Dynamic Pricing as Congestion Control in ATM Networks*

Jon M. Peha
Carnegie Mellon University

Abstract

Several researchers have recently advocated dynamic pricing mechanisms like the *smart market*. This paper explores how dynamic state-dependent pricing and explicit congestion control can both be used to avoid and alleviate congestion. We show that dynamic pricing has significant advantages for heterogeneous traffic, although it reduces raw throughput somewhat. When propagation delay is non-trivial, a *slow-reacting* version of dynamic pricing is preferable. This paper also advocates use of novel *stream-oriented* best-effort ATM services, with which a stream's arrival process is declared to the network before transmission begins and then policed, although there are no performance guarantees and best-effort streams are never blocked. With this approach, applications have incentive to decrease traffic burstiness, and to reveal important information about their packet streams, making mechanisms like slow-reacting dynamic pricing more practical.

1 Introduction

Many telephone networks, cable TV networks, and computer networks like the Internet will become *integrated-services networks*, which are networks that offer multiple services to support diverse traffic, e.g. one service for telephony, and another for VCR-quality movies. Each service may have a different price. This paper addresses *usage-based pricing*, where price depends on how a customer uses the network, e.g. how many packets are sent and when. Non-usage-based revenue sources such as flat monthly fees and government subsidies are outside the scope of this paper.

Some goals for usage-based pricing are the same as those for important traffic control algorithms. First, pricing is a mechanism for resource allocation. Instead of explicitly assigning resources to specific packet streams, pricing signals that those who derive a value below current price should not use a service, thereby allocating resources to more valuable streams. Second, pricing provides incentives to adjust user behavior, which is an alternative to explicitly constraining user behavior. Usage-based pricing can induce users to change factors like transmission rate, burstiness, or the time of day of their transmission. It can also induce users to reveal information about their traffic, like a stream's value, its performance requirements, or its arrival process. Indeed, without price incentives, users would never reveal that a given stream could tolerate significant delays or losses, making it impossible to use traffic control approaches like [1, 2, 3, 4], which allow the network to improve performance and carry more traffic [5]. Thus, all users can benefit from usage-based pricing [6, 7, 8, 9]. Note that these benefits can be realized without actually exchanging money. For example, an enterprise or military network can allocate resources and provide incentives by distributing abstract credits, which are redeemable for network services.

*This work supported in part by Lucent Technologies and by AT&T.

This paper examines usage-based pricing and traffic control mechanisms working in concert, focusing on connection-oriented networks using rate-based flow control, as is the case in many asynchronous transfer mode (ATM) networks. This paper will show that pricing mechanisms can motivate a network designer to modify, supplement, or even replace some traffic control mechanisms such as reactive congestion control and admission control. One particular issue we address is whether price should vary dynamically (or "responsively" [8]) to follow changes in network state (and not just time-of-day) as a number of researchers have proposed, just as traffic control algorithms might dynamically vary the bounds on the rate that packets are allowed to enter the network. Such a pricing mechanism may convince users not to transmit when the network is becoming congested. Dynamic state-dependent pricing will appear to users as prices that fluctuate randomly, which is not generally a desirable property. Indeed, it is pointless in applications where humans rather than computers directly determine when transmissions are made. It is hard to imagine a person watching long-distance prices fluctuate wildly on a meter, waiting for the perfect moment to make a telephone call. However, there are applications where dynamic state-dependent pricing makes sense. For example, a user wandering the World Wide Web (WWW) might instruct her browser to suspend operation, to retrieve text but not images, or to simply decrease the data rate, when prices are high. Similarly, a videoconference application may automatically allow video resolution to degrade or switch from color to black and white when prices are high [10]. If prices should vary with network state, how quickly? If you are watching a video changing back and forth between color and black-and-white, you would prefer price to change slowly, but if pricing is a form of congestion control, it would be better if prices could change rapidly. Other factors are also important. In particular, we will consider the impact of propagation delay across the network on the utility of dynamic state-dependent pricing.

Another issue to be addressed is whether the price for transmitting an individual packet should be independent of the other packets in the packet stream, or whether the entire stream should somehow be considered when setting the price. This is somewhat analogous to the debate about whether or not traffic control algorithms should view each packet as an independent entity, as is the case in a datagram protocol like the Internet's IP. We will argue that ATM pricing should be based in part on the stream's arrival process even in some cases where the packet stream is sent *best effort*, i.e. without specific guarantees about delay performance.

