A Framework for the Study of Pricing in Integrated Networks

Colin Parris

Tenet Group, Computer Science Division, UC Berkeley and

International Computer Science Institute, Berkeley, CA 94720

Srinivasan Keshav

AT&T Bell Laboratories

600 Mountain Ave., Murray Hill, NJ 07974

Domenico Ferrari

Tenet Group, Computer Science Division, UC Berkeley and

International Computer Science Institute, Berkeley, CA 94720

ABSTRACT

Integrated networks of the near future are expected to provide a wide variety of services, which could consume widely differing amounts of resources. We present a framework for pricing services in integrated networks, and study the effect of pricing on user behavior and network performance.

We first describe a network model that is simple, yet models details such as the wealth distribution in society, different classes of service, peak and off-peak traffic, elasticity of user's demand, and call blocking due to budgetary constraints. We then perform experiments to study the effect of setup, per packet and peak load prices on the blocking probability of two classes of calls passing through a single node enforcing admission control. Some selected results are that a) increasing prices first increases the net revenue to a provider, then causes a decrease b) peak-load pricing spreads network utilization more evenly, raising revenue while simultaneously reducing call blocking probability.

Finally, we introduce a novel metric for comparing pricing schemes, and prove that for the most part, a pricing scheme involving setup prices is better than a pricing scheme without such a component.

C. Parris and D. Ferrari were supported by the National Science Foundation and the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement NCR-8919038 with the Corporation for National Research Initiatives, by AT&T Bell Laboratories, Hitachi, Ltd., Hitachi America, Ltd., the University of California under a MICRO grant, and the International Computer Science Institute. The views and conclusions contained in this document are those of the authors, and should not be interpreted as representing official policies, either expressed or implied, of the U.S. Government or any of the sponsoring organizations.

1. Introduction

Data networks are rapidly becoming as essential to modern life as water and electricity distribution networks. As with any other utility, it is reasonable for these networks to charge customers to cover facility and operating costs. This paper presents a first attempt to deal with pricing issues for integrated networks.

There are several reasons why studying data network pricing is important. First, the introduction of higher speed networks has made it feasible to provide higher bandwidth services even over long haul lines. However, higher speed facilities cost a lot more than low-speed (voice grade) facilities. Thus, from the point of view of the service provider, cost recovery is quite important. A pricing scheme that leads to low network utilization could cause severe financial losses (particularly for new services, whose users could be rather price-sensitive). Second, the network will probably provide a spectrum of qualities of service (QOS). While pricing connections with a single QOS has been extensively studied in the phone network, the issues raised by multiple classes of service are new, interesting and practically important. Third, the pricing scheme allows the network to manipulate user demand to optimize network usage (for example, by peak load pricing). Thus, pricing can be thought of as an important tool for congestion control at the time scale of call establishment [5]. Pricing, in combination with congestion control schemes that operate at faster time scales, can lead to networks that provide better service to everyone.

In this paper we concentrate on two aspects of pricing of network services: a) the effects of prices on network utilization, revenue, and blocking probability and b) a comparison of different pricing schemes. We first describe the current system of pricing in networks and review previous work. Next, we present a simplified problem that retains some of the features of the real world, but is amenable to analysis. Simulation experiments are used to observe the effects of pricing. Finally, we present a new way to compare pricing schemes, and illustrate the methodology with an example.

2. Previous Work

The pricing structure in the current data networks is complex and *ad hoc*, and we are not aware of any publicly available information regarding the basis for these prices. As an example of current pricing strategies, in the TCP/IP Internet, regional networks are supported by governmental funding, and local

(LAN) costs are borne by participating organizations. Thus, end users typically pay only for their LAN costs and the membership fees of the regional network (which are usage insensitive). This system works reasonably well, since the majority of the costs are borne by government agencies, and network costs to end users are relatively small.

Pricing in the telephone network has been better studied, but the telephone network offers only a single quality of service, and the costs of providing service are better known. So, it is fairly simple to set a price such that the costs (plus some profit) are recovered [4]. While the basic pricing structure in telephone networks is simple, the pervasive regulatory structures and the *de facto* monopoly situation has meant that most of the research in pricing for telephone networks has been to study the effect of regulation rather than the effect of pricing on network efficiency [1]. In our work, we ignore regulatory constraints and consider only the performance effects of pricing.

In recent work, Cocchi et al. [2] have studied pricing in a reservationless network. They assume that each user can be characterized by a utility function, and can request a QOS by setting bits in its packets. They show that quality sensitive pricing is more efficient (in the Pareto sense) than a flat pricing scheme. The approach is strongly influenced by the reservationless assumption. Since resources cannot be reserved, users may suffer QOS degradation. Then, the only way to measure the 'goodness' of a pricing scheme is to measure the net satisfaction from the network, which means that user utility functions have to be assumed. This is undesirable because of the difficulty in determining a valid user utility function. Hence, this approach does not seem appropriate for a reservation oriented virtual circuit environment, where QOS guarantees can be made.

3. Assumptions in our work

We assume that the network is capable of making QOS guarantees, and that the QOS is negotiated at connection setup. A user presents the network with a set of QOS requirements. Based on the admission control scheme, the request is either accepted or rejected. Rejected users may try again with a degraded QOS request, or at a time when the network is less heavily loaded, or the network may inform the user about QOS requests that could be accepted.

We assume that the network charges a price for each connection. The revenue that is gained is used to offset costs.

For simplicity, we assume that QOS requests can be grouped into some small number of classes, such as those for interactive video communication, bulk data transfer, and so on. A user is characterized by a triple (*Class*, *Duration*, *Money*). *Class* is the class of service requested. *Duration* is the maximum duration of the connection. *Money* is the amount of money that is available for the connection (the budget). The user presents a request with its class type to the network, which responds with either (*Accept*, *Price*) or (*Reject*).

What are the things that can be priced? The price can be made sensitive to

- the amount of resources that are reserved,
- the number of cells or packets,
- the duration of connection,
- the time of day,
- the priority of connection termination,
- the quality of service.

