Optimal Pricing for Multiple Services in Telecommunications Networks Offering Quality-of-Service Guarantees

Neil J. Keon, *Member, IEEE,* and G. Anandalingam, *Senior Member, IEEE*

***Abstract*—We consider pricing for multiple services offered over a single telecommunications network. Each service has quality-of-service (QoS) requirements that are guaranteed to users. Service classes may be defined by the type of service, such as voice, video, or data, as well as the origin and destination of the connection provided to the user. We formulate the optimal pricing problem as a nonlinear integer expected revenue optimization problem. We simultaneously solve for prices and the resource allocations necessary to provide connections with guaranteed QoS. We derive optimality conditions and a solution method for this class of problems, and apply to a realistic model of a multiservice communications network.**

***Index Terms*—Economics, network design, pricing, quality of service (QoS).**

I. Introduction

IN THIS PAPER, we derive a nonlinear mathematical programming model for determining optimal prices for multiple services with guaranteed quality of service (QoS). We solve this model using an auction algorithm. The traditional distinction between voice networks, data networks, and cable TV networks is fast becoming obsolete. In the future, multiple communications services will be available to users over a single network. These different services will vary in terms of bandwidth requirements and tolerate different quality limitations such as loss of data and delay in transmission. Our model presents a way for network providers to set prices for these services, and allocate resources such that these QoS requirements are guaranteed while expected revenues are maximized.

A. Overview

We consider the problem of using pricing and resource allocation to manage multiple services networks with QoS guarantees. Although the arrival rate of connection requests can be fixed by setting the price, the observed arrivals, and thus, the observed revenue, are assumed to be a stochastic process. Also, the model we propose in this paper separates demand into *classes*, defined by service type as well as origin and destination. Our model maximizes *expected* revenue in order to obtain optimal prices and the resources needed to guarantee QoS for each service class.

The "resources" we consider are *buffer* space for temporarily storing users' data traffic, and *bandwidth* for transmitting the data through the network. All connections in a particular service class (e.g., voice, data, video) share the same buffer and bandwidth allocation. The resource allocations must be sufficient to satisfy two QoS guarantees which are defined for every service class individually. The first QoS guarantee is a restriction on the probability of data being lost in the network. Data is lost when resources allocated to a service class based on *expected* traffic calculations are insufficient to either transmit (using bandwidth) or store (using buffer space) all the data that is actually sent to the network by all users in a service class. The second QoS guarantee is a limit on the delay, or time the data is buffered, during transmission by the network. The two QoS criteria constrain the problem in terms of buffer space and bandwidth that must be allocated to each service class for a given level of demand.

The tradeoff between resource allocations among various service classes is complicated due to the fact that as the number of connections in a class increases we observe economies of scale in bandwidth allocations. This is due to two phenomena: smoothing of data traffic in buffers and statistical multiplexing gains. Economies of scale appear by decreasing the marginal resource allocations to any service class, i.e., by decreasing the resources required to supply one additional connection.

We construct a distributed search method called the "auction algorithm" to optimize the revenue by choosing prices, subject to constraints on the resource allocations implied by finite resources and QoS guarantees. Our search method takes advantage of several optimality properties we show for the resource allocations to each class. Using these properties, we reduce the search space for the overall problem and calculate optimal solutions.

B. Related Literature

We will briefly classify the related literature into two broad categories. First, we will discuss the bulk of the literature, which deals with pricing in best effort networks, such as the Internet. There has also been some recent work related to pricing connections for networks with QoS guarantees.

1) Pricing for Best Effort Service: The most celebrated packet-switched network, the Internet, offers *best effort service*, which is prone to unpredictable congestion and delays by definition. The current flow control scheme in the internet is called transmission control protocol (TCP) [2], [34]. Modifications to this scheme have been suggested that would allow

Manuscript received February 19, 2001; revised April 3, 2002; approved by IEEE/ACM Transactions on Networking Editor S. Low. This work was supported in part by the National Science Foundation under Grant NCR-9612781.

N. J. Keon is with the Cox School of Business, Southern Methodist University, Dallas, TX 75275 USA (e-mail: nkeon@mail.cox.smu.edu).

G. Anandalingam is with R. H. Smith School of Business and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: ganand@rhsmith.umd.edu).

Digital Object Identifier 10.1109/TNET.2002.808409

multiple service classes, and induce users to behave fairly and efficiently, through simple or randomized packet marking mechanisms in the network [10], [13], [14], [20]. Usage-based schemes, which charge based on the actual resources used, have also been proposed for these types of networks [8], [29].

Priority pricing is another suggestion for allowing multiple services over best effort networks. The best known work discusses a second bid auction, whereby it is incentive compatible, or in the user's best interests to truthfully reveal their true valuation of service in terms of a priority [22]–[24]. Another bidding paradigm, which results in a Nash equilibrium among users, was suggested for the Internet in [5]. A similar approach, again requiring users to bid for service, was proposed for available bit rate service, the best effort service offered in asynchronous transfer mode (ATM) networks [6].

A recent game theoretic model requiring users to choose from among several routes, each with its own delay, has been shown to have a stable equilibrium solution where the relative prices induce the desired operating point of the network operator [16], [17]. An interesting problem of regulating the arrival of jobs presented to the network is discussed in [26], and the informational requirements are relaxed, using an adaptive on-line method in [25]. Finally, recent results have been offered, based on dynamic programming that suggest a static price schedule results in network performance that is near optimal compared to congestion-dependent pricing [27].

2) Pricing With QoS Guarantees: Typically, a network with guaranteed QoS, such as the current voice network, must employ a call admission policy in order to satisfy the guarantees to users. In an alternative approach to call admission, it has been suggested that users guarantee their own QoS by purchasing the required bandwidth and buffer resources for their desired QoS directly from the network [19], [21]. In [31], an analysis of a market-based methodology is offered as evidence that pricing schemes can offer efficient resource allocations in connection-oriented networks offering QoS guarantees (see also [33]). These models feature a conventional view of pricing per connection and announcing the price to the public. The paper by de Veciana *et al.* [32] combines the conventional practice of prices being fixed and announced with the alternative view of allowing users to guarantee their own QoS by purchasing bandwidth and buffers directly. Resources are shared among users, incorporating multiplexing gains. On-line negotiation is also suggested as a framework for connection setup and allocation of network resources, using effective bandwidth as a base for pricing in [11] and [12]. In a related paper that deals with single link point-to-point ATM networks, the pricing problem was formulated as a constrained optimal control problem, and solved using a three-stage solution procedure [35]. The limitation of the approach is computational tractability.

Our paper belongs to this class of work, but uses a novel optimization model based on nonlinear mathematical programming. We use an optimal pricing scheme to moderate the demand for connection requests from different classes of service. Each "class" of service is defined by both the type of service (e.g., voice, data, video), and by where the telecommunications flow originates and where it is destined to. Thus, our paper tries to take into account the use of resources in the entire network. Based on the connection requests for each class of service at the specified prices, the model estimates optimal buffer and bandwidth allocations to satisfy QoS requirements. Another contribution of this paper is the algorithm we propose to obtain the optimal solutions.

C. Organization of the Paper

This paper is organized as follows. In Section II, we will describe the network model that we are considering and examine three important features, call blocking, loss probability, and delay, in some detail. Section III contains the complete formulation of the expected revenue maximization model, and presents optimality properties. Section IV provides details of the auction algorithm used to solve the problem. We present the implementation of the algorithm to an extensive example in Section V. We end the paper with concluding remarks in Section VI.

II. Model for Optimal Pricing With QoS Guarantees

In this section, we will give an overview of the model that we will be using to derive optimal prices for the multiple services in a telecommunications network where QoS is guaranteed. We will first describe an example network that will help us focus our thoughts on the particular features that we model, present the characteristics of the model, and then derive the particular mathematical constructs of these characteristics. In Section III, we will present the complete mathematical model, and in Section IV, we will present details of the algorithm to compute the optimal multiservice prices, based on the optimality conditions of the mathematical model.

A. The Network

We will use the ring-type network shown in Fig. 1, only to *highlight* the particular features that we are modeling. Note, however, that all features we consider in our model are found in any type of network. At each switch, there are inputs and outputs of telecommunications traffic. Traffic inputs are from many different *types* of service that either originate at that switch, or else arrive at the switch from other originating nodes through an incoming bandwidth pipe. Traffic outputs from the switch either go along an outgoing bandwidth pipe, or directly to a user connected to that switch.

Without loss of generality with respect to the network characteristics, we assume that buffer space is used at originating switches to collect and smooth data traffic, subject to delay constraints. The data traffic is then transported in bandwidth pipes through the rest of the path to its destination. We allocate aggregate quantities of buffer and bandwidth to connections grouped by service type i, origin j, and destination k. A service *class* is, therefore, denoted by the triple (i, j, k). Each service type i has traffic characteristics defined by *average* bandwidth r_i and *peak* bandwidth R_i. For example, voice and video have different average and peak bandwidth requirements.

B. Overview of the Model

The model is derived from the point of view of the service provider. The service provider sets prices for the different multiple services in order to maximize revenue subject to a set of

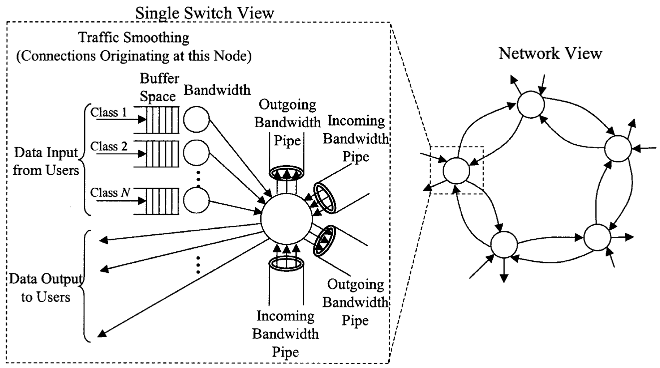

Fig. 1. Network (ring) service model at an individual switch.

constraints and conditions that are necessary to ensure both QoS and flow balance. The problem that the service provider solves can be summarized as follows.

Maximize expected revenue: This is a product of prices charged for each service and the demand for each service, taken to be an arrival rate of connection requests *that can be accommodated* at that price, subject to the following constraints.

- *Limited capacity*: The network switches have limited capacity (bandwidth).
- *Limit the blocking of connection requests*: Given bandwidth capacity constraints, there is a limited number of connections that can be supported in order to guarantee QoS, and some requests may have to be blocked. We set a limit on the blocking probability for each service class. Note that the capacity constraint and the limitation on blocking probability will in turn affect the number of connection requests that can be accommodated.
- *Limit the probability of packet loss*: Different services can tolerate different levels of packet loss and still guarantee QoS. We set a limit on the "equivalent capacity" (defined in detail later) allocated to each service in order to limit the probability of packet loss.
- *Limit the maximum delay*: Different services can tolerate different levels of delay and still guarantee QoS. We set a limit on the maximum allowable delay for each service.

