Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet

Peter C. Fishburn and Andrew M. Odlyzko
AT&T Labs - Research, 180 Park Avenue, Florham Park, NJ 07932
fish@research.att.com, amo@research.att.com

Abstract. The simple model on which the Internet has operated, with all packets treated equally, and charges only for access links to the network, has contributed to its explosive growth. However, there is wide dissatisfaction with the delays and losses in current transmission. Further, new services such as packet telephony require assurance of considerably better service. These factors have stimulated the development of methods for providing Quality of Service (QoS), and this will make the Internet more complicated. Differential quality will also force differential pricing, and this will further increase the complexity of the system.

The solution of simply putting in more capacity is widely regarded as impractical. However, it appears that we are about to enter a period of rapidly declining transmission costs. The implications of such an environment are explored by considering models with two types of demands for data transport, differing in sensitivity to congestion. Three network configurations are considered: (1) with separate networks for the two types of traffic, (2) with a single network that provides uniformly high QoS, and (3) with a single physical network that provides differential QoS. The best solution depends on the assumptions made about demand and technological progress. However, we show that the provision of uniformly high QoS to all traffic may well be best in the long run. Even when it is not the least expensive, the additional costs it imposes are usually not large. In a dynamic environment of rapid growth in traffic and decreasing prices, these costs may well be worth paying to attain the simplicity of a single network that treats all packets equally and has a simple charging mechanism.

1 Introduction

The Arpanet, which evolved into today's Internet, was a research project that did not provide for any payment mechanisms and treated all packets on an equal "best-effort" basis. The Internet has (with minor exceptions) inherited these properties. Packets are basically still treated equally. Charging usually is only for the bandwidth of the connection to the Internet and is independent of the amount of data sent and received. (See [11], especially [10], for a survey of the economics of the Internet.) These features, which provide for extreme simplicity in both operation and economics, have contributed to the spectacular growth of the Internet.

Although there have been persistent criticisms about the lack of Quality of Service (QoS) provision on the Internet, and about the charging scheme, thus far they have not been sufficiently convincing to modify the system. However, there are signs that change is imminent. Dissatisfaction with endemic congestion on the public Internet, which makes even Web surfing annoying, and the need to provide QoS for novel applications that are delay-sensitive, such as packet telephony and videoconferencing, are leading to demands for differential treatment of packets. Similar demands are coming from the corporate side. Private line networks use the same IP (Internet protocol) technology, are far larger in aggregate than the public Internet [3], and have been providing high QoS largely through low utilization levels [13]. However, with demand for bandwidth rising, corporate network managers are also demanding tools such as prioritization to ensure higher efficiency of network usage. Differential service quality will inevitably force introduction of more complicated pricing schemes than the present one, since it will be necessary to prevent all traffic from being sent on the highest quality level. The departure from the simple network operations and charging mechanisms of the Internet would represent at least a partial victory for the "Bell-heads" in the infamous controversy over networking [16].

The "Net-head" approach to the problems of poor service has been to provide greater bandwidth and keep the charging algorithm simple. This solution is used universally in LANs (local area networks) and has worked for corporate and research networks in the past. The objection to the "Nethead" approach is that it is too expensive, at least for the public Internet, since more than two decades of experience have shown that any bandwidth gets saturated quickly.

Data transport is a serious constraint on ISPs (Internet service providers), as it accounts for about half of the total cost of long-haul networks. While new optical fiber technologies led to a dramatic drop in rates for leased lines in the 1980s and early 1990s, prices have been increasing recently as a consequence of scarcity of supply and rapidly growing demand (see [3, 15] for examples). Network operators have been lowering their cost per unit of bandwidth by moving to higher capacity lines (see Section 2 for data and discussion of this issue) and by signing long-term leases. In an environment of rising prices, differential QoS and more sophisticated pricing schemes appear essential to meet the explosive data transport needs at affordable cost.

Do we have to give up on the simple operation and charging mechanisms of the current Internet? Both splitting of traffic into different QoS classes and complicated charging mechanisms impose heavy costs on developers of applications and network systems, and on network operators. Further factors in favor of simple fixed-fee charging mechanisms come from customer preferences (even those of large corporate customers), which often lead to higher revenues for service providers who use such pricing approaches [6]. An early 1988 paper by Anania and Solomon (published in [AnaniaS]) already presented several arguments for a simple flat-rate pricing approach to broadband networks.

Although simple flat rate pricing with uniform best-effort data transport is attractive, it has many defects. It provides a single level of service quality, and does not allow users to select what is best for their needs. Economists in general oppose it on the grounds that it leads to misallocation of resources. For fuller description of the arguments for abandoning the traditional Internet model, and for further references, see [11]. However, those arguments are based on experience with an environment that is likely to change drastically. (It is already an environment far removed from the traditional telecommunications world studied in [12], for example, and will diverge from it even further.) As mentioned above, long distance data transport prices have been rising in the last few years. The basic fiber optic network that carries both voice and data traffic was designed primarily for voice, and until a few years ago, most of the bandwidth was devoted to voice. In revenues, the network is still dominated by voice. However, the bandwidth devoted to data is already comparable to that used for voice [3], and data traffic is growing much more rapidly. Thus we can expect communications networks will grow rapidly and be increasingly dominated by data. Furthermore, WDM (wavelength division multiplexing) technology allows for expanding capacity without laying down more fiber (at least not on long distance routes), which is a very expensive process, especially when acquisition of rights-of-way is included. Within a few years, existing fiber will provide 100 or even 1000 times the bandwidth it did a couple of years ago, at modest additional cost. The main determinants of network costs will be the electronics needed to provide WDM and switching. However, in electronics "Moore's Law" reigns, with performance increasing while prices drop. What this means is that we are likely to enter an era in which the price of bandwidth continues dropping dramatically for a decade or more. The question is, what will this mean for service providers and consumers?