Ultimately, the choice of the pricing policy depends on what can be observed and accounted for. Because of this, any pricing scheme is likely to be unfair to someone. In our work, we are not as much concerned with fairness, as with the impact of pricing on network performance.

Since a pricing scheme, if studied in its entirety, is rather complex, we feel it is useful to study a simplified system that retains the basic structure of the problem, yet is amenable to analysis. We state the simplified problem below; in the rest of the paper, we will present an analysis.

We assume a network that provides only two classes of service, A and B, and that both classes require some fixed amount of bandwidth to be reserved for the connection. Let the class of service (or rather, type of service) refer to bandwidth only. In this case, the utilization is simply the fraction of bandwidth that is reserved. We will assume that a user wealth is bimodally distributed, with a fraction of users being ‘poor’ and the remainder ‘rich’. Once a connection is set up, a user can use all or part of its reserved bandwidth. However, when the user's budget limit is reached, the connection is dropped even if its duration has not been completely exhausted. The *Money* variable and this provision are intended to model, in a highly simplified way, the elasticity of demand.

What we aim to study are the following questions: How do different pricing schemes compare with respect to the blocking probabilities for class A and class B service? How do different pricing schemes compare with respect to total revenue generated ? How can pricing cause discrimination between users? How can peak load pricing be used in this framework? (for this last question, we will assume that some fraction of the demand is time-insensitive).

4. Notation

Users are indexed by i, and we assume that there are a total of n users requesting service (we will also assume that the mapping from users to connections is one-one). A user is represented by a triple (q = Class, T = Duration, m = Money).

Time is considered to be discrete, with intervals of unit length. We will assume that users enter and leave only at the end of an interval. The amount of revenue received by the network per interval is $R(.)$, and its cost per interval is $Cost(.)$. The total revenue is the sum of the revenues over all of the intervals. The utilization of the network in an interval is denoted by u, and is the ratio between the amount of bandwidth reserved for channels existing during that interval and the total bandwidth of the network during that interval. The capacity constraints of the network are expressed as $C(c_1, c_2, \cdots, c_k) = x$, where k is the number of classes of service, c_i is the number of users who have connections of class i, and x is 1 if the network can support all the connections, and 0 otherwise. For example, if a network can support one class 1 channel or up to two class 2 channels only, then $C(0,0) = C(1,0) = C(0,1) = C(0,2) = 1$, and $C(.,.)$ is 0 everywhere else.

A user request is denoted by $Req(q, T)$, and states the class of service and the maximum duration of the connection desired. If the network responds affirmatively, it specifies the price, which could have a fixed setup component and a variable component per interval, (m_f, m_v). If the user has sufficient money, m, to accommodate for the setup component of the price and allow a connection duration of at least 1 interval, the connection is accepted. There is partial acceptance since users whose budget, *m*, cannot support a connection for the required duration, *T*, are admitted into the network for the duration that they can afford. Connections may be rejected either because there is not enough capacity, or because the price of even a single interval exceeds the user's budget constraint. Only connections rejected due to lack of network capacity are considered blocked.

5. The Blocking Probability Criterion

How should the call blocking probability be evaluated? One way to do so is to determine (by simulation) the network utilization density curve (for example *Figure* 2). This curve represents network utilization versus the probability that such a utilization is achieved. Given the arrival distribution for each class of users, the amount of money they have, and the pricing scheme, this curve is well defined. It is clear that, as network services become more expensive, the number of users who can afford these services decreases, and the mass of the density curve shifts to the left.

This curve can now be used for several purposes. If some class, say class A, is for connections which utilize 10% of the network's bandwidth, then the blocking probability of a class A connection is simply the probability mass that lies to the right of 90% (= 100-10) utilization. Mathematically speaking, the blocking probability in the network is memoryless, in the sense that, for any particular value of utilization, the blocking probability for any connection is independent of the path used to reach that utilization. Hence the system can be represented by a Markov chain, and the density curve is simply the stationary distribution of a (continuous) Markov chain.

Under some simple conditions, the expected revenue derived from each class can also be computed from the density curve. For each point along the curve, we know the contribution from class A, and the contribution from class B (there may be several such combinations of A and B at each point). Now, if the pricing scheme is fair, then the price charged to A will be proportional to the fraction of bandwidth it consumes, and so the revenue gained at each utilization point is independent of the combination of A and B calls used to attain that utilization. Thus, for fair pricing schemes, the expected revenue is simply $\int_0^1 R(u)P(u)\,du$, where $R(u)$ is the revenue for utilization u, and $P(u)$ is the probability of that utilization.

If a pricing scheme is not fair, then the utilization per class has to be independently determined, and then the set of utilization curves can be used to find the expected revenue.

6. Simulations

6.1. Simulation Parameters and Methodology

For simulation purposes, the entire network is collapsed to a single node which is a composite of a switch and an output link. The bandwidth to be allocated is that of the output link. A single node is adequate for our analysis, since both call admission and collecting tolls can be performed at a single point at the entrance of the network. We do ignore the effects of routing, but for simplicity, in this paper we will only analyze a single node.

Calls from the two classes of service in our model, A and B, are assumed to request 1% and 5% of the total network bandwidth respectively. An analogy to these classes of service is that of a phone call, for Class A, and a low quality video conferencing session for Class B. All other performance parameters are identical for the two classes.

The node uses a real-time reservation-oriented call setup protocol to provide QOS guarantees to users. There is also an admission control scheme that accepts users' requests based on the class of service requested and the current utilization of the node. All pricing schemes interact with the admission control scheme in responding to users' requests. Details of the real-time protocol and the admission control scheme can be found in [3].

As mentioned previously, users are modeled by the triple (q = Class, T = Duration, m = Money). In this representation, q denotes the class of service required by the user. This is modeled as a bimodal distribution with a fraction β of users requesting Class A service and a fraction $1-\beta$ using Class B service. In the simulations given below β is 0.8.