The next section will describe different approaches to pricing. This will allow us to describe the issues addressed in this paper in more detail. The model of network and user behavior that we use to study these issues will be presented in Section 3, and the results achieved with different pricing and congestion control mechanisms is in Section 4. Section 5 presents the resulting conclusions.

0-7803-4198-8/97/$10.00 © 1997 IEEE

2 Taxonomy and Issues

A wide variety of pricing mechanisms have been proposed [11]. We begin by proposing a taxonomy that describes ways that packet streams might be admitted to the network, and corresponding pricing mechanisms: *guaranteed, packet-oriented best effort*, and *stream-oriented best effort*. Each of these three approaches may be appropriate for some types of traffic.

With some applications like telephony, it is preferable for a packet stream to be blocked than for it to be admitted and then experience unacceptable performance. Such traffic requires an a priori guarantee that performance requirements will be met. The *guaranteed* services meet this need by requiring calls to go through an *admission control* process before transmissions can begin. The application must first state the stream's packet arrival process and its performance requirements. If the requirements of this new stream and all existing streams cannot be met, then the call must be blocked. If the stream is admitted, a *policing* mechanism insures that the data rate and burstiness of the actual packet arrival process is no greater than that of the stated arrival process.

Prices should depend in part on the amount of resources consumed by each packet stream, and this is a function of the stream's average data rate, burstiness, performance objectives, and the blocking probability that is tolerable [12, 13]. We have proposed a framework in which the price of each service depends on all of these factors, and devised a method of determining optimal prices and optimal capacity [13]. Price for a guaranteed service in this framework also depends on time-of-day, and therefore expected network load, but not on actual load or network state. Once a guaranteed stream is admitted, the network is not able to reclaim the allocated resources, or change the price, until the application decides to terminate the call. Consequently, network state at the instant the call is admitted is far less important than expected load over the duration of the call. However, this argument holds only when performance and price guarantees are made.

The alternative to a guaranteed service is a *best effort* service, in which no a priori performance guarantee is asked for or given. We divide best effort services into two categories: *packet-oriented* and *stream-oriented*. With the former, each packet is handled independently, so network traffic control mechanisms have no idea what the packet arrival process is for a given stream. The price for a packet also does not depend on the characteristics of other packets in the stream. One serious limitation of this approach is that there is no disincentive for traffic to be bursty, even though it reduces throughput. This alone is a reason to discourage use of the packet-oriented best effort approach by pricing it high.

The simplest pricing mechanism of this kind is a fixed cost per packet or per bit. However, demand fluctuates randomly, causing underutilization when few users are willing to pay this price, and congestion when many are. This can be solved with a *smart market* mechanism [14], in which the user specifies the value of each packet, and the current *spot price* for transmitting across a given link is set such that the available capacity is just enough for all packets whose value exceeds that price. (In an ATM network, this value may be implicit in the virtual channel identifier.) This is an effective way to price best-effort services in a single-link network in which the queueing delay of the best-effort packets that are transmitted is unimportant to users [8, 13, 14]. However, there are additional issues in a network of queues. This paper will focus on one such issue, with others to be left for future work: unless propagation delay is negligible, it is impossible to determine the state of every link in a network, calculate appropriate prices, and advertise these prices throughout the network, before the state information becomes outdated. A distributed approach to pricing is needed, and traffic sources will inevitably decide whether or not to transmit based on slightly outdated information. Note that this limitation applies similarly to reactive congestion control, where sources inject packets into the network based on outdated information on the presence or absence of congestion.

Finally, we consider stream-oriented best-effort. As with guaranteed streams, applications must first declare their packet arrival process to the network, which is the most significant difference between a packet-oriented and stream-oriented service. Policing mechanisms will later penalize the stream if this declaration is not accurate. Thus, the network can charge more for bursty streams. Applications may also declare performance objectives. The network can then attempt to meet (and not greatly exceed) those objectives as in [1, 2, 3, 4], thereby allowing the network to improve performance and carry more traffic, but there is no firm guarantee that the objectives will actually be met, and stream-oriented best-effort calls are never blocked. Since revealing this information improves network efficiency, the application is rewarded with lower prices. Thus, the cost of transmitting a given stream via stream-oriented best effort is less than transmitting the same stream via packet-oriented best effort.

