LOCAL TELEPHONE COSTS AND THE DESIGN OF RATE STRUCTURES.

Bridger M. Mitchell

May 1981

NSF-DAR 80-12668

Rand/P-6647

This document has been approved for public release and sale; its distribution is unlimited.
LOCAL TELEPHONE COSTS AND THE DESIGN OF RATE STRUCTURES

BRIDGER M. MITCHELL[1]

The Rand Corporation

I. INTRODUCTION

A well-developed body of economic theory is available to guide the setting of prices for the multi-product regulated firm. Economic efficiency can be increased by designing rate structures that incorporate the basic principles developed from this theory. These principles call for provisionally pricing each of the firm's outputs at its marginal cost, testing to determine whether such rates satisfy a specified budget constraint (e.g., revenues = costs), and then suitably modifying the marginal-cost rates in order to satisfy the constraint. Most commonly, the trial rates produce insufficient revenue, and then rates must be raised according to the Ramsey rule—prices are increased above marginal costs in inverse proportion to the individual price elasticities of demand. This paper applies ratemaking theory to the design of rate structures for local telephone calls that efficiently reflect the costs of the local network.

The principal costs of supplying local telephone calls are embodied in the switching capacity of a local central office (exchange) and the

[1]Presented at the conference "Telecommunications in Canada: Economic Analysis of the Industry" in Montreal, March 4-6, 1981. I thank Patricia Danzon, Ed Park and William Vickrey for helpful discussions. This paper is based on research conducted under grant DAR-80-12668 from the National Science Foundation to The Rand Corporation.
trunking capacity that connects local offices together. (The dedicated local loop connecting the subscriber to the office is, of course, essential but its cost is independent of telephone usage). Several operating companies which are proposing to introduce local measured service are conducting special studies that will gather information about these costs, and how they vary with maximum loads. How should these costs, which relate to specific items of network equipment, be used to develop prices for telephone calls, which are commonly classified in terms of the hour at which they are placed and the distance between subscribers?

The markets for telephone services are characterized by high capacity-related equipment costs and very low variable (traffic) costs, the joint use of some equipment by several outputs, and the grouping together of several outputs that are charged a common price. The following sections develop a series of simple models that successively incorporate these basic elements. Throughout the paper I make several simplifying assumptions:

- the unit of output is a "call" of fixed duration, and there is a uniform rate of demand during any given period
- demand for a given output depends only on its own price, so that there are no temporal or spatial cross-elasticities
- all costs are due to providing capacity to meet the maximum rates of output and capacity can be constructed at constant returns to scale
- the costs of connecting subscribers to the telephone network via local loops are recovered in fixed monthly charges
II. A SINGLE EXCHANGE

All telephone calls originate and terminate in one exchange and this output is produced using a single component of switching capacity. There are only two commodities, $x^1$ and $x^2$, the number of calls made in two equal-length periods ("day" and "night"). This situation is a version of the well-known Boiteux-Steiner peak-load pricing problem in which a homogeneous resource with a maximum capacity is available to produce output in each period.

The economic structure of the one-exchange telephone call market is given by the rate of demand functions for the two periods

\begin{equation}
  x^1 = x^1(p^1), \quad x^2 = x^2(p^2),
\end{equation}

the required capacity

\begin{equation}
  K = \max_t \{x^t\},
\end{equation}

and total cost

\begin{equation}
  C = BK
\end{equation}

where $B$ is the per-call unit capacity cost.

The marginal cost of a commodity is the change in total cost that results from a one-unit increase or decrease in the production of that commodity. In the long run, capacity can be adjusted to meet maximum demand. Therefore if $x^1 > x^2$ at the observed prices, the marginal costs are

\begin{equation}
  mc^1 = \frac{\partial C}{\partial x^1} = B
\end{equation}

\begin{equation}
  mc^2 = \frac{\partial C}{\partial x^2} = 0.
\end{equation}
This basic model can be used to illustrate several approaches to pricing telephone service. Each rate structure can be evaluated in terms of its effect on economic welfare, measured by the sum of consumers' and producer's surplus

\[ W = CS + PS = \sum_{t} \int_{0}^{T(t)} C dt + \sum_{t} p_t x_t - \beta \max\{x_t\}. \]

