Chapter 4

Issues in the Pricing of Broadband Telecommunications Services

Bhaskar Chakravorti
William W. Sharkey
Padmanabhan Srinagesh
Bellcore

1. INTRODUCTION

In this chapter, we consider some of the long-term issues associated with the pricing of broadband telecommunications services. Future broadband digital networks will be based on a technological platform that will support voice, data, image, and video services. Policymakers and regulators have not yet reached a consensus on how these services will be deployed and priced in a competitive market. This chapter seeks to provide economic inputs that may be useful in developing appropriate technological and regulatory policies as these new services are deployed.

Consider first a very simple arithmetical exercise conducted by Robert Pepper of the FCC.[1] A naive, cost-based price structure would price every ATM cell alike, regardless of the traffic it represented. If each asynchronous transfer mode (ATM) cell is priced identically, and this price is chosen so that a local call costs a penny per minute, he argued that a 2-hour video movie (at 45 Mbps) would cost about $843.75. With compression and transmission at T1 speeds, the price for a movie would fall to about $30. This is unacceptably high when compared to substitutes such as videocassette rentals.

Pepper's solution was to suggest that every residential customer should be given an access line with sufficient bandwidth for a voice plus TV channel and be charged a flat rate equal to today's average expenditure on local calls and basic cable service (about $40). This pricing approach opens up some very lucrative arbitrage opportunities. An apartment building could install a PBX

(private branch exchange) and order a few access lines, each with the capacity of 672 voice circuits (45Mb/64Kb). These voice services could be resold to tenants for considerably less than $10 per month per tenant and generate huge profits for the reseller. Tenants could purchase video services from other vendors such as cable companies or use antennas to receive over-the-air broadcasts. This arrangement would be considerably less expensive than integrated access and would allow those who do not want cable television to benefit the most.

Arbitrage opportunities such as the one described here could arise if services with very different bandwidth requirements are served by the same technology and are therefore priced similarly. The history of telecommunications shows clearly that when arbitrage opportunities are made available, the market responds. WATS (wide area telephone service) resale, aggregation and resale of Multi-Location Calling Plans (MLCPs), and International Discount Telecommunications' "callback" service all arose from arbitrage incentives built into existing pricing structures. In a broadband context, large customers could use T1 service to send aggregated voice traffic from one location to another. T1 service provides bandwidth of 1.5 Mbs, which is sufficient to carry 24 simultaneous voice calls. Coincidentally, 1.5 Mbs is sufficient to carry compressed video transmissions of acceptable quality. This may suggest that a T1 line can provide an adequate transport mechanism to support video to the home. However, the cost of a T1 line (typically in the range of $500–$1,000 per month) makes it uneconomical for use in providing residential video service. In addition, T1 service allows for two-way transmission and channelizes the available bandwidth so that the 24 calls in progress can be separated at the receiving end. Neither of these functions is necessary for a residential video service. For both economic and technical reasons, T1 is not a platform that simultaneously supports the needs of aggregated voice and video service.

An alternative approach would be to provide a service that is designed to meet the specific economic and technical requirements of residential video entertainment. Asymmetric Digital Subscriber Loop (ADSL) may be such a solution. It provides one-way 1.5 Mbs transport into the home, is not channelized, and is based on different hardware than T1 service, allowing it to be tariffed at a rate that residences may be willing to pay. Furthermore, ADSL-based video service will not meet the technical requirements of aggregated voice transport. This suggests that a product line of services (T1 and ADSL in the prior example) that are designed to serve a spectrum of customer needs can be more effective than a single service.

There is no single "economically correct" model that can be used to price broadband services or to resolve the issues discussed earlier. In the following sections of this chapter we survey a significant body of relevant work that could be used to understand issues related to the pricing of broadband services. We consider the application of pricing methodologies in both partially regulated and in fully competitive markets, and focus attention to the issue of customer resale

and arbitrage in light of these economic models. Some of the specific pricing methodologies are presented in mathematical terms. We attempted, however, to convey the relevant ideas in nontechnical language at the beginning of every section. In a brief concluding section we indicate how the results in all of the sections can assist in developing a practical tariffing framework.

2. DEMAND-BASED PRICING

We first consider the standard economic approach to pricing in multiple product firms assuming profit maximization as an objective.[2] If a firm produces a single product, the determination of a profit maximizing price requires a knowledge of the demand and cost functions facing the firm. The demand function is simply a schedule of output quantities that the firm expects to sell at each conceivable price, and the cost function describes the total cost associated with each output level. At a sufficiently high price demand will be negligible so that total profits will be small (even though the profit per unit sold is large). As price is lowered, more units can be sold, and as long as the increased revenue exceeds the increased cost, profits to the firm will increase. The optimum profit maximizing output is the one in which the incremental (or marginal) revenue from an increase in sales exactly matches the marginal cost of an increase in output.

When a firm produces more than one output, similar principles apply, but the firm must now take into account the interactions on both the demand and cost side of increases in any one of its outputs. To describe the profit maximizing rule in this case it is necessary to introduce some mathematical notation. Suppose that $q = (q_1, \ldots, q_n)$ represents a vector of possible outputs for the firm and let $C(q)$ represent the total cost of producing the output vector q. If demands for each of the firm's products are independent, it is possible to write the inverse demand function $p_i = P_i(q_i)$, which expresses the amount that customers are willing to pay for the last unit produced when output is q_i.[3] If the firm is unregulated, profit maximization is achieved by equating marginal revenue with marginal cost in each market. The expression for marginal revenue is commonly expressed in terms of the elasticity of demand:

$$\eta_i = \frac{p_i\, dq_i}{q_i\, dp_i},$$

so that from the equality of marginal revenue and marginal cost one can derive the expression[4]:

$$p_i\left(1 + \frac{1}{\eta_i}\right) = \frac{dC}{dq_i}.$$

Because marginal costs are typically greater than zero, it follows that a monopolist will always choose a price at which marginal revenue is positive, which means

that demands will always be elastic (i.e., $\eta < -1$). If the demands are interrelated, the appropriate marginal revenue must be adjusted to reflect the effect of a change in the price of one product on the revenues that may be obtained in all other markets.

