On the Application of Peak-Load Pricing to Computer Services.

Kweku Ewusi-Mensah and Bennet P. Lientz
Graduate School of Management
University of California, Los Angeles

Technical rept.
Dec 77 122 p.

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited

This work was partially supported by the Information Systems Program, Office of Naval Research under contract N00014-75-C-0266, project no. NR 049-345.

N00014-75-C-0266
ABSTRACT

Computer networks are increasing in usage and availability. Concurrently, individual centers are dealing with increased levels of usage and service demands during prime hours. A model for addressing the peak load pricing problem for individual centers as well as networks is presented. The model is based on each center independently maximizing a specified objective function. Different price schedules are used for local and remote users over distinct time periods. Implementation of the model is discussed.
1. **Introduction**

There has been an increasing use and availability of network services. One type of network is one in which independent centers offer services through a network supplied by a service organization. In such a network, and for individual centers, the demand for services may be high during prime hours -- resulting in service problems and degraded performance. Characteristics of network services have been presented in Cotton [3], Smidt [24], and Sharpe [22]. Intensifying the problem is the frequent financial inability to provide resources to accommodate peak loads.

The above discussion leads to an interest in peak load pricing models for computer services. Economic problems have been cited by several sources including [6] and [11]. The peak load problem has been pointed out in [3]. Peak load problems arise in part because uniform rates over time create sharp demand peaks ([4], [22]). Nielsen [16] pointed out the need for flexible pricing. For network settings, the amortization and cost assignment of network services to remote users has been discussed in [22]. The effects of pricing to control demand, rationalize planning, and provide for comparison by users has been cited ([3]). Singer, Kanter, and Moore [23] and Gotlieb [7] have viewed prices as a rationing mechanism rather than as a tool for cost recovery. Cotton [3] has pointed out some of the objectives in pricing and its relationship to organizational objectives.

Given the recognition of the problems in pricing, there have been several approaches proposed for pricing. Smidt [25] examined the problems associated with average prices and considered patterns of user behavior. Marchand [14] developed a priority pricing model. Selwyn [20] and Nunamaker and Whinston [18] have proposed pricing discrimination to adjust demands for the organization. Shaftel and Zmud [21] developed
a mathematical programming method to obtain a price schedule.
Kriebel and Mikhail [11] developed a pricing model based on profit maximization to allocate resources in the presence of captive market demand. Kahn [10] has presented a justification for peak load prices.


Having reviewed previous work, we can characterize some aspects of the peak load problem. Each demand in a peak period makes a proportionate contribution to the incurrence of capacity costs over the long run (Khan [10]). This should be reflected in the price. Off-peak usage which is less and inelastic imposes no such demand.

The period of peak load may shift over time. With high elasticity of separate demands by users and a high peak period cost, the result may be excess capacity during previous peak periods and congestion in previously non-peak periods. A second factor leading to shifts is changes in the demand structure.

The peak load pricing model is presented in section 2. Section 3 discusses implementation.
2. Peak-Load Pricing Model

Several assumptions are made in the model presented in this section. First, users are assumed to be aware of different price structures available to them (see Lientz [12] for particular comparisons of user costs). Second, only one general service is considered (e.g., general time-sharing or remote job entry-batch). No provision is made for specialized services at a particular center. Such services would over time be provided to a captive market. Because of the single service, users can switch between services for most small-intermediate applications. Because of these assumptions, pricing differentials are not discriminatory since they will reflect the relative values of services in different periods to both users and management (Minasian [15]).

A third assumption is that cost of remote service communications is independent of distance. This means that all remote users are charged a fixed rate per unit (e.g. packet). This assumption can be weakened as will be noted later. Related to this is the fourth assumption that remote users are charged for communications services.

Several assumptions will be made about the functions of price, demand, and cost. Price are assumed to be differentiable and increasing as a function of demand. Total cost of a center is assumed to be twice differentiable, convex, and increasing as a function of local demand, remote demand, and capacity.