It is instructive to consider microprocessors. Table 1 shows the last dozen years from the history of Intel. For each year, the microprocessor listed is the most powerful model introduced that year, with the price the one available at the end of that year. (All dollar figures are in nominal dollars, and the prices are for orders of 100 or 1000 chips at a time.) The processing power, in mips (millions of instructions per second) is an imperfect measure of the computing power of processors. Still, it illustrates how the power of state of the art microprocessors has been growing at an exponential rate, while their prices have remained about constant. At the same time, revenues and profits have increased. Over the period illustrated by Table 1, computing power has grown over 60% per year, with prices of the most powerful available processors rather stable, while Intel's revenues have grown about 30% per year. A similar scenario appears to be realistic for high bandwidth communication networks in the next decade. What we explore are the implications of this kind of environment for the provision of QoS on the Internet. If available capacity doubles each year, or every two years, while total costs increase much more slowly, so that the price per unit of bandwidth decreases rapidly, it might make sense to provide uniformly high QoS for everybody and avoid the complexities of the schemes that are being considered.

Existing work on QoS, surveyed in the book [4], does not contain any projections of the degree to which the different proposals for providing QoS will lower network utilization. The relation between utilization of network capacity and perceived quality of service is a complex one. It is possible to have a lightly utilized network that delivers horrible service, but in general the lower the utilization rate, the better the service. Further, many networks, such as corporate Intranets, are already providing QoS largely through low utilization rates [13]. High-quality experimental networks such as vBNS also have very low utilizations. These networks are still operated on the "best-effort" basis, with no explicit guarantees (but with sophisticated traffic engineering tools). Congestion episodes are infrequent enough for this to be acceptable. In general, no matter how a network is engineered, lowering the traffic load on it will result in better service. The routers and switches are fast enough already that if congestion does not cause buffers to fill up, the quality is sufficient for all anticipated demands.

In this work, we will assume as a first approximation that

Table 1: Intel and its microprocessors. For each year, lists the most powerful general purpose microprocessors sold by Intel, its computing power, price at the end of the year (in dollars), and Intel's revenues and profits for that year (in millions of dollars).

year	processor	mips	price	revenue	net profit
86	386 DX (16 MHz)	5	300	1265	-173
87	386 DX (20 MHz)	6		1907	248
88	386 DX (25 MHz)	8		2875	453
89	486 DX (25 MHz)	20	950	3127	391
90	486 DX (33 MHz)	27	950	3922	650
91	486 DX (50 MHz)	41	644	4779	819
92	DX2 (66 MHz)	54	600	5844	1067
93	Pentium (66 MHz)	112	898	8782	2295
94	Pentium (100 MHz)	166	935	11521	2266
95	Pentium Pro (200 MHz)	400	1325	16202	3566
96				20847	5157
97	Pentium II (300 MHz)	600	735	25070	8945

improved QoS is associated directly with low utilization levels. Although schemes like those in [4] can increase the efficiency of the network, whether it has just a single best-effort service, or several classes of service, it is hard to incorporate them into an economic model until more is known about their performance.

To explore potential futures for QoS on the Internet with and without differential pricing, we will assume two types of demands in our models. One is for transport that is delay insensitive, such as many bulk file transfers or even email. The other is for transport of information that is sensitive to delay, such as packet telephony, or even some Web browsing. (In effect we will thus be considering Class of Service models for the Internet, and not the more involved QoS ones.) We refer to delay insensitive demand, or to its users, as type A, and to delay sensitive demand, or *its* users, as type B. Within a given time period, each type has a *potential volume* or potential demand, which is the total Internet transfer volume the type would use if the transfer charge or price were essentially zero. We denote their potential volumes by V_A and V_B, or simply by V as a general designation.

We will vary the ratio of V_A and V_B, but only within a narrow range, near equality. The justification for this is that in current data networks, the volumes of data sent over the congested public Internet and over the uncongested private line networks are comparable. If V_A were much larger than V_B, then clearly it would be best to send all data over an uncongested network designed for type A traffic. On the other hand, if V_B were much larger than V_A, the case for a separated network or a two-tiered network would be much stronger.

Because real use will be price sensitive, the actual volume carried for a user type during the period is modelled by $P(x)V$, where x is the price per unit of volume and $P(x)$ is the probability that a potential user will subscribe to the service at price x. We refer to P as the *demand function* and assume that $P(0) = 1$ and that $P(x)$ decreases toward 0 as x increases. An approximate but revealing measure of customer satisfaction is the *demand satisfaction* expressed as the percent of potential volume that customers subscribe to during the period, i.e., $100P(x)$. This should not be confused with the utilization of available network capacity since, for example, a channel that carries priority data may have a high demand satisfaction yet provide very good QoS because its transport capacity substantially exceeds the priority demands. Several forms will be considered for P to account for the possibility that our conclusions may depend on assumptions about the demand function.

Three network configurations are examined for provision of service to types A and B, as follows:

1. physically separate networks are used for each of A and B, with each network having its own cost, QoS, and price characteristics;

2. a single network is used for both A and B, with one price for all users that is constructed to provide the high QoS desired by B;

3. a single network is used for A and B, but the types are logically separated by software that differentiates between them and allows different QoS and prices for the two.

We refer to (1) as the *separated* network, to (2) as the *one-price* network, and to (3) as the *two-tiered* network. We assume for (3) that the types use logically separated channels and ignore techniques such as those in [4] that can lead to greater efficiencies, as when low-priority traffic is used to fill gaps in high-priority traffic. We also ignore the large increases in utilization rates that can be gained by exploiting different time-of-day patterns of use (which are discussed in detail in [14]). The main conclusion of our models is that factors of two in price or utilization do not matter much in an environment of increasing demand and falling prices.

The advantage of (3) over (1) is that the unified network can take advantage of economies of scale. We will not consider the added costs of providing for logical separation of the two traffic types on a two-tiered network.

As is shown in [13], current data networks are an inefficient amalgam of the separated network and the one-price network. They do resemble a separated network, with the public Internet operating in a congested mode with relatively high utilization rate (although lower than that of the switched voice network), while corporate networks have very low utilization rates. However, this is not the separated network of our model, since all corporate data, whether it is sensitive to delay or not, travels over underutilized networks, while all public Internet traffic goes over congested links. Thus we have two separate one-price networks.

The economic models used to determine prices for the three configurations we study are based on providers' costs and revenues. Costs include ongoing operational costs, depreciation and other overhead charges, and a reasonable rate of return or profit that might be limited by competition or regulatory constraints. We assume for each period that total cost is a function of actual volume carried, as described in the next section.