In this paper, we evaluate the effectiveness of dynamic state-dependent pricing as a method of allocating resources to the most valuable streams, and as a method of reacting to congestion. In the process, we observe how these pricing mechanisms work in combination with and instead of traditional reactive congestion control mechanisms, i.e. mechanisms that treat all streams the same. We determine whether these dynamic pricing approaches are effective in the face of significant propagation delay between the traffic sources and a congested link. To simplify the problem, we assume there is at most one potentially congested link; this assumption will be relaxed in future work. We will also assume here that all sources are roughly the same distance from the congestion. The effectiveness of state-dependent pricing obviously depends on how much state information is known. We will also show how useful the information provided to the network with stream-oriented best effort traffic can be.

3 The Model

At a given time, there are N best-effort packet streams running through the potentially congested link. Routing is not affected by transient congestion, as is appropriate for ATM networks and some datagram networks. Each of the N sources will choose to transmit when and only when the value to the user of packets in that stream exceeds the current price. Otherwise, the service is deemed too expensive. All N streams have the same arrival process, and all packets in a given stream are equally valuable, but no two streams have the exact same value per packet to their users. Thus, it is always possible to set price such that i streams have value greater than current price for any $i \leq N$.

We do not to presume to know the actual distribution of value in typical networks, and it can differ from network to network. Consequently, we will consider several distributions for value. Without loss of generality, we number the streams in increasing order of value. In each case, the value V_i for stream $i : 1 \leq i \leq N$ is proportional to i^a, and we will consider scenarios where $a = 0$, .5, 1, and 2. The exact values are scaled such that the average value $\sum_{i=1}^{N} V_i/N$ across all classes is 1. ($a = 0$ really means a is

negligible, so no two streams have identical value.)

Each of the N streams alternates between on- and off-states. While on, they transmit at a constant rate, and while off, they do not transmit. There is a maximum rate at which a source can transmit, which would always be its transmission rate in the absence of congestion control. The durations of on-periods and off-periods are independent and exponentially distributed. These durations are not affected by pricing or congestion control mechanisms, although the amount of information that can be transmitted during an on-period is. This would be a reasonable approximation for a variety of applications. For example, web browsers and video applications may decrease resolution when price is high. This is equivalent to saying that aggregate traffic is a combination of a guaranteed stream of minimum data rate and a best-effort stream that is active when price is low enough. As another example, a background distributed computation may be suspended when the price of communications services becomes high.

The propagation delay between each data source and the potentially congested link is P in each direction. At this link is an intelligent agent capable of sending congestion control or pricing messages back to the sources. When sending these messages, the goal is to maximize the total value derived from the network (which economists call *social welfare.*) The value derived from a given packet stream is the product of the value per packet and the throughput of that stream. The total value derived from the network is the sum of the values derived from all streams. Note that total value does not depend on the actual revenue transferred from consumers to the network in the form of usage-based fees.

Value depends on throughput, and the problem of maximizing throughput is relatively simple unless the network is subject to the phenomenon of congestion, as most real networks are. When there is the potential for congestion, as load increases, so does inefficiency, so load eventually reaches a point at which throughput peaks and then declines. We will use the following model for congestion. When load is below a certain threshold, throughput through the link equals the arrival rate. Any time instantaneous load exceeds that threshold, instantaneous throughput decreases linearly from the maximum with slope $-m$. For example, consider a 150 Mb/s link that becomes congested when load hits .8. j sources are each transmitting at rate R. When $jR \leq .8 \cdot 150 = 120$ Mb/s, then throughput is jR. When $jR > 120$ Mb/s, throughput is $120 - m(jR - 120)$, evenly split among the j sources. This is obviously an approximation, but it is a reasonable representation of any congestion-prone network, and it has been shown to be appropriate in some important cases [15]. In [15], each packet consists of 10 cells, and the loss of one or more cells in a packet means that the entire packet must be retransmitted, as can occur in ATM networks.

