Optimal Pricing for Integrated-Services Networks with Guaranteed Quality of Service[&]

by

Qiong Wang[*]

Jon M. Peha[^]

Marvin A. Sirbu[#]

Carnegie Mellon University

Chapter in *Internet Economics*, edited by Joseph Bailey and Lee McKnight, MIT Press, 1996

Also available at http://www.ece.cmu.edu/afs/ece/usr/peha/peha.html

---

[&] The research reported in this paper was support in part by the National Science Foundation under grants NCR-9210626 and 9307548-NCR. Views and conclusion contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. Government.

[*] Doctoral Candidate. Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213. Tel: +1 412 268 5617. Email: qw22@andrew.cmu.edu.

[^] Assistant Professor of Electrical and Computer Engineering, Engineering and Public Policy. Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. Tel: +1 412 268 7126. E-mail: peha@ece.cmu.edu.

[#] Professor of Engineering and Public Policy, Industrial Administration and Electrical and Computer Engineering. Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213. Tel: +1 412 268 3436. E-mail: sirbu@cmu.edu.

Abstract

Integrating multiple services into a single network is becoming increasingly common in today's telecommunications industry. Driven by the emergence of new applications, many of these services will be offered with guaranteed quality of service. While there are extensive studies of the engineering problems of designing integrated-services networks with guaranteed quality of service, related economic problems, such as how to price services offered by this type of network, are not well understood.

In this chapter, we analyze the problem of pricing and capacity investment for an integrated-services network with guaranteed quality of service. Based on an optimal control model formulation, we develop a 3-stage procedure to determine the optimal amount of capacity and the optimal price schedule. We show that pricing a network service is similar to pricing a tangible product, except that the marginal cost of producing the product is replaced by the opportunity cost of providing the service, which includes both the opportunity cost of reserving and the opportunity cost of using network capacity. Our findings lays out a framework for making investment and pricing decisions, as well as for the analysis of related economic tradeoffs.

The analysis in this chapter assumes an integrated-service network with fixed-length data units such as Asynchronous Transfer Mode (ATM) network. The same approach can be used to analyze variable packet length IP networks offering guaranteed quality of service through the use of protocols such as Resources reSerVation Protocol (RSVP).

Keywords. integrated-services, opportunity cost, optimal price, ATM

§1. Introduction

The economics of providing multiple types of services through a single network is a question of growing significance to network operators and users. As a result of the rapid development of packet-switching technology, it is becoming increasingly efficient to provide different telecommunication services through one integrated-services network instead of multiple single-service networks, such as telephone networks for voice communications, cable networks for broadcasting video, and the Internet for data transfer. In a packet-switched integrated services network, any piece of information, regardless of whether it is voice, image, or text, is organized as a stream of packets and transmitted over the network. By controlling the packet transmission rate and packet delay distribution of each packet stream, the network can use a single packet transmission technology to provide a variety of transmission services, such as telephony, video, and file transfer.

While integrating multiple services into a single network generates economies of scope, heterogeneous services complicate pricing decisions. For example, users watching High-Definition Television (HDTV) through the network require up to tens of megabits per second (Mbps) transmission capacity while users who make phone calls only send/receive tens of kilobits per second (kpbs); telnet users require mean cell transmission delay to be kept below a few tens of milliseconds but e-mail senders will tolerate longer delay; web browsing generates a very bursty cell stream while constant-bit-rate file transfer results in a very smooth cell stream; to carry a voice conversation with acceptable quality, under certain encoding schemes, packet loss rate, *i.e*. the percentage of packet that are allowed to miss a maximum delay bound (usually 30-50 ms for voice conversation), should not exceed 5%, while to carry video service, packet loss rate should be kept much lower (Peha,1991).

Asynchronous Transfer Mode (ATM) technology emerges as an appropriate basis for integrated-services networks. ATM networks have the capability to meet the strict performance requirements of applications like voice and video, and the flexibility to make

efficient use of network capacity for applications like e-mail and web browsing. The use of fixed-length packets (cells) also facilitate the implementation of high-speed switches. As a result, telephone and cable TV networks will adopt cell switching technology, as they expand the range of services that they offer. The Internet has already begun to offer new services like telephony, but without the guarantee of adequate performance that telephone customers have come to expect. Eventually, the Internet will also employ protocols that differentiate packets based on the type of traffic that they carry, and guarantee adequate quality of service appropriate for each service. This could be done by adopting ATM technology, or by adding the capability to guarantee performance through use of protocols like the Resource reSerVation Protocol [RSVP] (Zhang et al, 1993). This paper will focus on ATM-based integrated-services networks, as the technology is available today, but trivial extensions would enable the same approach to be applied to any integrated-services networks which offers quality of service guarantees.

Since there are great differences among the services offered by ATM networks, one might ask whether the prices of these service should also differ, and if so, how? There is some literature on how to price a network that offers heterogeneous services. Cocchi *et al* (Cocchi *et al*, 1993) study the pricing of a single network which provides multiple services at different performance levels. They give a very impressive example which shows that in comparison with flat-rate pricing for all services, a price schedule based on performance objectives can enable every customer to derive a higher surplus from the service, and at the same time, generate greater profits for the service provider. Dewan, Whang and Mendelson *et al* (Dewan and Mendelson, 1990; Whang and Mendelson,1990) developed a single queuing model in which the network is formulated as a server (or servers) with limited capacity, and consumers demand the *same* service from the server but vary in both willingness to pay for the service and tolerance for delay. Based on that model, they discussed the optimal pricing policy and capacity investment strategy. MacKie-Mason and Varian (Mackie-Mason and Varian, 1994) suggest a spot-price model for Internet pricing.

In their model, every Internet packet is marked with the consumer's willingness to pay for sending it. The network always transmits packets with higher willingness to pay and drops packets with lower willingness to pay. The network charges a spot price that equals the lowest willingness to pay among all packets sent during each short period. The major benefit of this approach is it provides consumers with an incentive to reveal their true willingness to pay, and based on that information, the network can resolve capacity contention in transmitting packets in a way that maximizes social welfare. In the work by Gupta *et al* (Gupta, Stahl, and Whinston, 1996), priority-based pricing and congestion-based pricing are integrated. In their pricing model, services are divided into different priority classes. Packets from a high-priority class always have precedence over packets from a low-priority class. The price for each packet depends not only on the packet's priority level, but also on the current network load.