The formula for the optimal pricing rule for a multiple product monopolist is a special case of the so-called *Ramsey pricing rule*, which could be applied whether or not the firm is regulated. Where a monopoly firm seeks to maximize its profits without any constraints on its level, a regulated firm may have as its objective the maximization of social surplus[5] subject to a budget constraint that is imposed by the regulatory process. The Ramsey pricing rule in the case of independent demands is given by the formula:

$$\frac{p_i - \frac{dC}{dq_i}}{p_i} = -\frac{k}{\eta_i}.$$

The number k is chosen to satisfy the budget constraint, where $k = 1$ corresponds to unconstrained profit maximization, $0 < k < 1$ corresponds to budget constrained pricing when there are increasing returns to scale so that prices in excess of marginal cost are required to recover total costs, and $k < 0$ corresponds to budget constrained pricing under decreasing returns to scale.[6]

The Ramsey pricing rule is generally accepted by economists as an appropriate methodology for pricing of heterogeneous outputs. In the pricing of broadband telecommunications services, however, it may not be appropriate to assume that demand functions are independent. These outputs can be either substitutes or complements for one another, and it is necessary for either a profit or surplus maximizing firm to take account of the relevant cross-elasticities of demand. Although the simple formulae defined earlier no longer apply, the derivation of profit and surplus maximizing prices is well understood theoretically and can be readily implemented given appropriate data. These data include estimates of the appropriate marginal costs and estimates of both own-price and cross-price elasticities of demand.[7] This information may be difficult to obtain, particularly for new services that would be offered on a broadband network.

There are two potential drawbacks to the Ramsey pricing methodology in addition to the informational requirements noted earlier. First, Ramsey prices may be perceived as inherently unfair and therefore politically nonviable in a regulatory environment. This follows because the rule requires that the markup of price above marginal cost should be the greatest in those markets in which the elasticity of demand is the least. From the point of view of overall economic efficiency, this rule makes perfect sense because customers with inelastic demands will curtail their consumption less than would customers in more elastic markets. If such a pricing methodology had been applied to traditional telephone services, access to the network and local usage would have borne a significantly larger share of common costs than interexchange toll. It is unlikely that such an outcome would have been accepted by state and local regulators.

The second potential difficulty with Ramsey pricing is that it does not account for the possibility of competition in one or more of the firm's markets. It may well happen that markets with inelastic demands are also served by active or potential competitors, who could profitably beat the Ramsey price. In a fully deregulated marketplace, the presence of competition does not pose any particular difficulties. In this case, the properly interpreted Ramsey pricing rule would take account of the increased elasticity in markets in which competitive forces were most vigorous and accordingly set prices in these markets close to marginal cost.

3. COST-BASED PRICING

For regulated firms, there are theoretical justifications supporting the use of a Ramsey pricing approach as outlined in the previous section. However, as a practical matter, regulated firms are often expected to set prices on the basis of fully distributed costs. In this section we describe a method of cost-based pricing that is, in some sense, the most reasonable among the various possible methods of cost-based pricing.

Cost-based pricing takes as given the vector q of customer demands. Rather than attempting to find the outputs q that maximize profit, or social surplus, cost-based pricing seeks to determine prices p_i that allocate the total cost $C(q)$ in a fair and consistent manner. In this section we demonstrate one method by which a pricing rule can be derived by means of technical properties, or axioms, that one might impose on the set of all conceivable pricing rules.[8] Although this section contains more mathematical notation than most other sections, the mathematics is included only for a precise statement of results. The reader can obtain a general understanding of the methodology without necessarily following the details of the mathematical derivations.

The cost-based pricing approach might be utilized in a regulatory framework, when regulators must consider whether it is in society's interest to allow telephone companies to deploy broadband networks that are capable of delivering broadband services. Because voice, video, and data services will share a substantial amount of common plant and equipment in a broadband network, an important input into any pricing approach is a sensible procedure for the allocation of such common costs. Currently accepted cost allocation methodologies, however, do not imply that every bit must be priced identically. In this section we briefly describe how one such cost-based pricing methodology is defined in the current economics literature.

Cost allocation methodologies can be defined by enumerating *properties* that a reasonable person might want to impose on the set of all possible pricing rules. These properties include ordinary accounting restrictions that are noncontroversial, as well as properties that seek to ensure that the pricing rule is perceived as fair. One set of properties that has been extensively studied is the following:

Property 1 (Cost Sharing): Revenues should exactly recover total cost. We note that total cost includes a payment to equity holders in the firm, which is required in order to allow them to earn a "fair rate of return" on their investment.

Property 2 (Monotonicity): If an increase in the output of a service unambiguously increases total cost, that service should be assigned a positive price.

Property 3 (Additivity): If it is possible to additively decompose the total cost of producing a set of outputs into two or more component cost functions, then the pricing rule should be additive over the component functions.

Property 4 (Consistency): If two commodities have exactly the same effect on total cost, they should be charged exactly the same price.

Property 5 (Rescaling Invariance): If units of measurement are changed, then prices should be rescaled in the natural way.

It has been demonstrated[9] that these five properties define a unique pricing rule, which has a natural interpretation as an average of marginal costs. The so-called *Aumann–Shapley pricing rule*, which the earlier properties define, is given by the formula:

$$p_i^{AS}(\bar{x}) = \int_0^1 \frac{\delta C}{\delta x_i}(t\,\bar{x})\,dt$$

which represents the price assigned to output i when the aggregate output vector $\bar{x}$ is produced and $C(x)$ represents the cost of producing any output x. Thus one sees that the Aumann–Shapley price for output i is the average of the marginal costs of producing an additional unit of output i, as outputs are expanded along the path from 0 to $\bar{x}$.[10] We note that it is possible to define other axiomatic pricing rules that are related to Aumann–Shapley pricing. For a full discussion of these rules, the reader is referred to the papers by McLean and Sharkey cited in the footnotes.

