraint inactive ($\gamma_{jt} = 0$; $t = 1, 2$) then substitution in (3a) gives

$$p_{jt} + \sum_{k=1}^{2} q_{jk}\frac{\partial p_{jk}}{\partial q_{jt}} = (1 + \lambda_j)\frac{\partial T_{jt}}{\partial q_{jt}}; \quad t = 1, 2$$

which simplifies to

$$MR_{jt} = (1 + \lambda_j)\frac{\partial T_{jt}}{\partial q_{jt}} + \frac{p_{jt}}{\varepsilon_{jkt}} > \frac{\partial T_{jt}}{\partial q_{jt}} + \frac{p_{jt}}{\varepsilon_{jkt}} \tag{16}$$

and

$$p_{jt} = \varepsilon_{jkt}\left\{MR_{jt} - (1 + \lambda_j)\frac{\partial T_{jt}}{\partial q_{jt}}\right\}$$

$$< \varepsilon_{jkt}\left(MR_{jt} - \frac{\partial T_{jt}}{\partial q_{jt}}\right) \text{ for } \lambda_j > 0 \text{ and } t \neq k; \quad t, k = 1, 2 .$$

The analogous results for non-local services derived from (3b) are

$$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \frac{p'_{jt}}{\varepsilon'_{jkt}} \tag{16'}$$

and

$$p'_{jt} < \varepsilon'_{jkt}\left(MR'_{jt} - \frac{\partial T_{jt}}{\partial q'_{jt}}\right)$$

for $\lambda_j > 0$ and $t \neq k$; $t, k = 1, 2$.

3. We let period $t = 1$ be the peak period and thus make the capacity constraint binding in that period such that

$$\gamma_{jt} > 0 \text{ for } t = 1$$
$$= 0 \quad t = 2 .$$

Now if the budget constraint is either active or inactive ($\lambda_j \geqslant 0$) then substitution in (3a) yields the following

$$p_{jt} + \sum_{k=1}^{2} q_{jk}\frac{\partial p_{jk}}{\partial q_{jt}} = (1 + \lambda_j)\frac{\partial T_{jt}}{\partial q_{jt}} + \gamma_{jt}$$

which on applying the integrability condition simplifies to

$$MR_{jt} - \frac{p_{jt}}{\varepsilon_{jkt}} = (1 + \lambda_j)\frac{\partial T_{jt}}{\partial q_{jt}} + \gamma_{jt} .$$

Substituting for γ_{jt} from (9) in the above equation gives

$$MR_{jt} = (1 + \lambda_j)\left\{\frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right\} + \frac{p_{jt}}{\varepsilon_{jkt}} \tag{17}$$

and

$$p_{jt} = \varepsilon_{jkt}\left\{MR_{jt} - (1 + \lambda_j)\left(\frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right)\right\}; \quad t \neq k;$$
$$t, k = 1, 2 .$$

When the budget constraint is inactive ($\lambda_j = 0$) we get

$$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p_{jt}}{\varepsilon_{jkt}} \tag{18}$$

and

$$p_{jt} = \varepsilon_{jkt}\left\{MR_{jt} - \left(\frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right)\right\}; t \neq k; \quad t, k = 1, 2 .$$

However, when the budget constraint is active ($\lambda_j > 0$) we get from (17) the results

$$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p_{jt}}{\varepsilon_{jkt}} \tag{19}$$

and

$$p_{jt} < \varepsilon_{jkt}\left\{MR_{jt} - \left(\frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right)\right\}; \quad t \neq k;$$
$$t, k = 1, 2 .$$

The corresponding results for non-local services are for inactive budget constraint ($\lambda_j = 0$)

$$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p'_{jt}}{\varepsilon'_{jkt}} \tag{18'}$$

and

$$p'_{jt} = \varepsilon'_{jkt}\left\{MR'_{jt} - \left(\frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right)\right\}; \quad t \neq k;$$
$$t, k = 1, 2 .$$

For active budget constraint ($\lambda_j > 0$) we have

$$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p'_{jt}}{\varepsilon'_{jkt}} \tag{19'}$$

and

$$p'_{jt} < \varepsilon'_{jkt}\left\{MR'_{jt} - \left(\frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}\right)\right\}; \quad t \neq k;$$
$$t, k = 1, 2 .$$

A summary of the results for both cases is given in **Table 1**.