In the optimal pricing models mentioned above, the fact that different applications may have different performance objectives was usually *not* considered. For example, Dewan, Whang and Mendelson's work (Dewan and Mendelson; 1990l; Whang and Mendelson, 1990) assumes that the consumer's willingness to pay depends only on expected mean delay, and Mackie-Mason (Mackie-Mason,1996) assumes that consumers do not care about delay—only whether or not their packets are eventually transmitted. There is no way, for example, to accommodate a service that would impose a maximum delay limit. These formulations also do not consider the case of heterogeneous data rate and burstiness. Consequently, pricing policies developed in these studies can not be applied in ATM integrated-services networks in which services differ from each other in terms of performance objectives and traffic pattern (data rate and burstiness). Some of these factors are discussed in the paper by Cocchi *et al* (Cocchi *et al*, 1993), however, they do not discuss procedures for designing an optimal pricing scheme. Gupta *et al* (Gupta, Stahl, and Whinston, 1996) consider different services which are divided into different

priority classes, however, none of these services can be guaranteed a given performance objective under their pricing scheme.

In this paper, we examine the optimal pricing problem for ATM integrated-services networks. In our approach, the optimal price for each service is determined from the demand elasticity for the service, as well as the opportunity cost of providing the service. The opportunity cost is determined by the required performance objectives and traffic pattern of each service. Since demand for network services usually changes with time of day, we will develop a time-varying price schedule (i.e. price as a function of time of day) instead of giving a single price for each service.

The rest of our paper is organized as follows: in section 2, we present service models for different services offered by an ATM integrated-services network. In section 3, we formulate an optimal pricing model and discuss how to solve it using a 3-stage procedure. We discuss the procedure in detail in section 4. Conclusions and future work are discussed in Section 5.

§2. The Network Service Model

Network capacity is often sub-additive, leading to conditions of natural monopoly for an integrated-services network operator. In the model which follows, we assume a single profit-maximizing monopolist is operating the network. In this chapter, we consider only a point-to-point single link network. This frees us from network routing details and allows us to focus our attention on the economic principles for designing pricing policy. The capacity of that link is denoted as C_T, whose unit is the maximum number of cells that can be transmitted over the link per unit of time.

The network is used for providing multiple services. Quality of service is measured by the distribution of cell delay time, where lost cells are considered as being delayed infinitely. A service will be labeled as a "guaranteed" service if during each session, the

network makes a commitment to meet some pre-specified delay objectives. These guarantees are typically expressed in stochastic, not absolute terms, e.g. no more than 5% of the cells will be delayed for more than 30 milliseconds; or the average delay will be less than 200 milliseconds. If no such guarantee is made, the service is considered as best-effort service. Telephone calls, High Definition Television (HDTV), and interactive games typically require some type of guaranteed service, while e-mail is usually specified as a best-effort service.

In our pricing model, the network service provider attempts to maximize profit which is the sum of profits from guaranteed services and best-effort service. In establishing a tariff for network service, one might charge for access independent of any usage; capacity reservation for guaranteed services, and actual usage. In this chapter, we assume dedicated access is priced at average cost, and the cost of all shared network facilities is recovered through a combination of reservation and usage prices. This assumption allows us to consider reservation and usage prices independent of access prices.

§2.1 Service Model for Guaranteed Services

Guaranteed services differ from each other significantly in terms of performance objectives, traffic pattern (data rate and burstiness), and call duration distribution. For example, HDTV service has a much stricter performance objective and 500-times higher mean data rate than telephone service. An HDTV session can take hours to complete, while telephone calls usually last only minutes. The transmission rate of the former (if the data stream is compressed) is also much burstier than that of the latter, which may not be compressed.

In this chapter, we assume the network offers N guaranteed services. Within the same service category i (i=1,N), calls require the same performance objective, exhibit the same inter-cell arrival statistics, and have call duration drawn from the same distribution.

We assume the price for a call using guaranteed service is determined by service type i, call starting time, and service duration. For a call of service i which begins at time t, $p_i(t)$ is the price which will be charged for each unit of time that the call lasts. A consumer will be charged a price equal to $p_i(t)$ times the call duration if the call starts at time t. We shall also assume that for calls of a given service, call duration is independent of price.

We define $\lambda_i[p_i(t),t]$ as the arrival rate of calls for service i given that the price of a call which starts at t will be $p_i(t)$ throughout the call[1]. We also assume that at any given price, $p_i(t)$, and any given time t, call arrivals are Poisson, i.e. the number of calls arriving within any period is independent of the number of calls which arrived within previous periods. Note that we have also assumed no cross-elasticity of demand between different services, which may not be realistic. We leave that enhancement for future paper.

To meet guaranteed performance objectives, the network can only carry limited numbers of calls simultaneously. These numbers are determined by performance objectives and traffic patterns of each service. To avoid accepting more calls than it can handle, ATM integrated-services networks enforce an admission policy by which the

---

[1] The consumer thus expected to pay $\frac{p_i(t)}{r_i}$ if call length has a mean value of $\frac{1}{r_i}$. It is more typical for a provider to define a price schedule $R_i(t)$ where a call is charged $R_i(t)$ at each instant it is in progress. Our formulation of $p_i(t)$is related to $R_i(t)$ by $\frac{p_i(t)}{r_i} = \int_t^{+\infty} R_i(\tau) e^{-r_i(\tau - t)} d\tau$ when call length is exponentially distributed.

network monitors the current network load and decides whether an incoming call should be admitted or rejected (Peha, 1993). This process is shown in Figure 1. For the purpose of this chapter, we assume calls are not queued if they can not be admitted immediately.

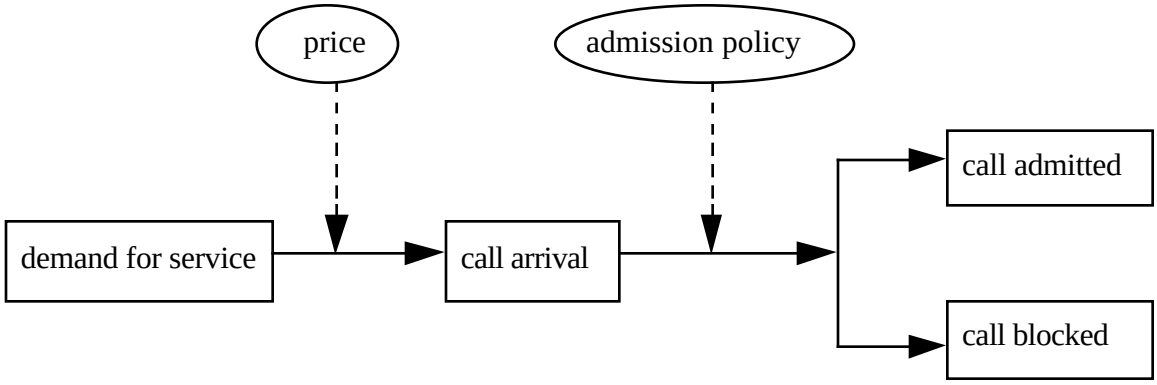

Figure 1

Call Admission Process for Guaranteed Services

For each service i (i=1,N), we assume the call duration is exponentially distributed with departure rate r_i. Define $q_i(t)$ as the number of calls underway of service i at time t, and $\overline{q_i}(t)$ as the expected value of $q_i(t)$. Under the assumptions we made about call arrival and departure processes, the rate of change of $\overline{q_i}(t)$ should follow:

$$\frac{d\overline{q_i}}{dt} = (1-\beta_i)\lambda_i(p_i,t) - r_i\overline{q_i}(t)\overline{q} \qquad i = 1, N \tag{2-1}$$

where β_i is the blocking probability.

Since both call arrival and departure are stochastic, unless the network has an infinite amount of capacity, there will always be a possibility of blocking calls. A high blocking probability gives consumers an unpleasant experience with the network and reduces the demand eventually, but a lower blocking probability also means more capacity will lay idle most of the time. From a network operator's perspective, the blocking probability should be kept at a desired level at which any marginal revenue increase from increasing demand by reducing blocking probability can no longer offset the marginal loss

from letting more capacity lay idle. Values of desired blocking probability are usually determined during the process of making long-term capacity investment decision. In the following, we will show how keeping blocking probability at the desired level will affect short-term pricing decisions:

Suppose the network only offers one service; then the blocking probability at each time can be determined by:

$$\beta = \frac{\frac{\rho^H}{H!}}{\sum_{i=0}^{H} \frac{\rho^i}{i!}} \qquad (2\text{-}2)$$

where β is the blocking probability, H is the maximum number of calls that can be carried by the network, and ρ is the product of call arrival rate and expected call duration.

Let $\tilde{\beta}$ be the blocking probability that the network operator desires to maintain. From (2-2), H can be uniquely determined by the desired blocking probability $\tilde{\beta}$ and the network load ρ, i.e. $H=d(\rho, \tilde{\beta})$. In other words, to keep blocking probability at a desired level under given load ρ, the network should be designed to carry H calls. This requirement can be translated into a demand for network capacity: define $\theta(H)$ as the amount of capacity needed to carry H calls, and $\alpha(\rho, \tilde{\beta})=\theta[d(\rho, \tilde{\beta})]$. $\alpha(\rho, \tilde{\beta})$ can be interpreted as the amount of capacity needed to keep blocking probability at $\tilde{\beta}$ when the network load is ρ.

Since at each time, the network load is related to the expected number of calls in progress by $\rho = \frac{\overline{q}}{1-\tilde{\beta}}$, we can also express the amount of capacity needed as a function of expected number of calls in progress as $A(\overline{q}, \tilde{\beta}) = \alpha(\frac{\overline{q}}{1-\tilde{\beta}}, \tilde{\beta})$. $A(\overline{q}, \tilde{\beta})$ increases with $\overline{q}$. $A(\overline{q}, \tilde{\beta})$ is defined as the amount of capacity required to carry an average of $\overline{q}$ calls with blocking probability $\tilde{\beta}$. If capacity required exceeds total capacity C_T, the network

either has to admit more calls than it can handle, thus failing to meet some quality of service guarantee, or exceed the desired blocking probability.

At each time *t*, $\overline{q}(t)$, the expected number of calls in progress is a function of previous and current prices. Therefore, in the short term, prices should be set such that the reserved capacity can never go above total capacity, i.e.:

$$A[\overline{q}(t), \tilde{\beta}] \le C_T \qquad \text{at all } t \qquad (2\text{-}3)$$

(2-3) defines the "*admissible region constraint*" (see Hyman et al, 1993; Tewari and Peha, 1995), which specifies the maximum number of calls that can be carried under a given amount of network capacity and a given blocking probability.

The definition of the admissible region constraint can be extended to a multiple services scenario, in which the reserved capacity is a function of the expected numbers of calls in progress for all services, which is shown below:

$$A[\overline{q}_1(t), \ldots\ldots, \overline{q}_N; \tilde{\beta}_1(t), \ldots, \tilde{\beta}_N(t)] \le C_T \qquad (2\text{-}4)$$

§2.2 Service Model for Best-effort Service:

Without a performance guarantee, cells of best-effort service will be put in a buffer and transmitted only when there is remaining capacity after the needs of guaranteed services have been met. If there is not enough buffer space for all incoming cells, some of them will be dropped.

In our model, users of best-effort service are charged on a per-cell basis. We assume all cells of best-effort service share a buffer of size B_s. The willingness to pay for sending each cell is revealed to the network. At each time *t*, the network sets a cut-off price, $p_b(t)$, which is a function of both current buffer occupancy and predicted willingness to pay values of future incoming cells. A cell will be accepted if and only if the willingness

to pay for that cell is higher than $p_b(t)$, and $p_b(t)$ will also be the price charged for sending that cell. Accepted cells will be admitted into the buffer as long as the buffer is not full. Once admitted into the buffer, cells will be eventually transmitted according to a sequence dictated by some scheduling algorithm, such as first-come-first-serve, or cost-based-scheduling (Peha, 1996).