T denotes the maximum duration of the connection requested by the user. Connection durations are exponentially distributed with a mean dependent on q. The mean of Class B service is 240 intervals, and that of Class A service is 36 intervals. We assume that all users send at the same fixed rate, thus the duration of a connection and the number of packets sent over the connection are easily related. The number of packets sent is the duration of the connection multiplied by the sending rate. As a result of this relationship, the price per packet and the price per interval can be used interchangeably, as users do not have duration times that are fractions of an interval.

m denotes user wealth, which is bimodally distributed with a fraction γ of users being 'poor' and a fraction 1- γ being 'rich'. A 'rich' user has a budget of $600 while a 'poor' user has $60. The budget of a user and the class of service requested are mutually independent.

User requests arrive according to a Poisson distribution with a rate of 1 request per interval. Users enter at the beginning of an interval and leave at the end of an interval. As the mean durations for class A and B connections are relatively long, in each interval there will be connections that were previously accepted and there may be connections that were accepted at the beginning of this interval. In some intervals there may be no connections present. In the Peak Load pricing simulations, user requests arrive according to a Poisson distribution, but the rate during the peak load period is 2 requests per interval, while that during the non-peak load period is 1 request per interval.

In each of the runs, 5000 user requests were simulated. Some of the data gathered on a per-interval basis are the revenue generated, the bandwidth utilized , and the requests submitted (as well as the breakdown into the number of requests accepted and the number of requests denied). Note that the parameter values chosen are not intended to reflect any current or future market conditions, their only value is to illustrate the methodology with quantitative examples.

6.2. Effect of Per Packet pricing

In *Per Packet* pricing users are charged only on the basis of the number of packets they send regardless of their class of service. Class B service permits the generation of 5 times as many packets per interval as Class A service. The first question studied is that of the general impact of the per-packet price on the total revenue and the average bandwidth utilization. To this end, simulations were conducted over a wide range of per-packet prices, and the results are recorded in Table I.

Table I shows that, as the per packet price increases, the revenue increases, a maximum is reached, and then the revenue decreases. The peak revenue is reached when the per packet price is about $50 per packet. This is easily explained, as at low per packet prices very little total revenue is generated, and at high per packet prices, users are unable to afford any services. It should be noted that at low per packet prices very little user discrimination is experienced (ie., both ‘rich’ and ‘poor’ users can gain access to the network) and the resulting average bandwidth utilization is very high. This can be seen when the per packet price is $1 and even $10. With a high per packet price, there is greater discrimination, as many requests are turned away (ie., all ‘poor’users) due to insufficient budgets, thereby decreasing the total revenue generated as well as the average utilization of the node. At an intermediate price (in this case $50 per packet), the total revenue is maximized, while the loss of poorer users leads to a node utilization of only 14.7%.

Table I - Per Packet Pricing.

$/packet	Total Revenue	Avg. Bandwidth Util (%)
1	78529	90.1
10	445980	58.6
20	516760	30.7
30	582630	24.1
50	596050	14.7
100	456100	5.6
250	348250	1.7

It should be noted that discrimination is directly related to the wealth distribution ratio γ. The larger the value of γ, the greater the number of ‘rich’ clients, hence the larger the total revenue generated at high per packet prices. Another significant observation is that, at high per packet prices, the blocking probabilities of both classes of service are zero. This is again due to discrimination. 80 % of the users are ‘poor’ and the budget of a ‘poor’ user is $60, so at per packet prices above $20 the fraction of ‘poor’ users at this node is insignificant, as their connection durations are short. In fact, with per packet prices above $60, the ‘poor’ users are not allowed access to the node. Hence, with per packet prices above $20, most of the node utilization is due to the ‘rich’ users, but, as they comprise only 20% of users, the network operates at a low utilization point, and hence blocking probabilities are zero.

Another area of interest is the determining an operating point based on a specific criterion such as the maximizing of revenue or the recovery of costs. To maximize revenue, it is clear from the earlier discussion that a per packet price of about $50 should be charged. If the goal is to recover costs, the situation needs a little more analysis. To facilitate the analysis, we will consider per packet prices whose blocking probabilities for any of the classes of service are greater than zero. For the purpose of this analysis, a blocking probability less than 0.001 % is identically zero; therefore per-packet prices above $10 are not considered.

Table II - Per Packet Pricing.				
$/packet	Total Revenue	Avg. Util (%)	Blk. Prob A (%)	Blk. Prob B (%)
1	78529	90.0	6.09	43.28
2	155938	91.0	4.69	39.86
4	286148	91.0	0.62	33.84
5	325140	81.0	0.56	17.246
6	384762	72.3	0.20	10.147
8	398864	61.0	0.002	2.992
9	411111	59.3	0.001	1.934
10	445980	58.6	0.001	1.832

A plot of blocking probabilities of Class A and B service, *Figure 1* shows that an increase in the per-packet price causes a decrease in the blocking probabilities. This decrease in blocking probabilities is due to the reduction in the duration of user connections thereby allowing other users easier access into the node. As Class A connections utilize less bandwidth, their blocking probabilities are less than those of Class B connections.

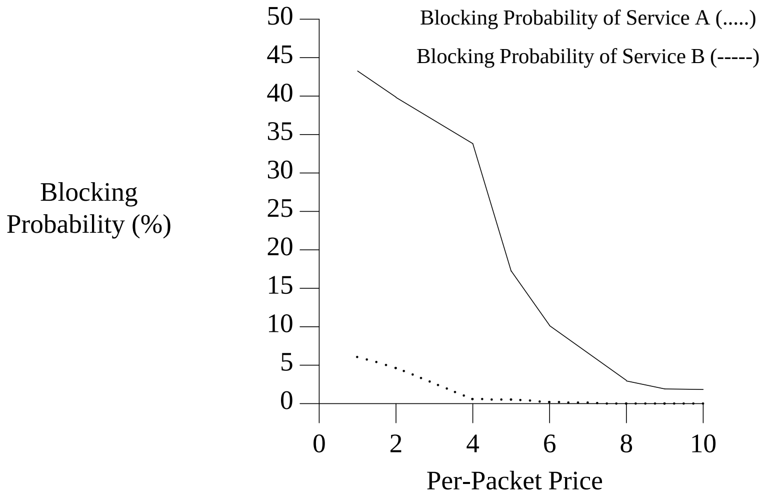

Figure 1
Blocking probabilities vs. per-packet price

If cost is to be recovered, then it is possible to operate at one of two operating points, corresponding to a high per packet price or a low per packet price, each with it own benefit. Note that at either operating point the revenue generated is approximately equal and is enough to recover costs. From a social welfare point of view it is better to choose the operating point corresponding to a low per-packet price, as both ‘rich’ and ‘poor’ users are given access to the node. At the operating point corresponding to the higher per-packet price, the blocking probability is very low (if at all present) due to the large amount of discrimination hence there is always access to the network. There will be a high blocking probability at the other operating point. The choice of operating point is dependent entirely on the goals of the network administrators, and in this case may not be market dependent.