A final assumption is that attention will be restricted to two periods (peak and non-peak). This has shown to be expandable to other cases ([11]) and is not overly restrictive. Erratic prices can merely add to the frustration of users in forecasting necessary service quality ([25]). Furthermore, highly sophisticated price fluctuations add to administrative costs.
As Eric [6] has indicated, there are several different objective functions that can be employed including:

- maximize revenue which is at least as great as cost
- maximize the sum of the centers' producers' and consumers' surplus of the user communities
- maximize profit of center

In what follows the first case will be used. It is exemplified by government and non-profit organizations. The mathematical development is similar in the other cases. The optimization problem based on the objective of revenue maximization is

\[
\text{maximize revenue} \quad (2.1)
\]

subject to

- minimum revenue constraint (or revenue-cost constraint)
- capacity constraint
- non-negativity constraint

Demand for services can be viewed as in microeconomics as the sum of all demands of users of different categories. Suppose there are $N$ centers. A center $(i)$ offers the service in period $t$ to local (remote) users at a unit price of $p_{it}$ ($p_{it}^r$); $t=0$ corresponds to the peak period; $t=1$ denotes non-peak time. Local user demands are denoted by $d_{it}$ which is a function of the vector of prices $p_t$, and is expressable in some form of machine units. Equivalently, the prices $\{p_{it}\}$ can be represented as functions of $\{d_{it}\}$. These are viewed as alternative standard forms by Hotelling [9]. This will be further simplified by assuming $d_{it}$ is a function of $p_{it}$ alone. This can be supported only in cases with small price fluctuations (which is the case here).
The production cost of a center \( TC_{it} \) to satisfy \( d_{it} \) is the sum of capacity costs \( (CA_{it}) \), network connection cost \( (NC_{it}) \), and variable operating costs \( (VC_{it}) \). Capacity costs are a function of expressed capacity \( (k_i) \) while variable operating costs are a function of capacity \( (k_i) \) and demand \( (d_{ij}) \). Bower [2] defines capacity costs and estimates this factor to be 70% of total costs. It is composed of the computer system (CPU, channels, I/O devices, memory, and other equipment), and staff costs (user relations, operations, etc.).

In a network setting when a center joins a network it leaves a monopolistic market and enters an oligopolistic market. This leads to interactive affects. Each center, though, is attempting to optimize its own objections.

Having presented the assumptions and discussed objective functions, demands, and prices, we can turn to the specific problem formulated in (2.1). Since the goal is revenue maximization, the center will devote resources to expanding computer resources. This can be motivated by a desire to remain competitive in the network.

With the notation defined thusfar, we have as the problem:

\[
\text{MAX} \quad (p_{it} d_{it} + p_{it}^* d_{it} + p_{2t} d_{2t} + p_{2t}^* d_{2t}^*) \quad (2.2)
\]

\[
d_{it}, \quad d_{it}^*
\]

with

\[
\sum_{t=0}^{1} p_{it} d_{it} + p_{it}^* d_{it}^* - TC_{it} (d_{it}, d_{it}^*, k_i) = M_i \quad (2.3)
\]

\[
d_{it} + d_{it}^* \leq k_i \quad (2.4)
\]

and \( d_{ij}, d_{ij}^*, p_{it}, p_{it}^*, k_i, \) and \( M_i \) non negative.

In (2.3) \( M_i \) is the minimum allowable profit at the center.
Applying lagrangian multipliers we obtain:

\[
\begin{align*}
\text{MAX} & \quad \sum_{t=0}^{1} p_{it} d_{it} + p'_{it} d'_{it} + \lambda_{i}\left(\sum_{t=0}^{1} p_{it} d_{it} + p'_{it} d'_{it} - TC_{it} - \gamma_{it} - \mu_{i}\right) \\
& + \sum_{t=0}^{1} \xi_{it} (k_{i} - d_{it} - d'_{it})
\end{align*}
\]

(2.5)