Discussion of results

The model dealt with two cases—independent and dependent demands between the periods. The independent demand case represents users with inflexible computer usage demands in either of the two periods and so may be unable to transfer some of their demands from one period to the other. The dependent demand case accommodates users with flexible usage demands who, given proper incentive, may be willing to transfer some of their demands from one period to another. In each of the above two cases, the following three solution possibilities were discussed.

Table 1 Summary of results for maximising profit

Period		Off-peak		Peak	
Demand		$\lambda_j = 0;\ \gamma_{jt} = 0$	$\lambda_j > 0;\ \gamma_{jt} = 0;\ t = 1, 2$	$\lambda_j = 0;\ \gamma_{jt} > 0;\ t = 1$ $= 0;\ t = 2$	$\lambda_j > 0;\ \gamma_{jt} > 0;\ t = 1$ $= 0;\ t = 2$
Independent	Local	$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}};\ t = 1, 2$	$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}};\ t = 1, 2$	$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j};\ t = 1, 2$	$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j};\ t = 1, 2$
	Non-local	$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}};\ t = 1, 2$	$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}};\ t = 1, 2$	$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j}$	$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j};\ t = 1, 2$
Dependent	Local	$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \frac{p_{jt}}{\varepsilon_{jkt}};\ t \neq k;$ $t, k = 1, 2$	$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}} + \frac{p_{jt}}{\varepsilon_{jkt}};\ t \neq k;$ $t, k = 1, 2$	$MR_{jt} = \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p_{jt}}{\varepsilon_{jkt}}$ $t \neq k;\ t, k = 1, 2$	$MR_{jt} > \frac{\partial T_{jt}}{\partial q_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p_{jt}}{\varepsilon_{jkt}}$ $t \neq k;\ t, k = 1, 2$
	Non-local	$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \frac{p'_{jt}}{\varepsilon_{jkt}};\ t \neq k;$ $t, k = 1, 2$	$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \frac{p'_{jt}}{\varepsilon_{jkt}};\ t \neq k;$ $t, k = 1, 2$	$MR'_{jt} = \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p'_{jt}}{\varepsilon_{jkt}}$ $t \neq k;\ t, k = 1, 2$	$MR'_{jt} > \frac{\partial T_{jt}}{\partial q'_{jt}} + \delta_{tl}\frac{\partial T_j}{\partial K_j} + \frac{p'_{jt}}{\varepsilon_{jkt}}$ $t \neq k;\ t, k = 1, 2$

1. This solution dealt with computer centres which have both unrestricted funds to provide services to users and also excess capacity to meet the usage demands of the user communities. This example is representative of newly installed computer systems which are in the process of attracting users.
2. For this solution, the computer centre is constrained by management to operate within the limits of the centre's budget constraint. But the centre may still have excess capacity to meet the computing needs of the user communities. This subcase may represent computer centres which have recently upgraded their system and therefore may have available excess machine capacity to meet any future increases in usage demands of users.
3. The third subcase dealt with computer centres which are not only operating within the limits of the budget constraint but also have usage demands in one period (the peak period or prime time) pressing on the limits of the system capacity. This solution is representative of computer centres with saturated system capacities which need to be fully upgraded.

The peak load pricing scheme formulated essentially involves the institution of different prices in different periods of usage. The general principle of peak load pricing of computer services that can be derived from the results of the model can be summarised briefly as follows:

1. Prices are set by the period of the day demand for computer services is made to the centre in the network. In the peak period or prime time, a relatively high price is charged to both local and non-local users to discourage usage and ease congestion in the system. But in the off-peak period or slack time a comparatively low price is charged to users to encourage usage of the system's idle resources. The motivation for instituting this type of pricing scheme is a need for a load levelling mechanism for smoothing out the pattern of demand of the computer centres in the network. In this way the quality of service at the centres is improved and effective and efficient utilisation of the system's expensive resources are encouraged.
2. No responsibility for capacity costs is imputed to those users or customers whose demand for computer services during off-peak hours do not add to system congestion by pressing on system capacity. This is because only peak period or prime time users cause the computer centres to incur additional capacity costs by creating a need for upgrading the system hardware and related software to ease system congestion due to usage at peak period or prime time.

3. Data collection and application of results

We now deal with the problem of the implementation of the pricing model in a 'real world' environment. Following De Salvia (1969), the first step in the analysis is to consider daily demand patterns of the different centres in the network to determine peak and off-peak periods for each centre. This is not a major empirical problem because current practices of many computer centres include load estimation by job class and differential pricing by service priorities (Kriebel and Mikhail, 1975).