Per-period revenue equals price times actual volume, i.e.,

$$\text{Revenue} = xP(x)V.$$

We then compute the actual price charged as the smallest x at which revenue equals cost. In doing this, we are not trying to maximize profit because it is already built into costs; we seek only to determine a reasonable price based on equality between costs and revenues. If no value of x satisfies the Revenue = Cost equation, then revenue is insufficient to cover cost at any price, and we refer to the configuration as *infeasible*. Although our models are based on equilibrium between revenue and cost rather than optimization schemes *per se*, we will compare prices, demand satisfactions, and revenues of the three network configurations to assess their performances with respect to each other.

We regard our models as informative but very rough approximations to an extremely complex environment and uncertain future. Explanations of aspects of cost, including economies of scale, effects on cost of enhanced QoS, and how costs may change over time in a competitive marketplace with rapidly increasing volume are described in the next section. Section 3 specifies the models more completely for the three network configurations in a static one-period scenario and describes solution procedures. Section 4 then extends the models to the dynamic scenario of a succession of periods in which potential volumes, costs, and implied prices change from period to period.

For computational simplicity in the dynamic analysis, we will assume that potential volume doubles from period to period. This assumption is made palatable by not fixing period lengths in advance. For example, two-year periods might be assume. See [3] for history and projections of growth patterns in data traffic. While voice traffic has been growing at around 10% per year, Internet traffic (measured in bytes) has been just about doubling each year in the 1990s, with the exception of 1995 and 1996, when it grew by factors of about 10 in each of those two years.

We have already mentioned that different market demand functions will be considered. A further accommodation for an uncertain future will be made by considering two very different patterns for changes in costs over time. The first is a *conventional pattern* in which costs change only because of the potential volume doubling from period to period. The second, which we refer to as the *dynamic pattern*, reflects not only the doubling assumption but also cost reductions driven by competition and technological advances. Dynamic-pattern revenues increase from period to period (except in one extreme scenario where they remain constant), but at a much slower rate than conventional-pattern costs. Both patterns are specified more completely in the next section.

As we will see in Section 4, the implications of our dynamic models depend on our different demand functions and cost patterns, but some trends emerge. For example, in comparisons between the separated network and two-tiered network, the prices for both A (ordinary service) and B (premium service) tend to be slightly higher for the separated network , whereas demand satisfactions are comparable. An anticipated finding is that dynamic-pattern costs drive prices substantially below those for conventional-pattern costs in all three networks. Another result that was not anticipated at the outset, is that the one-price network with its uniformly high QoS is competitive with the others under several assumptions. In regard to revenues (= costs), which are aggregated for A and B in the separated network, the highest revenues occur for either the separated network or the one-price network, whereas the lowest revenues occur for either the one-price network or the two-tiered network. The differences in the revenue picture are caused more by the different cost patterns than by the different demand functions. A more complete picture of these matters is given at the end of Section 4. The main conclusions, though, are that differences between the different networks are not great.

How can the one-price network be superior to the separated one? We show this with an example that simplifies our model by ignoring effects of price on demand. Suppose that type A and type B traffic are the same when measured in bytes, but that type B transmission requires much less congested networks, with capacity 4 times as large as that for type A. Suppose also that the cost of a network of bandwidth x is $x^{1/2}$. (Section 2 discusses cost formulas in detail.) Then the cost of the separated network is 3 ($= 1 + 4^{1/2}$), whereas that of the the one-price network is $8^{1/2} = 2.8284$, as the capacity has to be 8 times that of just the A network. Thus in this scenario, providing uniformly high QoS to everybody saves 6% of the cost. A much larger saving comes from having a single network, which makes life simpler for customers. On the other hand, there are also costs. For example, if there is no way to charge different prices for A and B traffic on a single network (as we will be assuming throughout the paper), then type A users will pay 1.4142 (half of total cost) instead of 1.0 for their own separate network, whereas type B users would see their charges drop from 2 to 1.4142. Thus different types of networks have varying impacts on social welfare. However, we argue that in the long run such costs might be bearable in the interests of simplicity. The reason is that rapidly decreasing costs of data transport mean everyone is as well off within one or two time periods as they would be with any other network solution.

The temporal aspects of technological change have a large impact on the marketplace. For example, for a long time, Intel microprocessors were slower, usually by at least a factor of 2, than comparably priced RISC chips. However, Intel was usually able to provide comparable price/performance ratio within two or three years. This, combined with the advantages of compatibility (i.e., lower costs to customers in upgrading) allowed Intel to increase its dominance in the processor business. Similar effects might favor simple schemes (such as the one-price network) over more efficient and socially optimal ones in data networks.

A summary of our study is provided in Section 5.

2 Economies of scale and other cost factors

Forecasting prices of telecommunications services has historically been a risky enterprise. As an example, we cite the paper [8]. Written in 1992 and published in 1993, it develops two models for leased line prices in the United States. Both models fit the data up to 1992 extremely well, and both models predicted a drop in prices of about 50% by 1998. Instead, prices have increased by approximately 50% since 1992, so they are about three times as high as predicted by Irvin's model. However, we feel that this was an anomaly, caused by unexpectedly high demand for data network bandwidth and little new growth in supply. At some point in the future, prices are likely to resume their decline.

There is no simple formula for costs of communication networks. It is almost always true that larger transfer volume or bandwidth purchases are less expensive per unit of volume or bandwidth than smaller ones, but even that is not always the case. For example, in April 1998, UUNet [17] was citing the following prices for dedicated Internet connections (not including the cost of local connections to the nearest UUNet site):

speed	price per month
56 Kbps	\$595
1.5 Mbps (T1)	\$1,795
45 Mbps (T3)	\$54,000

In this case, a 24-fold increase in bandwidth from a 56 Kbps line to a T1 incurs only a 3-fold increase in price, but the 28-fold increase in speed from a T1 to a T3 raises the cost by a factor of 30. This pricing may reflect scarcity of high-capacity lines, and possibly of handling the traffic from a T3 connection on a network that consists largely of T3 links. Similar linear pricing in bandwidth applies to speeds between T3 and OC3. (Sprint charges for these three speeds are \$897, \$2,062 and \$20,620, respectively, according to data at [Boardwatch], but these figures may not be strictly comparable to UUNet's because of special conditions and features.)