We will now describe each component of the model in greater detail using quantities illustrated in Fig. 1. We will present the complete mathematical formulation of the problem in Section II, and derive optimality conditions.

C. Deriving Demand

The quantity demanded by users, λ_{ijk}, is taken to be the arrival rate of connection requests for service i, with the connection originating at switch j and terminating at switch k. This quantity is determined by a demand function $f_{ijk}(p_{ijk})$ where p_{ijk} is the price charged per unit of time the connection is open. Based on current empirical results [1], [18], in our model we use the demand function

$$\lambda_{ijk} = a_{ijk} p_{ijk}^{-\varepsilon_{ijk}} \tag{1}$$

where ε_{ijk} is the (constant) elasticity of demand for service class (i, j, k). One important property of (1) is that marginal revenue with respect to prices will always be negative (meaning lower prices always increase revenue) provided that $\varepsilon_{ijk} > 1$. We assume elasticities in excess of unity for the remainder of the paper. Again, this is suggested by available research [1], [18].

D. Call Blocking and Expected Revenue

We are modeling a network offering connections with guaranteed QoS. Under these conditions, there are a limited number of connections that can be supported, and some requests may have to be blocked if resources are already reserved for other requests. We would like to minimize the blocking of connection requests, and capture this in our optimization model as a constraint.

For each service class, denoted by (i, j, k), we model the number of ongoing connections as an $M/GI/m/m$ queue. The capacity in the queueing model m refers to the maximum number of connections that we will allow to be admitted to a given service class, a variable we define as NCMAX_{ijk}. For consistency with this model, we state a number of assumptions, as follows. The arrival rate of connection requests for a service class given by λ_{ijk} (defined above) characterizes interarrival times that are independently and identically distributed exponential random variables. The average holding time of a connection, T_{ijk}, is known and the holding times of individual connections can occur according to any general distribution, but are independently and identically distributed for each connection. The traffic intensity of a class ρ_{ijk} is the product of arrival rate and average holding time, i.e., $\rho_{ijk} = \lambda_{ijk} T_{ijk}$.

The properties of the $M/GI/m/m$ queueing model are well known [4]. The probability of blocking a request within any service class, BP_{ijk}, is given by the Erlang B formula

$$\begin{aligned} \mathrm{BP}_{ijk} &= P\left(\mathrm{NC}_{ijk} = \mathrm{NCMAX}_{ijk}\right) \\ &= \frac{\rho_{ijk}^{\mathrm{NCMAX}_{ijk}}}{\mathrm{NCMAX}_{ijk}!} \Bigg/ \sum_{n=0}^{\mathrm{NCMAX}_{ijk}} \frac{\rho_{ijk}^{n}}{n!}. \end{aligned} \tag{2}$$

If the number of open connections NC_{ijk} is equal to the maximum number permitted for each class NCMAX_{ijk}, when a user's request for service is received, the connection will be blocked and lost to the system. Otherwise (i.e., $\mathrm{NC}_{ijk,} < \mathrm{NCMAX}_{ijk}$), the request will be admitted.

The expected number of busy connections is given by

$$E\left[\mathrm{NC}_{ijk}\right] = \rho_{ijk}\left(1 - \mathrm{BP}_{ijk}\right). \tag{3}$$

We define the expected rate of revenue generation attributable to a service class, REV_{ijk}, as the price charged per unit time for a single connection p_{ijk} multiplied by the expected number of ongoing connections $E[\mathrm{NC}_{ijk}]$

$$\mathrm{REV}_{ijk} = p_{ijk} E\left[\mathrm{NC}_{ijk}\right]. \tag{4}$$

E. Probability of Loss and Equivalent Capacity

In the previous section, we modeled the connection level traffic so we could quantify blocking probability for our model. Now we model the packet level traffic for the same set of connections, so we can quantify one of the QoS parameters, namely probability of lost packets. This will also allow us to relate the service classes to the necessary resource assignments, buffer space and bandwidth, to satisfy the QoS. At the switch,

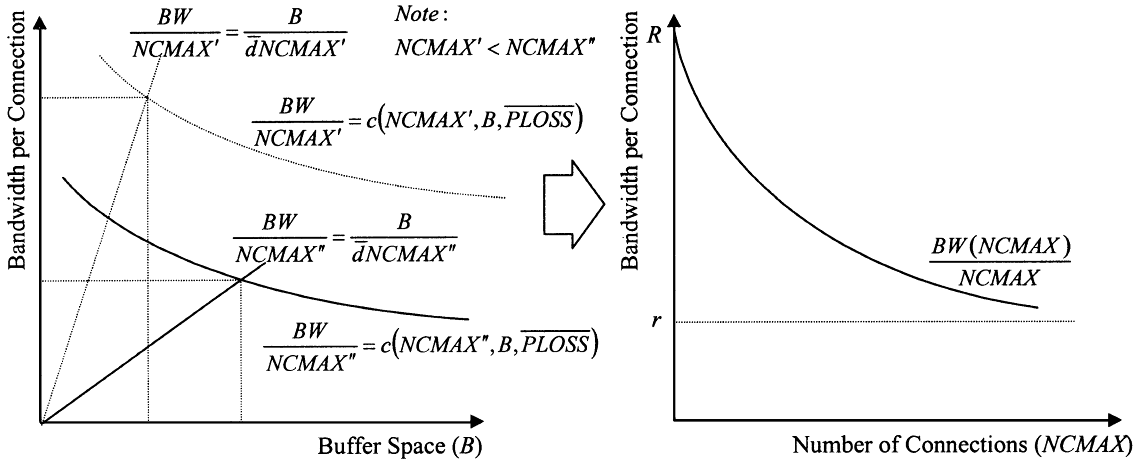

Fig. 2. Illustration of average bandwidth allocated per connection, using mapping from NCMAX to BW.

packets from all connections of the same class (i, j, k) share a first-in-first-out (FIFO) queue of size B_{ijk}. Up to NCMAX_{ijk} connections, with statistically identical data traffic, may share the allocated buffer at any time. Packet loss PLOSS_{ijk} occurs when there is buffer overflow.

To ensure that the guaranteed loss probability is satisfied for each service class, we require that the minimum bandwidth allocation per offered connection be at least equal to the *equivalent capacity*. Equivalent capacity c for a single source (connection) is defined as *the service rate of the queue that corresponds to a given (packet) loss* PLOSS_{ijk}[9], [28]. If bandwidth in excess of the equivalent capacity is assigned to a connection, the observed packet loss is less than PLOSS_{ijk}. There are two main effects that determine the equivalent capacity of a single source. *Effective bandwidth* refers to the fact that smoothing of data traffic in the buffer reduces the bandwidth required to achieve a prescribed loss rate of packets. *Multiplexing gains* refers to the gain in bandwidth required due to the statistical effect of mixing the data traffic of independent connections. We use the term *equivalent capacity* when referring to the combination of these two effects on bandwidth assignment (per connection) to a number of connections with guaranteed QoS. We now state some general properties of equivalent capacity.

In our model, all connections of a service class (i, j, k) share a single FIFO buffer. Consider a connection of a given service *type* i, which is distinct from service "class" where we include origin–destination information as well. The user is either sending data at a *peak* transmission rate R_i (i.e., the ON state) or is sending no traffic at all (i.e., the OFF state). The *average* rate of data transmission for the individual connection is given by r_i. The ON and OFF periods are assumed to be exponentially distributed, with the average length of an ON period given by b_i. QoS requirements for this type of data traffic model have been well studied, e.g., [9]. The traffic from all connections of the same type i is statistically identical, i.e., all connections classified as the same type have the same parameters r_i, R_i, and b_i.

Clearly, the equivalent capacity must be greater than or equal to the mean rate of data traffic r_i and less than or equal to the peak rate of data traffic R_i

$$r_i \leq c\,(\mathrm{NCMAX}_{ijk},\, \mathrm{B}_{ijk},\, \mathrm{PLOSS}_{ijk}) \leq R_i. \tag{5}$$

As Fig. 2 shows, equivalent capacity is decreasing in the amount of allocated buffer space B_{ijk}

$$\frac{\partial c\,(\mathrm{NCMAX}_{ijk},\, \mathrm{B}_{ijk},\, \mathrm{PLOSS}_{ijk})}{\partial \mathrm{B}_{ijk}} \leq 0. \tag{6}$$

As the number of connections for which we allocate resources NCMAX_{ijk} increases, equivalent capacity will decrease, reflecting multiplexing gains

$$\frac{\partial c\,(\mathrm{NCMAX}_{ijk},\, \mathrm{B}_{ijk},\, \mathrm{PLOSS}_{ijk})}{\partial \mathrm{NCMAX}_{ijk}} \leq 0. \tag{7}$$

Finally, as we allow greater loss probabilities, PLOSS_{ijk}, the equivalent capacity will decrease.

$$\frac{\partial c\,(\mathrm{NCMAX}_{ijk},\, \mathrm{B}_{ijk},\, \mathrm{PLOSS}_{ijk})}{\partial \mathrm{PLOSS}_{ijk}} \leq 0. \tag{8}$$

An in-depth analysis of equivalent capacity is outside the focus of this paper. Detailed discussions of buffering and multiplexing gains are given in [2], [9], [28], [32], and [34]. However, it should be noted that our approach remains unchanged for *any* derivation of equivalent capacity, which satisfies the general properties for equivalent capacity, given in (5)–(8) above. The inequalities admit special cases such as constant bit-rate traffic, with no smoothing or multiplexing gains. As we shall see in the next section, the data traffic model and equivalent capacity expressions in (5)–(8) are sufficient for formulating packet loss in our revenue-maximizing model.

It will become apparent later that we wish to solve buffer assignment and bandwidth assignment simultaneously. Therefore, we now illustrate some extensions of the equivalent capacity properties when maximum buffer delay is held constant, i.e., the buffer is sized according to a constant maximum delay given by $d = \mathrm{B}/\mathrm{BW}$. Fig. 2 shows the general relationship between equivalent capacity (average bandwidth per connection) and number of connections when maximum delay in the buffer is held fixed. This relationship is derived based on (5)–(8) above. (The service class indices (i, j, k) are omitted for clarity.) Note that equivalent capacity, as illustrated in Fig. 2, gives the bandwidth required to meet two QoS criteria, namely, loss probability PLOSS_{ijk} and maximum delay d_{ijk}.