An example of this is shown in Table I and II. Assume that the cost needed to be recovered during the period in question is $320,000; then, either $4 or $250 per-packet can be chosen as the operating point. If $4 is chosen, the utilization and blocking probability are larger than those at $250, but more users are allowed access to the node. The blocking probability of both Class A and B service at $250 is 0. The choice of the operating point at $4 is better, from a social welfare point of view, than that at $250, but this is done at the expense of network availability. The use of different parameter values will obviously generate different results, our endeavour here is to illustrate the point rather than provide a numerical answer.

In observing the probability density of utilization graph in *Figure 2*, a shifting to the left is noted as the per-packet price increases. This situation occurs as the price increases due to increased discrimination of users. With a price increase fewer users are admitted into the network, resulting in a lower bandwidth utilization; it should be noted that some utilizations are never attained at high per-packet prices.

6.3. Effect of Setup Price pricing.

In *Setup Price* pricing, users are charged a setup price for establishing a connection, and are then also charged for the number of packets sent. This setup price is assessed only at the beginning of the connection. The setup price is assumed to be independent of the class of service requested by the user, since in practice, setup simply involves writing data to a customer database, and the cost of this operation is independent of the resources consumed by a user. This scheme is identical to the per packet pricing scheme but with the inclusion of a setup charge. So, the Per Packet pricing scheme can be viewed as a Setup price pricing scheme with the setup price being $0.

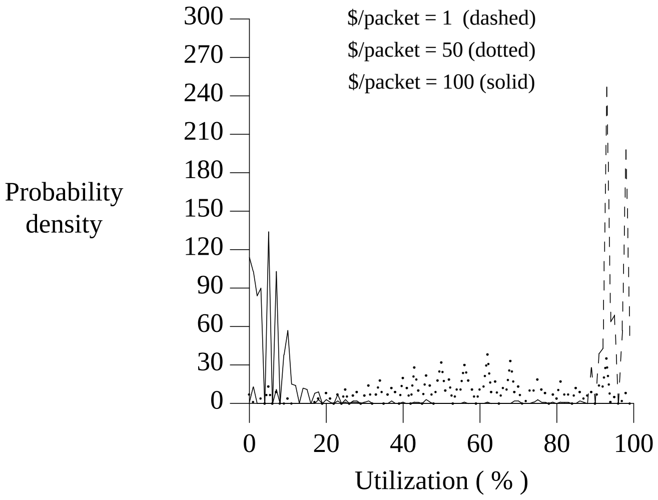

Figure 2
Probability density of bandwidth utilization

The setup price can be used to increase user discrimination, since it can be independent of the per-packet price. For example, a high setup price with a low per-packet price would only allow 'rich' users to access the network, but then they would pay a low price for the service (ie., a low per-packet price). We first examine the general trend of the total revenue and the average bandwidth utilization under a range of setup and per packet prices. In these simulations the setup price was set at 20 % of the per packet price (except in the case of $1/packet).

Table III shows that, as the per-packet and setup price increase, the total revenue increases until it reaches a maximum, and then decreases. The peak revenue is at $50, with the associated setup price of $10. As before, the larger setup and per-packet price introduce discrimination, especially at the $60 per packet price where all 'poor' users are denied access to the network. The general trend is similar to that of the Per Packet pricing scheme, which is not surprising, since the Per Packet pricing scheme is a special case of the Setup price scheme.

In comparing the Setup price scheme to the Per Packet pricing scheme (ie., Table I to Table III), we observe that the total revenue generated by the Per Packet pricing scheme in the $10 to $30 range is greater than that generated by the Setup price pricing scheme. This situation is not what would be expected, as the lower blocking probability caused by the inclusion of a setup charge should allow more users access to the node thereby, causing the revenue generated to be greater than or equal to that of the Per Packet pricing scheme.

The reason for this situation lies in the fact that the setup price is not a multiple of the per packet price. Assuming a per packet price of $20 (and a setup price of $2), a ‘poor’ user (whose budget is $60) in the Per Packet pricing scheme would be able to send at most 3 packets thereby paying $60, whereas in the Setup price scheme the user would be able to send at most 2 packets thereby paying $44. Thus the network loses $16 due to the value of the setup price. This loss of revenue in the per packet range of $10 to $30 causes the total revenue of the Per Packet pricing scheme to be greater than that of the Setup price pricing scheme.

Results of simulations with setup prices being multiples of the per-packet prices are shown in Table IV, where the total revenue generated is constantly larger than that generated in the Per Packet pricing scheme (Table II). The only exception is the per-packet price of $9, and this is due to the fact that $2 are lost by the Setup price scheme because of the setup price of $18. This accounts for the 1.98% drop. A simple calculation can corroborate this reasoning. On the average 4000 (80% of 5000) users are ‘poor’ and from each ‘poor’ user the node loses $2, for a total loss of $8000. The difference between the Per Packet pricing scheme total revenue ($411111) and that of the Setup price pricing scheme ($403425) is $7686. Considering that $8000 is an average value, the difference in revenues is close enough to verify the above reasoning.

In the per-packet range from $50 to $250, in both schemes the ‘poor’ users have very little impact on the node, and, in the setup price scheme the ‘rich’ users pay the setup price in addition to the per-packet price, thereby increasing the total revenue generated in this range.