A better view of transmission costs might be offered by examining leased line prices. In April 1998, the tariffed monthly rates for an approximately 300 air mile private line, with about 5 miles of local connections that are leased from a local phone company were about as follows:

speed	price per month
9.6 Kbps	\$1,150
56 Kbps	\$1,300
128 Kbps	\$3,000
256 Kbps	\$3,800
512 Kbps	\$5,100
1.5 Mbps (T1)	\$7,000
7.7 Mbps	\$37,000
45 Mbps (T3)	\$66,000

(In practice, long-term leases and bulk purchase discounts might reduce these costs by up to 50%, see [9], for example. It is worth noting that the local access connections account for about 60% of the cost of a 9.6 or 56 Kbps line and about 17% of a T1 or a T3.) The exact figures depend on distance [9], but this dependence has decreased greatly over time [3].

Using the leased line prices cited above, we can see that a moderately good fit for the cost of carrying a given volume in one time period at the most common speeds between 56 Kbps and 45 Mbps is obtained by making the cost proportional to the volume, raised to a power in the range of 0.5 to 0.7 that we denote by s and refer to as the *economy-of-scale parameter*. (In the general economics literature, $1/s$ is known as the *elasticity of scale,* and we are assuming it is constant.) Economies of scale can arise from reduced requirements for the multiplexing equipment needed to provide low speed links on a high-capacity network as well as lower costs of sales, administration, maintenance, and related operational costs. It is reasonable to suppose that the same s value will apply in the future for greater volumes. Although later examples assume a value of $s = 2/3$, we write our cost formulas for general s. (For comparison, [7] uses a value of $s = 1/2$.) Today, $s = 2/3$ applies only through T3 speeds, and charges for OC3 (155 Mbps) private lines are reportedly often higher than for equivalent capacity in T3 lines. However, as traffic grows, and new technologies are deployed, it is not unreasonable to expect that our cost formula will apply at higher bandwidths as well.

In particular, we will assume that the cost for demand type A in a period with potential volume V_A and demand probability $P(x)$ at price x is given by

$$\text{Cost for } A = [P(x)V_A]^s \ .$$

This applies to the separated network, where costs are scaled in units determined for the separated A case. Using the same scale, we assume that the cost for demand type B under similar conditions is

$$\text{Cost for } B = [\psi P(x)V_B]^s \ ,$$

where ψ, which we refer to as the *premium factor*, is a parameter that exceeds 1 to account for higher cost and enhanced QoS for type B users. Reasonable values for ψ might lie in the range of 2 to 4, judging by the comparison of different networks in [13]. For example, if $\psi = 2$, then the B part of the separated network is arranged to carry the same volume as the A part at twice the capacity. Single-period costs for one-price and two-tiered networks have related forms that are described in next section.

The preceding costs apply to an initial period, which can be taken to be the present or some other base period. The conventional pattern for costs, in which costs change from period to period only as a function of the doubling of potential volume, implies that costs t periods in the future from the base period will be

$$[P(x)2^tV_A]^s = [P(x)V_A]^s 2^{ts} \quad \text{for} \quad A$$

and

$$[\psi P(x)2^tV_B]^s = [\psi P(x)V_B]^s 2^{ts} \quad \text{for} \quad B$$

in the separated network.

However, competition and technological advances along with rising demand may lead to substantially lower costs than those given by the conventional pattern. Among other things, developments in WDM mean that fiber capacity is not a limiting factor. Instead, the electronics that connect end users to the fiber are becoming the main obstacle, and improvements in optical and silicon technology are likely to induce rapid decreases in the price/performance ratio. Although prices of connections of a fixed speed might not drop dramatically, the bulk of the data transport capacity that is purchased is likely to cost far less per unit of volume than at present. (That is the pattern seen in prices of microprocessors and DRAMs.) We model such effects in our dynamic pattern for costs by dividing the conventional next period cost by $\sqrt{2}$, a factor that accumulates exponentially over time. For example, the present cost of $[P(x)V_A]^s$ for A in the separated network translates into the dynamic-pattern cost of

$$[P(x)2^tV_A]^s 2^{-t/2} = [P(x)V_A]^s 2^{t(s-1/2)}$$

t periods in the future, which is substantially less than the figure of $[P(x)V_A]^s 2^{ts}$ for the conventional pattern. We regard $\sqrt{2}$ as a fairly drastic dynamic factor, representing an extreme case for unit cost reduction. For example, if $s = 1/2$ then total cost remains the same as potential volume doubles.

Because period lengths are flexible, we allow for varying rates of decrease in unit cost as time progresses. If period length is one year and $s = 2/3$, the conventional pattern presumes a yearly decrease of about 20% in unit cost, and the dynamic pattern presumes a yearly decrease of about 44% in unit cost. If period length is two years and $s = 2/3$, the yearly unit cost decreases are 10% for the conventional pattern and 22% for the dynamic pattern.

3 One-period static analysis

This section discusses our models for a fixed period in which A has potential volume V_A, B has potential volume V_B, and both have demand function P. As before, ψ is the premium factor for higher QoS and $s < 1$ is the economy-of-scale parameter. The example later in this section takes $V_A = V_B$, $\psi = 3$ and $s = 2/3$. The next section considers other arrangements for V_A, V_B, ψ and s.

Let x, y, and z denote the prices for type A in the separated network, for type B in the separated network, and for both types in the one-price network, respectively. The costs for these networks are as follows:

$$\begin{aligned} \text{separated: } A \text{ cost} &= [P(x)V_A]^s \\ B \text{ cost} &= [\psi P(y)V_B]^s \\ \text{Total} &= [P(x)V_A]^s + [\psi P(y)V_B]^s \\ \text{one-price: Cost} &= \{\psi[P(z)V_A + P(z)V_B]\}^s \\ &= [\psi P(z)]^s (V_A + V_B)^s \ . \end{aligned}$$

For one-price, ψ applies to both A and B because this network offers the premium service to both types.