The average bandwidth allocated per connection, as a result of the mapping from B to BW given by (5)–(8) is illustrated by the downward sloping curves in the left panel above. An increase in the number of connections served, from NCMAX' to NCMAX'', results in the downward shift of the average bandwidth assignment per connection. The straight lines represent the relationship between the buffer B and the bandwidth BW for fixed maximum delay, i.e., $\mathrm{BW} = d \times \mathrm{B}$. Since the figure is drawn in a space representing average bandwidth per connection, $\mathrm{BW}/\mathrm{NCMAX}$, the shift outward in the slope of the lines results from the same increase in the number of connections served, from NCMAX' to NCMAX''. Collecting the solutions to such a system of two equations in two unknowns yields the downward sloping curve in the right panel of Fig. 2.

Fig. 2 is conceptual. In our numerical example, we will use the equivalent capacity results from [9] to calculate the equivalent capacity per connection summarized as

$$c\,(\mathrm{NCMAX}_{ijk},\, \mathrm{B}_{ijk},\, \mathrm{PLOSS}_{ijk}) = \min\left(\hat{c},\, \frac{\mu + \alpha'\sigma}{\mathrm{NCMAX}_{ijk}}\right) \tag{9}$$

where

$$\hat{c} = \frac{\alpha R_i - \mathrm{B}_{ijk} + \sqrt{(\alpha R_i - \mathrm{B}_{ijk})^2 + 4\alpha \mathrm{B}_{ijk} r_i}}{2\alpha} \tag{10}$$

$$\alpha = \ln\left(\frac{1}{\mathrm{PLOSS}_{ijk}}\right) b_i \left(1 - \frac{r_i}{R_i}\right) \tag{11}$$

$$\mu = r_i \mathrm{NCMAX}_{ijk} \tag{12}$$

$$\sigma = \sqrt{r_i\,(R_i - r_i)\,\mathrm{NCMAX}_{ijk}} \tag{13}$$

$$\alpha' = \sqrt{-2\ln(\mathrm{PLOSS}_{ijk}) - \ln(2\pi)}. \tag{14}$$

The equivalent capacity per connection (9) is calculated as the minimum of two distinct approximations. The effective bandwidth is approximated by (10). The second term in the minimum expression in (9) reflects multiplexing gains. This approximation is based on the stationary bit rate. To calculate this expression, we need the additional expression (11), as well as the mean of the aggregate bit rate (12), the standard deviation of the aggregate bit rate (13), and an approximate inversion of the normal distribution (14). The calculations above separate equivalent capacity into regions dominated by either smoothing effects in the buffer or multiplexing gains. These expressions are one example of suggested computational methods for equivalent capacity that satisfy (5)–(8).

F. Delay

The second QoS parameter that we consider is the maximum allowable delay d_{ijk} for each service class. We ignore transmission delay and consider only delay in the buffer. Therefore, we can quite easily set bounds for the potential delay a packet may experience.

$$d_{ijk} \le \frac{\mathrm{B}_{ijk}}{\mathrm{BW}_{ijk}} \tag{15}$$

where B_{ijk} is the buffer space and BW_{ijk} is the bandwidth allocated to service (i, j, k). The maximum delay any packet may experience, given by (15), is simply the size of the allocated buffer space divided by the allocated bandwidth. The buffer is served on a FIFO basis.

III. Optimization Model

A. The Model

Given the derivations above, we are now in a position to present the complete mathematical formulation of the revenue maximization model that would yield optimal prices and resource allocations. We wish to solve for the price p_{ijk} for each service class denoted by (i, j, k), as well as the volume of service offered NCMAX_{ijk}. The volume of service offered refers to how many connections we can serve simultaneously, given the bandwidth BW_{ijk} reserved all along the path between j and k, and the buffer space B_{ijk} reserved at the origin switch. We assume that bandwidth is limited but buffer space is not. The network service is assumed to be offered using a connection admission policy, so that the QoS requirements for probability of loss PLOSS_{ijk} and delay d_{ijk} are satisfied within certain limits. We include the blocking probability BP_{ijk} resulting from the use of a connection admission policy to ensure that we have sufficient resources set aside for service class (i, j, k), such that a satisfactory proportion of connection requests can be admitted when they are requested.

The total expected revenue from all service connections in the network is given by $\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NS}} \sum_{k=1}^{\mathrm{NS}} (p_{ijk} E\{\mathrm{NC}_{ijk}\})$. Because the expected number of connections arrivals NC_{ijk} is a function of the arrival rate, the blocking probability, and the average holding time of a connection, we can replace the objective function with $\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NS}} \sum_{k=1}^{\mathrm{NS}} (p_{ijk} f\{\lambda_{ijk},\, \mathrm{BP}_{ijk},\, T_{ijk}\})$.

Thus, the revenue optimization model that will yield optimal prices and resource allocation is given by

(PNet)

$$\mathrm{Max}_{p_{ijk}, \mathrm{B}_{ijk}, \mathrm{BW}_{ijk}, \mathrm{NCMAX}_{ijk}} \times \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{NS} \sum_{k=1}^{NS} (p_{ijk} f\{\lambda_{ijk}, \mathrm{BP}_{ijk}, T_{ijk}\}) \tag{16}$$

subject to

$$\mathrm{BP}_{ijk}\,(\lambda_{ijk}\,(p_{ijk})\,, T_{ijk}, \mathrm{NCMAX}_{ijk}) \le \overline{\mathrm{BP}}_{ijk}, \quad 1 \le i \le \mathrm{N},\, 1 \le j \le \mathrm{NS},\, 1 \le k \le \mathrm{NS} \tag{17}$$

$$c\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right) \le \frac{\mathrm{BW}_{ijk}}{\mathrm{NCMAX}_{ijk}}, \quad 1 \le i \le \mathrm{N},\, 1 \le j \le \mathrm{NS},\, 1 \le k \le \mathrm{NS} \tag{18}$$

$$\frac{\mathrm{B}_{ijk}}{\mathrm{BW}_{ijk}} \le \bar{d}_{ijk}, \quad 1 \le i \le \mathrm{N},\, 1 \le j \le \mathrm{NS}, 1 \le k \le \mathrm{NS} \tag{19}$$

$$\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NS}} \sum_{k=1}^{\mathrm{NS}} \mathrm{BW}_{ijk} IVP_{ijkx} \le \overline{\mathrm{BW}_x}, \quad 1 \le x \le \mathrm{NS} \tag{20}$$

$$IVP_{ijkx} = \begin{cases} 1, & \text{if } x \in VP_{ijk} \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le i \le \mathrm{N},\, 1 \le j \le \mathrm{NS},\, 1 \le k \le \mathrm{NS},\, 1 \le x \le \mathrm{NS} \tag{21}$$

$$\lambda_{ijk} \ge 0,\, p_{ijk} \ge 0,\, \mathrm{BW}_{ijk} \ge 0, \mathrm{B}_{ijk} \ge 0 \tag{22}$$

$$\mathrm{NCMAX}_{ijk} \ge 0, \text{ integer} \tag{23}$$

where

REV_{ijk} rate of revenue generation associated with service class (i, j, k);

$c(.)$ equivalent capacity of a single connection;

VP_{ijk} set of all x in the path for class i originating at j and terminating at k, $1 \leq x \leq \mathrm{NS}$;

$\overline{\mathrm{BP}}_{ijk}$ maximum blocking probability for a connection from service class (i, j, k);

$\overline{\mathrm{PLOSS}}_{ijk}$ maximum packet loss probability for a connection from service class (i, j, k);

$\bar{d}_{ijk}$ maximum delay for a connection from service class (i, j, k);

$\overline{\mathrm{BW}_j}$ the capacity (bandwidth) at switch j.

In this formulation, the objective function (16) seeks to maximize the average rate of revenue generation from ongoing connections, which is given by price multiplied by the *expected* number of connections. The constraints restrict the performance measures and resource assignments to be within bounds set outside the problem as *network policy*. Budgets on each quantity are denoted by a bar overhead. Constraint (17) restricts the call-blocking probability for every class below some prescribed limit. Constraint (18) ensures the probability of loss for each service class PLOSS_{ijk} is satisfied. The buffer delay for each service class is constrained to be less than or equal to the limit given by the QoS guarantee in constraint (19). We have a bandwidth capacity constraint at each switch, (20), but we assume there is no capacity on allocated buffer space, and there are no link capacities. The capacity at each switch must accommodate all traffic originating at the switch as well as traffic originating elsewhere but routed through the switch. We assume the routes are known and fixed. There is an indicator function, (21), for every class (i, j, k), which indicates if any switch is included in the path. There are a number of nonnegativity constraints, given by (22). Finally, there is an integrality constraint on the maximum number admitted to each class NCMAX_{ijk} given by (23), since connections can only be admitted in discrete quantities.

We wish to call attention to one limitation of our formulation of the capacity constraints, which may overassign bandwidth along the paths followed. The effective bandwidth of the aggregate traffic for the service class (i, j, k) may be reduced after smoothing in the buffer at the originating switch. Our formulation assigns bandwidth all along the path based on the effective bandwidth at the originating switch and does not take into account the statistical properties of the output traffic at the originating switch. The simplicity of the formulation outweighs our concern regarding excess bandwidth assignments.

B. Optimality Properties

In general, the problem given in (PNet) is a nonlinear nonconvex mixed integer problem. However, there are a number of necessary conditions for an optimal solution, which we will use to search for a solution. First, we will discuss necessary conditions related to constraints (17) – (19) in problem (PNet). These constraints are relatively simple as they apply to each service class independently of all other service classes. Based on this first set of necessary optimal conditions, we will then state a pair of optimality conditions reflecting the marginal values of each service class (i, j, k) in the solution.

1) Resource Allocation to Individual Service Classes: There are a number of conclusions we can immediately draw about optimal solutions to the problem (PNet). First, we consider the resources that must be assigned to each individual service class.

Theorem: Given downward-sloping demand curves and plentiful buffer space at each originating switch, if an optimal solution to (PNet) does not coincide with the condition that marginal revenue equals zero, i.e., the partial derivatives of (1) with respect to prices are all less than zero at optimality, then the optimal solution to (PNet) must satisfy the properties given by (24)–(27), shown at the bottom of the page.

Proof: See the Appendix for proof.