When applied to broadband telecommunications services, the Aumann–Shapley pricing rule defines prices as a function of traffic characteristics such as the frequency of arrival, duration of the call, and the bandwidth requirement.[11] Because the costs associated with traffic intensities of services offered on a broadband network consist of congestion and delay for other services, these costs can also be reflected in the cost-based pricing approach. Let $q = (q_1, \ldots, q_n)$ represent a vector of n "service classes" (e.g., voice, video, data, etc.). Demands for service arrive at a transmission point consisting of k channels and, for simplicity, we assume that the arrival of a "call" of type i is a Poisson process so that q_i measures the probability that an additional call arrives at any instant of time. Calls con-

tribute to overall system congestion in two ways. First, each type i call has a duration, or size, that is exponentially distributed with mean r_i and variance r_i^2. Second, the cost of providing a service depends on the number of processors that are simultaneously required and the different protocols required in transmission. Thus, arriving calls also contribute to system congestion through the number, d_i, of simultaneous channels that are required for the duration of a type i call.

Several different cost functions can be constructed depending on the queue discipline and the buffer size. Queue discipline refers to the order in which arriving jobs are processed. Buffer size refers to the capacity of the system to hold jobs prior to the commencement of service or during service for store and forward applications. Let k represent the number of channels and let B represent the buffer size. Let $g(k,B)$ represent the cost of building a system with k channels and a buffer of capacity B. Typically, g is an increasing function of k and B. Given k, B, and the values of q_i, r_i, and d_i, let $\beta_i(k,B;q,r,d)$ be the blocking probability for a call of type i. Finally, let $w_i(k,B;q,r,d)$ represent the expected waiting time for a call of type i.

We next consider two models that may be used to define a cost function for a general telecommunications design problem:

Model 1: Let $\beta_1^*, \ldots, \beta_n^*$ and $w_1^*, \ldots, w_n^*$ be "acceptable" blocking probabilities and expected waiting times. In this model, the design problem is to minimize the system cost $g(k,B)$ of constructing a facility such that blocking and waiting costs are within acceptable limits. Solving this optimization defines a cost function C as a function of outputs $q = (q_1, \ldots, q_n)$, where each output i is characterized by its service time r_i and its bandwidth requirement d_i.[12]

Model 2: Let c_i be the value to a type i caller if his call is not blocked. Equivalently, c_i represents the economic loss associated with a blocked call of type i. Let γ_i represent the economic loss associated with a unit of time spent waiting in the queue. In this model the system designer wishes to maximize social surplus, which is equivalent to minimizing the sum of capacity cost plus blocking and waiting costs.[13]

A useful special case of Models 1 and 2 is one in which buffer capacity is equal to zero and server requirements are homogeneous, with $d_i = 1$ for all i. Then the blocking probability is the same for all call types and is given by the Erlang loss formula. The cost functions in Models 1 and 2 are then defined by integer optimization problems that can be solved in principle for any vectors of demand parameters q, r, and d. Naturally there are substantial computational difficulties associated with this approach, and specific pricing rules have, so far, been obtained only for even more specialized situations.[14]

Although it does not appear likely that pricing rules based on cost allocation procedures can fully resolve the Robert Pepper conundrum noted in the introduction, the cost allocation approach clearly indicates that average cost per cell pricing is overly simplistic. This follows because the cost function that appropriately models the cost of providing a variety of services depends on the full array of traffic characteristics that characterize the services. Because an Aumann–Shapley price is an average of marginal costs, the Aumann–Shapley pricing rule depends on a complex, and economically meaningful, set of demand parameters, rather than simply on the number of cells that are transmitted.

Cost-based pricing rules have been criticized by economists on several grounds. In their most elementary form, as presented earlier, these rules do not take any account of customer demand elasticities. Furthermore, despite their axiomatic foundations, cost-based pricing rules are inherently arbitrary from a purely economic perspective. That is, they ignore traditional concepts of economic efficiency that relate marginal benefits, or marginal revenues, to marginal costs of production. In addition, cost-based pricing rules are, by definition, unresponsive to competitive pressures that differ in different markets. Thus, cost-based pricing rules have the potential for inviting entry even in situations in which such entry would increase total industry costs. Finally, pricing rules based on cost allocation procedures do not take any account of the potential for customer arbitrage among services. In the remaining sections of this chapter we consider these issues in greater detail.

4. SUBSIDY FREE AND SUSTAINABLE PRICING

In an environment of free entry, demand-based pricing tends toward charging what the traffic will bear, whereas cost-based pricing leads to rules that are completely unresponsive to customer-demand elasticities. In a partially regulated but partially competitive environment, some degree of flexibility, but something less than full flexibility on the part of the regulated firm, appears to be called for as an alternative to either of the approaches previously discussed. The theories of cross-subsidization and of sustainable pricing seek to establish an appropriate degree of flexibility by identifying permissible bounds on prices for individual outputs and collections of outputs.

Thus, the theory of *subsidy-free pricing* is primarily an application of the techniques of cost-based pricing in situations in which there is competition in one or more of the regulated firm's markets. In traditional telecommunications pricing, the allocation of non-traffic-sensitive costs associated with the local loop has been a persistent issue. For these costs, it is well known (and well documented) in the economics literature that all cost allocations are inherently arbitrary, and that reliance on specific fully distributed cost allocation rules in a partially competitive environment can lead to undesirable outcomes, both for

telecommunications consumers and for the regulated firm. Nevertheless, there exist in the literature well-established procedures for identifying bounds on permissible cost allocations such that no group of consumers is disadvantaged by any other group of consumers. We consider these issues in this section.

The fundamental principle of the theory of subsidy-free pricing is that no group of customers should pay more for the outputs that it consumes than it would if served by a specialized firm devoted to its needs alone.[15] If it is assumed that each product of a multiproduct firm is consumed by a distinct group of customers, then subsidy-free pricing requires that no subset of customers pays more than the stand-alone cost of serving them. If S represents any subset of customer classes and q^S represents the outputs associated with S, then the subsidy-free conditions can be written:

$$\sum_{i \in S} p_i q_i \leq C(q^S).$$

In addition, the firm must continue to break even (including the return to equity holders) so that $\Sigma_{i \in N} p_i q_i \geq C(q)$. An equivalent way of defining subsidy-free prices is in terms of the "incremental cost" of serving any subset of consumers. According to this criterion, every group should pay at least the incremental cost of serving it so that:

$$\sum_{i \in S} p_i q_i \leq C(q) - C(q^{N-S}).$$

where $C(q^{N-S})$ represents the cost of serving all customers other than S. According to this approach, as long as every subset pays enough to cover its incremental cost, any remaining cost can be assigned arbitrarily to any group of customers without violating the principle of fairness implicit in the subsidy-free constraints.