When too many best-effort streams are in the on-state, the link is congested, and throughput is suboptimal. So is the total value derived from the network. When a stream is best effort, there is no guarantee that performance, data rate, or price, will not change. Thus, when congestion occurs, this agent at the congested link can use a congestion control mechanism to limit the transmission rate of all best-effort streams, as occurs with an ATM Available Bit Rate (ABR) service. Each source that is currently in the on-state will then transmit at the rate specified by this network agent. Another approach is to dynamically change prices. When this occurs, all best-effort streams for which the new price exceeds the value of the information will stop transmitting, and the others will continue to transmit at the maximum rate. It is also possible to use a combination of these mechanisms, where all streams whose value exceeds the current price will transmit at the rate set by the congestion control mechanism.

We consider three different congestion control mechanisms and three different pricing mechanisms, yielding a total of nine different approaches. The three congestion control mechanisms are signified by NC, SC, and FC. NC means that there is no congestion control, so sources always transmit at the maximum rate. SC means there is a slow-reacting congestion control mechanism, so the network may impose a maximum transmission rate on every source that is a function of the total number of streams currently passing through the link, and this maximum rate will change whenever a new call is initiated or an old one is terminated. However, the maximum rate does not change as streams move between the on- and off-states. This rate is selected to maximize the total value derived by the system. Of course, the SC approach would be difficult to implement without stream-oriented best-effort, because the network would not know the number N of streams with packet-oriented best effort. Instead, it would have to record the ever-changing packet arrival rates and do some kind of filtering to estimate the load. It would attempt to react quickly enough to notice when calls begin or end, but slowly enough not to react to temporary changes in packet arrival rate. This is far more complex and less accurate than what is possible with stream-oriented best effort, and it would lead to less stable prices and data rates.

Finally, FC means fast-reacting congestion control. In this case, the agent at the potentially congested link sends a message to all sources whenever the arrival rate to the congested link changes. This message indicates the maximum rate at which each source is allowed to transmit. Of course it takes one propagation delay for the message to reach the sources, and another delay P before it affects the traffic arrival rate at the congested link. In this period of $2 \cdot P$, some of the N streams might have gone from the on-state to the off-state, and vise versa. Thus, an FC mechanism at time t sets the maximum rate to maximize the expected total value derived from the system at time $t + 2P$, given the number of streams in the on-state at time t, the total number of streams N, and knowledge of the mean on- and off-periods. This knowledge is easily available with stream-oriented services, since it is declared, but it is more difficult to obtain with packet-oriented best effort, where it must be based on long-term statistics. FC is representative of current congestion control algorithms which do not discriminate among active streams.

The three pricing policies (NP, SP, FP) are analogous to the three congestion control policies (NC, SC, FC). With NP, there is no usage-based pricing, so all sources transmit when in the on-state. SP means price is a function of the number of streams N, and does not change as streams alternate between the on- and off-states. SP is far more practical with stream-oriented best effort, since the number of packet streams is known with stream-oriented, and is difficult to determine with packet-oriented. FP means that prices change as instantaneous arrival rate changes. FP is a bit more complex than the comparable congestion control approach - FC. With NP or SP, an FC congestion control mechanism can easily determine the number of streams in the on-state. However, with FP, the pricing mechanism knows that a given stream is currently in the on-state if and only if the value per packet of that stream is greater than the current price. Otherwise, the source would not be transmitting any way. At best, the network can know whether the stream was in the on-state at

the last time when price dipped sufficiently low for this source to transmit. As a result, our FP algorithm maintains a certain amount of historical information. In particular, at time t, it is assumed that the intelligent agent knows which sources were active at times $t - iP$ for any positive integer i. The optimal price also depends on which of the N streams are in the on-state, rather than just how many of them are. This makes our FP algorithm somewhat complicated; in reality, a less complicated and less effective version might be implemented.

4 Performance Results

We can see the value of dynamic state-dependent pricing by observing performance wth fast-reacting or slow-reacting pricing (FP or SP), as opposed to no usage-based pricing (NP). Also, if SP or SC look promising, it is another argument for stream-oriented best effort, which makes it easy for the network to know how many best-effort streams are passing through a given link and their arrival processes. For the sake of comparison, FC-NP is most representative of current networks (e.g. ATM ABR).

The results shown in this section with fast-reacting pricing (FP) were achieved via simulation, and the 95% confidence interval is, at worst, within 5% of the values shown. For the other approaches, results were achieved analytically, so they are exact.