The call-blocking probability for all service classes is set to its greatest permitted value for all service classes in (24). Incidentally, this is the only condition that pertains directly to prices. The bandwidth assigned to each service class, for the maximum number of connections admitted, will be equal to the corresponding equivalent capacity, as stated in (25). The third condition, (26), states that because connections have to be integer, allocation of the equivalent capacity for any additional connections would violate feasibility. This means that no further increases in bandwidth assignments are possible at an optimal (feasible) solution. Buffer space is assigned for all services originating at each switch, such that packets for each service may experience a delay up to the tolerated delay, (27), assuming that buffer space at each switch is plentiful.

$$\mathrm{BP}_{ijk}\left(\lambda_{ijk}\left(p_{ijk}\right), T_{jk}, \mathrm{NCMAX}_{ijk}\right) = \overline{\mathrm{BP}}_{ijk}, \quad 1 \leq i \leq \mathrm{N}, 1 \leq j \leq \mathrm{NS}, 1 \leq k \leq \mathrm{NS} \tag{24}$$

$$\frac{\mathrm{BW}_{ijk}}{\mathrm{NCMAX}_{ijk}} = c_{ijk}\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right), \quad 1 \leq i \leq \mathrm{N}, 1 \leq j \leq \mathrm{NS}, 1 \leq k \leq \mathrm{NS} \tag{25}$$

$$\begin{aligned}&\Big(\left(\mathrm{NCMAX}_{xjk}+1\right) c_{xjk}\left(\mathrm{NCMAX}_{xjk}+1, \mathrm{B}_{xjk}, \overline{\mathrm{PLOSS}}_{xjk}\right)\\ &\left.+\sum_{\substack{i=1\\ i\neq x}}^{\mathrm{N}}\sum_{k=1}^{\mathrm{NS}} \mathrm{NCMAX}_{ijk} c_{ijk}\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right)\right) > \overline{\mathrm{BW}_j}\\ &\qquad 1 \leq x \leq \mathrm{N}, 1 \leq j \leq \mathrm{NS}\end{aligned} \tag{26}$$

$$\mathrm{B}_{ijk} = \bar{d}_{ijk} \mathrm{BW}_{ijk}, \quad 1 \leq i \leq \mathrm{N}, 1 \leq j \leq \mathrm{NS}, 1 \leq k \leq \mathrm{NS} \tag{27}$$

The optimal resource allocation is to assign the appropriate effective bandwith to each service class, (25), and buffer space proportional to the equivalent capacity (27). Typically, buffer space is thought of as a parameter in the equivalent capacity calculation, while we are treating it as a variable (along with bandwidth), for which we solve a system of two equations in two unknowns. That is, we solve (25) and (27) for BW_{ijk} and B_{ijk}

$$\begin{aligned}\mathrm{BW}_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)&\\ =\mathrm{BW}:\frac{\mathrm{BW}}{\mathrm{NCMAX}_{ijk}}&\\ =c\left(\mathrm{NCMAX}_{ijk},\bar{d}_{ijk}\mathrm{BW}_{ijk},\overline{\mathrm{PLOSS}}_{ijk}\right).& \quad (28)\end{aligned}$$

Going back to Fig. 2 which shows equivalent capacity as a decreasing function of buffer space and the number of connections served, the optimality property (27) is illustrated with a linearly increasing function of buffer allocation, where both sides of the expression have been divided by NCMAX_{ijk}. As the number of connections served increases, the slope of the line becomes less steep. Based on this simple graphical analysis, the system of equations (25) and (27) must have a unique solution and be decreasing in terms of NCMAX_{ijk}. Note that if there are no smoothing effects of multiplexing gains, as with constant bit-rate traffic ($R_i = r_i$), then the allocation per connection is simply a constant value. In all other cases, the total allocation must, therefore, be increasing, but the allocation per connection may be decreasing, depending on the properties of the service class. This shows economies of scale in the optimal resource allocation.

Using another optimality property from Theorem 1, the call-blocking constraint will be binding by (24). We can solve for the the optimal arrival rate λ_{ijk} for a given NCMAX_{ijk}, using (1)

$$\rho^*_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)=\rho:\overline{\mathrm{BP}}_{ijk}=\frac{\dfrac{\rho^{\mathrm{NCMAX}_{ijk}}}{\mathrm{NCMAX}_{ijk}!}}{\sum_{n=0}^{\mathrm{NCMAX}_{ijk}}}\frac{\rho^n}{n!} \quad (29)$$

$$\lambda_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)=\frac{\rho^*_{ijk}}{T_{ijk}}. \quad (30)$$

The arrival rate λ_{ijk} is determined by the limit on call blocking. We must first calculate the maximum traffic intensity for NCMAX_{ijk}, given in (29), and then calculate the arrival rate using (30). Because the resource allocation tables are independent of demand, we have not yet related the arrival rate to a price. The bid tables, presented in Section IV, will relate the price p_{ijk} required to produce the desired arrival rate λ_{ijk}, as well as the marginal valuation at the volume of service provided NCMAX_{ijk}.

2) Resource Allocation Tradeoffs Between Service Classes: The optimality properties, from Section III-A-1, reduce the problem to choosing the optimal values for NCMAX_{ijk}. For now, consider a *linear relaxation* to (PNet), eliminating the integrality constraints on NCMAX_{ijk}. For solutions satisfying the optimality properties given by (24) – (27), the Karush–Kuhn–Tucker necessary conditions are trivially satisfied, with the following exception, which results from differentiating the set of constraints (20):

$$\frac{\partial\mathrm{REV}_{ijk}}{\partial\mathrm{NCMAX}_{ijk}}=\sum_{x=1}^{\mathrm{NS}}v_x\frac{\partial\mathrm{BW}_{ijk}}{\partial\mathrm{NCMAX}_{ijk}}IVP_{ijkx},$$
$$1\le i\le\mathrm{N},1\le j\le\mathrm{NS},1\le k\le\mathrm{NS}. \quad (31)$$

Recall that IVP_{ijkx} in (31) was defined as an indicator parameter in the problem (PNet). For each service class (i,j,k), we divide both sides of (31) by $\partial\mathrm{BW}_{ijk}/\partial\mathrm{NCMAX}_{ijk}$, and simplify the necessary optimality conditions

$$u_{ijk}=\sum_{x=1}^{\mathrm{NS}}v_xIVP_{ijkx},$$
$$1\le i\le\mathrm{N},\ 1\le j\le\mathrm{NS},\ 1\le k\le\mathrm{NS}. \quad (32)$$

The economic interpretation of (32) is as follows. Any service class must yield a marginal return per unit of bandwidth equal to the sum of the marginal values for all switches (v_x for switch x) along the route for the given class (i,j,k).

Incorporating the integrality requirement in NCMAX_{ijk}, we define marginal values per unit of bandwidth for increasing or reducing the number of connections in the solution, u^+_{ijk} and u^-_{ijk}, respectively

$$u^+_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)=\frac{\mathrm{REV}_{ijk}\left(\mathrm{NCMAX}_{ijk}+1\right)-\mathrm{REV}_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)}{\mathrm{BW}_{ijk}\left(\mathrm{NCMAX}_{ijk}+1\right)-\mathrm{BW}_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)} \quad (33)$$

$$u^-_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)=\frac{\mathrm{REV}_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)-\mathrm{REV}_{ijk}\left(\mathrm{NCMAX}_{ijk}-1\right)}{\mathrm{BW}_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)-\mathrm{BW}_{ijk}\left(\mathrm{NCMAX}_{ijk}-1\right)}. \quad (34)$$

The marginal valuations, u^-_{ijk} from (33) or u^+_{ijk} from (34), are simply the changes in expected revenue divided by the change in bandwidth allocation, for an increase or decrease of one in NCMAX_{ijk}. Note that by definition, $u^+_{ijk}(\mathrm{NCMAX}_{ijk})$ equals $u^-_{ijk}(\mathrm{NCMAX}_{ijk}+1)$.

Discrete approximations of the continuous necessary optimalty conditions (32) are

$$u^+_{ijk}\left(\mathrm{NCMAX}_{ijk}\right)\le\sum_{x=1}^{\mathrm{NS}}v_xIVP_{ijkx}\le u^-_{ijk}\left(\mathrm{NCMAX}_{ijk}\right),$$
$$1\le i\le\mathrm{N},1\le j\le\mathrm{NS},1\le k\le\mathrm{NS}. \quad (35)$$

In stating (35), we assume that marginal revenue is decreasing in NCMAX_{ijk}, i.e., $u^-_{ijk}>u^+_{ijk}$. By (35), it is not profitable to change the bandwidth allocated to any service class. The marginal value from an increased allocation is less than the sum of marginal values of bandwidth at switches along the path.

IV. Network Auction Algorithm

The search procedure for the global network solution seeks to identify the marginal valuations of bandwidth at all switches v_j that will maximize expected revenue. Based on the optimality conditions (24)–(26), it is clear that the search will have to test and adjust marginal values at each switch until a solution is found that allocates bandwidth fully at each switch in order to satisfy the optimality conditions.

A. Storage of Problem Data in Bid Tables

We can exploit the first set of necessary optimality properties (24)–(27) to simplify the search space for the the problem (PNet). These properties dictate that optimal allocations of BW_{ijk} can be calculated from (28) based on the values of NCMAX_{ijk}. There is a unique arrival rate of connection requests, λ_{ijk}, associated with NCMAX_{ijk}, according to the call-blocking property (29). In turn, prices p_{ijk} are related to values of NCMAX_{ijk} through the arrival rate λ_{ijk}, given by the demand functions (1). This gives us a reduced space of problem data, where all other problem variables are calculated as functions of NCMAX_{ijk}, which are integer-valued variables. Furthermore, we have defined marginal values of service classes u^-_{ijk} and u^+_{ijk} for any value of NCMAX_{ijk} in (33) and (34).

We can calculate all the variables referred to above off-line and summarize the search space of the problem in tables indexed by the integer-valued variables NCMAX_{ijk} (Table I). We call Table I a *bid table* because the marginal values u^+_{ijk} (or u^-_{ijk}) represent the maximum amount a rational agent would bid (or accept) per unit of bandwidth to add (or remove) a connection of service class (i,j,k). Note that the bid table contains much less information than all the feasible values of NCMAX_{ijk}, BW_{ijk}, B_{ijk}, and p_{ijk}, thus, making it much faster to solve (PNet).

B. Retrieval of Problem Data From Bid Tables

For any given set of marginal bandwidth valuations v_j, we look up the "bid" and the associated number of connections and resource allocations. Strictly speaking, we select NCMAX_{ijk} for a given v_j according to the following rule:

$$\mathrm{NCMAX}_{ijk}(v_1, v_2, \ldots, v_{\mathrm{NS}}) = \max\left(\mathrm{NCMAX}_{ijk} : u^+_{ijk} < \sum_{i=1}^{\mathrm{NS}} v_x IVP_{ijkx} < u^-_{ijk}\right). \tag{36}$$

We search from the bottom of the bid tables and take the first (and largest) value of NCMAX_{ijk} for which $u^+_{ijk} < v_j < u^-_{ijk}$. For relatively small values of NCMAX_{ijk}, the u^-/u^+ values may be increasing due to large multiplexing gains. Since we are interested in revenue maximization, (36) selects the largest value of NCMAX_{ijk} which satisfies the optimality property (35). The simplest way to find this value is by looking up the bid table data starting from the bottom of the table. We will now present a method called the *auction algorithm* for ensuring that the chosen NCMAX satisfies the optimality conditions.