The Revenue = Cost equations for the preceding networks are

$$\begin{aligned} xP(x)V_A &= [P(x)V_A]^s \\ yP(y)V_B &= [\psi P(y)V_B]^s \end{aligned}$$

and

$$zP(z)(V_A + V_B) = [\psi P(z)(V_A + V_B)]^s \ .$$

In the first equation, $xP(x)$ for the forms we use for P increases to a maximum and then decreases for larger x, whereas $P(x)^s$ on the right side decreases from 1 at $x = 0$ and approaches 0 as x gets large. If the single-peaked curve for $xP(x)V_A$ lies beneath the decreasing curve for $[P(x)V_A]^s$, i.e., if $xP(x)V_A < [P(x)V_A]^s$ for all $x \geq 0$, then the A part of the separated network is infeasible. Otherwise, there will typically be two x values, say $x_1 < x_2$, where the curves cross. We take x_1 as our price solution to $xP(x)V_A = [P(x)V_A]^s$ because it gives a lower price, higher revenue, and greater utilization than x_2. Similar remarks apply to the other Revenue = Cost equations.

We introduce a new parameter for the two-tiered network. It is the ratio $\lambda \geq 1$ of the higher to the lower price in this network, i.e., $\lambda = y/x$ when y is the premium price and x is the ordinary price. When λ is not made explicit, the two-tiered Revenue = Cost equation is

$$xP(x)V_A + yP(y)V_B = [P(x)V_A + \psi P(y)V_B]^s \; .$$

Unlike the one-price case, ψ applies here only to the premium service because of the two-tiered structure. In keeping with the rationale of a two-tiered network, we regard this network as feasible only if the preceding equation holds for some (x, y) with $y \geq x$. We note also that costs could be increased slightly for the two-tiered network because of the additional costs of network operators, as well as those of users, who have to adjust to a more complicated pricing scheme. However, we do not believe that this matters very much since the models are approximate in the first place.

A feasible two-tiered network offers more freedom of choice than the others because it typically has a continuum of (x, y) solutions to the Revenue = Cost equation in which y increases as x decreases in moving away from the equal-prices solution where $x = y$. We have found that two-tiered revenue is often greatest when x and y are close together, but note also that $x = y$ defeats the purpose of a two-tiered network. We shall therefore regard $\lambda = y/x$ as a control variable subject to policy decision. Reasonable values for λ range from about 2 to 4, so that the premium service costs about two to four times as much as the ordinary service per unit volume. Our use of λ also eases the computational burden of solving the Revenue = Cost equation since, when λ is given, we need only solve for x and then obtain y from $y = \lambda x$.

When λx is substituted for y in the preceding two-tiered equation, it becomes

$$x[P(x)V_A + \lambda P(\lambda x)V_B] = [P(x)V_A + \psi P(\lambda x)V_B]^s \; .$$

As for the other networks, the solution is taken as the smallest x that satisfies the equation when it is feasible.

We consider three forms for the demand function P in the example that follows. They are

$$P_1(x) = e^{-x^2} \quad \text{for} \quad x \geq 0 \, ,$$
$$P_2(x) = \frac{e^{-x}}{1+x} \quad \text{for} \quad x \geq 0 \, ,$$
$$P_3(x) = \frac{1}{1+x^4} \quad \text{for} \quad x \geq 0 \, .$$

Figure 1 illustrates the differences between the three. P_1 and P_3 begin high for small x, decrease rapidly as x gets into a mid-range, and have very narrow tails. P_2 begins its descent immediately, levels off sooner than P_1 and P_3 and has a fat tail. When prices are low, P_2 is much more sensitive than the others to small price changes. This is the most important difference between them because most of the solutions we have seen for our networks have prices well below 1.

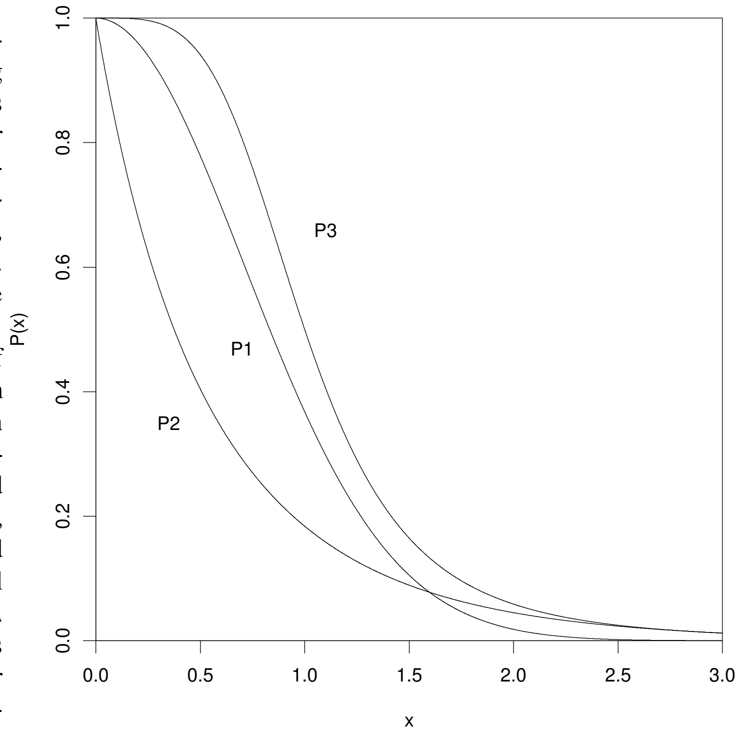

Figure 1: Three demand functions

We now turn to an example with parameter values $\psi = 3$, $s = 2/3$, and $\lambda \in \{2, 4\}$. The example has six scenarios in the 2-by-3 cross classification {low potential volume, high potential volume} $\times$ $\{P_1, P_2, P_3\}$. With $V_A = V_B$, we set the low potential volume for each of A and B at $V_1 = 32$, and set the high potential volume at $V_2 = 64V_1$.