1. **Flat-rate pricing**

Set \( p_1 = p_2 = 0 \). The total (usage) cost of local telephone service is recovered by increasing the fixed monthly charge per subscriber. Such flat-rate pricing seems inefficient. But because prices are zero, equipment to measure the number of calls is not needed, and the resulting saving in resources can outweigh the gains of per-call charges. Nevertheless, in order to focus on the design of usage-sensitive rate structures I will neglect measurement costs in this paper.[1]

2. **Average-cost pricing**

Set \( p_1 = p_2 = p^* = \text{average cost} = C/(x^1 + x^2) \). Charging a positive price \( p^* \) per call is seemingly more efficient than flat-rate pricing.

In period 1 capacity is a scarce resource; the reduced demand due to the positive price will reduce calling and therefore capacity and total costs. But calling will also be reduced in period 2, even though excess capacity is available. Compared to flat-rate pricing the net result can be either a gain or loss in welfare.

[1]For a comparison of benefits and costs under flat versus measured rates in a simple case, see Mitchell, 1980.
In Fig. 1 the reduction in calling from flat-rate levels, \( x^1(0) \) and \( x^2(0) \), to average-price levels, \( x^1(p^*) \) and \( x^2(p^*) \), reduces capacity costs by \( \beta[x^1(0) - x^1(p^*)] \), shown by areas \( S^1 + T + U \). At the same time consumer surplus is reduced by \( S^1 \) in period 1 and \( S^2 \) in period 2. Thus, as illustrated in Fig. 1a, welfare is increased if \( S^1 + T + U > S^1 + S^2 \).

If the demand curve is linear, \( S^1 = T \), and a welfare gain occurs when \( S^1 + U > S^2 \). In contrast, Fig. 1b shows a relatively more elastic off-peak demand. In this case \( S^1 + U < S^2 \); average-cost pricing imposes greater welfare losses in the off-peak market than it achieves in net savings in the peak period.

3. Peak-load pricing with a firm peak

Set \( p^1 = \beta \), \( p^2 = 0 \). In period 1, the marginal cost of an additional call is the marginal cost of increasing capacity, \( \beta \). So long as demand in period 2, at a zero price, is less than period 1 output, the marginal cost of an additional call in period 2 is zero. These rates are optimal. Moreover, because capacity is produced at constant returns to scale, average cost and marginal cost are equal per unit of peak output. Therefore, these marginal-cost prices exactly recover total costs.
(a) Increased welfare: \( S^1 + U > S^2 \)

(b) Reduced welfare: \( S^1 + U < S^2 \)

Fig. 1 — Welfare effects of average-cost pricing
4. Peak-load pricing with a shifting peak

Set \( p^1 > 0, p^2 > 0, p^1 + p^2 = \beta \). If the previous rate structure, with \( p^2 = 0 \), would cause the period 2 demand to exceed period 1 demand the result is a "shifting peak." In this case a positive period 2 price is necessary to equalize demands \((x^1 = x^2)\) in both periods. The optimal rates are those that simultaneously (a) bring about this joint peak, and (b) sum to the marginal costs of capacity.

In the joint peak case the marginal cost of a commodity depends on whether its output is increased or decreased. An increase of 1 unit of either \( x^1 \) or \( x^2 \) requires adding a full unit of capacity and therefore has a marginal cost of \( \beta \); but a decrease in either output permits no saving in capacity and has a zero marginal cost. However, the optimal prices of the joint peak case may be interpreted as the marginal opportunity cost of output in each period when capacity is fixed.[2] The opportunity cost of supplying a marginal call in period 1 is the value of the most valuable alternative that must be foregone—the withdrawal of one period-1 call worth \( p^1 \) from some other subscriber. Similarly, the opportunity cost in period 2 is \( p^2 \). And the sum of subscribers' marginal valuations of capacity, \( p^1 + p^2 \), must equal the marginal cost of expanding capacity, \( \beta \). Except where explicitly noted below, I assume hereafter that a "firm peak" exists at the rate structures under consideration.