C. Bounds on Marginal Value of Bandwidth

For convenience in describing the search procedure, we define bounds on the *optimal set of marginal valuations* of bandwidth for all switches, as follows:

$$\{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{\mathrm{NS}}\} < \{v^*_1, v^*_2, \ldots v^*_{\mathrm{NS}}\} < \{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{\mathrm{NS}}\} \tag{37}$$

TABLE I
DEFINITION OF BID TABLE

$NCMAX_{ijk}$	BW_{ijk} eq. (28)	B_{ijk} eq. (27)	λ_{ijk} eq. (30)	p_{ijk} eq. (1)	u^+_{ijk} eq. (33)	u^-_{ijk} eq. (34)
1	$BW_{ijk}(1)$	$B_{ijk}(1)$	$\lambda_{ijk}(1)$	$p_{ijk}(\lambda_{ijk}(1))$	$u^+_{ijk}(1)$	$u^-_{ijk}(1)$
2	$BW_{ijk}(2)$	$B_{ijk}(2)$	$\lambda_{ijk}(2)$	$p_{ijk}(\lambda_{ijk}(2))$	$u^+_{ijk}(2)$	$u^-_{ijk}(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮

where

v^*_j optimal marginal valuation at switch j;
$\underline{v}_j$ lower bound on the value of v^*j at switch j;
$\bar{v}_j$ upper bound on the value of v^*j at switch j.

Clearly, using the bid table and (37), we can guarantee that a *set of marginal values* is either an upper bound or a lower bound on an optimal set of marginal values

$$\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NS}} \sum_{k=1}^{\mathrm{NS}} \mathrm{BW}_{ijk}(\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{\mathrm{NS}}) IVP_{ijkx} > \overline{\mathrm{BW}}_x \forall x \tag{38}$$

$$\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NS}} \sum_{k=1}^{\mathrm{NS}} \mathrm{BW}_{ijk}(\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{\mathrm{NS}}) IVP_{ijkx} < \overline{\mathrm{BW}}_x \forall x. \tag{39}$$

A set of marginal values that undervalues the bandwidth at all switches results in an overassignment of bandwidth at all switches and is infeasible. Thus, we can use such a set of marginal values as a *lower bound* on the set of optimal marginal values. Likewise, a set of marginal values which overvalues the bandwidth at all switches will be feasible and can be used as an upper bound on an optimal set of marginal values. Note that while we offer bounds on the optimal value of v^*_j in (37), we claim only to be seeking bounds for a local optimal solution to the problem (PNet), which is nonconvex and contains integer-valued variables.

D. Search Procedure

Our search method begins with arbitrary bounds on the optimal marginal valuation of bandwidth at each switch which satisfy (38) and (39). In each iteration, we decrease the Euclidean distance between the two sets of bound and change the direction of the line segment between the two sets of bounds in the marginal value space. When the upper and lower bounds are very near to each other, and we have a feasible solution that fully assigns capacity at all switches, we terminate the search and obtain a near-optimal solution, where capacity is fully assigned and revenue is very close to optimal. The infeasible solution given by the lower bounds on marginal value of bandwidth at each switch will also provide a bound on the distance from a local optimal solution.

The search algorithm is as follows.

```
Step 1 Initialization: For initialization, we simply
 require an initial set of bounds on the optimal
 marginal values as defined in (38) and (39).
```

Step 2 *Identifying Undervalued (and Overvalued) Bandwidth*: We wish to identify the best feasible and worst infeasible solutions that lie along the line segment between the current upper and lower bounds on marginal value. The bid table data is calculated according to optimality properties on resource assignments, and the constant elasticity of demand model implies higher revenue for lower prices and more connections provided. Therefore, the *best feasible* solution along the line segment between the upper and lower bounds on marginal value is where the largest number of connections are provided by assigning the maximum possible capacity. This, in turn, is obtained at the lowest feasible marginal values. Similarly, the *worst infeasible* solution along the line segment is found with the highest marginal values along that line, for which the capacity assignment is infeasible at every switch. We solve the following line search problems to obtain these two cases:

(P-*feasible*)

$$\mathrm{Max}\alpha \tag{40}$$

subject to

$$\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}}\mathrm{BW}_{ijk} \cdot (\bar{v}_1-\alpha m_1, \bar{v}_2-\alpha m_2, \ldots, \bar{v}_{\mathrm{NS}}-\alpha m_{\mathrm{NS}})\, IVP_{ijkx} < \overline{\mathrm{BW}}_x, \quad \forall x. \tag{41}$$

(P-*infeasible*)

$$\mathrm{Max}\beta \tag{42}$$

subject to

$$\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}}\mathrm{BW}_{ijk} \cdot (\underline{v}_1+\beta m_1, \underline{v}_2+\beta m_2, \ldots, \underline{v}_{\mathrm{NS}}+\beta m_{\mathrm{NS}})\, IVP_{ijkx} > \overline{\mathrm{BW}}_x \forall x \tag{43}$$

where

$$m_j = \frac{\bar{v}_j - \underline{v}_j}{\sum_{x=1}^{\mathrm{NS}} \bar{v}_x - \underline{v}_x} \tag{44}$$

$$0 < \alpha, \quad \beta < \sum_{x=1}^{\mathrm{NS}} \bar{v}_x - \underline{v}_x. \tag{45}$$

The line segment between the current set of lower bounds and upper bounds is given by (45). The maximum revenue feasible solution along this line is given by the set of marginal values $\bar{v}_j - \alpha^* m_j$. The minimum revenue infeasible solution is given by $\bar{v}_j + \beta^* m_j$. We will use these solutions in the next step of the algorithm to determine which bounds should be changed before the next iteration of the algorithm.

Step 3 *Adjusting Lower and Upper Bounds on Marginal Value*: Along the line segment between the upper and lower bounds, we wish to identify the switch for which the bandwidth which is most undervalued. That is, the switch (or switches) for which marginal value of bandwidth is "too low" at the solution α^* to the revenue maximization problem (P-*feasible*), yielding the most assigned bandwidth. Similarly, we wish to select the switch (or switches) for which bandwidth is most overvalued, or the switch for which the lower bound is "too high" at the solution β^* to the problem (P-*infeasible*), yielding the least assigned bandwidth. These values are defined mathematically as follows:

$$x_{\mathrm{undervalued}} = \operatorname*{argmin}_x \sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}}\mathrm{BW}_{ijk} (\bar{v}_1-\alpha^* m_1, \ldots, \bar{v}_{\mathrm{NS}}-\alpha^* m_{\mathrm{NS}})\, IVP_{ijkx} \tag{46}$$

$$x_{\mathrm{overvalued}} = \operatorname*{argmax}_x \sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}}\mathrm{BW}_{ijk} (\underline{v}_1+\beta^* m_1, \ldots, \underline{v}_{\mathrm{NS}}+\beta^* m_{\mathrm{NS}})\, IVP_{ijkx}. \tag{47}$$

The determination of which switches are over- or undervalued in terms of their marginal values of bandwidth is the same as simply choosing the switch with the highest assigned bandwidth from the solution to (P-*feasible*) and that with the lowest assigned bandwidth to (P-*infeasible*). For example, if one switch is 100% assigned at the solution to (P-*feasible*) while all others are less than 50% assigned, the marginal value of bandwidth at the fully assigned switch is relatively too low for an optimal solution, since capacity should be fully assigned at all switches. Note that ties are permitted so that there may be more than one under- or overvalued switch.

To effectively raise the marginal value of bandwidth when solving (P-*feasible*), we will increase the lower bound on marginal value at that switch. When we then repeat the line search to solve (P-*feasible*) in the next iteration, a higher marginal value will result at that switch relative to the other switches. Similarly, we will decrease the upper bound on marginal value for the switch that has the lowest infeasible assignment at the optimal solution to (P-*infeasible*), e.g., lower the upper bound at a switch with 101% assignment when all others are $\sim 150\%$ assigned. This effectively lowers the marginal value calculated in the nest iteration where we solve (P-*infeasible*) again.

$$\bar{v}_{x_{\mathrm{overvalued}}} := \bar{v}_{x_{\mathrm{overvalued}}} - \alpha^* m_{x_{\mathrm{overvalued}}} \tag{48}$$

$$\underline{v}_{x_{\mathrm{undervalued}}} := \underline{v}_{x_{\mathrm{undervalued}}} + \beta^* m_{x_{\mathrm{undervalued}}}. \tag{49}$$

We raise the lower bound on marginal value for the undervalued switch to the value given by the solution to (P-*infeasible*), as given in (49). Similarly, we decrease the upper bound on marginal value for the overvalued switch by taking the value in the solution to (P-*feasible*), as given by (48). Note the use of an assignment operator, ":=" in (49) and (48). The value to the left of ":=" is the new value being calculated, while the value to the right of ":=" is based on the previous quantities.

Step 4 *Termination Test*: When the line search between the bounds yields a solution to (P-*feasible*) where capacity is fully assigned at all switches, we terminate the algorithm, and assign the values found in the line search problem (P-*feasible*) as the solution.

If

$$\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}}\mathrm{BW}_{ijk} \cdot (\bar{v}_1 - \alpha^* m_1, \ldots, \bar{v}_{\mathrm{NS}} - \alpha\alpha^* m_{\mathrm{NS}}) IVP_{ijkx} \approx \overline{\mathrm{BW}}_x, \quad \forall x \tag{50}$$

then

$$v_j^* = \bar{v}_j - \alpha^* m_j, \quad \forall j. \tag{51}$$

Terminate the search.

Else

Go to Step 2.

There are a couple of things to note about this iterative algorithm. First, the algorithm is really a *bisection search* as described in Step 3, where the gap between the lower and upper bounds reduces at every step. It is well known that bisection searches are guaranteed to *converge*. We also show this numerically (see Fig. 4). Next, because our search space is contained in the bid tables described earlier, which are indexed by integer variables NCMAX_{ijk}, our search space is "lumpy." As such, "fully assigned" means an arbitrarily chosen level of capacity utilization such as 99% assigned capacity at every switch. This is why we use the notation approximately equal, $\approx$, rather than requiring strict equality, $=$.