We consider the separated and one-price networks first. The Revenue = Cost equations for A separate, B separate, and the one-price network are, for $V = V_1$,

$$\begin{aligned} xP(x)V_1 &= [P(x)V_1]^{2/3} \\ xP(x)V_1 &= [3P(x)V_1]^{2/3} \\ xP(x)(2V_1) &= [3P(x)(2V_1)]^{2/3} \end{aligned}$$

respectively. These simplify to

$$\left.\begin{array}{ll} P_1: & x^3e^{-x^2} \\ P_2: & x^3e^{-x}/(1+x) \\ P_3: & x^3/(1+x^4) \end{array}\right\} = \left\{\begin{array}{ll} 1/32 & A \text{ separate} \\ 9/32 & B \text{ separate} \\ 9/64 & \text{one-price.} \end{array}\right.$$

The right-hand sides of these equations are multiplied by 1/64 to obtain the corresponding equations for V_2.

Table 2 shows approximate solution values in terms of price x, demand satisfaction S and revenue R. The one-price network price in each row is midway between the prices for A and B in the separated network, P_2 induces slightly higher prices than P_1 and P_3, and prices drop dramatically with high volume. The ratios of premium service to ordinary service prices for the separated network lie between 2 and 3.5.

In all cases, demand satisfaction is substantially higher for P_1 and P_3 than P_2, and aggregated demand satisfaction for the separated network is about the same as one-price satisfaction.

Revenues are obviously higher for the high volume cases, but the high-to-low ratios are smaller than the 64-fold increase in volume because of economies of scale. Moreover, because P_3 implies greater propensity to subscribe than P_2, and P_1 implies greater propensity to subscribe than P_2 for all prices in the table, revenues run highest for P_3, next highest for P_1 and lowest for P_2. There is significantly less difference proportionately between revenues at high volume than at low volume.

We now bring the two-tiered network into the picture with x the cheaper two-tiered price and $y = \lambda x$ the premium service price. The Revenue $=$ Cost equation noted earlier for the two-tiered network reduces to

$$\frac{xe^{-x^2} + \lambda x e^{-(\lambda x)^2}}{[e^{-x^2} + 3e^{-(\lambda x)^2}]^{2/3}} = \frac{1}{(32)^{1/3}} \quad \text{for} \quad P = P_1, \quad V = V_1$$

$$\begin{aligned} &\frac{xe^{-x}/(1+x) + \lambda x e^{-\lambda x}/(1+\lambda x)}{[e^{-x}/(1+x) + 3e^{-\lambda x}/(1+\lambda x)]^{2/3}} \\ &\quad = \frac{1}{(32)^{1/3}} \quad \text{for} \quad P = P_2, \quad V = V_1 \\ &\frac{x/(1+x^4) + \lambda x/(1+(\lambda x)^4)}{[1/(1+x^4) + 3/(1+(\lambda x)^4)]^{2/3}} \\ &\quad = \frac{1}{(32)^{1/3}} \quad \text{for} \quad P = P_3, \quad V = V_1\ . \end{aligned}$$

The right sides of these are multiplied by $1/(64)^{1/3} = 1/4$ for V_2.

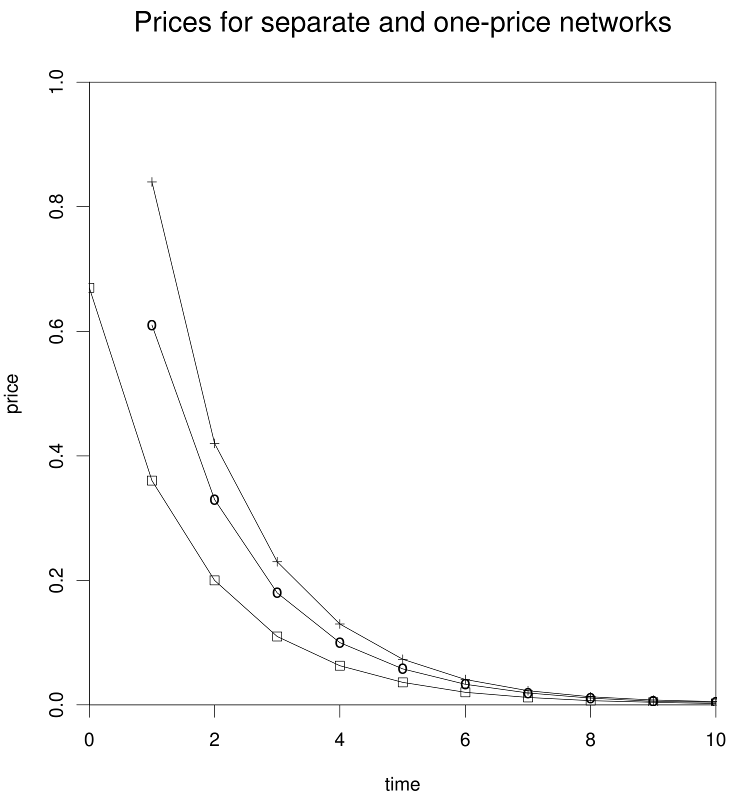

Figure 2: Evolution of prices for A users in a separate network (line with squares), for B users in a separate network (line with crosses), and in a one-price network (line with circles), for the scenario of Table 7.

Table 3 shows the two-tiered results to the right of the one-price results. Comparisons between $\lambda = 2$ and $\lambda = 4$ for the two-tiered case reveal little difference in demand satisfaction or revenue. In each row, the two prices for $\lambda = 4$ (e.g., .18 and $4 \times .18 = .72$) surround the prices for $\lambda = 2$ (e.g., .28 and $2 \times .28 = .56$). Without exception, the one-price network price is greater than the average of the two two-tiered prices for a given category (V_i, P_j, λ), and can be greater than the larger of these two as for $\lambda = 2$ in rows 1 and 2. Finally, the one-price network has uniformly higher revenue and uniformly lower demand satisfaction than the two-tiered network.