The Kuhn-Tucker conditions for optimization can be applied to (2.5). The various subcases will be considered within the context of independent and dependent demands between periods. The general cases that will be referenced are:

\[
d_{it} \geq 0; \quad (1 + \lambda_{i})(p_{it} + \sum_{k=0}^{1} d_{ik}) \frac{\partial p_{ik}}{\partial d_{it}} - \lambda_{i} \frac{\partial TC_{it}}{\partial d_{it}} - \gamma_{it} \leq 0 \quad (2.6)
\]

\[
d'_{it} \geq 0; \quad (1 + \lambda_{i})(p'_{it} + \sum_{k=0}^{1} d'_{ik}) \frac{\partial p'_{ik}}{\partial d'_{it}} - \lambda_{i} \frac{\partial TC_{it}}{\partial d'_{it}} - \gamma_{it} \leq 0 \quad (2.7)
\]

\[
k_{i} \geq 0; \quad -\frac{\partial TC_{it}}{\partial k_{i}} + \gamma_{it} + \phi_{it} \leq 0 \quad (2.8)
\]

\[
a_{i} \geq 0; \quad \sum_{t=0}^{1} p_{it} d_{it} + p'_{it} d'_{it} - TC_{it} - M_{i} \geq 0 \quad (2.9)
\]

\[
\gamma_{it} \geq 0; \quad k_{i} - d_{it} - d'_{it} \geq 0 \quad (2.10)
\]

Case 1 -- Independent Demands between Periods

Here cross-effects are zero so that the partial derivatives of \(p_{ik}\) and \(p'_{ik}\) with respect to \(d_{it}\) and \(d'_{it}\) are zero (\(k=t\)). Several subcases must be considered. In all subcases we delete the development for remote services since it is analogous to that for local users.

Subcase 1.1 profit constraint inactive; capacity constraint inactive, then \(\lambda_{it}\) and \(\lambda_{i}\) are zero. The expression (2.6) becomes

\[
p_{it} d_{it} + \frac{\partial p_{it}}{\partial d_{it}} = 0 \quad \text{for } t=0, 1 \quad (2.11)
\]
Substituting for elasticity we have
\[ p_{it} (1 - \frac{1}{\xi_{it}}) = 0 \quad \text{for } t = 0, 1 \quad (2.12) \]

Since \( p_{it} \) is non-zero, \( \xi_{it} \) must be zero.

Let \( MR_{jt} \) be the marginal revenue so that \( MR_{jt} = p_{it} \left( 1 - \frac{1}{\xi_{it}} \right) \).

From (2.12) we have that \( MR_{jt} \) is zero if and only if \( \xi_{ij} \) is unity. This case corresponds to the situation in which users cannot move freely between periods (independent demands) with no limitations on capacity or funding. This occurs when there is a substantial capacity increase and demand has only started to take advantage of the capacity.

Subcase 1.2 profit constraint active \((\lambda_i > 0)\) and capacity constraint inactive.

The expression (2.6) becomes
\[ p_{it} + d_{it} \frac{dp_{it}}{d\lambda_{it}} - \frac{\lambda_i}{1 + \lambda_i} \cdot \frac{dTC_{it}}{d\lambda_{it}} = 0 \quad (2.13) \]

which reduces to
\[ \lambda_i \frac{\xi_{it}}{(1 + \lambda_i)(\xi_{it} - 1)} \frac{dTC_{it}}{d\lambda_{it}} = 0 \quad (2.14) \]

or
\[ MR_{it} = \frac{\lambda_i}{1 + \lambda_i} \frac{dTC_{it}}{d\lambda_{it}} < \frac{dTC_{it}}{d\lambda_{it}} \quad (2.15) \]

Further manipulation of (2.15) yields
\[ \lambda_i = MR_{it} / \left( \frac{dTC_{it}}{d\lambda_{it}} - MR_{it} \right) \quad (2.16) \]

In (2.16) \( \lambda_i \) is infinite if and only if \( MR_{it} \) is equal to \( dTC_{it} / d\lambda_{it} \). This subcase corresponds to the previous situation (subcase 1.1) except that there are budget limitations.