The final line search problem (P-*infeasible*) yields an upper bound on the objective value. That is, for the local optimum given by (51) and the solution to (P-*infeasible*), we can calculate the expected revenues and state the duality gap for the local near-optimal solution is given as follows:

$$\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{NS}}\sum_{k=1}^{\mathrm{NS}} \left(\mathrm{REV}_{ijk}\left(\underline{v}_1 + \beta^* m_1, \ldots, \underline{v}_{\mathrm{NS}} + \beta^* m_{\mathrm{NS}}\right) - \mathrm{REV}_{ijk}\left(\bar{v}_1 + \alpha^* m_1, \ldots, \bar{v}_{\mathrm{NS}} + \alpha^* m_{\mathrm{NS}}\right)\right) > 0 \tag{52}$$

V. Example Problem

We will now present a simple real-world problem with two service classes and show how to set optimal prices and resource allocations using the model presented in Section III along with the search method presented in Section IV.

A. Network Structure and Service Classes

Consider a single network offering voice and video connection to users via a set of five switches. We consider a bidirectional ring network (Fig. 3). The voice connections can be either local, i.e., routed through a single switch, or long distance to anywhere in the network, i.e., routed from any origin to any destination switch. The video connections originate from a single switch (labeled "1" in the figure), where a video server is available and may be routed to any switch in the network, including the originating node. The connections are all routed using the minimum-hop routing.

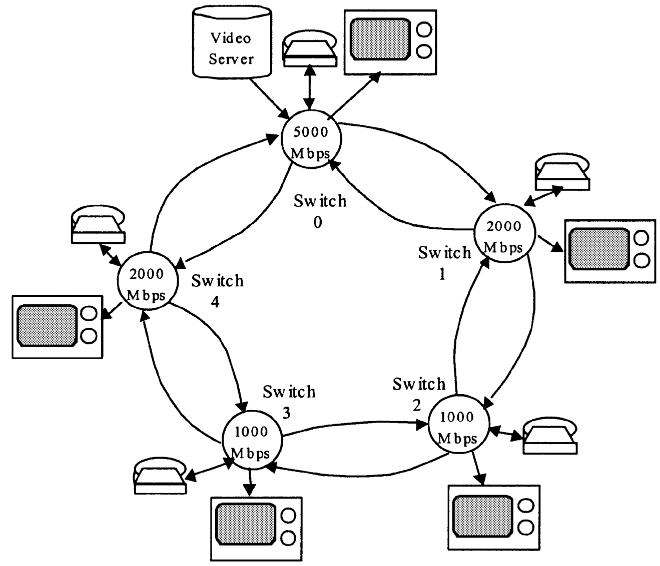

Fig. 3. Two service networks with a single video server.

Switch capacities are chosen as follows. Switch "0" is included in the path for every video connection and has 5000 Mb/s capacity, or five times as much capacity as switches "2" and "3," which are each included in one fifth of the paths for video connections. Similarly, switches "1" and "4" have two times the capacity of "2" and "3." The network services offered are voice and video. The service class definitions are given in Table II, which also provides traffic data.

Voice service is a reflection of traditional voice service, which does not have a high peak bandwidth, and has relatively short periods of bursts. In terms of QoS, relatively high loss rates of packetized voice may be acceptable, but large delays cannot be tolerated. Voice connections are typically of short duration (e.g., $T_{\mathrm{voice}} = 3.0$ min), which results in a lower average number of connections in use for any arrival rate of requests, as defined by (3). Video is intended to reflect bursty sources of data traffic, where the bursts, such as action scenes, can be a high peak and continue for prolonged periods. The peak and mean data rates are defined conservatively, based on MPEG-1 trace data in [29]. The length of the idle/busy period is chosen based on a suggestion in [34] that video sessions see busy periods in the order of 10 s. For users with buffer space at the client end, a modest delay in the transmission of packets is acceptable. Some packet loss can be concealed, but a higher degree of reliability is required than for voice.

We assume constant elasticity of demand for both services, with the elasticity values taken from [18] (Table III). Note that the elasticity of video is higher than that of voice. This means that the revenue associated with video connections increases

TABLE II
SERVICE CLASS DEFINITIONS. DATA TRAFFIC PARAMETERS AND OTHER PARAMETERS

(a) Data Traffic Parameters

	Mean. Rate (Mbps)	Peak. Rate (Mbps)	Average Burst (s)
Service Class	r_i	R_i	b_i
Voice	0.032	0.064	1.0
Video	1.0	10.0	10.0

(b) Other Parameters

	Packet Loss Probability	Blocking Probability	Delay (s)	Average Holding Time (s)
Service Class	$\overline{PLOSS}_i$	$\overline{BP}_i$	$\overline{d}_i$	T_i
Voice	10.0×10^{-5}	0.01	0.0	3.0
Video	10.0×10^{-9}	0.01	5.0	30.0

TABLE III
DEMAND FOR SERVICES

Service Class	Demand
Voice ($\forall j,k$)	$\lambda_{voice,jk} = 1.0p^{-1.05}$
Video ($j = 1$)	$\lambda_{video,jk} = 2.0p^{-2.50}$
Video ($j \neq 1$)	0

more rapidly as the price is lowered than for voice connections. We have chosen the scaling factor in the functions to be 1 and 2, respectively. This means that at the price 1 (or a normalized price), the arrival rate of requests for video is twice as high, which reflects the higher value of a connection with video than that with simply voice.

B. Illustration of Solution Algorithm

Figs. 4 and 5 provide details of the search steps when we implemented the algorithm for the example problem above. As we can observe, the search starts with lower and upper bound values at vectors **0** and **1**, respectively, and, through a number of bisection steps, converges to the "best" (i.e., near-optimal) solutions at iteration 8. All other variables also converge. To clarify exactly how the tight bounds are generated by the search method, we explain the step through an iteration in the Fig. 4 example, before continuing with the interpretation of the solution to the same example.

Single Iteration of the Search Algorithm: At iteration "0", Fig. 4 shows the starting upper and lower bounds on node marginal values. Recall from (32) that the marginal revenue per unit of bandwidth (u^+) for a given service class must offset the sum of marginal value(s) of bandwidth (v) at every along the path assigned to the service class. At each estimate of node marginal value, bandwidth assignments, etc., will be obtained from the bid table. Thus, at the upper bounds on node marginal values, resource assignments will be very small and *feasible*. Conversely, at the lower bounds there will be *infeasible* overassignment of resources. This is shown in Fig. 5.

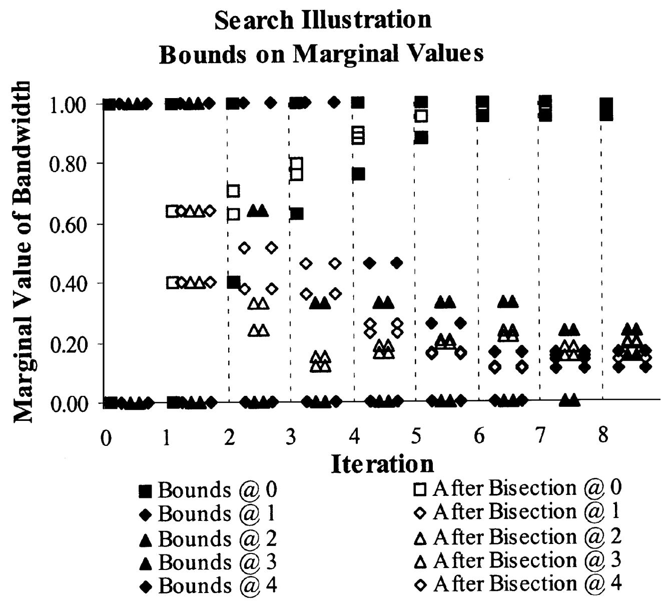

Fig. 4. Search algorithm steps: Convergence of marginal values.

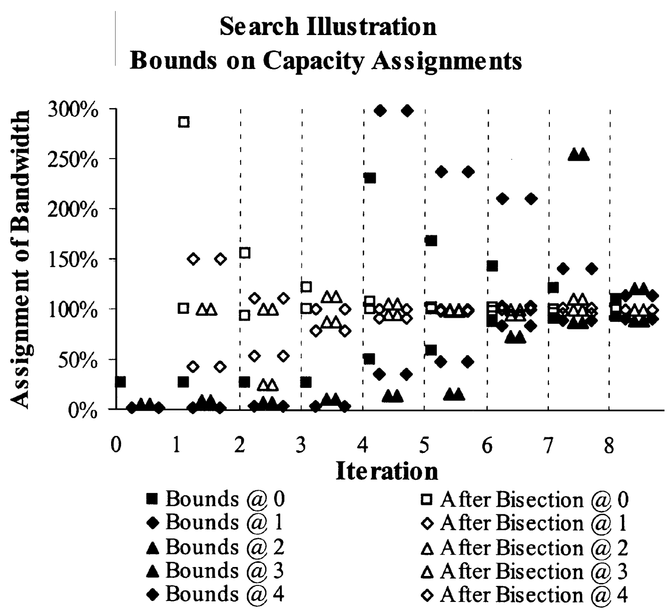

Fig. 5. Search algorithm steps: Convergence of capacity assignments.

At the initial upper bounds (1,1,1,1,1), bandwidth assignments at each node are (26.42%, 2.66%, 5.32%, 5.32%, 2.66%) of available capacity. At the lower bounds on node marginal values (0, 0, 0, 0, 0), the assignments are (899.32%, 988.73%, 1137.74%, 1137.74%, 988.73%) of available capacity. (These are not shown because they are outside the scale of Fig. 5).

At iteration "1," a *bisection* along the line segment joining (0, 0, 0, 0, 0) and (1, 1, 1, 1, 1) is obtained. If feasibility is not important, the maximum distance found along that line segment yields node marginal values to be (0.40, 0.40, 0.40, 0.40, 0.40); see Fig. 4. In this case, the bandwidth assignments are (285.73%, 150.39%, 100.01%, 100.01%, 150.39%) of capacity (shown in Fig. 5). When feasibility has to be maintained, the bisection search yields node marginal values of (0.64, 0.64, 0.64, 0.64, 0.64), for which the bandwidth assignments are (99.98%, 43.04%, 8.10%, 8.10%, 43.04%); see Fig. 5.

