Title

Connection establishment in high-speed networks

Journal
IEEE Journal on Selected Areas in Communications, 13(7)

Authors
Jordan, Scott
Jiang, Hong

Publication Date
1995-09-01

DOI
10.1109/49.414635

Peer reviewed

Connection Establishment in High-Speed Networks

Scott Jordan, *Member, IEEE,* and Hong Jiang, *Student Member, IEEE*

***Abstract*— The evolving view of connection establishment for connection-oriented services in high-speed networks such as ATM involves a contract negotiation process between a user agent and a network agent. The first stage consists of separate roles for the user and the network. The user agent must characterize the information streams that will be transmitted and the performance parameters that define the desired quality of service for that user. Similarly, the network agent must determine the network's resources and its capabilities to accommodate various mixes of service types. The second stage involves negotiations between multiple network and user agents, in which the parties agree to set up connections to transmit the agreed information streams in a manner to guarantee the agreed qualities of service, and at agreed prices. In this paper, we focus on these two stages that together form the connection establishment process. After this process, during the connection, the network must police the user to determine compliance with the information stream characteristics, and must implement flow control, service priority mechanisms and packet multiplexing disciplines as necessary to guarantee the quality of service.**

I. Introduction

EMERGING networks such as asynchronous transfer mode (ATM) will attempt to provide guaranteed performance to variable bit rate (VBR) services. Data networks have generally provided VBR services, but only on a best effort basis. Telephone networks have generally provided guaranteed performance, but only to circuit-switched services.

In ATM, user information streams are organized into small fixed-length packets called cells. These cells are sent through shared transmission lines and routed by switches with shared buffers. This resource sharing through cell-switching increases network utilization levels and gives the network the flexibility of providing potentially unlimited classes of service. However, it will unavoidably lead to congestion and degradation of quality of service (QoS) if no admission control or flow control is enforced to restrict the number of users and the amount of network resources consumed by each user.

The design of control mechanisms depends on the characterization of user information streams, the desired QoS, and network resource management. Allocation of bandwidth and buffers among different traffic streams with different QoS should be accomplished in an efficient and fair manner, especially when users' demand exceeds the network's capacity. In a centralized approach, the network can decide the most efficient way to allocate resources and then enforce its decision on individual users and on network controls at various network levels. However, this strategy requires that the network collect user information, such as traffic characteristics and user valuation of each service, to compute the optimal solution. This solution can result in heavy traffic to the central node, a large computation burden on the central node, and lack of reliability. Therefore, much research has recently been focussed on distributed approaches to resource allocation.

All the problems discussed above are increasingly considered as part of connection establishment. The evolving view of connection establishment for connection-oriented services involves a contract negotiation process between a user agent and a network agent. The first stage consists of separate roles for the user and the network. The user agent must characterize the information streams that will be transmitted and the performance parameters that define the desired QoS for that user. Similarly, the network agent must determine the network's resources and its capabilities to accommodate various mixes of service types. The second stage involves negotiations between multiple network and user agents, in which the parties agree to set up connections to transmit the agreed information streams in a manner to guarantee the agreed QoS, and at agreed prices. In this paper, we focus on these two stages that together form the connection establishment process. After this process, during the connection, the network must police the user to determine compliance with the information stream characteristics, and must implement flow control, service priority mechanisms and packet multiplexing disciplines as necessary to guarantee the QoS.

In Section II, we first review recent approaches to characterization of a user's information stream and of a user's desired QoS. Promising contributions include multiparameter characterizations of burstiness, burstiness curves, and effective bandwidth. Section III reviews recent approaches to determining a network's capabilities to simultaneously accommodate a mix of different service types and QoS. Such analysis has focussed on connection-level resource sharing and on call admission policies that differentiate services by resource usage and required performance. Third, in Section IV, we review recent proposals for a distributed negotiation process between multiple network and user agents. Groundbreaking proposals have recently discussed the role of service prices as a resource allocation mechanism.

In Section V, we explore ways to integrate various proposals from previous sections into a complete distributed connection establishment procedure which encourages network efficiency through optimal resource allocation. This involves forming linkages between a traffic model, QoS metrics, service discipline, resource management architecture, statistical multiplexing, admission control, user and network objectives, and resource allocation in a manner such that the individual

Manuscript received September 30, 1994; revised April 1, 1995. This work was supported by the Ameritech Foundation.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA.