This one-exchange model of the peak-load pricing problem yields clear-cut guidelines for ratesetting:

- price should be highest in the period with the maximum demand;
- price should exclude capacity costs in a period that has excess capacity;
- optimal prices are equal to marginal costs.

This conventional economic wisdom is an extensive abstraction from the complexities of actual regulated industries. When expanded to include fuel costs, it is perhaps most nearly applicable to the pricing of electricity, an industry in which the bulk of the fixed resources take the form of central generating and transmission capacity which is needed by all consumers.

However, the technology of the telephone industry corresponds less accurately to this paradigm. Instead, capacity is distributed throughout the network in a large number of separate facilities, each of which is available to serve only certain types of calls. To better characterize these aspects of telephone technology, I examine successively more detailed models.
III. SEVERAL ISOLATED EXCHANGES

In each exchange, subscribers place and receive calls only within the exchange. If each exchange has its own rate structure, the pricing problem is that of the previous model. But in practice, a single rate structure must be designed for an entire group of exchanges—for example, all exchanges within one state.[1]

To illustrate this case, it is sufficient to consider just two exchanges, A and B, with demands

\[ x_A^t(p_A^t), \quad x_B^t(p_B^t) \quad t = 1, 2 \]

and capacities

\[ K_A = \max \{x_A^t\}, \quad K_B = \max \{x_B^t\} \]

The total cost of local telephone usage is

\[ C = B_A K_A + B_B K_B. \]

Restricted Rate Structures

Because the exchanges are grouped the rates must satisfy the restrictions

\[ p_A^1 = p_B^1, \quad p_A^2 = p_B^2. \]

[1] In discussing this paper William Vickrey points out that such restrictions on the rate structure would be avoided if the telephone company could signal the price to the subscriber at the time he placed his call. Indeed, such dynamic pricing, when combined with equipment to automatically forward one-way messages, promises substantial improvements over static time-of-day rate structures.
Of course, a common rate structure for both exchanges that is based either on flat rates or on an average price that applies in both periods will satisfy these restrictions. These cases are much like those considered for a single exchange.

Optimal restricted peak-load rate structures can be determined by maximizing the welfare function (5) subject to the pricing restrictions (9). In general, the optimal prices are weighted averages of the marginal costs of the individual commodities

$$p^t = \left[ \frac{b^t_A}{(b^t_A + b^t_B)} \right] mc^t_A + \left[ \frac{b^t_B}{(b^t_A + b^t_B)} \right] mc^t_B$$

where

$$b^t_j = \partial x^t_j / \partial p^t \quad j = A, B$$

It is important to note that the weights for the marginal costs are composed of the slopes of the demand curves, not the number of calls. When a common price must be charged for two commodities with differing marginal costs, some loss of efficiency must result. For example, suppose the common price were set equal to the marginal cost in market A. Then the gap between this price and marginal cost in market B would cause a distortion given by the familiar welfare triangle with area proportional to the slope of the demand curve in that market. Bringing the price closer to $mc_B$ will reduce that loss but create one in market A. The best balance of gain and loss depends on the demand changes in each market, as shown by equation (10).

A key result of restricting the admissible set of rate structures is that the optimal pricing rules can no longer be stated solely in terms of cost data, i.e., set price equal to marginal cost. Instead, as shown in equation (10), demand data, in the form of slopes or elastici-
ties, are commingled into the pricing rule. Two types of peak-load cases need to be considered.

1. Same peak period in each exchange

With maximum demand in period 1 in both exchanges, marginal costs are

\[ mc^1_A = \beta^A, \quad mc^1_B = \beta^B \]

(12)

\[ mc^2_A = mc^2_B = 0. \]

Thus, the optimal rates are

\[ p^1 = \left[ \frac{b^1_A}{b^1_A + b^1_B} \right] \beta^A + \left[ \frac{b^1_B}{b^1_A + b^1_B} \right] \beta^B \]

(13)

\[ p^2 = 0. \]

The requirement that the exchanges be grouped for ratemaking imposes a particular type of data aggregation. Although there are four separate commodities, the admissible rate structure distinguishes only two types of output—total period-1 demand and total period-2 demand (the number of daytime calls and the number of nighttime calls throughout the state). For this case the optimal price for period-1 calls is a weighted average of the per-unit capacity costs in each exchange; in period 2 each exchange has idle capacity and the price is therefore zero.