Subcase 1.3 Period \( t=0 \) is peak period and capacity constraint is active in that period.
In this subcase $y_{it}$ is positive ($i=0$) and zero ($i=1$). For $\lambda_i > 0$ we have from (2.6)
\[ (1 + \lambda_i) (p_{it} + c_{it} \frac{\partial p_{it}}{\partial d_{it}}) - \lambda_i \frac{\partial TC_{it}}{\partial d_{it}} - y_{it} = 0 \tag{2.17} \]
or
\[ MR_{it} = (\lambda_i \frac{\partial TC_{it}}{\partial d_{it}} + y_{it}) / (1 + \lambda_i) \tag{2.18} \]
But from (2.8) we have
\[ y_{it} = \delta_{tu} \lambda_i \frac{\partial TC_{in}}{\partial k_i} \tag{2.19} \]
Where $\delta_{tu}$ is the Kronaecker-delta function. Combining (2.18) and (2.19) gives
\[ MR_{it} = (\lambda_i / (1 + \lambda_i)) (\frac{\partial TC_{it}}{\partial d_{it}} + \delta_{tu} \frac{\partial TC_{in}}{\partial k_i}) \tag{2.20} \]
or
\[ p_{it} = (\lambda_i \xi_{it} / (1 + \lambda_i) (\xi_{it} - 1)) (\frac{\partial TC_{it}}{\partial d_{it}} + \delta_{tu} \frac{\partial TC_{in}}{\partial k_i}) \tag{2.21} \]
and
\[ MR_{it} > \frac{\partial TC_{it}}{\partial d_{it}} + \delta_{tu} \frac{\partial TC_{in}}{\partial k_i} \]
If the profit constraint is inactive (2.17) becomes
\[ MR_{it} = p_{it} (1 - 1/\xi_{it}) = 0 \tag{2.22} \]
which is the same result as subcase 1.1. In general, with both $\xi_{it}$ and $\lambda_i$ positive there are constraints on both budget and capacity.

Case 2. Dependent Demands between Periods

In this case cross-effects are non-zero ($\frac{\partial R_{it}}{\partial d_{it} \epsilon_{jkt}}$). As in the previous cases we will consider several subcases and will explore only local usage. Results will be summarized.
Subcase 2.1. Profit constraint inactive; capacity constraint inactive.

Expression (2.6) becomes

$$\frac{p_{it} + p_{it}}{p_{it}} \frac{g_{it}}{g_{it}} + p_{it} \frac{d_{it}}{d_{it}} \frac{d_{it}}{d_{it}} = 0$$

(2.23)

for $t+k, k_t = 0, 1$. Using Hotelling's results for integrability,

$$\frac{d_{it}}{d_{it}} = \frac{d_{it}}{d_{it}}$$

for $t+k, k_t = 0, 1$. This can be used in (2.23) to give

$$MR_{it} = p_{it} \frac{1}{\xi_{it}}$$

(2.24)

Expression (2.24) is equivalent to

$$\frac{p_{it}}{\xi_{it}} \left( 1 - \frac{1}{\xi_{it}} - \frac{1}{\xi_{it}} \right) = 0$$

so that with $p_{it}$ non-zero we have

$$\xi_{it} + \xi_{it} = \xi_{it}$$

Subcase 2.2. Profit constraint active ($\lambda_i > 0$); capacity constraint inactive.

Substituting into expression (2.6) and simplifying gives

$$(1+\lambda_i) (MR_{it} + p_{it} / \xi_{it}) = \lambda_i \frac{\partial T_{cit}}{\partial d_{it}}$$

(2.25)

t+k, k_t = 0, 1

so that

$$\frac{p_{it}}{\xi_{it}} > \xi_{it} \left( MR_{it} - \frac{\partial T_{cit}}{\partial d_{it}} \right)$$

(2.26)

In the limit as $\lambda_i$ tends to infinity $\frac{\partial T_{cit}}{\partial d_{it}}$ tends to

$$MR_{it} - \frac{p_{it}}{\xi_{it}}, t+k, k_t = 0, 1.$$
Subcase 2.8. Period \( t=0 \) is peak period. Then \( \gamma_{it} \) is positive (\( t=0 \)) or zero (\( t=1 \)). Expression (2.6) becomes

\[
(H_{ki})(MR_{it} - \rho_{it}/\varepsilon_{it}) = \lambda_i \frac{\partial TC_{it}}{\partial x_{it}} + \gamma_{it}
\]  

(2.27)

\( t+k, k,t = 0,1 \). From (2.8) we have

\[
\gamma_{it} = \delta_{tu} \lambda_i \frac{\partial TC_{iu}}{\partial k_i}
\]  

(2.28)

\( k,t = 0,1 \). Combining these results gives

\[
\rho_{it} = (\lambda_i \varepsilon_{it} \varepsilon_{ikt}/(1+\lambda_i) (\varepsilon_{it} + \varepsilon_{ikt} - \varepsilon_{ikt} - \varepsilon_{ukt})(\frac{\partial TC_{it}}{\partial x_{it}} + \delta_{tu} \frac{\partial TC_{iu}}{\partial k_i})
\]  

(2.29)

With an inactive profit constraint (2.29) becomes

\[
\rho_{it} = MR_{it} \varepsilon_{ikt}
\]  

(2.30)
3. Discussion of Implementation

This section addresses problems associated with data collection and analysis in order to implement a peak load pricing model. The first step in data collection is to determine peak and off-peak demand periods for each center. This can be done by analyzing job accounting data (see [4] for example).