However, because peak load pricing has not been directly considered in the literature on pricing computer services, there may arise some problems in collection of data necessary to implement fully the results of the proposed model. In cases where only total cost data are available, that may be used to estimate individual operating costs incurred in the separate periods. The errors which might thus be introduced by a failure to consider all costs in this case is, according to De Salvia, an underestimation of marginal costs. Conversely, errors which overestimate the individual costs will provide an overestimation of marginal costs.

The following types of data will be needed to implement the results of the model fully.

1. *Daily usage data*
That is jobs processed or machine utilisation/output over time. This will enable the analyst to determine clearly what the daily usage pattern is and thus help to determine the peak and off-peak periods.

2. *Operating cost data*
What is needed is the cost of providing the services (i.e. machine utilisation/output) over the same time period as in 1. This will be used to derive estimates of the operating cost incurred in the separate periods (i.e. peak and off-peak periods).

3. *Capacity cost data*
The cost of the computer and other capital equipment used in providing the services. The maximum volume of service the computer is capable of handling. This is necessary so an estimate of the maximum capacity requirements for each period can be derived based on the demand requirements determined in 1 above.

4. *Revenue and profit*
Total revenue and profit, if any, the centre receives for services it provides to users over the time period specified in 1.

5. *Demand elasticity data*

Two types of elasticity data will be needed:

(*a*) Own-price elasticity of demand within each period which will indicate what the peak and off-peak period elasticities may be.

(*b*) Cross-price elasticities of demand between the peak and off-peak periods.

Once the data collection problem has been resolved, the peak load pricing problem then essentially becomes one of applying a charge equivalent to a short-run marginal cost to off-peak users and a charge equal to long-run marginal cost to peak periods. It is our claim that if demand in the two periods is sufficiently price elastic, then this type of price structure would serve to expand the off-peak user population and diminish peak consumption. It should also be noted that the new pricing system will tend to be self-correcting in the following sense. The proposed rate structure will invariably give rise to additional information or data which will permit subsequent improvements to be made in the proposed rates. Thus after an initial period of experimentation or study, a reasonably stable system of prices will be achieved. However, as De Salvia argues, a peak load pricing scheme cannot be expected to be as stable as rates which ignore the peak load problem. This is due to the fact that users may be constantly engaged in finding ways of avoiding peak usage in order to reduce the cost of the computing service they receive. However, because of the effort involved in the evaluation process, the search may be expected to be limited.

4. Management uses of results

Some management uses of the results are:

1. To enable management to fix more rationally their prices to reflect the hours of consumption on the part of users.
2. To enable management to charge higher prices at times when consumption in any time period would tend to rise above the level of capacity of the computer system. This will hopefully help to lower the corresponding portion of the demand curve during those peak hours.
3. In the same manner, lower prices can be charged to increase the demand for services at those hours when the computer system is heavily underutilised. On the other hand, if consumption at the lower rates tends to rise above the desired level, which might necessitate incurring new capacity cost to meet with the demand, enough higher rates can be charged to keep consumption at that level.
4. After all these pricing policies have been implemented and the congestion cost to users in the system is still beyond an acceptable level, that is, substantial numbers of users experience longer turnaround time or slower response times, then it will be necessary for management to consider extensions to the system's capacity to cope with the rising user demands. Failure to take this remedial action might cause the centre to lose the business of some of their users with sufficiently inelastic demands for turnaround time or response times.

Some benefits that users in the network can also derive from the results are:

1. Users can be informed not only of the prices centres charge for a unit of service, but equally important, how much it costs them in time as a result of congestion in the system, especially during peak hours of usage.
2. Users can effectively shop around for lower cost for the unit of service offered at any centre so as not to be confined to any one centre in the network. For example, users in New York City may find it advantageous in the first three hours of their working day (which may be part of their peak period) to transfer most of their computing work to Los Angeles and vice versa for Los Angeles users during the last three hours of their working day. This will be cost-effective if the users can determine the overall cost of their services, including transmission costs and delays due to congestion in the system to be cheaper than what their local centres offer them.

5. Summary and conclusion

In this paper we have focused attention on the problem of peak loads in computer centres and networks. In analysing the problem, we dealt specifically with private commercial centres whose main objective is profit maximisation under budget and capacity constraints. The results obtained describe what the various pricing schemes may be under different operating conditions. The peak load pricing formulation presented provides a load levelling mechanism for reducing demand at congested centres in a network and hence improve response and turnaround times during computer prime time.