Table 2: Prices, demand satisfactions, and revenues for separated and one-price networks

		A separate			B separate			Separated network totals		One-price networks		
		x	S	R	x	S	R	S	R	x	S	R
V_1	P_1	.33	90	9.4	.82	51	13.4	71	22.8	.58	71	26.6
(Low)	P_2	.40	48	6.1	1.4	10	4.6	29	10.7	.85	23	12.6
	P_3	.32	99	10.0	.71	80	18.1	90	28.1	.53	93	31.6
V_2	P_1	.079	99	160	.17	97	331	98	491	.13	98	527
(High)	P_2	.084	85	146	.18	70	266	78	412	.14	76	444
	P_3	.079	100	162	.16	100	336	100	498	.13	100	536

Table 3: Prices, demand satisfactions, and revenues for one-price and two-tiered networks

		One-price network			Two-tiered network $\lambda = 2$			Two-tiered network $\lambda = 4$		
		x	S	R	x	S	R	x	S	R
V_1	P_1	.58	71	26.6	.28	83	21.8	.18	78	19.2
(Low)	P_2	.85	23	12.6	.36	40	12.5	.28	37	10.9
	P_3	.53	93	31.6	.27	96	24.4	.17	92	23.2
V_2	P_1	.13	98	527	.065	99	397	.040	99	400
(High)	P_2	.14	76	444	.071	82	349	.044	82	339
	P_3	.13	100	536	.066	100	407	.040	100	406

4 Dynamic Analysis

We present results of our dynamic analysis primarily for the separated and one-price networks to keep matters fairly simple. The results for the two-tiered network in the dynamic case are similar to those in the preceding section in comparison to the other networks, and their trends over time are similar to the trends described in this section. For example, for an array of parameters, $\lambda = 2$ and $\lambda = 4$ have very similar demand satisfactions and revenues although their prices, x and λx, obviously differ. The cheaper two-tier price at $\lambda = 4$ is about 60% of the cheaper price at $\lambda = 2$, so the premium price at $\lambda = 4$ is about 20% higher than the premium price at $\lambda = 2$. Revenue comparisons show a general pattern in which a two-tiered network either has the lowest revenue or the middle revenue of the three networks.

For the separated and one-price networks, we begin our dynamic process at period $t = 0$ with low potential volumes, and run each network through 11 periods. In our initial runs, which are partly shown in Tables 4 through 7, we took $V_A = V_B$, with a value of 4 at $t = 0$ and $2^t(4)$ for $t \geq 1$. These tables also use $\psi = 3$ and $s = 2/3$. An infeasible situation is shown by asterisks.

Tables 4 through 7 consider the "low" and "high" demand functions P_2 and P_3 (see Figure 1) along with the conventional cost pattern C_I and the extreme dynamic pattern C_{II} of rapidly decreasing unit cost. The tables pertain to (P_2, C_I), (P_2, C_{II}), (P_3, C_I) and (P_3, C_{II}), respectively.

We considered changes in ψ, s, and V_A and V_B to see how much they affect the nature of the results shown in the tables. The specific changes include $\psi = 2$, $s = 1/2$, $(V_A, V_B) = (4, 16)$ and $(V_A, V_B) = (16, 4)$ for the initial period. We comment on these briefly after noting aspects of Tables 4–7.

Revenue. Except for very low potential volume, a provider who offers either the A service or the B service for comparable potential volumes in the separated network makes more money from the premium B service. A provider who offers one of the two main network configurations shown in the tables earns more with the one-price network, but the difference between the two is not great in any case.

Separated versus one-price prices. The one-price network price always falls between the A-separate and B-separate prices. It tends to be about midway between the separated network prices when C_I applies, and is closer to the A-separate price when C_{II} applies.

Table 4: P_2, C_I

t	A separate x	S	R	B separate x	S	R	Separated totals S	R	One-price network x	S	R
0	1.3	13	.63	*	*	*	6	.63	*	*	*
1	.79	25	1.60	*	*	*	13	1.60	*	*	*
2	.55	37	3.28	*	*	*	19	3.28	1.4	10	4.61
3	.40	48	6.15	1.4	10	4.61	29	10.8	.85	23	12.6
4	.30	59	11.0	.85	23	12.6	40	23.6	.59	35	26.3
5	.23	65	19.0	.59	35	26.3	50	45.3	.43	46	49.9
6	.18	71	32.3	.43	46	49.9	59	82.2	.32	55	89.8
7	.14	77	53.8	.32	55	89.8	66	144	.24	63	156
8	.11	81	88.9	.24	63	156	72	245	.19	70	266
9	.084	85	146	.19	70	266	78	412	.14	76	444
10	.066	88	237	.14	76	444	82	682	.11	81	732

Table 5: P_2, C_{II}

t	A separate x	S	R	B separate x	S	R	Separated totals S	R	One-price network x	S	R
0	1.3	14	.63	*	*	*	6	.63	*	*	*
1	.47	42	1.6	*	*	*	21	1.6	1.1	17	2.9
2	.23	65	2.4	.59	35	3.3	50	5.7	.43	46	6.2
3	.12	79	3.1	.28	60	5.3	69	8.3	.21	67	9.0
4	.066	88	3.7	.14	76	6.9	82	10.6	.11	81	11
5	.036	93	4.3	.077	86	8.5	90	12.8	.061	89	14
6	.020	96	4.9	.043	92	10	94	15.0	.034	94	16
7	.012	98	6.0	.024	95	12	97	17.7	.019	96	19
8	.007	99	7.1	.014	97	14	98	21.0	.011	98	22
9	.004	99	8.1	.008	98	16	99	24.3	.006	99	24
10	.002	100	8.2	.005	99	20	99	28.4	.004	99	33

Demand satisfaction. As time passes, demand satisfactions approach 100%. The approach is much more rapid for C_{II}. In either case, the forces that drive down unit cost make the service affordable to virtually every potential user.

The main trends noted above and in the preceding section do not change substantially when other values of the parameters are used. In most cases, we are near a full-utilization scenario of 100% by $t = 10$, so there is no significant difference among P_1, P_2, and P_3 for larger t. Revenues at such a time are a bit higher for the one-price network, but the difference is not great. There is clearly an advantage in price for priority users with the one-price network, which penalizes ordinary users by about 30% or higher prices than in the separated network. However, the lower A-separate price is approximately equal to the single price for the one-price network one or two periods hence, so in a dynamic world the ordinary users do not fare too badly and might even become attracted to the QoS provided by a one-price network.

Fig. 2 shows the prices from Table 7 for the separate and one-price networks. It shows graphically how quickly the the prices on the one-price network get reduced to the levels of the separate network for A users.

5 Summary

Our purpose has been to compare three network configurations for data transmission over the Internet when user demands are divided into delay-sensitive and delay-insensitive demands. Prices for the demand types were based on transfer volume and determined by equality between network costs and revenues. Dynamic uncertainties were accounted for by considering alternative futures for demands and costs, including economies of scale for costs and possible effects of competition and technological advances.