Given the broad range of capacity assignments (from 8.10% to 99.98% of node capacity in the feasible case and from 100.01% to 285.73% of node capacity in the infeasible case), we have an intuition that the bounds on the node marginal value either underassigns and overassigns bandwidth at certain nodes in a relative sense. Node 0 with an infeasible assignment of 285.73% of capacity in this iteration is most overassigned, or what we now call "undervalued." Therefore, we raise the marginal value for node 0 relative to all the other nodes by resetting the lower bounds from (0, 0, 0, 0, 0) to (0.40, 0, 0, 0, 0). Likewise, the assignment at the tightest feasible solution is most underassigned or "overvalued" at nodes 2 and 3, where only 8.10% of available capacity is included in the current feasible solution. Therefore, we lower the relative marginal value of nodes 2 and 3 by resetting the upper marginal value bounds from (1, 1, 1, 1, 1) to (1, 1, 0.64, 0.64, 1). The new bounds are passed to the algorithm for the next iteration and define the bounds at the beginning of iteration 2 in Fig. 4. The bisection search along the line segment between bounds on marginal values proceeds as above until convergence.

C. Solution Interpretation

The solution is found when the algorithm converges to a set of marginal values on individual node bandwidth, which occurs after eight iterations in Fig. 4. The final marginal values of the nodes selected for the near-optimal solution, i.e., the final solution to (P-*feasible*), as discussed in Step 4 of Section IV, are (0.97, 0.14, 0.20, 0.20, 0.14). The expected revenue at this solution is 11 046.03. The duality gap, calculated by comparing the solution to (P-feasible) with the solution to (P-infeasible), is roughly 0.24%; i.e., the best feasible solution we found was *within* 0.24% of the optimal solution. The bandwidth assignments at every switch are in excess of 99%; the criteria we used for termination in S4, i.e., the optimality conditions, are, more or less, satisfied. Given the marginal values above, the optimal prices are simply looked up in the bid tables and are given in Table IV.

The prices reflect relative scarcity of bandwidth at switch 1, due to the location of the video server. Relative to switch 1, local (same origin and destination) voice connections are priced inexpensively elsewhere in the network. The long-distance connections must yield much higher marginal values than local connections to be profitable, since the long-distance connections must outweigh the marginal value of local bandwidth all along the path of the connection. High marginal values correspond to higher prices and relatively smaller allocations. The video connections at switch 1 are the least expensive. Video connections at switch 1 use only resources at switch 1, while all other video connections must use this resource plus more resources at other switches. Consequently, the volume of video connections at switch 1 must also be the highest anywhere in the network. However, the cheaper video service comes at the price of expensive voice service through this switch, because the voice connections over this switch must use resources with a very high marginal value relative to the other switches.

TABLE IV
OPTIMAL PRICE PER UNIT TIME FOR A CONNECTION OF SERVICE (i, j, k)

Service Type (i)	Origin (j)	Destination (k) 1	2	3	4	5
Voice	1	0.76	0.88	1.06	1.06	0.88
	2	0.88	0.10	0.24	0.40	1.00
	3	1.06	0.24	0.14	0.29	0.40
	4	1.06	0.40	0.29	0.14	0.24
	5	0.88	1.00	0.40	0.24	0.10
Video	1	1.65	1.91	2.32	2.32	1.91

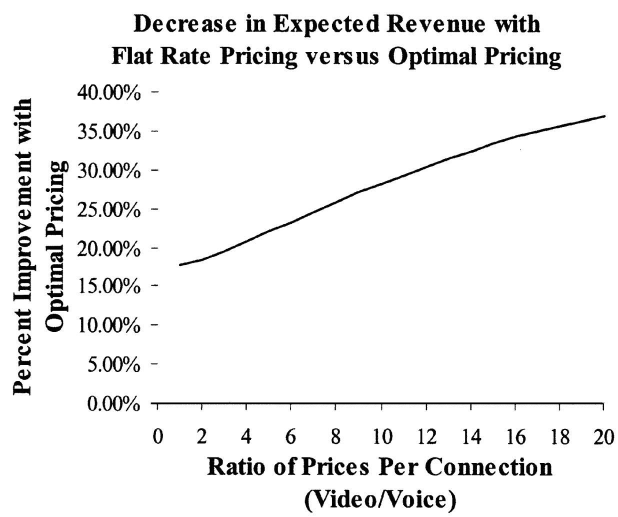

Fig. 6. Comparison of per-hop pricing with optimal pricing.

D. Comparison to Flat-Rate Pricing Approaches

To evaluate the usefulness of our approach, we compared optimal pricing with two flat-rate pricing approaches, per-hop and per-connection pricing. In the current U.S. voice market, pricing per connection regardless of origin and destination is very popular. Pricing-per-hop is another flat-rate pricing scheme that has been proposed, and mimics the use of prices proportional to link use.

For both of these pricing approaches, we need only choose a price for each service type in the network, a total of N prices. Because we have only two service types, we assume a ratio between the voice price and video price and then simply solve for the smallest voice price that achieves feasibility at all switches. We will use the bid tables with the resource allocations satisfying optimality properties and find the smallest feasible flat rate price for voice, subject to the assumed ratio between voice and video prices. Results are summarized in Figs. 6 and 7.

The optimal pricing mechanism provided greater revenue than either flat-rate pricing approach. The revenue increase was upward of 20% in all cases. In general, pricing per hop produces higher revenue than pricing per connection in the reasonable ranges for relative video and voice prices. In this range, the expected revenue is relatively insensitive to the ratio between the prices for the two services. The pricing per connection shows the highest profit when voice and video connections are charged at the same rate. The prices essentially fill the network bandwidth with video connections, which are

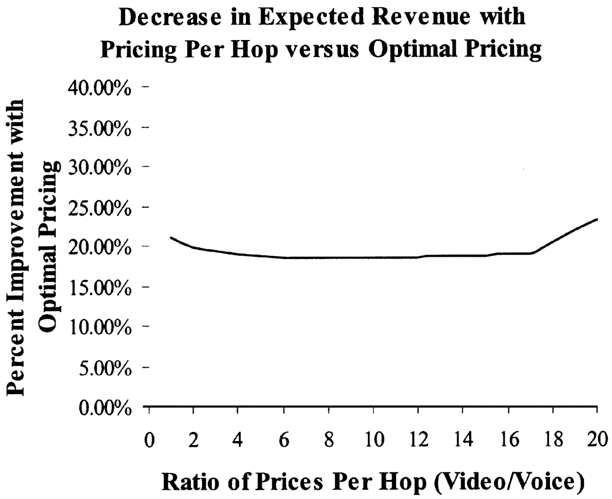

Fig. 7. Comparison of per-connection pricing with optimal pricing.

highly elastic and generate significant revenue as prices are lowered. However, users are unlikely to accept a voice rate as high as would be necessary for a video connection. For this reason, the solutions with larger ratios between the video price and voice price are probably more practical. These solutions earn significantly less revenue.

VI. Concluding Remarks

This paper presents a mathematical programming model for optimal pricing, and bandwidth and buffer allocation for multiple services with QoS guarantees in a connection-oriented network. We also present a novel solution methodology. We show, using numerical experiments, that flat-rate pricing (whether by connection or hop) is inferior to the multiservice pricing obtained from our model. Our model provides a very powerful mechanism for pricing multiple services in communications networks.

In order to implement our pricing scheme in a practical real-world setting, the first decision to be made is whether one uses a centralized architecture or a decentralized one. In either case, a processor has to use demand (i.e., connection request) information to make projections of future demand for the different types of services. In a centralized architecture, the master switch will directly get information about capacity availabilities at each switch. Using this and the demand projections, the master switch will calculate node marginal values for each switch, and will determine the prices to be set and the resources to be allocated for each service class, updating it at reasonable time intervals. The model presented in this paper is particularly useful in this setting.

In a decentralized architecture, *each switch* will also have to calculate the prices and resources needed for each class of service. In order to have globally optimal solutions, each switch will have to communicate information pertaining to itself to the rest of the switches. Depending on whether the communication is sent only to neighboring switches, or to all of the switches, the actual calculations at any point in time might differ. The model we have presented in this paper will still work, if all of the decentralized switches use it for pricing and resource allocation, but obtain perfect instantaneous information from the other switches on node marginal values and demand for bandwidth. However, if there is partial information and/or a significant time lag for the communications protocol to yield up-to-date perfect information, a simultaneous implementation of the model by all switches might not yield a globally optimal solution.

We are currently working on multiple service pricing mechanisms for decentralized networks. We are also working on obtaining good demand projections, which is a further limitation of this paper that assumes that all parameters of the stochastic arrival process is known. The beginning of such work was reported in [15].

Appendix
Optimality Theorem

We argue for each of the optimality properties in the theorem individually.

Property 1: Assume there is an optimal solution with prices and resource allocations such that the call-blocking constraint is nonbinding, for at least one service class

$$\mathrm{BP}_{ijk}\left(\lambda_{ijk}\left(p_{ijk}\right), T_{ijk}, \mathrm{NCMAX}_{ijk}\right) < \overline{\mathrm{BP}}_{ijk}. \tag{53}$$

It follows that for this class the network operator could allow a higher rate of connection requests and still satisfy the call-blocking constraint

$$\begin{aligned}&\exists\lambda'_{ijk}\left(p'_{ijk}\right) > \lambda_{ijk}\left(p_{ijk}\right)\\&\quad: \mathrm{BP}_{ijk}\left(\lambda'_{ijk}\left(p'_{ijk}\right), T_{ijk}, \mathrm{NCMAX}_{ijk}\right) \le \overline{\mathrm{BP}}_{ijk}.\end{aligned} \tag{54}$$

Demand is downward sloping, and marginal revenue with respect to prices is less than zero by assumption, so that the price corresponding to the higher rate of requests must be lower and revenue must be higher

$$\frac{d\lambda_{ijk}\left(p_{ijk}\right)}{dp_{ijk}} < 0, p'_{ijk} < p_{ijk} \tag{55}$$

$$\mathrm{BP}_{ijk}\left(\lambda_{ijk}\left(p_{ijk}\right), T_{ijk}, \mathrm{NCMAX}_{ijk}\right) = \overline{\mathrm{BP}}_{ijk}. \tag{56}$$

Therefore, for (53) an optimal solution cannot exist. The optimal solution must be such that call-blocking is binding [see (24)].

Assume there exists an optimal solution such that the allocation of bandwidth to a particular class is greater than the equivalent capacity for that class

$$c\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right) < \frac{\mathrm{BW}_{ijk}}{\mathrm{NCMAX}_{ijk}}. \tag{57}$$

There must exist a feasible allocation of less bandwidth, according to constraint (18) in (PNet) (the smallest feasible assignment is equal to equivalent capacity at the maximum permitted loss probability) for the particular class such that the QoS is still satisfied

$$\begin{aligned}&\exists\mathrm{BW}'_{ijk} < \mathrm{BW}_{ijk}\\&\quad: c\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right) = \frac{\mathrm{BW}'_{ijk}}{\mathrm{NCMAX}_{ijk}}.\end{aligned} \tag{58}$$

The reassignment given by (58), where equivalent capacity c_i is minimized by selecting the loss probability equal to its maximum permitted value, may make it possible to assign bandwidth for an additional connection, lower the price of such a class, and increase revenue. Even if admitting more connections is not possible due to the reassignment in (58), revenue cannot decrease from the reassignment. Therefore, the assumption that bandwidth is excess of effective bandwidth, (57), makes no sense and the optimal solution is invalid. We must have assigned bandwidth equal to equivalent capacity [see (25)].