Table III - Setup Price Pricing.

$/packet	Setup Cost	Total Revenue	Avg. bandwidth Util
1	1	79123	91.0
10	2	443496	48.0
20	4	444676	26.0
30	6	481986	19.0
50	10	607010	14.0
100	20	484280	6.1
250	50	360150	1.6

Another interesting comparison can be made by observing the revenue vs. utilization graphs of the two schemes *(Figure 3A, 3B)*. Both of the curves increase monotonically with utilization; however, in the setup price scheme the curve is actually a scatter plot. This is due to the different setup charges of the different combinations of connections that produce the same utilization.

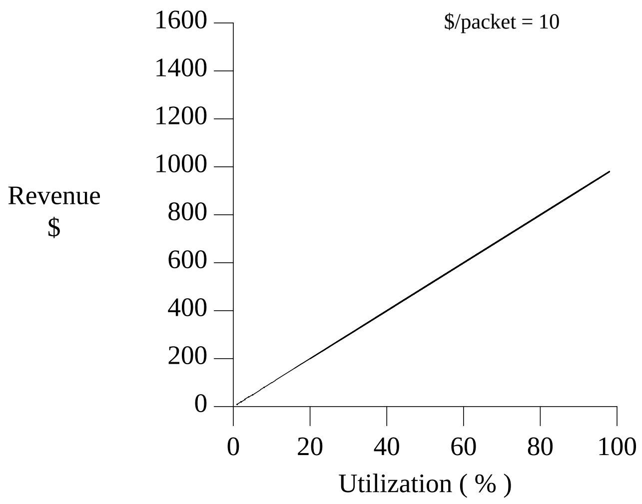

Figure 3A
Revenue vs. utilization for the Per Packet pricing scheme

The use of setup price can be deemed "unfair" in the sense that users with short conversations subsidize users with long conversations as both pay the same setup price. However, if there is an actual cost to set up a connection (we assume that the actual setup cost is $0), then this cost subtracted from the setup price will cause the scatter plot to become a straight line, similar to that seen in the Per Packet scheme.

We now analyze the effect of the setup price on the blocking probability of each class of service. In this case, the setup price is twice the per packet price and we have chosen small per-packet prices to facilitate the analysis. We can see from Table IV that the additional discriminatory effect introduced by the setup charge (ie., the duration of connections of the ‘poor’ users are reduced) causes the blocking probabilities to be lower than experienced with per packet pricing. This is also evidenced by the resulting lower average utilizations associated with the setup price scheme as compared with that of the Per Packet scheme.

From this result we can conjecture that, if the blocking probability of the node is too high, the inclusion of a setup price into the Per Packet pricing scheme would cause a reduction in the blocking probability. There may also be a corresponding reduction in the total revenue. Of course, another way to reduce blocking probability is to increase the capacity of the node. The actual value of the setup price introduced and the

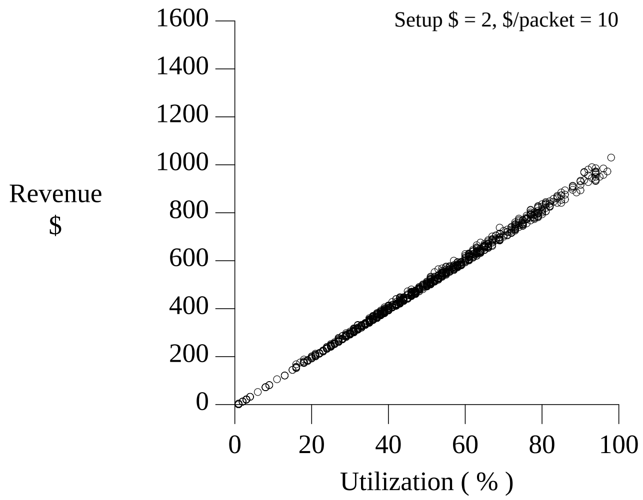

Figure 3B
Revenue vs. utilization for the Setup price pricing scheme

corresponding reduction in blocking probability and total revenue are heavily dependent on the wealth distribution among the users.

Table IV - Setup Price Pricing.					
$/packet	Setup Cost	Total Revenue	Avg. Util (%)	Blk. prob A (%)	Blk. prob B (%)
1	2	90946	91.3	4.384	36.013
2	4	170158	91.0	0.562	34.719
4	8	309672	81.0	0.030	24.170
5	10	364225	73.1	0.011	11.425
6	12	388038	64.0	0.002	2.867
8	16	439968	53.8	0.001	0.468
9	18	403425	46.2	0.000	0.871
10	20	447140	44.0	0.000	0.852

Figure 4 shows the probability density of utilization for the Setup price scheme at three different setup prices. The density curve shifts to the left as the setup price is increased, due to the additional discrimination introduced by the setup price. The larger the price, the fewer the users allowed access to the node or the shorter the duration of their connections. As the curve shifts to the left, there is a corresponding reduction in blocking probability.

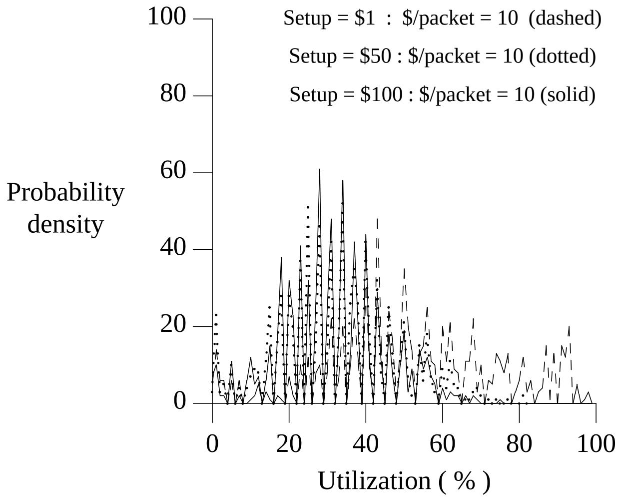

Figure 4
Probability density of utilization (Setup Price pricing scheme)

6.4. Effect of Peak load pricing

In *Peak Load Pricing* there are specified periods of time designated as *peak* and *off peak*, respectively. A period consist of a number of consecutive intervals characterized by a specific value of the arrival rate of requests. During the peak period, requests arrive at a rate larger than during off-peak periods. There are an equal number of peak and off-peak periods in each run. There are specific prices associated with each period, and the peak period price is obviously higher than the price associated with the off peak period. For the purpose of these simulations, we assume that prices are fixed during each period and that conversations beginning during any period are charged at the rate associated with that period for the duration of the connection. In each period there is a setup price and a per-packet price. The setup price is the same for both peak and off peak periods. During peak periods the per-packet price is 3 times that of the off peak period per-packet price.