Because the weights for the capacity costs are the slopes of the demand curves in each period, not the number of calls, this rate structure will not (except by chance) exactly recover total costs when capacity costs vary by exchange. To satisfy the revenue constraint (without resorting to a fixed charge), one or both prices must be adjusted. The best feasible rate structure would modify these prices, taking the demand elasticities in each market at each period into account. As a
result, a positive off-peak price could be efficient if demand is relatively inelastic in that period.

2. Peak periods vary by exchange

Suppose that in exchange A the maximum demand occurs in the first period, whereas in exchange B demand is maximal in period 2. In this case marginal costs are

\[ mc_1^A = \beta_A, \quad mc_1^B = 0 \]

(14)
\[ mc_2^A = 0, \quad mc_2^B = \beta_B \]

and the optimal rates are

\[ p_1^1 = \frac{b_1^A}{(b_1^A + b_1^B)} \beta_A, \quad p_2^2 = \frac{b_2^B}{(b_2^B + b_2^A)} \beta_B. \]

(15)

In period 1, the price is a fraction of the marginal capacity cost in exchange A. The relationship for period 2 price is similar. In each case the proportions depend on the demand slopes of the commodities in each exchange. Again, these optimal prices will not generally satisfy the budget constraint and the best feasible prices would modify these rates on the basis of demand elasticities.

**Quantity-weighted marginal costs**

A feasible method of meeting the budget constraint is to construct the prices using quantity weights in place of slope weights in the previous formulas. Let

\[ \theta_t^A = \frac{x_t^A}{x_t^A + x_t^B}, \quad \theta_t^B = \frac{x_t^B}{x_t^A + x_t^B} = 1 - \theta_t^A \quad t = 1, 2 \]

\[ \theta_t^B = 1 - \theta_t^A \]

\[ \theta_t^B = 1 - \theta_t^A \]

\[ \theta_t^B = 1 - \theta_t^A \]
be the proportions of the grouped outputs that occur in each exchange in period $t$. For case 1 (same peak period) set

$$p_1^1 = \theta_A^1 \beta_A^1 + \theta_B^1 \beta_B^1, \quad p_2^2 = 0.$$  

For case 2 (different peak periods) set

$$p_1^1 = \theta_A^1 \beta_A^1, \quad p_2^2 = \theta_B^2 \beta_B^2.$$  

These rates, based only on quantity information, can be given an informative interpretation in terms of suitably defined marginal costs.

**Marginal cost of a group of commodities**

When commodities are grouped it is not immediately apparent just what the "marginal cost" of the aggregate is. To define its marginal cost we must specify how each of the components of the group changes when the group itself changes by one "unit."

One plausible definition is to specify that the quantities of each commodity in the group vary proportionately. Thus for a change $dx^t$ in the group quantity let the components change by

$$dx^t_A = \theta_A^t dx^t, \quad dx^t_B = \theta_B^t dx^t.$$  

The marginal cost of the grouped output in period $t$ is then

$$mc^t = \frac{\partial C}{\partial x} = \theta_A^t mc^t_A + \theta_B^t mc^t_B \quad t = 1, 2.$$  

Thus, the group marginal cost is defined as the quantity-weighted average of the individual commodity marginal costs.

With constant returns to scale, the rates (equation (17) or (18)) based on this measure of marginal costs are feasible. And in one special case they will be optimal—when the individual commodities that make up a group have the same elasticities of demand. To see this,
write the equation (15) for the optimal restricted rates in terms of elasticities:

\[
 p^t = \frac{\eta^t_{A} x^t_{A}}{\eta^t_{A} x^t_{A} + \eta^t_{B} x^t_{B}} mc^t_{A} + \frac{\eta^t_{A} x^t_{A}}{\eta^t_{A} x^t_{A} + \eta^t_{B} x^t_{B}} mc^t_{B}
\]

When \( \eta_A = \eta_B \), the weights for the terms \( mc^t_{A} \) and \( mc^t_{B} \) are just the quantity weights \( \theta^t_{A} \), \( \theta^t_{B} \). In this case, commodities are homogeneous in terms of demand, and the optimal pricing rule requires only cost data.
IV. A NETWORK OF EXCHANGES

Each exchange has intra-exchange calling as in the previous model. In addition, there are inter-exchange (AB) calls which make use of capacity in the originating and terminating exchanges and also require a third capacity component—trunking facilities that connect the exchanges. The key feature of this model is the introduction of joint production, which occurs when local exchange switching capacity is shared by two different commodities.