We first consider the case where all N streams have roughly the same value (i.e., $V_i \propto i^0$). Figure 1 shows total value, which in this case equals system throughput, as a function of N. (The channel is 150 Mb/s, and throughput is maximized when arrival rate is 120 Mb/s. When arrival rate exceeds 120 Mb/s, throughput is degraded at slope $m = 1$. The average duration of on- and off-periods is 100 ms and 200 ms, respectively, as might be reasonable for best-effort streams that enhance resolution for variable-bit-rate video. The maximum rate for a single source is 10 Mb/s.) Figure 1 shows that if there is no congestion control or dynamic pricing (NC-NP), then the congestion phenomenon is strong; value increases for small N but decreases for large N, eventually approaching 0. However, if there is dynamic pricing or dynamic congestion control or both, even if it is slow-reacting, value is not degraded as N grows large. Thus, dynamic state-dependent pricing could conceivably replace congestion control. The figure also shows that fast-reacting congestion control (FC) is somewhat more effective than dynamic pricing, since FC-NP always outperforms SC-FP. This is always the case when the objective is to maximize throughput (i.e. $V_i \propto i^0$). This can be explained as follows. A congestion control approach instructs all N streams to transmit at a given rate if they are in the on-state. There is uncertainty because the exact number of sources that will be in the on-state $2P$ from now is not known exactly, and if it is either too high or too low, throughput is degraded. A pricing approach would instead encourage the i "most valuable" streams to transmit at full rate if they are in the on-state. Since $N > i$, there is less variance in arrival rate with congestion control than with pricing, so expected throughput is slightly greater with congestion control.

We now consider a case where the value V_i of stream i is proportional to i (i.e., $V_i \propto i^1$), so there is more reason to use pricing. Figure 2 shows the throughput achieved by each of 45 streams. The parameters are the same as in Figure 1, but with $N = 45$. (The parameters in this curve will serve as our default assumptions in the rest of this section unless otherwise specified.) When there is no pricing (NP), all streams have the same throughput, and that throughput is best with fast-reacting congestion control (FC) and worst with no congestion control (NC). With slow-reacting pricing (SP), throughput is close to the maximum for those willing to pay the price to transmit and 0 for the rest. However, with fast-reacting pricing (FP), many streams get throughputs between 0 and 10, because they are able to transmit when and only when many of the more valuable streams are in the off-state. We also show an optimal curve, which is achievable only if the agent at the congested link can predict the future perfectly, or equivalently, if the propagation delay is 0. Its shape is similar to the FP curve.

Figure 3 shows the total value derived by the network as a function of the number of streams N when streams may not have the same value. Results are shown for $V_i \propto i^0, i^{.5}, i^1$, and i^2. In Figure 3-a, the propagation delay $P = 1$ ms, and in Figure 3-b, propagation delay $P = 10$ ms. From both figures, it is clear that dynamic pricing (FP or SP) is far more effective than NP when streams are not all equally valuable (i.e. $a > 0$), as one would expect, and the more the value of the various streams can differ, the more useful dynamic pricing is. We also see in Figure 3-a that FC-FP outperforms FC-SP in each scenario where pricing helps. In Figure 3-b, while FC-FP is still better than FC-SP, the difference is much smaller. Moreover, the implementation of a pricing mechanism that requires prices to be constantly recalculated would be much more complicated than a pricing mechanism in which prices are only calculated when a new stream begins or an old one terminates. All else being equal, users would also prefer a system in which price changed more slowly. If performance is comparable, as it appears in Figure 3-b, there may be an opportunity to use the simpler scheme.

Since propagation delay is clearly an important factor in the relative effectiveness of these schemes, Figure 4 shows total value as a function of propagation delay with our default parameters. FC-FP is a somewhat useful approach (relative to FC-SP) when propagation delay is 5 ms or less, but with a propagation delay greater than 10 ms, it only slightly outperforms FC-SP. A metropolitan-area network can have propagation delays of a few ms, but this is not reasonable for a wide-area network. Within the continental U.S. alone, propagation delays can exceed 30 ms, so there is no point in using a fast-reacting pricing mechanism if FC-SP is possible. Of course, this conclusion depends on some of our other parameters. An obvious assumption to examine is that the average on- and off-periods are 100 and 200 ms, respectively, yielding an average time between the beginning of successive on-periods of 300 ms. It is the ratio of this number to propagation delay that really matters. For FC-FP to be beneficial at 30 ms instead of 5 ms, one need only change the mean time between successive on-periods from 300 ms to 1800 ms. In this case, an FC-FP pricing mechanism would be calculating prices on the order of seconds, which is not excessive. In fact, with on and off periods this long, it would not be a problem to initiate the transmission of a new stream-oriented best effort stream for every on-period, and terminate it every off-period, which would cause FC-FP and FC-SP to yield identical performance.