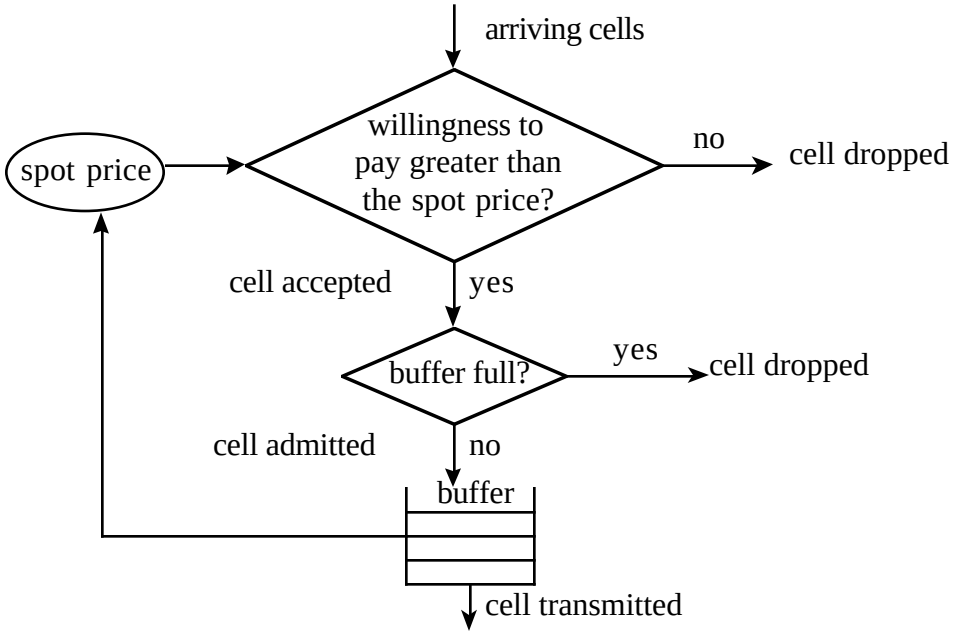

Figure 2
Service Model for Best-effort Service

If we assume that at time t, the arrival process of cells of best-effort service is Poisson with expected value $\lambda_b(0,t)$; the acceptance of cells is also Poisson with expected value $\lambda_b[p_b(t),t]$. Define $s_b(t)$ as the instantaneous transmission rate of best-effort service at that time, then:

$$s_b(t) \le C_T - s[q_1(t), q_2(t), \ldots, q_N(t)] \qquad (2\text{-}5)$$

where $s[q_1(t), q_2(t), \ldots, q_N(t)]$ is the instantaneous transmission rate of all guaranteed services, which is a function of numbers of calls in progress. Equation (2-5) implies the instantaneous transmission rate of best-effort service can not exceed the total bandwidth left after transmitting all guaranteed services.

If one accepts the assumptions that: 1) accepted cells constitute a Poisson random process; 2) the instantaneous transmission rate depends on the bandwidth left by guaranteed services, which is also random; and 3) the buffer size is limited, there is a possibility that even accepted cells (i.e. cells with willingness to pay higher than the cut-off price) can be dropped because the buffer can become temporarily full. Define $\upsilon(t,\Delta t)$ as the number of cells actually admitted into the buffer during the interval $[t,t+\Delta t)$; then the instantaneous admission rate can be defined as:

$$\omega_b(t) = \lim_{\Delta t \to 0} \frac{\upsilon_b(t,\Delta t)}{\Delta t} \tag{2-6}$$

$\omega_b(t)$ is a random variable and we assume its expected value is $\varpi_b(t)$, then

$$\varpi_b(t) \le \lambda_i[p_b(t),t] \tag{2-7}$$

Define $q_{b}(t)$ as the number of cells in the buffer at time t, then:

$$\frac{dq_b(t)}{dt} = \varpi_b(t) - s_b(t) \tag{2-8}$$

and

$$s_b(t) \le q_b(t) \le B_s \tag{2-9}$$

§3. The Optimal Pricing Policy

In this section, we will discuss the profit-maximizing pricing policy for network operators. We formulate an optimal control model to derive the pricing policy, and discuss how to solve this model through a 3-stage procedure.

§3.1 The Optimal Pricing Model

Assume a network operator wants to maximize total profit over a period composed of multiple identical business cycles (such as days). The cycle length is T. Her rational behavior would be to choose a price schedule for each type of guaranteed service $p_i(t)$, and

best-effort service, $p_b(t)$, and the amount of bandwidth C_T to maximize the following objective:

$$\int_0^T \{\sum_{i=1}^{N}(1-\tilde{\beta}_i)\frac{\lambda_i[p_i(t),t]}{r_i}p_i(t)+\varpi_b(t)p_b\}dt - K(C_T) \quad (3\text{-}1)$$

under constraints:

$$\frac{d\overline{q}_i}{dt} = (1-\tilde{\beta}_i)\lambda_i(p_i,t) - r_i\overline{q}_i,\ \overline{q}_i \geq 0 \quad i=1,N \quad (3\text{-}2)$$

$$A[\overline{q}_1(t),.....,\overline{q}_N;\tilde{\beta}_1(t),....,\tilde{\beta}_N(t)] \leq C_T \quad (3\text{-}3)$$

$$\frac{dq_b(t)}{dt} = \omega_b(t) - s_b(t) \quad (3\text{-}4)$$

$$\text{when } q_b(t) = B_s \qquad \omega_b(t) \leq s_b(t) \quad (3\text{-}5)$$

$$0 \leq q_b(t) \leq B_s \quad (3\text{-}6)$$

$$s_b(t) \leq C_T - s[q_1(t),q_2(t),...,q_N(t)] \quad (3\text{-}7)$$

$$\text{when } q_b(t) = 0 \qquad \omega_b(t) \geq s_b(t) \quad (3\text{-}8)$$

$$q_i(0)=q_{i0},\ i=1,N \quad (3\text{-}9)$$

Interpretations of these constraints are the same as discussed in section 2, and definitions of variables can found in both section 2 and in the following list:

Variables of guaranteed services:

N	number of different services;
$p_i(t)$	unit price for service i, as a function of call starting time t;
$\lambda_i(p_i,t)$	call arrival rate of service i at time t, when price is p_i;
r_i	call departure rate of service i;
$q_i(t)$	number of calls of service i in progress at time t;
$\overline{q}_i(t)$	expected value of $q_i(t)$;
$s[q_1(t), q_2(t), ..., q_N(t)]$	total data rate of all guaranteed services at time t;
$\overline{s}[\overline{q}_1(t),\overline{q}_2(t),...,\overline{q}_N(t)]$	average total data rate of all guaranteed services at time t;
$\tilde{\beta}_i(t)$	desired blocking probability for service i at time t;

Variables describing best-effort service:

$p_b(t)$	price for admitting one cell into the buffer at time t;
$q_b(t)$	queue length of best-effort service at time t;

$s_b(t)$	cell transmission rate at time t;
$\lambda_b[p_b(t),t]$	cell accepting rate, i.e. arrival rate of cells with willingness to pay higher than $p_b(t)$;
$\omega_b(t)$	admission rate of cells at time t;
$\varpi_b(t)$	expected value of $\omega_b[p_b(t), t]$;
Other variables:	
T	duration of business cycle;
C_T	total bandwidth;
$K(C_T)$	amortization of capacity investment cost over one cycle;
B_S	buffer size.