To consider a very simple example, let the cost function be given by $C(q_1,q_2) = f + c_1q_1 + c_2q_2$, where f represents a fixed cost of production, and c_1 and c_2 represent constant marginal costs. Such a cost function is the simplest kind of function in which issues of cost allocation arise. In this case a price vector $p = (p_1, p_2)$ is subsidy free whenever $c_i \leq p_i \leq c_i + f/q_i$ for each service i. A slightly more complicated but also more realistic example is one in which stand-alone cost functions are given as follows:

$$\begin{aligned} C(q_1, 0) &= f_1 + c_1q_1 \\ C(0, q_2) &= f_2 + c_2q_2 \\ C(q_1, q_2) &= f_{12} + c_1q_1 + c_2q_2. \end{aligned}$$

In this case the subsidy-free constraints require that $c_1 + \frac{f_{12} - f_2}{q1} \leq P_1 \leq c_1 + \frac{f_1}{q_1}$, and that a similar constraint holds for p_2.

In a free entry environment, in which all potential entrants have access to the same technology, embodied in the cost function $C(q)$, subsidy-free prices also correspond to "sustainable" prices. Sustainable prices are defined as prices that do not invite entry when the industry is a "natural monopoly" (i.e., total costs are minimized when one firm produces the industry output). This is easily seen by referring to the stand-alone test for cross-subsidization. If the stand-alone test does not hold for a particular subset S of consumers, then it would be possible for an entrant to choose alternative prices $p_i' < p_i$ for each customer i and still make positive profits. Of course, this kind of entry is more likely to occur if entry barriers are extremely low and customers are highly responsive to possible small price differences or conditions that may not apply in telecommunications markets. Nevertheless, the theory of subsidy-free pricing defines a framework for pricing in the presence of competitive pressures that may be a useful consideration as a firm faces the complex issue of pricing broadband services.

5. NONLINEAR PRICING AND THE ARBITRAGE ISSUE

A nonlinear price structure is one in which a consumer's bill is not proportional to the amount he or she purchases. Billing structures consisting of a fixed monthly fee and a fixed usage charge per unit are nonlinear, as a doubling of the units purchased will not result in a doubling of the bill. The price structures for most telecommunications services are nonlinear.

A brief history of the forces responsible for the widespread use of nonlinear prices provides some insight into the arbitrage possibilities that may arise if broadband services are offered. Nonlinear pricing rules have been adopted in the past as a consequence of a regulated telephone company's need to offer volume discounts to its large users. This need arises from the fact that the costs of networks (or the facilities that comprise them) are largely fixed, and the variable costs associated with providing service on a network that is in place are comparatively small. A customer with a sufficiently high level of use will find a tariff structure such as MTS (Message Toll Service) with usage-sensitive charges more expensive than a dedicated facility. Thus, competitive entry into interexchange telecommunications was initially limited to large firms that formed private networks. It is worth stressing that the alternative available to large users involved large fixed costs and no usage-related costs, and that this alternative was typically available on a point-to-point basis. AT&T initially sought to prevent bypass to private facilities by pricing private-line services attractively. As AT&T private lines were provided out of existing facilities that were installed to meet future demand growth, the additional cost of providing these lines was close to zero. The choice between MTS and private lines resulted in a nonlinear price structure, as large users on private lines paid a smaller price (on average, and for additional calls) than did those on MTS.

Volume discounting of switched services was developed along similar lines. MCI introduced its Execunet tariff with the intention of sharing facilities across medium to large users whose traffic was not concentrated in a few routes. AT&T developed WATS service to appeal to the customers who might find Execunet better than either MTS or private-line services. The widespread availability of WATS-like services led to the first major wave of aggregation and resale. It was relatively easy for WATS resellers to set up operations based on inexpensive PBXs that allowed subscribers to dial in, authenticate themselves, and then dial out on a WATS line that connected them to their called party over the public-switched network. The extent of this resale market was probably unanticipated by AT&T. At its peak, the resale market consisted of more than 1,000 WATS resellers. Many have since gone out of business. A factor that probably played a part in the contraction of this industry was the decision by AT&T and the other long distance companies to flesh out their product line by offering a range of options to medium-sized customers. ProAmerica (later ProWATS) and other new products such as Reach Out America and MCI's Friends and Family have reduced the difference in the unit prices paid by large and medium customers. Nevertheless, these differences persist, and some resellers continue to serve niche markets.

As noted in Briere (1990, p. 219),[16] arbitragers are shifting their focus to profitable opportunities created by Multi-Location Calling Plans (MLCPs). These tariffs are designed to meet the needs of customers with offices in many locations, none of which is large enough to benefit from the volume discounts in tariffs such as AT&T's Megacom or ProWATS. The MLCP allows the firm to enroll all its locations in the plan and compute its discount based on the total volume at all locations. MLCP resellers take advantage of this tariff by aggregating customers into collections with enough aggregate volume to benefit from the volume discounts and by jointly applying for an MLCP account. This business is estimated to amount to more than $1.6 billion per year.

The ability of large users to use an alternative access provider implies that local telephone companies can successfully compete only by developing volume discounts aimed at the very largest users. Moreover, in the presence of competition from resellers, economic theory suggests that in order to compete effectively, a provider must flesh out the product line to reach large, medium, and small users. The use of volume discounts is likely to remain important in the broadband environment. It is likely that large suppliers of video services will face bulk tariffs for switched bandwidth that offers them connectivity to their subscribers. Nonlinear pricing theory implies that unless a range of pricing options suitable for medium-sized users is developed, resellers may have the incentive to enter and compete by reselling services intended for video providers.