Each user request is associated with an *elasticity* which determines if a request arriving during a peak period can be deferred to a subsequent off peak period. A request with elasticity 0 cannot be deferred and *vice versa*. In the simulations, elasticity is modeled as a random boolean variable with probability 0.2 of being 1. If a user makes an elastic request during a peak period and the request can be completed (ie., the user can pay for the entire duration requested), the user is admitted and charged at the peak load per-packet price. However, if the user has an insufficient budget, the user is deferred to an off peak period and allowed the maximum duration that its money can pay for up to the requested duration.

To make a suitable analysis of the effect of this pricing scheme control, simulations of baseline cases were run. In these control simulations, the environment remained the same, in that the arrival rate during peak and off peak periods were the same. However, all users had elasticities of 0 and the per-packet price was the same during both periods. This per-packet price was 2 times that of the off peak period price (ie., the average of the peak and off peak per packet price). The setup price was the same in all the peak load experiments. The results of the simulations over a range of setup and per packet prices are given in Table V. In the PLP (Peak Load Pricing simulations) section of the table, the $/packet price is the off peak period per-packet price.

In both of the simulation groups in Table V, as before, the general trend is that the total revenue increases until a maximum is reached, and then decreases. In the control simulations (NPLP - Non Peak Load Pricing simulations), the total revenue is less than that generated during the PLP (Peak Load Pricing) simulation group, for the first three simulations. This is due to the elasticity of the users: during the peak period the per packet prices are $3, $15, and $45, respectively, and elastic users whose budgets do not allow completion of the conversation are deferred to an off peak period which would provide them with a longer connection duration. As a result of these deferrals, the Peak Load Pricing scheme shows a larger total revenue generated.

Table V - Peak Load Pricing.

Exper.	$/packet	Setup Cost	Total Revenue	Avg. Util	Max. Util
NPLP	2	1	221523	55.2	99.0
NPLP	10	2	449474	24.1	98.0
NPLP	30	6	477450	7.8	64.0
NPLP	50	10	622880	5.9	45.0
NPLP	100	20	456360	2.4	25.0
NPLP	200	40	304680	0.7	9.0
PLP	1	1	225407	67.3	99.0
PLP	5	2	451789	24.5	80.0
PLP	15	6	506031	9.8	54.0
PLP	25	10	477070	6.2	33.0
PLP	50	20	396872	2.7	21.0
PLP	100	40	363780	1.2	15.0

In the latter three simulations, the peak load prices are $75, $150, and $200, respectively, which introduces a high level of discrimination, so that ‘poor’ users with no elasticity are denied access to the node. The ‘poor’ users with elasticity encounter per-packet prices of $25, $50, and $100, which may permit a brief conversation duration at best. These two scenarios cause the total revenue generated by the Peak Load Pricing scheme to be less than that of the Non Peak Load scheme in this range.

An important consequence of Peak Load Pricing is that user requests are dispersed over a longer time, which results in a lower maximum utilization and a more even distribution among the utilization. An example of these characteristics can be seen in *Figure 6* and Table V. The maximum utilizations of the PLP simulations usually range from about 4% to 10 % lower that those of the NPLP simulations. The exceptions to this occur in the first and last simulations. In the first simulation, the low value of the peak load per-packet price, $3 (as compared to a per-packet price of $2) is so affordable that few elastic clients take advantage of their elasticity; therefore, deferral, which is the major characteristic of Peak Load Pricing, is seldom done. In the last simulation, the large peak load per-packet price of $300 is paid by a few of the ‘rich’ users who cannot be deferred, thus substantially increasing the total revenue, while with the per-packet price of $200 in the NPLP simulations, the ‘rich’ users need pay no more than $200 per packet. The maximum utilization increases due to deferral by elastic users who cannot afford the $300 per-packet peak period price for the entire duration of their conversation but can afford the $100 per-packet off peak period price. In the NPLP simulation, there is no possibility of paying less than $200, and those users who do not have sufficient money are denied access to the node, hence the corresponding reduction in maximum utilization and total revenue.

The trend of the blocking probabilities can be seen by examining the maximum utilizations. As the maximum utilizations experienced with Peak Load Pricing are generally less than those with per packet pricing, the blocking probabilities experienced with Peak Load pricing will generally be less than those of Non Peak Load per packet pricing.

7. Comparing Pricing Schemes

How should two pricing schemes be compared? A pricing scheme has two effects: a) it generates revenue for the service provider, b) by discriminating against poorer users, it reduces the average utilization, and hence the blocking probability of a call. Loosely speaking, we claim that one pricing scheme is

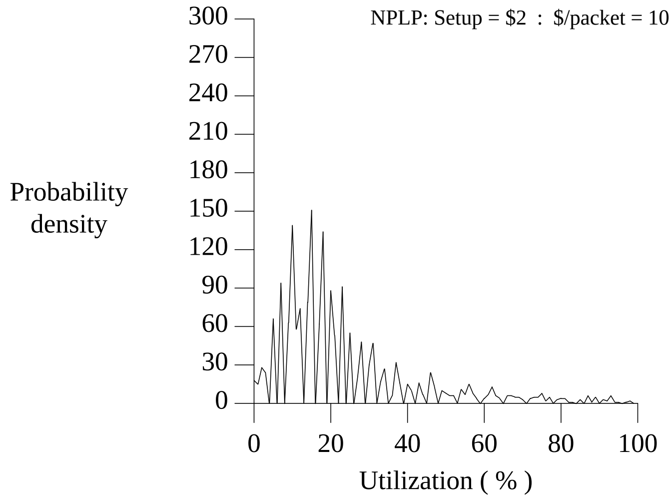

Figure 6 (a). Probability density of utilization: Non Peak Load Pricing, $10/packet