In (3-1), $(1-\beta_i)\lambda_i(p_i,t)dt$ is the expected number of calls of service i that will be admitted during the period $[t,t+dt)$. Multiplying this number by the unit price, $p_i(t)$, and expected call duration, $\frac{1}{r_i}$, yields the expected revenue from all calls of service i admitted in that interval. At time t, the network also charges a price for each cell of best-effort service that enters the buffer, and $\varpi_b(t)dt$ is the expected number of cells that will enter the buffer at that time. Thus $\varpi_b(t)p_b(t)dt$ is the expected revenue from best-effort service at t. The total expected profit is calculated by summing up expected revenue from all services, accumulated over all time in $[0,T]$, minus the amortized capacity cost. At this point, we assume zero discount rate for simplicity.

§3.2 The Solution: A 3-stage Procedure

Though it would be ideal to solve the model defined in (3-1) - (3-8) directly to get the analytical form of the optimal pricing trajectory $(p_i(t),p_b(t))$ and the optimal amount of bandwidth (C_T), it is mathematically intractable. Therefore, we construct a three-stage procedure to find a near-optimal solution. At each stage, we will make some simplifying assumptions, or treat some variables as constants, and solve part of the problem. The solution obtained at one stage will be used either as an input to the next stage or as a

feedback for modifying assumptions made in the previous stage. This process is iterated until prices stabilize at a near-optimal level.

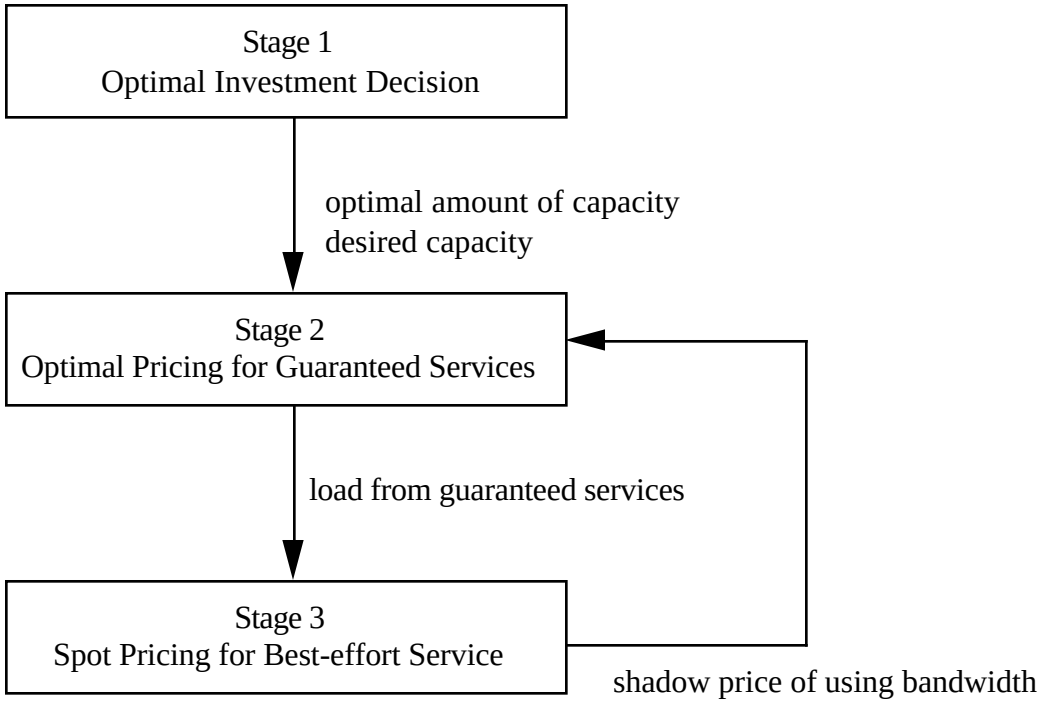

Figure 3
The 3-stage Procedure

The 3-stage procedure is defined as follows: at stage 1, we solve a long-term optimal investment problem to find the optimal amount of total bandwidth (C_T), as well as the desired blocking probability, $\tilde{\beta}_i(t)$, which we expect will vary with time of day. Using these values as inputs, we develop the optimal pricing policy at the second stage. The result shows that the optimal price for a service should be a function of the opportunity cost of providing that service. The opportunity cost is determined by both the service characteristics and the shadow prices of reserving/using network bandwidth. We give trial values to shadow prices and set up a price schedule for each guaranteed service accordingly. Based on these price schedules, the traffic load from guaranteed services can be determined. Under a given traffic load from guaranteed services, at the third stage, we

formulate a more precise model to describe the cell flow of best-effort service at each moment. The spot price for best-effort service is then derived to maximize the revenue from best-effort service. From these spot prices, we can then decide the instantaneous value of using network bandwidth. This information is used as feedback to the second stage for adjusting the trial value of shadow prices we previously calculated, so the price schedule for guaranteed services can be refined. The process is iterated until both the price schedule for guaranteed services and the spot price for best-effort service stabilize.

In the next section, we will discuss the implementation details at each stage, and interpret the economic implications of our results.

§4 Implementation of the 3-stage Procedure

§4.1 Stage 1: Optimal Investment

At this stage, we formulate and solve an optimization problem to determine the optimal amount of total bandwidth, C_T, and the desired blocking probability of each guaranteed service at each time, $\tilde{\beta}_i(t)$, i=1,N. The formulation of the problem is as follows:

Divide [0,T] into M time intervals, each lasting w_m, (m=1,M). Take the average arrival rate $\lambda_{im} = \frac{\int_{[m-1,m)} \lambda_i(\tau) d\tau}{w_m}$ as the arrival rate for all time during that interval. λ_m is determined by price p_{im}. We also assume that calls admitted during the interval [m-2,m-1] will have no influence on traffic load within the interval [m-1,m]. β_{im} is the blocking probability during the interval [m-1,m], which is a function of network loads within that interval.

At this stage we ignore blocking due to finite buffer space for best-effort traffic. Then the expected cell acceptance rate equals the expected cell admission rate, i.e. $\lambda_{bm}(p_{bm}) = \varpi_{bm}$. In other words, all cells with willingness to pay higher than the cut-off

price are assumed to be able to enter the buffer. To keep the queue length in the buffer reasonably short, we assume the expected cell admission rate equals the expected cell transmission rate, i.e. $\varpi_{bm} = \bar{s}_{bm}$. Consequently, $\lambda_{bm}(p_{bm}) = \bar{s}_{bm}$.