Another parameter worth exploring is maximum rate per stream, which we had previously assumed to be 10 Mb/s. (This is equivalent to changing the maximum link throughput, since it is the ratio of these two numbers that matters.) Figure 5 shows total value as a function of the maximum rate with a propagation delay of 10 ms and $N = 45$ streams. The effects of increasing maximum rate are similar to the effects of increasing the number of streams as shown in Figure 3. It is more effective to have dynamic

usage-based pricing (SP and FP) than not (NP), except where all streams are of roughly equal value ($a = 0$). Fast-reacting pricing slightly outperforms slow-reacting pricing. Figure 6 shows the case where the maximum rate of each stream is varied, but the number of sources is also varied so that the average load on the link remains fixed. Here we see that if maximum rate becomes extremely large, so the number of streams becomes quite small (e.g. 4 streams at 120 Mb/s), then fast-reacting pricing becomes much more effective relative to slow-reacting pricing. To see the reason, let j be the number of streams that are in the on-state and transmitting at maximum rate when the link's throughput is maximized. Price is set such that the k most valuable streams will choose to transmit if they are in the on-state, $k \geq j$. If j and k are large, the number of sources in the on-state at any given time will be fairly close to j. If k and j are small, the coefficient of variation of the number of streams in the on-state is greater. Thus, fast-reacting pricing, which adjusts to such changes, becomes more effective. However, in the near term, it seems unlikely that there will be much use of such high-data-rate applications. In the long term, high-data-rate streams will become more common, but link capacities will also increase, so it still may not be the case that a small number of streams can consume all of a link's capacity. It therefore remains to be seen whether this effect will limit the effectiveness of slow-reacting pricing.

5 Conclusion

We have evaluated the network's ability to maximize total system value (social welfare) despite potential congestion, through use of a variety of mechanisms. These include slow- and fast-reacting congestion control, and slow- and fast-reacting pricing. We have seen that, in many ways, dynamic pricing is an alternative to congestion control, and vise versa, since both allow the network to avoid and alleviate congestion. Dynamic state-dependent pricing has an important additional advantage; it allocates resources to the more valuable streams, so pricing is more effective when value varies significantly from stream to stream. Congestion control was found to be somewhat more effective than pricing if all streams are of comparable value, so there is a throughput penalty for pricing. (In the vocabulary of [8], maximizing economic efficiency means degrading network efficiency.)

In most cases, when fast-reacting congestion control is used and propagation delay is significant, there is little difference between fast-reacting pricing and slow-reacting pricing. This would certainly be the case in a wide-area network. The one notable exception where the fast-reacting approach does much better is if there are a small number of streams capable of transmitting at very high data rates, consuming much of the congested link's capacity. Slow-reacting pricing is also more attractive to users, it is easier to implement, and less communications capacity is spent on the exchange of pricing information. Consequently, in many cases, slow-reacting pricing will prove preferable. The attraction of fast-reacting pricing in previous work comes in part from the fact that the scenarios considered have involved negligible propagation delay. Of course, other issues must still be addressed to determine the practicality of dynamic state-dependent pricing, whether it is slow-reacting or fast-reacting.

Slow-reacting dynamic pricing is only possible if the network can determine how many streams are passing through a given congested link. This is difficult if the network is not explicitly informed when best-effort streams begin and end. This is one piece of evidence supporting our assertion that networks should offer *stream-oriented best-effort* services and corresponding pricing. With this approach, a customer would be charged based on both the duration of a stream and the number of packets sent, and both these prices would be affected by the declared arrival process. Although these services offer no performance guarantees, they provide price incentives for users to indicate the arrival processes of their streams and the performance objectives before transmissions begin. Slow-reacting pricing is just one of the schemes that becomes practical when the network learns the number of streams on any link, average data rates, and burstiness. Sophisticated traffic control approaches like [1] can also be used more extensively, thereby allowing the network to meet given performance objectives while carrying more traffic. Stream-oriented best effort service also allows price disincentives for bursty traffic.