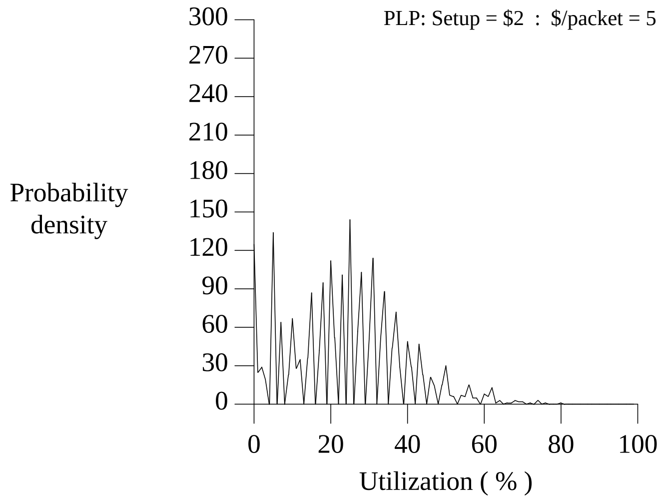

Figure 6 (b). Probability density of utilization: Peak Load Pricing, $5/packet

better than another if for the same revenue generated, the blocking probability is lower. The idea is that for the better scheme, while the provider still recovers costs, the users are better able to access the network.

We now discuss the idea further. Note that for a given pricing scheme, as prices are raised, the revenue increases, then decreases, and the blocking probability keeps decreasing till it reaches 0. Suppose a network provider has a operating cost that is to be recovered. This cost can be recovered by operating either at a high price, where the blocking probability is nearly 0, or at low cost, where there is some chance of blocking. In the interests of social welfare, that is, allowing poorer users to gain access to the network, we consider only the lower of the two operating points. Then, for each pricing scheme, there is a utilization

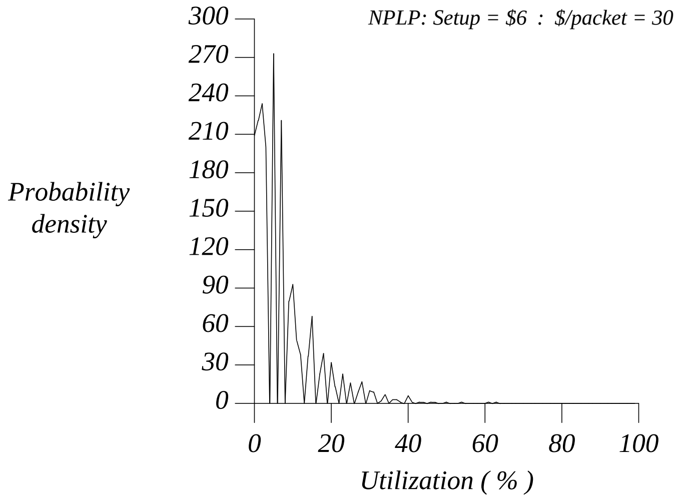

Figure 6 (c). Probability density of utilization: Non Peak Load Pricing, $30/packet

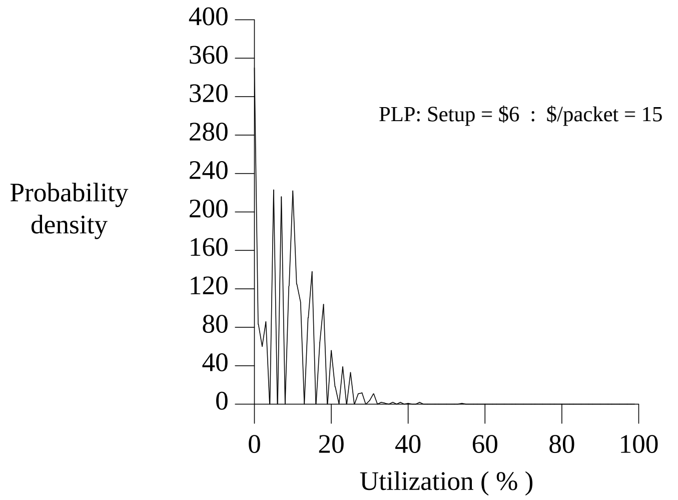

Figure 6 (d). Probability density of utilization: Peak Load Pricing, $15/packet

value at which costs are recovered, and the same utilization corresponds to a particular blocking probability for each class. We call this the breakeven blocking probability (BBP) of that class (*Figure* 7). This represents the amount of degradation of service users from that class must suffer so that the network will have enough revenue to sustain its operations. Since the blocking probability must be low in order for users to gain utility from the network (*irrespective* of their utility functions), we define the figure of merit of a pricing scheme as its breakeven blocking probability. From the point of view of users from that class, one pricing scheme is better than an other if its BBP is lower.

In order to obtain the revenue and blocking probability curves as functions of utilization, we more