The network operator controls p_{im}, p_{bm}, and C_T to maximize total profit, i.e.

$$\max_{p_{im}, p_{bm}, C_T} \sum_{m=1}^{M} w_m [\sum_{i=1}^{N} \frac{(1-\beta_{im})\lambda_{im}(p_{im})p_{im}}{r_i} + p_{bm}\lambda_{bm}(p_{bm})] - K(C_T) \quad (4\text{-}1)$$

$$\text{s.t.} \qquad \bar{q}_{im} = \frac{(1-\beta_{im})\lambda_{im}(p_{im})}{r_i} \quad (4\text{-}2)$$

$$\dot{\beta}_m = A(\bar{q}_{im}, i = 1, N, C_T), \quad (4\text{-}3)$$

where $\vec{\beta}_m = (\beta_{1m}, ..., \beta_{Nm})$

$$\bar{s}[(1-\beta_{im})\bar{q}_{im}, i = 1, N] + \lambda_{bm} \le C_T, \qquad m=1, M \quad (4\text{-}4)$$

This is an optimization problem with $(N+1)M+1$ controlling variables. It can be solved either by non-liner optimization techniques or generic algorithms such as simulated annealing. The resulting C_T and β_m will be considered as optimal values for the total amount of capacity and for blocking probability in each period.

The solution we have obtained so far is not truly optimal because we have made several simplifications. One simplification is that we assume the traffic load in any period has no influence on the traffic load in succeeding periods. We have also ignored the fact that the arrival rate may change continuously over time within each period by using a single value λ_b as the arrival rate for all time in a period $[m\text{-}1,m)$. Both simplifications will cause inaccuracy in our results. Interestingly, the effects of these two simplifications depend on how we divide $[0,T)$ into different intervals. If we divide $[0,T)$ into longer intervals, i.e. w_m is larger, the effect of not considering the relationship between traffic load in different periods will be smaller and the effect of ignoring the change of arrival rate within a period is more serious. If we choose a smaller w_m, the effects will go in the opposite direction.

Therefore, w_m should be chosen to minimize the total negative effect of these two simplifications[2].

§4.2 Stage 2: Optimal Pricing

We now allow the arrival rate to change continuously over time, consider the dependency of traffic load at different times, and derive the optimal pricing policy at this stage. We will still keep the assumption that for best-effort service, cell admission rate equals cell transmission rate at all times, and ignore blocking of best-effort traffic. As a result, $\lambda_b(p_b,t)$, the arrival rate of cells for which the willingness to pay is above the cut-off price $p_b(t)$ at time t, is used both as the average rate of cell admission into the buffer and the average rate of cell transmission out of the buffer at time t for best-effort service in the problem formulation.

Given the amount of bandwidth (C_T) and optimal blocking probability ($\tilde{\beta}_i(t)$, i=1,N) calculated at stage 1, we can simplify the optimal pricing model defined in (3-1) - (3-9) as follows:

$$\underset{p_i(t),p_b(t)}{\text{maximize}} \int_0^T \{\sum_{i=1}^{N}[1-\tilde{\beta}_i(t)]\lambda_i(p_i,t)\frac{p_i}{r_i}+\lambda_b(p_b,t)p_b\}dt \qquad (4\text{-}5)$$

subject to:
$$\frac{d\bar{q}_i}{dt}=[1-\tilde{\beta}_i(t)]\lambda_i(p_i,t)-r_i\bar{q}_i(t) \qquad i=1,N, \qquad (4\text{-}6)$$

$$A[\bar{q}_1(t),\ldots..,\bar{q}_N;\tilde{\beta}_1(t),\ldots.,\tilde{\beta}_N(t)]\le C_T \qquad (4\text{-}7)$$

$$\lambda_b(p_b,t)+s[\bar{q}_1(t),\ldots,\bar{q}_N(t)]\le C_T \qquad (4\text{-}8)$$

$$q_i(0)=q_{i0},\ i=1,N \qquad (4\text{-}9)$$

---

[2]It is preferable to choose a larger w_m if call arrival rate is stable over time, and call duration is long, and a smaller w_m if arrival rate is sporadic and call duration is short.

We assume that the optimal solution exists for this pricing model. The optimal solution to equation (4-5) through (4-9) must obey the following proposition, which yields the optimal pricing policy:

Proposition: The optimal pricing policy:

Suppose $p_i^*(t)$, $p_b^*(t)$ are the optimal solutions to the pricing model defined in (4-5)-(4-9), then:

(1) $$p_i^*(t) = \frac{\varepsilon_i(p_i^*,t)}{1+\varepsilon_i(p_i^*,t)} * h_i(t) \quad \text{and} \quad p_b^*(t) = \frac{\varepsilon_b(p_b^*,t)}{1+\varepsilon_b(p_b^*,t)} * l_2(t) \qquad (4\text{-}10)$$

if $l_2(t) > 0$ and $h_i(t) > 0$, i=1,N

or

(2) $$p_i^*(t) = \frac{\varepsilon_i(p_i^*,t)}{1+\varepsilon_i(p_i^*,t)} * h_i(t) \quad \text{and} \quad p_b^*(t) = p_b^0(t) \qquad (4\text{-}11)$$

if $l_2(t) = 0$ and $h_i(t) > 0$, i=1,N

or

(3) $$p_i^*(t) = p_i^0(t) \quad \text{and} \quad p_b^*(t) = p_b^0(t) \qquad (4\text{-}12)$$

if $h_i(t) = 0$, i=1,N

where: $p_i^0(t)$maximizes $p_i(t)\lambda_i(p_i,t)$, $p_b^0(t)$maximizes $p_b(t)\lambda_b(p_b,t)$,

$$\varepsilon_i(p_i^*,t) = \frac{\partial \lambda_i}{\partial p_i} * \frac{p_i^*}{\lambda_i}, \qquad \varepsilon_b(p_b^*,t) = \frac{\partial \lambda_b}{\partial p_b} * \frac{p_b}{\lambda_b} \qquad (4\text{-}13)$$

$$h_i(t) = \int_t^T [\frac{\partial A}{\partial \overline{q}_i} l_1(\tau) + \frac{\partial \overline{s}}{\partial \overline{q}_i} r_i e^{-r_i(\tau - t)} d\tau \qquad i=1,N \qquad (4\text{-}14)$$

$l_1(t)$ is the Lagrangian multiplier of constraint (4-7),

$l_2(t)$ is the Lagrangian multiplier of constraint (4-8).

In §4.2.1 below, we discuss the economic implications of this policy. How to decide the optimal pricing schedule for guaranteed services based on the policy is discussed in §4.2.2.

§4.2.1 Economic implications

The pricing policy shown in (4-10) is designed for situations in which the network capacity is tightly constrained. If the network operator prices services without considering capacity constraints, for guaranteed services, either the network can not meet performance requirements, or some services will experience a blocking rate beyond the designed value.