Property 2: Assume that a solution is optimal and possesses the following property for at least one local service class, e.g., (x, j, j):

$$\left((\mathrm{NCMAX}_{xjj} + 1)\, c\left(\mathrm{NCMAX}_{xjj} + 1, \mathrm{B}_{xjj}, \overline{\mathrm{PLOSS}}_{xxjj}\right) + \sum_{\substack{i=1 \\ i \neq x}}^{\mathrm{N}} \sum_{k=1}^{\mathrm{NS}} \mathrm{NCMAX}_{ijk} c_{ijk}\left(\mathrm{NCMAX}_{ijk}, \mathrm{B}_{ijk}, \overline{\mathrm{PLOSS}}_{ijk}\right) \right) \leq \overline{\mathrm{BW}}_j. \tag{59}$$

For the particular service class (x, j, j) we can allocate resources for an additional connection and tolerate a higher arrival rate of requests at a lower price, while satisfying all constraints

$$\exists \lambda'_{xjj}\left(p'_{xjj}\right) > \lambda_{xjj}\left(p_{xjj}\right), p'_{xjj} < p_{xjj} \quad \text{: all constraints satisfied} \tag{60}$$

Similar to the above, a lower price results in increased revenue. Therefore, (59) is false and capacity must be fully allocated [see (26)].

We first assume that there exists an optimal solution that satisfies the following property for a single service class (i, j, k), which contradicts condition (27) in Theorem 1

$$\mathrm{B}_{ijk} \bar{d}_{ijk} < \mathrm{BW}_{ijk}. \tag{61}$$

For this service class, as we increase the buffer allocation B_{ijk}, the equivalent capacity c, or minimum bandwidth assignment per connection, falls with the increased buffer space

$$\frac{\partial c(\mathrm{B}_{ijk})}{\partial \mathrm{B}_{ijk}} < 0. \tag{62}$$

Therefore, there exist buffer and bandwidth allocations different from those given in (61), for which the bandwidth assignment is lower

$$\exists \mathrm{B}'_{ijk} > \mathrm{B}_{ijk}, \mathrm{BW}'_j = c_j\left(\mathrm{B}'_j\right) < \mathrm{BW}_j : \mathrm{B}'_{ijk} = \bar{d}_{ijk} \mathrm{BW}'_j. \tag{63}$$

If the difference in bandwidth is sufficiently large, there may be a revenue increasing solution, by (26), meaning the current solution cannot be optimal. On the other hand, if the bandwidth allocation does not yield an improvement in revenue, the revenue cannot be decreased by the reallocation described above. Therefore, the optimal solution must satisfy (27).

Acknowledgment

The authors would like to thank R. Guerin, Y. Korilis, and V. Chandru and the referees of this journal for comments and suggestions

References

[1] M. Aldebert, M. Ivaldi, and C. Roucolle, "Telecommunications demand and pricing structure: An econometric analysis," in *Proc. 7th Int. Conf. Telecommunications Systems: Modeling and Analysis*, Nashville, TN, Mar. 1999, pp. 255–267.

[2] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[3] D. Bertsekas, "Auction algorithms for network flow problems: A tutorial introduction," *Comput. Organizat. Applicat.*, vol. 1, pp. 7–66, 1992.

[4] U. N. Bhat and I. V. Basawa, *Queueing and Related Models*. New York: Oxford Univ. Press, 1992.

[5] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer networks: Motivation, formulation, and example," *IEEE/ACM Trans. Networking*, vol. 1, pp. 614–627, Dec. 1993.

[6] C. Courbetis, V. Siris, and G. D. Stamoulis, "Integration of pricing and flow control for available bit rate services in ATM networks," in *Proc. Globecom*, London, U.K., 1996, pp. 644–648.

[7] A. A. Economides and J. A. Silvester, "Multiobjective routing in integrated services networks: A game theory approach," in *Proc. IEEE INFOCOM*, 1991, pp. 1220–1225.

[8] R. J. Edell, N. McKeown, and P. P. Varaiya, "Billing users and pricing for TCP," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1162–1175, Sept. 1995.

[9] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, July 1991.

[10] R. J. Gibbens and F. P. Kelly. (1999) Resource pricing and the evolution of congestion control. [Online]. Available: http://www.statslab.cam.ac.uk/~frank/evol.html

[11] H. Jiang and S. Jordan, "The role of price in the connection establishment process," *Eur. Trans. Telecommun. Related Technol.*, vol. 6, no. 4, pp. 421–429, July/Aug. 1995.

[12] S. Jordan and H. Jiang, "A pricing model for high speed networks with guaranteed quality of service," in *Proc. IEEE Infocom*, Mar. 1996, pp. 888–895.

[13] F. P. Kelly, "Tariffs and effective bandwidths in multiservice networks," in *Proc. Int. Test Conf.*, 1994, pp. 401–410.

[14] F. P. Kelly, A. K. Maullo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, 1998.

[15] J. N. Keon, "Pricing in telecommunications networks offering multiple services and quality of service guarantees," Ph.D. dissertation, Univ. Pennsylvania, Systems Eng. Dept., Philadelphia, PA, Aug. 2000.

[16] Y. A. Korilis and A. A. Lazar, "On the existence of equilibria in noncooperative optimal flow control," *J. ACM*, vol. 42, pp. 584–613, May 1995.

[17] Y. A. Korilis, T. A. Varvarigou, and S. R. Ahuja, "Incentive-compatible pricing strategies in noncooperative networks," in *Proc. IEEE INFOCOM*, vol. 2, 1998, pp. 439–446.

[18] S. Lanning, D. Mitra, Q. Wang, and M. Wright, "Optimal planning for optical transport networks," *Philosph. Trans. Royal Soc. London A*, vol. 358, no. 1773, pp. 2183–2196, Aug. 2000.

[19] S. H. Low and P. Varaiya, "A new approach to service provisioning in ATM networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 547–553, Oct. 1993.

[20] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Networking*, vol. 7, pp. 861–865, Dec. 1999.

[21] S. H. Low, "Equilibrium bandwidth and buffer allocations for elastic traffics," *IEEE/ACM Trans. Networking*, vol. 8, pp. 373–383, June 2000.

[22] J. K. Mackie-Mason and H. Varian, "Pricing the internet," in *Public Access to the Internet*, B. Kahin and J. Keller, Eds. Cambridge, MA: MIT Press, 1995, pp. 269–314.

[23] J. K. Mackie-Mason and H. R. Varian, "Pricing congestible network resources," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1141–1149, Sept. 1995.

[24] J. K. Mackie-Mason and H. Varian, "Some economics of the internet," in *Networks, Infrastructure and the New Task for Regulation*, W. Sichel and D. L. Alexander, Eds. Ann Arbor, MI: Univ. Michigan Press, 1996, pp. 107–36.

[25] Y. Masuda and S. Whang, "Dynamic pricing for network service: Equilibrium and stability," *Managem. Sci.*, vol. 45, no. 6, pp. 857–869, 1999.

[26] H. Mendelson and S. Whang, "Optimal incentive-compatible priority pricing for the $M/M/1$ queue," *Oper. Res.*, vol. 38, no. 5, pp. 870–883, 1990.

[27] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of network services," *IEEE/ACM Trans. Networking*, vol. 8, pp. 171–184, Apr. 2000.

[28] H. G. Perros and K. M. Elsayed, "Call admission control schemes: A review," *IEEE Commun.*, vol. 34, pp. 82–91, Nov. 1996.

[29] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," in *Proc. 20th Annu. Conf. Local Computer Networks*, Minneapolis, MN, 1995, pp. 397–406.

[30] S. Shenker, "Service models and pricing policies for an integrated services internet," in *Public Access to the Internet*, B. Kahin and J. Keller, Eds. Cambridge, MA: MIT Press, 1995, pp. 315–337.

[31] P. Thomas, D. Teneketzis, and J. K. Mackie-Mason, "A market-based approach to optimal resource allocation in integrated-services connection-oriented networks," *Oper. Res.*, vol. 50, no. 4, pp. 603–616, 2002.

[32] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in wide-area ATM networks using effective bandwidths," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1081–1089, June 1995.

[33] G. de Veciana and R. Baldick, "Resource allocation in multiservice networks via pricing: Statistical multiplexing," *Comput. Networks ISDN Syst.*, vol. 30, pp. 951–962, 1998.

[34] J. Walrand and P. Varaiya, *High-Performance Communication Networks*. San Francisco, CA: Morgan Kaufmann, 1996.

[35] Q. Wang, J. M. Peha, and M. A. Sirbu, "The design of an optimal pricing scheme for ATM integrated services networks," *J. Electron. Publish., Special Issue Internet Economics*, vol. 2, no. 1, 1996.

Neil J. Keon (M'00) received the B.Comm. degree with joint honors in economics and finance from McGill University, Montreal, QC, Canada, in 1992, the M.Sc. degree in systems science from the University of Ottawa, Ottawa, ON, Canada, in 1995, and the M.S.E. and Ph.D. degrees in systems engineering from the University of Pennsylvania, Philadelphia, in 1999 and 2000, respectively.

Since 2000, he has been with the Cox School of Business, Southern Methodist University, Dallas, TX, where he is currently an Assistant Professor of information technology and operations management.

Dr. Keon is a member of the Institute for Operations Research and Management Science (INFORMS).

G. Anandalingam (M'85–SM'90) received the B.A. degree in electrical sciences from Cambridge University, Cambridge, U.K., in 1975 and the Ph.D. degree in operations research with economics from Harvard University, Cambridge, MA, in 1981.

He has been with the Institute for Systems Research, University of Maryland, College Park, since 2001, and is currently the Ralph J. Tyser Professor of Management Science at the Robert H. Smith School of Business, University of Maryland. From 1987 to 2001, he was the National Center Professor (Engineering School), and Professor of Operations and Information Management (Wharton School) at the University of Pennsylvania, Philadelphia. He has published more than 70 papers on telecommunications and networking issues, and a book on telecommunications network design to be published by Kluwer Scientific. He is an Associate Editor of *Operations Research, Management Science, Telecommunications Systems, Network and Spatial Economics*, and *Computers and Operations Research*. He is on the Advisory Boards of a number of information technology startup companies.

Dr. Anandalingam is a Senior Member of the Institute for Operations Research and Management Science (INFORMS).