Table 6: P_3, C_I

	A separate			*B* separate			Separated totals		One-price network		
t	x	S	R	x	S	R	S	R	x	S	R
0	.67	83	2.2	*	*	*	42	2.2	*	*	*
1	.51	94	3.8	*	*	*	47	3.7	1.2	32	6.2
2	.40	98	6.3	1.2	32	6.2	65	12.5	.71	80	18.1
3	.32	99	10.0	.71	80	18.1	90	28.1	.53	93	31.6
4	.25	100	16.0	.53	93	31.6	96	47.6	.42	97	51.8
5	.20	100	25.4	.42	97	51.8	99	77.2	.33	99	83.2
6	.16	100	40.4	.33	99	83.2	100	124	.26	100	133
7	.13	100	64.5	.26	100	133	100	198	.21	100	212
8	.10	100	102	.21	100	212	100	314	.16	100	336
9	.079	100	162	.16	100	336	100	497	.13	100	536
10	.063	100	258	.13	100	536	100	794	.10	100	852

Table 7: P_3, C_{II}

	A separate			*B* separate			Separated totals		One-price network		
t	x	S	R	x	S	R	S	R	x	S	R
0	.67	83	2.2	*	*	*	42	2.2	*	*	*
1	.36	98	2.8	.84	66	4.5	82	7.3	.61	88	8.6
2	.20	100	3.2	.42	99	6.5	99	9.7	.33	99	10.4
3	.11	100	3.6	.23	100	7.4	100	11.0	.18	100	11.8
4	.063	100	4.0	.13	100	8.4	100	12.4	.10	100	13.3
5	.036	100	4.6	.073	100	9.3	100	14.0	.058	100	14.8
6	.020	100	5.1	.041	100	10.5	100	15.6	.033	100	16.9
7	.012	100	6.1	.023	100	11.8	100	17.9	.019	100	19.5
8	.007	100	7.2	.013	100	13.3	100	20.5	.011	100	22.5
9	.004	100	8.2	.008	100	16.3	100	24.6	.006	100	24.6
10	.002	100	8.2	.005	100	20.5	100	28.7	.004	100	32.8

The three network configurations investigated were a separated network for the demand types, a single one-price network that provides high QoS to all users, and a two-tiered network that logically distinguishes between types. Dynamic analysis showed that network comparisons can be sensitive to demand and cost scenarios, no network is obviously superior to the others, and as t gets large the trends are pretty well fixed. In terms of prices, the premium-service one-price network benefits delay-sensitive users but penalizes delay-insensitive users, and the two-tiered network usually gives a modest advantage over the separated network to both types. The largest revenues occur either for the separated network or the one-price network. Demand satisfaction percentages for the three are comparable, with no network uniformly superior to the others. Potential user participation approaches 100% as time passes, and this happens quickly when unit costs and prices decrease rapidly. Even the delay-sensitive users see their prices and demand satisfactions approach what they could obtain on a separate network within one or two time periods.

Acknowledgements: We thank Dave Belanger, Chuck Kalmanek, Tony Lauck, and Clem McCalla for helpful comments.

References

[1] L. Anania and R. J. Solomon, Flat–the minimalist price, pp. 91-118 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[2] *Boardwatch* magazine, ⟨http://www.boardwatch.com/⟩.

[3] K. G. Coffman and A. M. Odlyzko, The size and growth rate of the Internet. Available at ⟨http://www.research.att.com/∼amo⟩.

[4] P. Ferguson and G. Huston, *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* Wiley, 1998.

[5] P. C. Fishburn and A. M. Odlyzko, Dynamic behavior of differential pricing and Quality of Service options for the Internet, extended version, available at ⟨http://www.research.att.com/∼amo⟩.

[6] P. C. Fishburn, A. M. Odlyzko, and R. C. Siders, Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars, *First Monday,* vol. 2, no. 7 (July 1997), ⟨http://www.firstmonday.dk/⟩. Also to appear in *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property,* D. Hurley, B. Kahin, and H. Varian, eds., MIT Press. Available at ⟨http://www.research.att.com/∼amo⟩.

[7] J. Harms, From SWITCH to SWITCH∗ - extrapolating from a case study, *Proc. INET'94,* pp. 341-1 to 341-6. Available at ⟨http://info.isoc.org/isoc/whatis/conferences/inet/94/papers/index.html⟩.

[8] D. R. Irvin, Modeling the cost of data communication for multi-node computer networks operating in the United States, *IBM J. Res. Develop.* vol. 37 (1993), pp. 537-546.

[9] B. Leida, A cost model of Internet service providers: Implications for Internet telephony and yield management, M.S. thesis, department of Electr. Eng. and Comp. Sci. and Technology and Policy Program, MIT, 1998. Available at ⟨http://www.nmis.org/AboutNMIS/Team/BrettL/contents.html⟩.

[10] J. K. MacKie-Mason and H. R. Varian, Economic FAQs about the Internet, pp. 27–62 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. A version is available in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[11] L. W. McKnight and J. P. Bailey, eds., *Internet Economics,* MIT Press, 1997. Preliminary version of many papers available in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[12] B. M. Mitchell and I. Vogelsang, *Telecommunications Pricing: Theory and Practice,* Cambridge Univ. Press, 1991.

[13] A. M. Odlyzko, Data networks are lightly utilized, and will stay that way. Available at ⟨http://www.research.att.com/∼amo⟩.

[14] A. M. Odlyzko, The economics of the Internet: Utility, utilization, pricing, and Quality of Service. Available at ⟨http://www.research.att.com/∼amo⟩.

[15] J. Rendleman, Connectivity crunch stymies IT access to high-speed lines, *PCWeek,* April 13, 1998, pp. 1 and 20. Available at ⟨http://www.zdnet.com/pcweek/news/0413/13t1.html⟩.

[16] S. G. Steinberg, Netheads vs. Bellheads, *Wired,* 4, no. 10 (Oct. 1996), pp. 144-147, 206-213. Available at ⟨http://www.wired.com/wired/4.10/features/atm.html⟩.

[17] UUNet Access Services, available at ⟨http://www.us.uu.net/html/access_services.html⟩.