For best-effort service, if the number of cells admitted exceeds the number of cells transmitted, the queue would grow without bound. Our proposition shows that under these scenarios, the network operator's optimal strategy is to attach an opportunity cost to each service ($h_i(t)$ for guaranteed service i, and $l_2(t)$ for best effort service), and price a network service in the same way as pricing a tangible product, except that the marginal production cost should be replaced by opportunity costs.

We now explain the rationale for using $h_i(t)$ as the opportunity cost for providing guaranteed service i, and $l_2(t)$ as the opportunity cost for providing best-effort service, starting by explaining the Lagrangian multipliers of the two capacity constraints. The economic implication of the Lagrangian multiplier of a resource constraint is the maximum value that can be derived from having one more unit of the constrained resource, i.e. the shadow price of consuming one unit of that resource. In our case, $l_1(t)$, $l_2(t)$ are shadow prices of reserving and using one unit of bandwidth, respectively. Since we measure the bandwidth in terms of the number of cells that can be sent per unit of time, at time t, when one cell of best-effort service is sent, one unit of bandwidth is consumed. Therefore, the unit opportunity cost for best-effort service at time t is just the shadow price of using one unit of bandwidth at that time, i.e. $l_2(t)$.

To meet performance requirements for guaranteed services, the network needs to reserve some capacity each time a call is admitted. At each moment, part or all of reserved bandwidth will actually be used by guaranteed services. Consequently, the opportunity cost should include two components: the opportunity cost of reserving the bandwidth, and the opportunity cost of using it. In our formulation, at time t, the former equals the shadow price for reserving one unit of bandwidth, $l_1(t)$, times the marginal increase of the amount of reserved bandwidth for admitting one more call, $\frac{\partial A}{\partial \overline{q}_i}$, and the latter equals the shadow price for using one unit of bandwidth, $l_2(t)$, times the marginal increase of bandwidth usage

which results from admitting one more call, $\frac{\partial \bar{s}}{\partial \bar{q}_i}$. The total opportunity cost for a call is thus the sum of these two components, accumulated over all time. Since the service duration is an exponentially-distributed random variable, the total cost, $h_i(t)$, is estimated by taking mathematical expectation, using the distribution function of the call duration ($r_i e^{-r_i t}$).

Equation (4-10) is appropriate when the number of guaranteed calls that can be admitted while meeting performance requirements is still limited, but there is more than enough capacity to carry the cells from all guaranteed calls that are admitted, as well as all of the best-effort traffic that the network wants to carry. This situation might occur, for example, if the guaranteed calls are extremely bursty, or their performance requirements are extremely strict. i.e.

$$\lambda_b(p_b, t) + \bar{s}[\bar{q}_1(t), \ldots, \bar{q}_N(t)] < C_T$$

As a result, at time t, the shadow price of using the bandwidth, $l_2(t)$, equals 0, and the optimal pricing policy should follow (4-10), i.e. the network operator should set price to maximize total revenue from best-effort service without considering the constraint on data rate.

Equation (4-11) specifies the pricing policy for the situation when there is an excessive amount of bandwidth. In this case, even if the network operator maximizes revenue without considering capacity constraints, she can still meet performance objectives for all services, keep blocking probability below the desired level, and have more transmission capacity for best-effort service than what is needed. As a result, both the opportunity costs for guaranteed services and the opportunity cost for best-effort service equal zero (i.e. $h_i(t) = 0$, $l_2(t) = 0$). This only happens when capacity is not constrained for both reservation and use for all time, or in other words, the capacity is over provisioned. Since we have assumed that the capacity, C_T, is set at the optimal level in stage 1, this cannot occur.

§4.2.2 The optimal pricing schedule for guaranteed services

As shown in (4-10), (4-11), the optimal price for guaranteed services depends on the ε_i, which is the demand elasticity, $\frac{\partial A}{\partial q_i}$ which reflects traffic characteristic and performance requirements, as well as $l_1(t)$, $l_2(t)$, the shadow prices for reserving and using bandwidth, respectively, i.e. :

$$p_i(t) = \frac{\varepsilon_i}{1+\varepsilon_i} h_i(t)$$

where $h_i(t) = \int_t^T [\frac{\partial A}{\partial \bar{q}_i} l_1(\tau) + \frac{\partial \bar{s}}{\partial \bar{q}_i} l_2(\tau)] r_i e^{-r_i(\tau - t)} d\tau$

To find $p_i(t)$, values of $l_1(t)$, $l_2(t)$ need to be determined. At this point, we assume the values of $l_2(t)$ have been estimated and given as $\hat{l}_2(t)$. (This prior estimation will be modified by the feedback from stage 3). We then set $l_1(t)$ to the trial value $\hat{l}_1(t)$, and construct the following procedure to find the optimal value for $p_i(t)$, as well as to modify the estimate of $l_1(t)$

1) Calculate the optimal pricing schedule for guaranteed services by :

$$\hat{h}_i(t) = \int_t^T [\frac{\partial A}{\partial \bar{q}_i} \hat{l}_1(\tau) + \frac{\partial \bar{s}}{\partial \bar{q}_i} \hat{l}_2(\tau)] r_i e^{-r_i(\tau - t)} d\tau \text{ and } \hat{p}_i(t) = \frac{\varepsilon_i}{1+\varepsilon_i} \hat{h}_i(t)$$

2) The call arrival rate of guaranteed services i at time t is then $\hat{\lambda}_i[\hat{p}_i(t), t]$. Given $\hat{\lambda}_i(\hat{p}_i(t), t)$ and the total amount of bandwidth, C_T, the expected number of calls in progress, $\hat{a}_i(t)$, and the blocking probability, $\hat{\beta}_i(t)$, can be determined.

3) If $l_1(t)$ is underestimated, $\hat{p}_i(t)$ will be lower than its optimal value, so call arrivals will be higher than the optimal level, which leads to the situation that blocking probability is higher than the desired level, i.e. $\hat{\beta}_i(t) > \tilde{\beta}_i(t)$ at some t. If $l_1(t)$ is over-estimated, $\hat{p}_i(t)$ will be lower than its optimal value and $\hat{\beta}_i(t) < \tilde{\beta}_i(t)$.

4) Increase or decrease $\hat{l}_1(t)$ by Δl_1, depending on whether it is over or under estimated. Go to 1) to calculate $p_i(t)$.