Demands are

\[ x^t_A = x^t_A(p^t_A) \]
\[ x^t_B = x^t_B(p^t_B) \]
\[ x^t_{AB} = x^t_{AB}(p^t_{AB}) \]

The capacity constraints are

\[ K_A = \max \{x^t_A + x^t_{AB}\} \]
\[ K_B = \max \{x^t_B + x^t_{AB}\} \]
\[ K_{AB} = \max \{x^t_{AB}\} \]

where I assume each inter-exchange call requires the switching capacity of an intra-exchange call in each of the two exchanges as well as inter-exchange trunking. Total costs are then

\[ C = \beta_A K_A + \beta_B K_B + \beta_{AB} K_{AB} \]
The optimal prices are obtained from the Kuhn-Tucker conditions of the mathematical program. The prices are

\[
\begin{align*}
    p_A^1 &= \mu_A^1, \\
    p_A^2 &= \mu_A^2, \\
    p_B^1 &= \mu_B^1, \\
    p_B^2 &= \mu_B^2, \\
    p_{AB}^1 &= \mu_A^1 + \mu_{AB}, \\
    p_{AB}^2 &= \mu_B^2 + \mu_A^2,
\end{align*}
\]

(25)

where \( \mu_j^t \) is the dual variable in period \( t \) for capacity of type \( j \). The central result is that even when there are as many prices as commodities, the technological interdependence of the separate markets destroys the simple correspondence between maximum demands and maximum prices. However, with firm peaks, the optimal prices are equal to the marginal costs of the individual commodities.

For example, suppose that exchange A and inter-exchange calls are day peaking (\( x_A^1 > x_A^2 \) and \( x_{AB}^1 > x_{AB}^2 \)) and exchange B is night peaking (\( x_B^1 < x_B^2 \)). The optimal prices will be

\[
\begin{align*}
    p_A^1 &= \beta_A, \\
    p_A^2 &= 0, \\
    p_B^1 &= 0, \\
    p_B^2 &= \beta_B, \\
    p_{AB}^1 &= \beta_A + \beta_{AB}, \\
    p_{AB}^2 &= \beta_B.
\end{align*}
\]

(26)

Thus the inter-exchange calls should pay positive prices in both periods, not only in their peak (\( t=1 \)) period. Moreover, despite the

fact that $AB$ calling is highest in period 1, $p^{2}_{AB}$ could mathematically exceed $p^{1}_{AB}$, although this is unlikely in the particular example of local and inter-exchange calls.

**Restricted rate structures**

Here we reach the "realistic" case for telephone ratemaking. In practice, rates might well be restricted to be the same for all intra-exchange calls at a given time of day throughout a region or state, with separate rates applying for inter-exchange calls. Frequently, however, the same "off-peak" percentage discount is applied to both types of calls. In this case the restrictions are

$$p^{t}_{A} = p^{t}_{B}, \ t = 1, 2$$

$$(27) \quad p^{2}_{AB} = \lambda p^{1}_{AB} \quad \text{where} \quad \lambda = p^{2}_{A}/p^{1}_{A} = p^{2}_{B}/p^{1}_{B}.$$ 

Effectively, there are three rate parameters—the mean levels of the intra-exchange and inter-exchange rates and the percentage discount in the off-peak period.