We have discussed some of the potential problems one is likely to encounter in the collection of data and subsequent application of the results of this work. The problems are largely attributable to the fact that because most computer centres have never considered peak load pricing as proposed in this study, the data necessary for full implementation of the results may not exist in a form that is most desired. However, these difficulties notwithstanding, we are persuaded that the results obtained provide a good approximate solution to the problem. The results also serve to demonstrate the feasibility of applying peak loads pricing based on economic principles to the computer network services industry. This provides a considerable improvement in the operating performance of the computer network services industry over the often arbitrarily determined administrative alternatives in current use (Kriebel and Mikhail, 1975).

References

BAILEY, E. E. (1972). Peak-Load Pricing Under Regulatory Constraint, *Journal of Political Economy*, Vol. 80 No. 4, pp. 662-679.

COTTON, I. W. (1975). Microeconomics and the Market for Computer Services, *Computing Surveys*, Vol. 7 No. 2, pp. 95-111.

DE SALVIA, D. N. (1969). An Application of Peak-Load Pricing, *J. of Business*, Vol. 42, pp. 458-476.

ERIC, M. J. (1975). An Economic Model of a Computer Network, Program in Info. Tech. and Tel., Report No. 18, Stanford University, 1975.

EWUSI-MENSAH, K. (1978). Peak Load Pricing of Computer Network Services, Ph.D. Dissertation, Graduate School of Management, University of California, Los Angeles.

HIRSHLEIFER, J. (1958). Peak Loads and Efficient Pricing: Comment, *Quarterly Journal of Economics*, Vol. 72 No. 3, pp. 451-462.

HOTELLING, H. (1935). Demand Functions with Limited Budgets, *Econometrica*, Vol. 3, pp. 66-78.

KRIEBEL, C. H. and MIKHAIL, O. I. (1975). Dynamic Pricing of Resources in Computer Networks, in *Logistics*, M. Geisler (ed.). North-Holland,Tims Studies in Management Science, pp. 105-124.

LEHMAN, M. M. (1972). Computer usage control, *The Computer Journal*, Vol. 16 No. 2, pp. 106-110.

LITTLECHILD, S. C. (1970). Peak-Load Pricing of Telephone Calls, *Bell J. Econ. and Mgmt. Sci.*, Vol. 1 No. 2, pp. 191-210.

NIELSEN, N. R. (1968). Flexible Pricing: An Approach to the Allocation of Computer Resources, AFIPS Fall Joint Computer Conference, pp. 521-531.

NIELSEN, N. R. (1970). The Allocation of Computer Resources—Is Pricing the Answer, *CACM*, Vol. 13 No. 8, pp. 467-474.

PRESSMAN, I. (1970). A Mathematical Formulation of the Peak-Load Pricing Problem, *Bell J. Econ. and Mgmt. Sci.*, Vol. 1 No. 2, pp. 304-312.
SHAFTEL, T. L. and ZMUD, R. W. (1974). Allocation of Computer Resources Through Flexible Pricing, *The Computer Journal*, Vol. 17 No. 4, pp. 306-312.
SHARPE, W. F. (1969). *The Economics of Computers*, Columbia University Press, New York.
SMIDT, S. (1968). Flexible Pricing of Computer Services, *Management Science*, Vol. 14 No. 10, pp. 518-600.
STEINER, P. O. (1957). Peak Loads and Efficient Pricing, *Quarterly Journal of Economics*, Vol. 71, pp. 585-610.
WILLIAMSON, O. E. (1966). Peak Load Pricing and Optimal Capacity Under Indivisibility Constraint, *American Economic Review*, Vol. 56, pp. 810-827.

Book reviews

Introduction to Computer Data Processing Second Edition by Margaret S. Wu, 1979; 521 pages. (*Harcourt Brace Jovanovich*, £10·35)

This is a typically large (521 pages), well produced and substantial American text book. The artwork and pictures are of a high quality but the 'pop art' cover and chapter openings do not help a book which is supposed to give a reasonable overall view of the computer and its applications. Another unusual feature is confining the material to only half the page, irrespective of whether there is any artwork present or not.

The book has been designed for an introductory data processing course without specifying the sort of course or level of student. It is primarily based on the old batch mainframe approach and covers various aspects of these fairly comprehensively. However, for the ever-increasing numbers who never come across this stage in the development of computers, the book does not contribute very much, particularly as it still has a chapter on punched card procedures. A basic aim was focusing on 'the fundamental concepts of computing' but it has 124 pages describing various programming languages and techniques such as flowcharting. The book has some useful illustrations and diagrams but, apart from this, I feel that there are many more suitable text books available covering this field.