This paper is the first step in investigating the utility of dynamic pricing. Simply by considering propagation delay, we have shown that fast-reacting pricing is of limited value. Future work must relax more assumptions, e.g. examine networks with multiple bottlenecks and more diverse streams.

References

[1] M. A. Lynn and J. M. Peha, "Priority Token Bank Scheduling in a Network of Queues," *Proc. IEEE International Conference on Communications ICC-97*, June 1997, pp. 1387-91.

[2] J. M. Peha and F. A. Tobagi, "Cost-Based Scheduling and Dropping Algorithms to Support Integrated Services, " *IEEE Trans. Commun.*, vol. 44, no. 2, Feb. 1996, pp. 192-202.

[3] J. Hyman, A. A. Lazar, and G. Pacifici, "Real-Time Scheduling with Quality of Service Constraints," *IEEE J. Select. Areas Commun.*, vol. 9, No 7, Sept. 1991, pp. 1052-63.

[4] D. Lee and B. Sengupta, "Queueing Analysis of a Threshold Based Priority Scheme for ATM Networks," *IEEE/ACM Trans. Networking*, vol. 1, no. 6, pp. 709-17, Dec. 1993.

[5] J. M. Peha, "Heterogeneous Criteria Scheduling: Minimizing Weighted Number of Tardy Jobs and Weighted Completion Time," *Computers and Operations Research*, vol. 22, no. 10, Dec. 1995, pp. 1089-1100.

[6] R. Cocchi, S. Shenker, D. Estrin, L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Example," *IEEE/ACM Trans. Netw.*, vol. 1, no. 6, pp. 614-27, Dec. 1993.

[7] J. F. MacKie-Mason and H. Varian, "Pricing Congestible Network Resources," *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, Sept. 1995, pp. 1141-9.

[8] J. F. MacKie-Mason, L. Murphy, and J. Murphy, "The Role of Responsive Pricing in the Internet," *Internet Economics*, J. Bailey and L. McKnight eds., MIT Press, 1997, pp. 279-303.

[9] A. Gupta, D. O. Stahl, and A. B. Whinston, "A Priority Pricing Approach to Manage Multi-Service Class Networks in Real-Time," *Internet Economics*, J. Bailey and L. McKnight eds., MIT Press, 1997, pp. 323-52.

[10] K. Danielsen and M. Weiss, "User Control Nodes and IP Allocation," *Internet Economics*, J. Bailey and L. McKnight eds., MIT Press, 1997, pp. 305-21.

[11] S. Jordan and H. Jiang, "Connection Establishment in High-Speed Networks," *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, Sept. 1995, pp. 1150-61.

[12] J. M. Peha and S. Tewari, "The Results of Competition Between Integrated-Services Telecommunications Carriers," accepted to appear in *Information Economics and Policy*.

[13] Q. Wang, J. M. Peha, and M. Sirbu, "Optimal Pricing for Integrated-Services Networks with Guaranteed Quality of Service," *Internet Economics*, J. Bailey and L. McKnight eds., MIT Press, 1997, pp. 353-76.

[14] H. Varian and J. K. MacKie-Mason, "Pricing the Internet," *Proc. Public Access to the Internet*, B. Kahin and J. Keller, eds., Englewood Cliffs, NJ: Prentice Hall, 1995.

[15] J. M. Peha, "Retransmission Mechanisms and Self-Similar Traffic Models," *Proc. IEEE/ACM/SCS Communication Networks and Distributed Systems Modeling and Simulation Conference*, Jan. 1997, pp. 47-52.

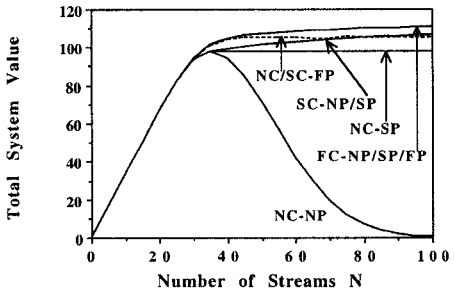


Figure 1: Total value (system throughput) vs. number of streams N. $V_i \propto i^0$. Propagation delay $P = 10$ ms. Mean on-period = 100 ms. Mean off-period = 200 ms. Max rate per source = 10 Mb/s. Max link throughput = 120 Mb/s. $m = 1$.