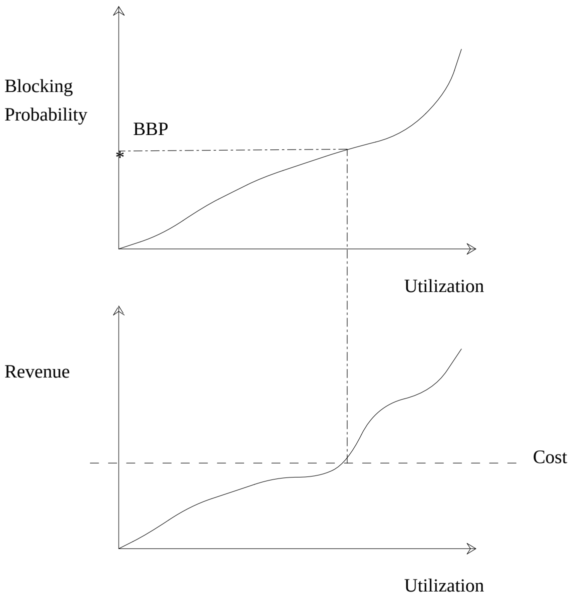

Figure 7: Breakeven blocking probability

generally define utilization to be a function of the variables which determine either the revenue or the blocking probability in the system. Since revenue may, in general, depend on a number of variables, each of them must be accounted for in the utilization. That is, the simulation must vary each of parameters that may affect the utilization, and compute the average utilization over the entire range of parameters. This can introduces some difficulties. In the context of our model, however, since revenue is gained only from setup and duration prices, the situation is simpler. We simply vary the number of connections over the duration of the simulation, and observe the revenue, the average utilization, and the blocking probability of the connection (from the utilization density curve). Then, both the blocking probability vs. utilization and revenue vs. utilization curves are well defined.

An example of this methodology will be illustrated for comparing the Per Packet and Setup Price pricing schemes. Only the blocking probabilities corresponding to Class B service are considered. As average utilization is the common independent variable in the revenue vs average utilization and blocking probability vs average utilization curves, it is sufficient to plot the revenue vs blocking probability as the per packet price is increased. The data in Tables II and IV is used to construct the plot in *Figure 8*.

With increasing per packet prices the operating point on the curve moves from left to right, with the Setup price curve generally below the Per Packet price curve. For a chosen revenue it is clearly seen that the blocking probability of the Setup price scheme is generally lower than that of the Per Packet price

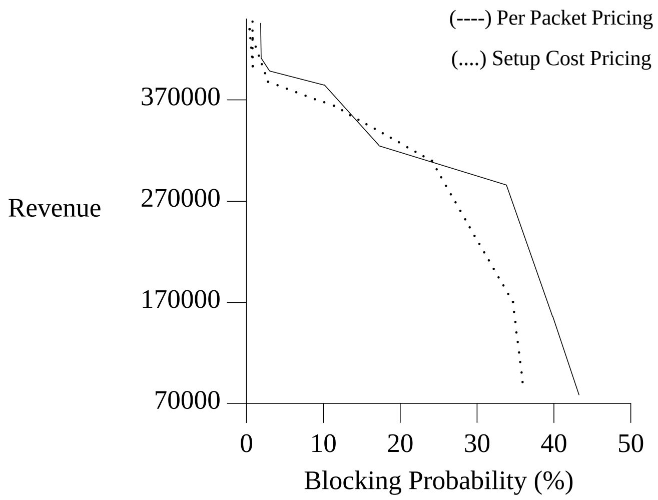

Figure 8
Revenue vs. blocking probability for the two pricing schemes

scheme. Thus, for the most part, the Setup price scheme is preferred.

The only exception to this occurs in the 15% to 25% blocking probability range (revenue in the $320,000 to $350,000 range). Here, the blocking probability of the Per Packet pricing scheme is less than that of the Setup price pricing scheme. It should be noted that in this revenue range, we are comparing the Per Packet pricing scheme with a per packet price of $5-$6, while the per packet price in the Setup price pricing scheme is $4-$5. As our point of view is that of obtaining the recovery cost at the least blocking probability, we are somewhat indifferent to the per packet price paid by the users, and in this range the Per Packet scheme would be the preferred scheme.

8. Discussion and Conclusion

We have begun an investigation of the difficult problem of pricing services in a reservation-oriented integrated network. The issues we have considered in this first paper are:

- the influences of the pricing scheme on network utilization, revenue, and blocking probability, and
- the definition of a criterion for comparing different pricing schemes.

Our main conclusions are that

- Given a pricing scheme, increasing prices first increases and then decreases net revenue, but the blocking probability continuously decreases.

- As prices increases, network utilization decreases.
- The effect of setup prices is to increase revenue and decrease blocking probability. However, setup prices are unfair in that short conversations subsidize longer conversations.
- Peak load pricing reduces the peak utilization and the blocking probability of all classes of traffic, and increases revenue by spreading demand over peak and non-peak periods. Thus, it is a useful tool in controlling user demand to prevent congestion.
- According to our criterion, Setup pricing is better than Per Packet pricing, for the most part.

Our study has been done under a number of simplifying but restrictive assumptions, which we plan to relax at least partially in future work. For instance:

- the user model, in which money *m* and *elasticity* are introduced to represent (quite coarsely) the behavior of users when confronted with varying prices, i.e.,the influence of pricing on demand;
- the network model, in which only one of the resources of a real network appears, i.e., link bandwidth;
- the network model, in which only a single node is considered; the extension of single node analysis to that of a group of nodes with the associated routing effects needs to be considered.

These and other coarse approximations of reality have nevertheless resulted in a simulation model whose behavior has qualitatively resembled the one that is to be expected of an integrated services network. They have allowed us to present our criterion for comparing various pricing schemes using blocking probability and revenue, and the effects of the pricing schemes on network performance. Thus, we believe that our network model and proposed methodology for evaluating pricing schemes lay a firm foundation for exploring these issues in future work.

9. References

1. J. H. Alleman and R. D. Emmerson, in *Perspectives on the Telephone Industry: the Challenge for the Future*, Ballinger Publishing Company , 1989. (Papers presented at a conference, Local Exchange Pricing: the Challenge of the Future, Nov 16-18 1987, New Orleans).

2. R. Cocchi, D. Estrin, S. Shenker and L. Zhang, A Study of Priority Pricing in Multiple Service Class

Networks, *Proc. ACM SigComm 1991*, September 1991.

3. D. Ferrari and D. Verma, A Scheme for Real-Time Channel Establishment in Wide-Area Networks, *IEEE J. on Selected Areas in Communications*, April 1990.

4. J. Hazlewood, Optimum Pricing as Applied to Telephone Service, *Review of Economic Studies 12*, 2 (1951), 67-78.

5. S. Keshav, Congestion Control in Computer Networks, *PhD thesis* , University of California, Berkeley , August 1991.