A. A. MOELWYN-HUGHES (Leicester)

Computer System Reliability by Roy Longbottom, 1980; 321 pages. (*John Wiley*, £11·50)

How did the Reviews Editor know that the system for which I am responsible was not performing as it should? That I was critical of its reliability or rather the lack of it! It was thus timely to be asked to review a book on computer system reliability.

The book is a thorough and detailed account of factors which impinge upon unit and system reliability and thus upon methods of calculating and predicting reliability figures and serviceability ratios. It is not until the end of Chapter 11 that the writer addresses 'quality of service from a user's point of view', where he rightly states that the user requires some estimate of what to expect and thus a basis of complaining if service levels are unacceptable. The writer appears not to distinguish the manager of the computer facility from the end user of the service that is provided. It is the reviewer's experience that certainly for an online system end users can be relatively tolerant of a system having an apparently low serviceability ratio provided that the number and frequency of interruptions to service are minimised. As a manager, I have observed a relatively high (acceptable?) serviceability ratio when the end users are complaining of frequent interruptions to the service. Thus end users of a service appear to look for consistency, apparently either consistently good or consistently bad, provided it is predictable! The book makes no attempt to analyse the psychology of the end user; however, it does an excellent job of discussing the nature of computer systems and analysing their performances. It commences with a discussion of 'failures', proceeds through 'reliability variations over time', 'quality assurance', 'environment' and 'software', all as factors impinging upon a total system. The author is then in a position to discuss in separate chapters 'fault symptoms', 'down time and maintenance', 'serviceability', 'maintainability' and 'reliability calculations'. Roy Longbottom is head of the large scientific systems branch of the Central Computer Agency and has been able to draw upon considerable experience and statistics of a wide range of processors, systems and components to illustrate his text in tabular and graphical form.

The author rightly spends time defining terminology as interpreted by the manufacturer and as seen and experienced by the user. He clearly distinguishes time to repair a fault from investigation times undertaken by the manufacturer and incidents observed by the user. We all know of the time and trouble caused by intermittent faults and their investigation as against the relatively short time lost to diagnosis and repair once a fault has become solid. The user of course is only too conscious of time lost to system dumps and analysis to aid engineering or software investigation and then time to recover the operational situation following an incident. The author rightly contrasts down time seen by the user as against that reported by a manufacturer. This text is very useful to any reader, be he designer, manufacturer, service engineer or user, to get a clear understanding of records he may wish to keep and methods of analysis and report. Indeed, the author dedicates two chapters to practical reliability and serviceability calculations and an appendix gives a program for a programmable calculator.

The other main content of the book is a detailed discussion on acceptance trials with particular emphasis on procedures followed by the CCA. Also reported are the procedures advised by the General Service Administration (GSA) of the American Government. An appendix details a set of exerciser programs written in FORTRAN which have been used as part of acceptance trials.

The book requires detailed reading and study to understand all that it has to say, requiring time which is recommended as a worthwhile investment. If, as a result, users and manufacturers have a better mutual understanding of reliability and serviceability leading to users demanding, and manufacturers achieving, better performance figures, then the author will have done the industry a service.

A. H. WISE (Leicester)

Fundamentals of Fortran Programming, Second Edition, by R. C. Nickerson, 1980; 450 pages. (*Prentice-Hall* £7·75 paper)

This is an extraordinarily thorough introduction to FORTRAN and to programming aimed at students in a wide range of disciplines. The language described is based on the 1966 standard but many common extensions and some FORTRAN 77 features are also covered, always carefully delineated by 'some versions have . . .'. Good program structure is discussed but not pushed too hard.

The material is said to have been class-tested at San Francisco State University and the author seems to have anticipated almost every possible question from students. This completeness makes the text suitable for self-teaching but also makes it remarkably slow; by page 82 only simple READ, WRITE and FORMAT have been covered. Arrays come in at page 278 and subprograms at page 358. The author sticks to his title by not getting as far as unformatted input/output or more esoteric things like EXTERNAL, COMPLEX or P formats, or FORTRAN 77 facilities like direct-access input/output and he omits some commonly used statements, possibly on the grounds that their use may be considered poor practice, such as alternate returns, ENTRY, ASSIGN and EQUIVALENCE. The fundamentals however get exhaustive treatment.

D. T. MUXWORTHY (Edinburgh)