In principle, the mathematical program for the welfare-maximizing network prices can be solved for any specified constraints on the rate structure. For a small problem—such as this example—this is quite feasible. But for realistic situations, the dimensions of the problem are substantially greater. $M$, rather than two exchanges, must be considered. Because the rate of demand varies over both the daily and weekly cycle, $N$ distinct periods must be analyzed. And there are several levels of inter-exchange calls, conventionally grouped according to distance bands.
V. EVALUATING THE EFFICIENCY OF TELEPHONE RATE STRUCTURES

A practical approach is to use the structure of the demand and cost model to evaluate the welfare effects of alternative rate structures without attempting to achieve a global optimum. This approach should be undertaken at two levels.

1. A Given Rate Structure

A particular rate structure specifies a grouping of commodities into time periods, distance bands, and perhaps geographic areas. The quantity-weighted marginal costs of each grouped output can be calculated by proportionately incrementing demands of each commodity in the group. (For example, if the peak period is 8 a.m. - 5 p.m., weekdays, the traffic load curve at those hours can be increased by a constant percentage). By calculating the "Ramsey number" of each group k at current prices and output levels

\[ R_k = \eta_k(p_k - m_c k)/p_k \]

the group marginal costs can be compared with prices. [1] If rates are optimal (given the rate structure), all of the Ramsey numbers will be equal. If not, welfare can be increased by raising rates for groups with low Ramsey numbers and reducing rates for high \( R_k \) values.

2. Alternative Rate Structures

Some redesigning of the rate structure may yield welfare gains at least as large as those achievable by adjusting rate levels. Two closely related questions must be investigated--the number of different

---

prices to charge, and the particular commodities to be included in each
group. For example, local telephone rates might be limited to two price
levels throughout the week, with the particular hours that peak prices
apply determining which telephone calls are grouped together.

Guidance for grouping commodities is provided by two results from
the earlier analysis:

1. If, within each group, all commodities have the same marginal
cost, then a group price equal to the common marginal cost will
be (first-best) optimal. This will be true even if the commo-
dities have different elasticities of demand.

2. If, within each group, all commodities have the same elasticity
of demand, then price should be equal to the quantity-weighted
average of the commodity marginal costs.

For the telephone network, the marginal costs of several commodi-
ties will be similar when they (a) have the same peak period, and (b)
use equipment that has similar unit capacity costs. As for demand elas-
ticities, they will perhaps be similar when exchanges are grouped by
type of customer.

These general considerations suggest that efficient grouping will
be promoted by combining commodities according to similarities in both
marginal costs and demand elasticities. For example, alternatives to a
proposed 8 a.m. - 5 p.m. peak period could be considered by comparing
both the demand elasticities and marginal costs at, say, 6, 7, and 8
p.m. with those in earlier hours. Hours that clearly follow the earlier
elasticity and cost pattern readily suggest an extended period for the
time-of-day rate structure. A mixed pattern of elasticities and costs, however, would require evaluating different combinations of grouped hours.

To evaluate a change in the number of prices in the rate structure additional data are needed. Practical restrictions on the number of separate rates are presumably due to the "transactions costs" the subscriber must bear to cope with an increasingly detailed structure of rates, and the additional administrative complexity for the telephone company of calculating, defending and revising such rates. Measurements of transaction costs are not readily available. However, one can demonstrate the size of the efficiency gain that could be realized by adding an additional rate, or on the other hand, the efficiency cost of simplifying the rate structure in a specified manner. These values can then be compared to subjective assessments of the hassle of coping with rate structures of differing complexity.
VI. SUMMARY

The design of appropriate rate structures for local telephone calls should be determined by the technology and cost characteristics of the local network. Apart from the equipment dedicated exclusively to serve each subscriber, nearly all of the costs of local telephone service are due to providing capacity sufficient to meet maximum demands. Thus, some form of peak-load pricing is desirable. A uniform average-cost price at all hours may be less efficient than a flat-rate tariff which charges nothing per call, even if metering were costless.

Switching and trunking capacity is distributed throughout the network and jointly used by different types of calls. As a result, optimal prices may be positive when demand is below the maximum level, and the highest rate need not occur at the hour of peak demand.

Realistic rate structures can have only a limited number of separate prices, requiring that individual commodities be aggregated into groups. An efficient rate structures will combine hours and routes that have similar marginal costs and demand elasticities.
REFERENCES