Another issue is whether voice, video, data, image, and multimedia traffic will each be tariffed independently, or whether packages of switched transmissions services that support multiple applications will be offered. Price structures that offer

separate discounts for voice and data (for example) will appeal to customers who have large volumes for either data or voice or both, but not to those who have the same total volume of use yet moderate volumes of each use. Discounts based on total volume across all uses may induce users with high volume in one use, but relatively lower total use, to seek specialized service from competing providers. Whether discounts should be targeted at specific applications or offered based on total use across all applications is a question for further study.

Another important question concerns the form in which volume discounts are offered. Should volume discounts be offered to customers who presubscribe to the appropriate plan and pay one-time installation charges and high monthly fees, or should incremental discounts be offered automatically to customers whose use exceeds prespecified levels? The former approach places the risk of making the wrong choice of plan on the customers. Although telephone company revenues would appear to be higher under this approach, resellers who purchase bulk service from the telephone company and aggregate users in order to minimize risk of usage variation may offer highly effective competition, thus giving the telephone company an incentive to introduce automatic discounts.

We have argued that the development of nonlinear price structures has been motivated largely by competition for large users, but that pricing methodologies should also take account of the potential entry by resellers. Much of the theory of nonlinear prices considers the efforts of a monopolist to segment his or her market through the use of selective discounts.[17] A recent paper by Mandy considers the sustainability of these prices in a competitive market.[18] The main result of the paper is that nonlinear prices will not be sustainable. Competitive firms will seek to reduce market share among those groups paying a low average price by raising the price to these groups, and they will compete for market share among groups paying a high average price by lowering the price to them. This will unravel the nonlinear price, and all groups will pay the same price in equilibrium.

This result is critically dependent on the assumptions that all firms have the same cost structure, that there are no marketing costs, and that there are no quality differentials across firms. The specific ways in which these assumptions are violated will determine the form of nonlinear pricing that could emerge in competitive equilibrium. A clear understanding of cost differences across firms and the marketing costs associated with reaching consumers with limited information is therefore a topic in need of additional research.

6. PRIORITY PRICING AND INTERRUPTIBLE SERVICE

There is now a large literature on the optimal design of product lines in which any one product of the line can substitute for other members of the line. This literature has significant implications for the pricing of services aimed at different applications, although none of the theories specifically consider broadband services.

Consider the provision of applications such as telephony, electronic mail, video services, remote access to host computers, and distributed processing. These applications have differing requirements for underlying network attributes such as security, bandwidth, lost cells, delay, and delay variation. Thus, voice can tolerate a cell-loss probability on the order of 10^{-4}, whereas interactive compressed video requires that the loss probability be on the order of 10^{-10} for acceptable Quality of Service (QOS). Delay and delay variation criteria for voice and video transmissions are roughly 10 msecs. File transfer can often sustain delays on the order of 10 sec while meeting QOS.[19] Therefore, there is considerable heterogeneity in applications' needs for network attributes.

In addition, different customers have differing willingness to pay for these attributes. In the European context, if the network owner integrates vertically into the provision of applications such as electronic mail and video services (i.e., provide content as well as distribution), and if the technical interfaces presented to the customers do not allow for easy substitution across services, then each service can be priced in accordance with the theory of multiproduct monopoly. This theory has been well studied.[20] QOS for each service can be ensured through the use of appropriate congestion control systems and through related resource reservation schemes. This approach would be sustainable only if the network provider could ensure that a line purchased for a particular application (say video conferencing) is not used instead for another application (say tying together two PBXs). As the price per cell is not constant across applications, this kind of arbitrage may be attractive to some customers. An open question is: Should regulators impose heavy penalties for "misuse" of service offerings on the grounds that arbitrage will not allow for cost recovery through efficient market segmentation? What arguments would support this position?

More difficult choices must be made in the current U.S. context, in which line-of-business requirements may preclude network providers from integrating vertically into all stages of production in the information services industry. It is possible that the network operator may then be limited to providing access and transport services alone. Even though all applications run on the same network, their differing requirements for QOS can be supported by offering a product line of access and transport services, with each product in the line offering different qualities in dimensions such as cell delay, cell loss, and priority. Under this theory, differential prices could then be justified on the basis of differential costs associated with different QOS. An important question is: Can a self-selection scheme be used to segment the market in an economically efficient way? The literature on product-line pricing provides a useful framework for the analysis of this issue. Relevant papers include those by Mussa and Rosen,[21] Srinagesh and Bradburd,[22] and Srinagesh, Bradburd, and Koo.[23]

One theme in these papers is that efficient cost recovery requires that the offered spectrum of qualities be wider than a narrow technical analysis would suggest. Mussa and Rosen showed that this expansion of the product set required that the

highest quality application be provided with undistorted quality. All other applications should face quality below that provided in a fully competitive market. Srinagesh and Bradburd showed that there are plausible circumstances in which it is optimal to provide lowest quality applications with undistorted quality and to provide superfluous quality to all higher quality applications. Srinagesh, Bradburd, and Koo developed an alternative model in which it pays the firm to offer undistorted quality to consumers in the middle of the spectrum, with quality degradation of lower qualities and quality enhancement of higher qualities. The factor that determines the characteristics of the optimal product line is the correlation between marginal and total utilities across customers. Primary market research on the distribution of willingness to pay across the potential customer population is therefore a critical input in determining the optimal product line.

Many congestion control strategies currently under discussion, such as call control procedures for the setup of virtual connections and flow control procedures, provide a basis for the implementation of product-line pricing of services.[24] Most preventive congestion schemes for connection-oriented networks are based on the notion of a *traffic descriptor*[25] that captures the (statistical) effect of the call on congestion (or network utilization). Examples of traffic descriptors are peak bandwidth requirements, peak to average bandwidth ratios (or burstiness), and duration of burstiness.[26] The literature on congestion has not yet provided a definitive description of this important variable. Call control schemes typically formulate conditions under which a call with a particular traffic descriptor should be accepted.

The traffic descriptor can also be used as a market segmentation mechanism. In particular, we can conceive of different grades of service as being defined in terms of the treatment by the network of calls with different traffic descriptors. Although the engineering view of congestion control focuses on the issue of fairness in handling calls, the economic view would focus on treating different calls differently, with higher priority given to higher priced calls. This scheme could be the basis for successful (in an economic efficiency sense) product-line pricing if traffic descriptors correlated well with willingness to pay. A more general view would make call connection parameters one element of a broadly defined QOS measure.