Problems arise in data collection later when individual cost elements must be identified and associated with peak periods. Difficulties arise because of lack of past detailed data. Costs can then be underestimated (see De Salvia [4] for a discussion of this). Similar problems have been encountered in the electric utility industry (meters are being installed to measure usage by time of day). The cost elements that should be identified for each time period include operating costs (costs of providing services), capacity costs (hardware and other capital expenditures), revenue, demand elasticity data, and cross price elasticities of demand between peak and off-peak periods. These last two elements may be estimated from previous data or obtained by simulation. As an example, an individual center was considered. Data was obtained for a six-month period batch jobs. The classification of data included the number of jobs by job class on an hourly basis over the period. Cost and revenue estimates were obtained from the center's statement of income and expenses for the same six-month period. Monte Carlo simulation was used to estimate elasticity coefficients. Usage data was analyzed for patterns (using histograms) and analysis of variance methods were applied. A three-way analysis of variance without replication was carried out for month, hour, and job class to determine the effect of hour on job class, difference in usage.
volume by month and interactions between any two factors. The Statistical Analysis System Software was used ([26]). The results using the F-test revealed that the most significant effects on job volume are hour and job class.

In analyzing the pattern of data the peak period was from 10 AM to 6 PM. Tests were made to confirm periodicity. The usage itself was found to be closely fitted by a Fourier polynomial. An alternative approach would be to employ a modified empirical distribution function.

As indicated previously, Monte Carlo simulation was employed to estimate the price and cross-price elasticities of demand between peak and non-peak periods. The purpose was to determine the effect on usage by period for a different price schedule. A constrained optimization problem was defined as:

\[
\text{Min } \sum_i \left( \frac{T_{it}}{D_t} \right) / N_t
\]

with

\[
\sum_i C_{it} p_t \leq B_t
\]

In (3.1) \( T_{it} \) is the actual turnaround time of job \( i \) in period \( t \), \( D_t \) is the desired turnaround time, and \( N_t \) is the number of jobs in period \( t \). In (3.2) \( C_{it} \) is the actual processing time (for job accounting), \( p_t \) is the unit price, and \( B_t \) is a budget constraint. The objective function is then the weighted turnaround time.

Data was collected for 82,240 jobs over a six-month period (13,706.64 average/month). The standard deviation of monthly usage was 1829.16. Proportion of usage in peak hours (10 AM - 6 PM) was .535. The next steps were:
A. Generate pseudo random numbers from \( N(13,706.67, (1829.16)^2) \) distribution for monthly usage and generate total usage \( (N_t) \) for period.

B. Input prices \( p_1 \) and \( p_2 \), and differences in price change of the prices \( \Delta p \) and solve the optimization problem (3.1) and (3.2) to obtain the value \( B_t \).

C. Iterate the steps with \( p_1 \) (or \( p_2 \)) replaced by \( p_0 \) \( (p_0 + \Delta p (p_1 - \Delta p/2)) \) with a fixed \( B_t \) to generate \( N_t \).

D. Computer demand elasticity and cross-elasticity for each period using

\[
E_{ij} = \frac{q_i (N_i - N_i^*)}{q_i (\beta - \beta^*)}, \quad i,j = q_1
\]

(3.3)

The process is then repeated (starting at B) for a specified sample size and the mean and standard deviations calculated for the elasticities. The overall procedure is repeated for new values of \( p_1 \) and \( p_2 \).