The process is iterated until $\hat{\beta}_i(t) = \tilde{\beta}_i(t)$ or is within a tolerable error band.

The price schedule for guaranteed services is based on the given estimates of $l_2(t)$, i.e., the shadow price for using the bandwidth. This estimate was given arbitrarily at the beginning, and needs to be modified by using feedback from the third stage.

§4.3 Spot Pricing

Given the prices for guaranteed services obtained at the second stage, the distribution of available capacity for best-effort service as a function of time can be determined as $C_T - s[q_1(t),\ldots,q_N(t)]$. At each instant, the network operator will set $p_b(t)$, the spot price for admitting cells of best-effort service into the buffer to maximize:

$$\int_t^T p_b(t) * \omega_b(t) dt \tag{4-15}$$

under constraints:

$$\frac{dq_b}{dt} = \omega_b(t) - s_b(t) \tag{4-16}$$

$$s_b(t) \le C_T - s[q_1(t),\ldots,q_N(t)] \tag{4-17}$$

$$0 \le q_b(t) \le B_s \tag{4-18}$$

$$\text{when } q_b(t) = B_s \qquad \omega_b(t) \le s_b(t) \tag{4-19}$$

$$\text{when } q_b(t) = 0 \qquad \omega_b(t) \ge s_b(t) \tag{4-20}$$

Given $\omega_b(t)$, $s_b(t)$ are random variables with complicated distributions, the problem in (4-15)-(4-20) can not be solved directly. However, through simulation, we can design heuristic rules that indicate how the spot price, $p_b(t)$, should be set based on current buffer occupancy and the expected distribution of willingness to pay of cells arriving in the future.

As soon as the spot price, $p_b(t)$, is determined, a new estimate of $l_2(t)$ can be constructed. This can be done by using the proposition above that defines the optimal pricing policy. Equation (4-10), i.e. $p_b^*(t) = \frac{\varepsilon_b(p_b^*,t)}{1+\varepsilon_b(p_b^*,t)} * l_2(t)$ applies when the bandwidth is fully used, and Equation (4-11), i.e. $l_2(t)=0$ applies otherwise. The new estimate can then be used as feedback to revise the optimal pricing schedule for guaranteed services.

The optimal pricing policy is reached by iterating the second and the third stages until both the price schedule for guaranteed services and the expected spot price for best-effort service stabilize.

5. Conclusions and Future Work

In this chapter, we discuss the optimal pricing policy for Integrated-service networks with guaranteed quality of service based on ATM technology. By formulating the pricing decision as a constrained control problem and developing a three stage procedure to solve that model, we find there is great similarity between the optimal pricing policy for network services and the optimal pricing policy for conventional products. We demonstrate that under capacity constraints, the service provider should consider the opportunity cost incurred by serving a customer. This opportunity cost should be used to determine the price of a network service in the same way as the marginal production cost is used to determine the price of a conventional product. We derive the mathematical expressions that calculate opportunity costs for different services offered by a single integrated-services network, and explain the implications of these expressions.

Though our procedure is designed for maximizing the service provider's profit, a similar approach can as well be used to maximize other objectives, such as social welfare.

Note the pricing policy developed in this paper optimizes the profit for providing integrated services under the assumption that the demand for each service is independent of prices of any other services. In future work, we will relax that assumption and consider the cross-elasticity effect among services. Even in the absence of cross-elasticity effect, the price of one service can also affect the demand for another service if the network adopts a three-part tariff pricing scheme, under which users are not only charged for each service based on reservation and usage, but also pay a flat subscription fee (e.g. an access charge). In this case, the network operator may maximize profit by setting reservation or usage

prices for each service different from the optimal values derived in this chapter. As another example, in the presence of positive network externalities, it can be optimal to price access below average cost, recovering the balance from the increased demand for usage which results from a larger network population. Our paper considers neither three-part tariff nor positive demand externalites. The design of an optimal pricing schedule with the consideration of these factors is an interesting issue that remains to be explored.

References:

Cocchi, R., Shenker, S., Estrin, D., and Zhang, L. (1993), "Pricing in computer networks: motivation, formulation, and example," *IEEE/ACM Transactions on Networking*; vol. 1, no. 6 (December), pp. 614-27.

Dewan, S. and Mendelson, H. (1990), "User Delay Costs and Internal Pricing for a Service Facility," *Management Science*, vol. 36, no. 12, pp. 1502-1517.

Gupta, A., Stahl, D. O., and Whinston, A. B. (1996), "Priority Pricing of Integrated Services Networks," *Internet Economics*, Chapter 12, forthcoming.

Hyman, J. M., Lazar, A. A., and Pacifici, G. (1993), "A Separation Principle between Scheduling and Admission Control for Broadband Switching," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4 (May), pp. 605 - 616.

Kamien, M. I. and Schwartz, N. L. (1981), "*Dynamic Optimization: the Calculus of Variations and Optimal Control in Economics and Management*," North Holland.

Mackie-Mason, J. K. and Varian, H. R. (1994), "Pricing the Internet," in "*Public Access to the Internet*," B. Kahin and J. Keller, Eds. Edglewood Cliffs, N. J. Prentice-Hall.

Pappas, J. L. and Hirschey, M. (1987), "*Managerial Economics*," 5th ed. Dryden Press.

Peha, J. M. and Tobagi, F. A. (1996), "Cost-Based Scheduling and Dropping Algorithms To Support Integrated Services," *IEEE Transactions on Communication*s, vol. 44 No. 2 (February), pp. 192-202.

Peha, J. M. (1993), "The Priority Token Bank: Integrated Scheduling and Admission Control for an Integrated-Services Network," *Proceedings of IEEE International Conference on Communications*, ICC-93, Geneva, Switzerland, pp. 345-51.

Peha, J. M. (1991), "*Scheduling and Dropping Algorithms to Support Integrated Services in Packet-Switched Networks*," Ph.D. Dissertation, Technical Report No. CSL-TR-91-489, Computer Systems Laboratory, Stanford University.

Tewari, S. and Peha J. M. (1995), " Competition Among Telecommunications Carriers That Offer Multiple Services," *Proceedings of 23rd Telecommunications Policy Research Conference*, Solomon Island, Maryland.

Whang, S. and Mendelson, H. (1990), "Optimal Incentive-Compatible Priority Policy for the M/M/1 Queue," *Operations Research*, vol. 38, pp. 870-883.

Zhang, L., Deering, L., Estrin, D., Shenker, S., and Zappala, D. (1993), "New Resource Reservation Protocol," *IEEE Network*, vol. 7, no. 5 (September), pp. 8-18.