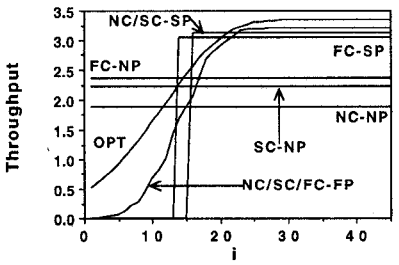


Figure 2: Throughput of each stream i. $V_i \propto i^1$. Propagation delay $P = 10$ ms. $N = 45$ streams. Mean on-period = 100 ms. Mean off-period = 200 ms. Max rate per source = 10 Mb/s. Max link throughput = 120 Mb/s. $m = 1$.

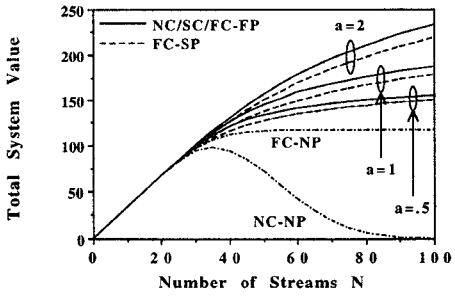
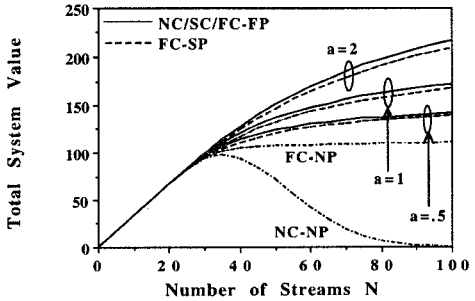


Figure 3: Total value vs. number of streams N. $V_i \propto i^a$. Mean on-period = 100 ms. Mean off-period = 200 ms. Max rate per source = 10 Mb/s. Max link throughput = 120 Mb/s. $m = 1$.
(a): Propagation delay $P = 1$ ms. (above)
(b): Propagation delay $P = 10$ ms. (below)



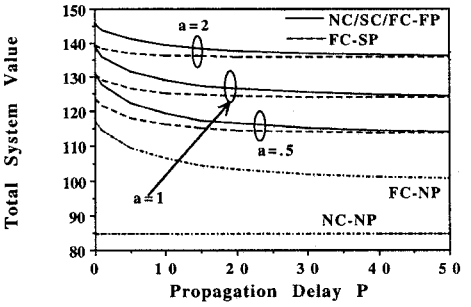


Figure 4: Total value vs. propagation delay P. $V_i \propto i^a$. $N = 45$ streams. Mean on-period = 100 ms. Mean off-period = 200 ms. Max rate per source = 10 Mb/s. Max link throughput = 120 Mb/s. $m = 1$.

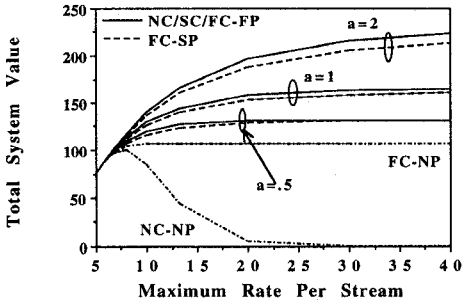


Figure 5: Total value vs. max rate per source. $V_i \propto i^a$. Propagation delay $P = 10$ ms. $N = 45$ streams. Mean on-period = 100 ms. Mean off-period = 200 ms. Max link throughput = 120 Mb/s. $m = 1$.

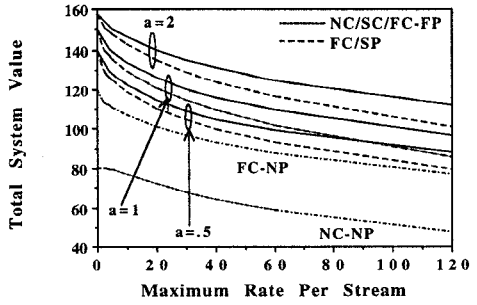


Figure 6: Total value vs. max rate per source. (Number of sources) · (Max rate per source) = 4800 Mb/s. $V_i \propto i^a$. Propagation delay $P = 10$ ms. Mean on-period = 100 ms. Mean off-period = 200 ms. Max link throughput = 120 Mb/s. $m = 1$.