Another alternative may be to directly mark high price cells with a priority marker.[27] In this scheme, the issue of priority would not be handled only during call setup, but also during the progress of the call itself. One advantage of this procedure is that it will work for a network that handles both connection-oriented and connectionless traffic. Yet another alternative, suggested by Egan,[28] is to use the signaling system to indicate high or low prices based on network congestion and to allow the customers to modulate offered load in response to the price signal. Egan also suggests the use of interruptible service contracts that block (at least partly) some large users' low-priority traffic during congested periods.

In conclusion, we stress two points. It is important to understand the elements of QOS that matter for customer satisfaction if effective market segmentation strategies are to be implemented. It is also important that switch design (buffer management and call acceptance protocols) be guided by the economics of market segmentation.

7. INCENTIVE PRICING WITH INCOMPLETE INFORMATION

In general, the price of a product should take into account: (a) the costs of manufacturing and supplying it, (b) the customers' willingness to pay for it, and (c) the market structure of the industry in which the product belongs and the prices and output choice made by competitors and the potential for entry into the industry. The choice of an optimal price depends on the firm's knowledge about the many economic parameters underlying these factors. Typically, this knowledge is incomplete. In the absence of a mechanism to gather or elicit information that enhances the firm's knowledge base, the price structure may be less efficient than prices based on full information. These issues arise in the pricing of broadband services.

To give an idea of the kind of mispricing that can occur because of incompleteness of information, we focus on one of the factors previously listed. We assume that the network provider is fully informed about its own costs and technological parameters and about its competitors; however, it cannot directly verify the willingness to pay by customers. An obvious approach to bridging this gap is for the network provider to conduct market surveys to ascertain these values for the different services to be offered on the broadband platform. The information gathered from such surveys is likely to determine not only the tariffing of the services, but also the level of investment in the new fiber-optics network.

Typically, a customer's willingness to pay for a service is determined by the benefits derived from the service. Consider the following highly simplified scenario. Suppose that a network provider contemplates building a broadband network capable of providing new services to a group of 100 subscribers. Each customer i obtains a private benefit of B_i, which could be expressed in terms of dollars and can be interpreted as customer i's "true" maximum willingness to pay. Suppose that the cost of building the network is \$1 million, and that this cost has to be fully recovered by a one-time increase in current rates. Suppose that the sum of the true benefits B_i far exceeds the cost of building the network, $B_1 + \ldots + B_{100} > \$1$ million; hence, the network clearly has positive value. On the other hand, it is reasonable to expect that for any i, $B_i < \$1$ million; hence, no customer will find it worthwhile to finance the network by him- or herself.

Suppose that the network's marketing representative plans to conduct a census of all customers in order to obtain an estimate of willingness to pay for new services offered on the broadband network. Each customer is asked to give an

indication of the value that he or she places on the proposed network by choosing a number on a scale from 0 to 10. The cost of the network will be allocated across customers as a function of the numbers that are reported. However, if all customers report the minimum valuation, then the network is not built. If the network is built, then, of course, no one can be denied access to the services it provides, regardless of whether they indicated a willingness to share in the cost of its construction. Thus, if Customer i reports 7, Customer j reports 2, and the sum of the reports made by all customers is 500, then Customer i will be charged \$1 million times (7/500) and Customer j will be charged \$1 million times (2/500). On the other hand, a Customer k who reports 0 pays nothing.

Given the simple cost allocation scheme outlined, how do we expect the customers to report? From the standpoint of any individual customer, it is never rational to report any number other than 0. This is true regardless of what the other customers may have reported. To understand the rationale for this result, consider the following reasoning on the part of Customer i: "If no one else reports a positive number, then the network is not built, and my payoff (in dollars) is 0. But announcing a positive number, say R_i, would yield a negative payoff of $B_i - \$1$ million, because I would have to finance all of it. If the network is built (i.e., some other customers do report positive values), I will obtain a payoff of B_i if I report 0 and $B_i - R_i / (R_1 + \ldots + R_{100})$ for any report $R_i > 0$. So in every conceivable instance, it is a "dominant" strategy for me to report $R_i = 0$." Every customer rationalizes a report of 0 in this manner, and the network is not built, even though everybody could have benefited from its presence.

The example here is, of course, an extreme case. However, the main point it makes is applicable in all other situations involving incompleteness of information about the preferences of customers: The pricing or cost allocation rules will not be efficient.

The economics literature has expended a considerable amount of energy in tackling the problems associated with incompleteness of information.[29] This area is broadly referred to as the theory of mechanism design, and it finds applications not only to questions relating to allocating the costs of a project such as a broadband network, but also to the design of flow control algorithms for prioritizing users of the network once it is in place.

Once again, instead of giving a broad survey of the literature on mechanism design, we illustrate how the information gap is bridged by such mechanisms using a simple example. We consider the problem of flow control in a communications network, which involves allocating the usage of the network so that an optimal trade-off is reached between submission of jobs to the network and the congestion that results.

Suppose that the network is to be used by different types of customers ranked from those with the highest priority to those with the lowest priority. The priority levels are private information to the customers, and the network administrator cannot observe them. Also, as is evident, it is too costly to audit the customer to

obtain an accurate reading of the appropriate priority level. The differences in priority levels translate to differences in marginal utilities to the customers from being allocated a particular arrival rate onto the network and marginal disutilities from the delays generated due to congestion. We formalize the argument as follows.

Suppose that each customer accesses the network at a rate q_i. The aggregate effect of all the customers attempting to use the network is that it leads to a delay, denoted D. Each customer i is characterized by a utility function that is dependent on q_i and D given by $U_i(q_i, D)$. This function is increasing in the first argument and decreasing in the second for all customers; however, for all values of q_i and D, the absolute values of the partial derivatives $\delta U_i/\delta q_i$ and $\delta U_i/\delta D$ are higher for higher priority customers.