The results, using the data, are given in figure 3.1. Figure 3.2 (3.3) presents a graph of the elasticity of the peak and non-peak period \( (E_{ii}, E_{11}) \) (cross-price elasticities \( (E_{10}, E_{01}) \)). The low values of the prices contributed to the high increase in non-peak periods. The results show a decrease in usage in peak hours. The most sensitive elasticity is the cross-price elasticity in the non-peak period given a price in the peak period. The least sensitive is the opposite cross-price elasticity. This would be represented by some users' willingness to pay more for peak usage, while marginal users transfer the non-peak periods.
Figure 3.1. Results of Simulation

<table>
<thead>
<tr>
<th></th>
<th>Price</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak</td>
<td>.11 (old price)</td>
<td>7241.5</td>
<td>1139.9030</td>
</tr>
<tr>
<td></td>
<td>.14 (new price)</td>
<td>5697</td>
<td>894.2678</td>
</tr>
<tr>
<td></td>
<td>.15</td>
<td>5315.23</td>
<td>846.9824</td>
</tr>
<tr>
<td></td>
<td>.16</td>
<td>4983.132</td>
<td>817.269</td>
</tr>
<tr>
<td></td>
<td>.17</td>
<td>4670.898</td>
<td>734.0925</td>
</tr>
<tr>
<td></td>
<td>.18</td>
<td>4422.863</td>
<td>699.6345</td>
</tr>
<tr>
<td></td>
<td>.19</td>
<td>4210.066</td>
<td>698.1174</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>4010.699</td>
<td>660.4673</td>
</tr>
<tr>
<td></td>
<td>.21</td>
<td>3814</td>
<td>601.0786</td>
</tr>
<tr>
<td>non-peak</td>
<td>.11 (old price)</td>
<td>6294.464</td>
<td>990.7654</td>
</tr>
<tr>
<td></td>
<td>.095</td>
<td>7266.597</td>
<td>1129.089</td>
</tr>
<tr>
<td></td>
<td>.09</td>
<td>7696.363</td>
<td>1196.953</td>
</tr>
<tr>
<td></td>
<td>.085</td>
<td>8149.632</td>
<td>1257.303</td>
</tr>
<tr>
<td></td>
<td>.08</td>
<td>8667.832</td>
<td>1391.895</td>
</tr>
<tr>
<td></td>
<td>.075</td>
<td>9229.796</td>
<td>1430.806</td>
</tr>
<tr>
<td></td>
<td>.07</td>
<td>9803.664</td>
<td>1606.205</td>
</tr>
<tr>
<td></td>
<td>.065</td>
<td>10669.89</td>
<td>1898.985</td>
</tr>
<tr>
<td></td>
<td>.06</td>
<td>11513.56</td>
<td>1896.918</td>
</tr>
</tbody>
</table>
Figure 3.2 Elasticities

- Elasticity in non-peak period
- Elasticity in peak period

Figure 3.3 Cross-elasticities

- Cross-elasticity in non-peak period
- Cross-elasticity in peak period
Because of the lack of historical data and pricing changes, validation of the simulation was not performed. Sufficient data is now being collected using some of standard available job accounting systems. Some such systems permit cost variation by period as part of cost management.

The next step is to apply the preceding analysis to the operations of the computer center itself. Approximately 70% of the budget are capital related and 30% operation related. Because the simulation was restricted to batch jobs, the percentage of total cost allocated to this category is computed. The estimate used was 78% and is consistent with the literature. Using available data the average cost of capacity (operation) was computed to be $.23/.11) permit. Using average costs as surrogates to marginal costs, the estimate of costs in the peak period is $.34 (.23 + .11). For the non-peak period it is $.11. Allocation of capacity based on availability of hardware and support attributable to consumption.

4. Conclusion

The principles of peak load pricing have been applied to networks of computers as well as individual centers. Three distinct objective functions have been identified; that of revenue maximization has been explored in detail. Simulation and use of the model have been discussed.

Acknowledgement

The authors wish to acknowledge the helpful comments and suggestions of Professor Donald Erlenkotter.
References


## On the Application of Peak-Load Pricing to Computer Services

### Key Words

- Computer networks

### Abstract

Computer networks are increasing in usage and availability. Concurrently, individual centers are dealing with increased levels of usage and service demands during prime hours. A model for addressing the peak load pricing problem for individual centers as well as networks is presented. The model is based on each center independently maximizing a specified objective function. Different price schedules are used for local and remote users over distinct time periods. Implementation of the model is discussed.