The utilities of the customers obtained after a network is built translates into a commitment to pay for the network before it is built. Hence, the network provider's objective is to maximize the sum of the utilities $U_i(q_i, D)$ over all i. This is expected to maximize the aggregate amount that the network operator can expect to raise up front to build the network.

We can imagine the network provider requesting information on the values of $\delta U_i/\delta q_i$ and $\delta U_i/\delta D$ from customers and then adjusting the access rate for each customer in a way so that $\Sigma_i U_i(q_i, D)$ is maximized. Assuming that the network capacity constraint does not restrict choices, the optimal access rate $q^* = (q_1^*, \ldots, q_n^*)$ is achieved when $\Sigma_i U_i$ can be increased no further. This occurs if:

$$\frac{\delta}{\delta q_i}[U_i(q_i^*, D)] = 0 \text{ for all } i.$$

Of course, the network operator has no direct knowledge of the customers' true values to determine q^*.

One way to solve this problem is for the network operator to have a series of discussions with the users and to ask them for information on their utility functions. Based on these discussions, the permitted access rate for each user i, q_i, is adjusted. This may involve reducing the rate for some users j and increasing it for i. The network operator effectively becomes a clearinghouse for access rates. It can be shown that such adjustments can be made to follow a set of rules so that eventually the objective $\Sigma_i U_i(q_i, D)$ is maximized and the optimal rate q^* is achieved. Essentially, the users behave as if they were players in a game in which the objective is to maximize utility. The adjustment rules as a function of the series of discussions determine the reallocation of utility. It can be shown that as part of the discussions, the users will reveal information about their utilities in a way so that q^* can be determined.[30]

8. CONCLUDING COMMENTS

As was stated in the introduction, there is no single pricing methodology that can be recommended in all circumstances. In this chapter we attempted to outline the approaches that economic theory suggests to be most relevant in

pricing broadband telecommunications services. Four specific methodologies were considered: (a) demand-based or Ramsey pricing (including profit maximization as a special case), (b) cost-based pricing using the Aumann–Shapley pricing rule (or one of its variants), (c) nonlinear and product-line pricing, and (d) priority and interruptible service pricing. These rules are not mutually exclusive. For example, nonlinear pricing can be implemented by differentiating customers in a quality dimension in which priority and interruptibliliy of service are important attributes. Furthermore, priority of service can, in principle, be incorporated into the cost-based pricing methodologies because costs of serving different customer classes will depend on their priority level.

Whatever pricing methodologies are adopted, eventually they must be competitive. Therefore, on a per packet basis, a video packet will likely be heavily discounted relative to a voice packet. Voice signals might then be aggregated and packaged so as to resemble a video transmission, assuming that detection by traffic-distinguishing methods is imperfect. As a result, video channels could be used to carry voice traffic at a price that is significantly lower than that being charged by a local operating company for bulk transport of voice. This is one example in which the possibility of arbitrage or resale may have an impact on future pricing methodologies.

We also noted that a telecommunications pricing structure cannot be made arbitrarily complex, as is the case with the pricing of airline services. Telecommunications customers are not likely to tolerate a tariffing system that requires a translation by a computer or an agent (similar to a travel agent). An implication of the need for "simple" pricing rules is that classical economic pricing rules based on marginal cost may not be practical. The marginal cost of a signal in a telecommunications network is basically the cost imposed on the network due to the additional level of congestion. Congestion, in turn, depends on the rate of arrival of packets that constitute a call, the duration of the call, the number of channels used, and the composite effects these attributes have on an alternative call being blocked, delayed, or rerouted. Because the probabilities of such occurrences are constantly changing, the prices should, from an economic perspective, be changing in response. For practical reasons, the prices are set on a much coarser grid to accommodate the customers' aversion to complexity. Moreover, due to externalities between different services on the same broadband platform, the technological parameters do not fully account for the true economic costs. Such externalities are due to a variety of factors: Different services are substitutes for each other (such as voice versus email), and other services complement each other (for example, voice and information services such as electronic yellow pages). In addition, there are congestion externalities. In sum, it is arguable that practical limitations on the pricing structure for services over a broadband network may lead to suboptimal outcomes for both service providers and their customers.

Although this discussion is preliminary in nature, it suggests a useful starting point for telecommunications companies that must consider numerous other

factors in pricing in the early stages of broadband deployment. In addition to focusing on the traditional customer base, a parallel effort could be directed toward securing new revenue sources from nontraditional customers. Although it has been widely noted that the convergence of telecommunications and information processing technologies may have implications for traditional entertainment markets,[31] relatively little attention has been paid to the converse proposition—that revenues from nontraditional sources might benefit telecommunications providers under innovative pricing methodologies. In light of these noted difficulties associated with resale and arbitrage, these new revenue sources may play a central role in future approaches to pricing of telecommunications services.

ENDNOTES

1. Pepper, R., *Through the Looking Glass: Integrated Broadband Networks, Regulatory Policies and Institutional Change*, FCC Office of Plans and Policy Working Paper No. 24, 1988, p. 46.
2. An elementary exposition of the theory of multiproduct monopoly pricing is contained in Stigler, G., *The Theory of Price*, Third Edition, London: Macmillan, 1966, p. 211. A more rigorous treatment may be found in Tirole, J., *The Theory of Industrial Organization*, Cambridge, MA: MIT Press, 1989, p. 69.
3. When demands are not independent, the inverse demand function depends on all other quantities produced by the firm. We note that our analysis also ignores income effects, which is reasonable in an analysis of telecommunications demand functions.
4. Because total revenue in market i is given by $P_i(q_i)q_i$, profit maximization requires chosing output q_i such that:

$$\frac{d}{dq_i}[P_i(q_i)q_i] = \frac{d}{dq_i}[C(q_1, \ldots, q_n)].$$

5. Social surplus is defined as profits plus aggregate consumers' surplus, in which the latter is the difference between the amount that consumers would be willing to pay for a given output, given an all-or-nothing offer, and the amount they are asked to pay. This quantity is computed by integrating the appropriate demand function.
6. More detailed discussions of Ramsey pricing may be found in Sharkey, W. W., *The Theory of Natural Monopoly*, Cambridge, UK: Cambridge University Press, 1982; Brown, S. J. and D. S. Sibley, *The Theory of Public Utility Pricing*, Cambridge, UK: Cambridge University Press, 1986; Tirole, *op. cit.*; and, Mitchell, B. M. and I. Vogelsang, *Telecommunications Pricing: Theory and Practice*, Cambridge, UK: Cambridge University Press, 1991.
7. Cost functions appropriate for broadband telecommunications services are considered in R. P. McLean and W. W. Sharkey, "An Approach to the Pricing of Broadband Telecommunications Services," *Telecommunications Systems*, forthcoming. Demand elasticities may be determined from econometric techniques applied to survey data on anticipated demand or to actual demand for closely related services.
8. This approach, however, does not take account of political, regulatory, or public policy considerations.
9. See Miriam, L. and Y. Tauman (1982), "Demand Compatible, Equirable, Cost Sharing Prices," *Mathematics of Operations Research*, 7, 45–56. A somewhat different axiomatic characterization of Aumann–Shapley pricing is given in Billerica, L. J. and D. C. Heath (1982), "Allocation of

Shared Costs: A Set of Axioms Yielding a Unique Procedure," *Mathematics of Operations Research*, 7, 32–39.

10. A more complete discussion of the Aumann–Shapley pricing rule and its application to some specific cost functions is contained in R. P. McLean and W. W. Sharkey, "Alternate Methods for Cost Allocation in Stochastic Service Systems," unpublished manuscript.
11. Specific functional forms of such pricing rules are contained in the papers by McLean and Sharkey, *op. cit.*
12. Formally, the cost function C is defined by:

$$C(q,r,d) = \underset{k,B}{\text{Min}}\, g(k,B)$$

subject to

$$\beta_i(k,B;q,r,d) \le \beta_i^* \text{ for all } i$$
$$w_i(k,B;q,r,d) \le w_i^* \text{ for all } i.$$

13. Formally, the cost function $C(q,r,d)$ is given by the following expression:

$$C(q,r,d) = \underset{k,B}{\text{Min}}\, g(k,B)\left\{g(k,B) + \sum_{i\in N} c_i q_i \beta_i(k,B;q,r,d) + \sum_{i\in N} \gamma_i w_i(k,B;q,r,d)\right\}.$$

14. See papers by McLean and Sharkey, *op. cit.* The Erlang loss formula that defines the blocking probability is given by:

$$\beta(q,r,k) = \frac{(q\cdot r)^k/k!}{\sum_{i=0}^{k}(q\cdot r)^i/i!}.$$

Because the buffer capacity B is constrained to be equal to 0, there are no waiting costs, and the cost functions in Models 1 and 2 become, respectively:

$$C(q,r,d) = \underset{k=0,1,2,\ldots}{\text{Min}}\, g(k) \text{ s.t. } \frac{(q\cdot r)^k/k!}{\sum_{i=0}^{k}(q\cdot r)^i/i!} \le \beta^*$$

and

$$C(q,r,d) = \underset{k=0,1,2,\ldots}{\text{Min}}\, g(k) + (c\cdot q)\frac{(q\cdot r)^k/k!}{\sum_{i=0}^{k}(q\cdot r)^i/i!}.$$

15. This concept clearly is not intended to apply in all situations. Technically, it is appropriate in a technological environment of increasing returns to scale, in which additional consumption by one group of consumers does not necessarily make every other consumer worse off. In a decreasing returns situation (e.g., fishing from a common ocean or drawing water from a common aquifer), different principles must be applied. Even in an increasing returns environment, subsidy-free prices may fail to exist, as has been described in Sharkey, *op. cit.*
16. Briere, Daniel D., *Long Distance Service: A Buyer's Guide*. Norwood, NJ: Artech House, 1990.
17. A comprehensive survey can be found in Wilson, R., *Nonlinear Pricing*, London: Oxford University Press, 1993. The issues of risk are treated in Clay, K., D. S. Sibley, and P. Srinagesh (1992), "Ex Ante and Ex Post Pricing: Optional Calling Plans and Tapered Tariffs," *Journal of Regulatory Economics*, 4, 115–138.

18. Mandy, D. M. (1992). "Nonuniform Bertrand Competition," *Econometrica*, 60, 1293–1330.
19. See Hong, D. and T. Suda (1991, July), "Congestion Control and Prevention in ATM Networks," *IEEE Network Magazine*, p. 11.
20. See, for example, Brown and Sibley, *op. cit.*
21. Mussa, M. and S. Rosen (1978), "Monopoly and Product Quality," *Journal of Economic Theory*, pp. 301–317.
22. Srinagesh, P. and R. Bradford (1989), "Quality Distortion by a Discriminating Monopolist," *American Economic Review*, pp. 96–105.
23. Srinagesh, P., R. Bradford, and H-W. Koo (1992, June), "Bidirectional Distortion in Self-Selection Problems," *Journal of Industrial Economics*, 40, 223–228.
24. See Hong and Suda, *op. cit.* and Vakil, F. and H. Saito, "On Congestion Control in ATM Networks," *IEEE LCS Magazine*, forthcoming.
25. See Vakil and Saito, *op. cit.*
26. See Hong and Suda, *op. cit.*
27. See Hong and Suda, *op. cit.*, pp. 14–15.
28. Egan, B. (1987), "Costing and Pricing the Network of the Future," *Proceedings of the IEEE, ISS*, pp. 483–490.
29. For recent surveys of the field, see Moore, J., "Implementation in Environments with Complete Information," and Palfrey, T., "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design," both in *Advances in Economic Theory*, ed. by J.-J. Laffont, Cambridge: Cambridge University Press, 1993.
30. For details on how such an adjustment rule works, see Chakravorti, B., "Optimal Flow Control of an M/M/1 Queue with a Balanced Budget," *IEEE Transactions on Automatic Control*, forthcoming. A survey of this literature is contained in Sharkey, W. W., "Network Models in Economics," *Handbook of Operations Research and Management Science: Networks*, forthcoming.
31. See, e.g., Curtis, T. and K. Means, "Market Segmentation and the IBN Policy Debate," in *Integrated Broadband Networks: The Public Policy Issues*, ed. by M. C. J. Elton, Amsterdam: North-Holland, 1991.