IEEE Log Number 9413110.

0733–8716/95$04.00 © 1995 IEEE

parts interrelate. This discussion reveals several outstanding research problems, which we further explore in Section VI.

II. User Characterization

The quality of service of a call is often measured in terms of cell loss probability, delay jitter, and end-to-end delay. Different users can have very different valuations on the QoS specifications, varying from delay-sensitive voice applications to loss-sensitive data applications. However, user information streams are often not deterministic, and thus it is difficult to reserve resources for them. Therefore, characterization of individual traffic sources and the interaction among them is essential for understanding how the QoS can be delivered for individual users while maintaining high levels of network utilization. In this section, we briefly review some recent work on traffic source characterization.

A. Multiparameter Characterizations

Consider a single user's information flow into the network as a continuous stream of packets. Measure the instantaneous rate of this flow in packets per unit time. A simple way to describe relevant characteristics of this traffic stream is to specify a few statistics, such as peak rate, mean rate, bit rate variance, and/or maximum numbers of cells arriving during a period. A considerable body of research has shown that appropriate choice of statistics can provide significant information helpful to source characterization (cf. [1]). For instance, the ratio of peak rate to mean rate is often used as a measure of the burstiness of a traffic source. If a multiparameter characterized source is fed into a queue, however, it can be difficult to similarly characterize the output stream by a few parameters, or to analyze the resulting performance. In addition, this method occasionally fails to fully capture the nature of VBR traffic sources, and thus the admission control based on this method can cause either under-utilized or overloaded networks.

Random processes, such as Poisson processes and Markov chains, have long been used as more detailed traffic models. Recently, Markov-modulated Poisson processes (MMPP's) and Markov-modulated fluid flows (MMFF's) have garnered considerable attention as possible VBR traffic models. MMPP's model instantaneous packet arrivals as a Poisson process, while MMFF's assume constant arrivals during short time periods. Both models use a Markov chain to modulate the mean rate of packet arrivals over larger time periods. Recent research has had considerable success at choosing parameters for these models to capture the unique statistical characteristics of specific applications such as voice, video, and LAN traffic [2]–[5]. Furthermore, if each switch can also be described as a memoryless queueing system, then the output stream can often be similarly modeled, and the performance easily gauged. This approach, however, is often limited by the number of states allotted to describe a source, and can become cumbersome when multiple source types are present.

B. Effective Bandwidth

A recent approach that does not directly model the traffic source random process, but instead models a source's use of network resources at the user/network interface (UNI), involves the concept of effective bandwidth. A source is fed into a finite buffer served at a constant rate. If the loss is asymptotically exponential in the buffer length, then the source is said to have an effective bandwidth. The rate of the exponential decrease depends on the service rate and on the burstiness of the source.

The concept is usually used in reverse: in the range of small loss probabilities and large buffer lengths, a source must be served at a rate at least equal to its effective bandwidth in order to meet the corresponding loss criterion. The effective bandwidth is between the source's mean and peak rates, and is a function of the source's burstiness and of $\xi = (\log(\text{loss probability}))/(\text{buffer length})$.

Many traffic models have been found to have effective bandwidths. Hui [6], [7] first used a Gaussian approximation to compute the tail distribution of a buffer for a user information stream with burst arrivals. He then tightened the bound on the tail distribution using large deviation theory for the bufferless case. Kelly [8] proved the existence of effective bandwidth in an M/G/1 system with constraints on mean work load and in a GI/G/1 system with a constraint on tail probability. Gibbens [9] and Guerin [10] found expressions of effective bandwidth for on-off Markov fluid sources. They also noted that in some cases the accuracy of effective bandwidth is insensitive to buffer sizes provided the loss probability is very small, usually in the order of 10^9. Kesidis [11] obtained more general solutions to effective bandwidth for various traffic sources including constant rate, memoryless, discrete-time Markov, Markov fluid, and MMPP sources.

One nice property of many systems with effective bandwidths is that the effective bandwidth of multiplexed sources sharing the same buffer is equal to the sum of their individual effective bandwidths. Early results obtained by Anick [12] and Mitra [13] for statistically multiplexed Markov fluids showed that the tail distribution of the buffer can be approximated by the term with the largest negative eigenvalue of a matrix related to the source's generator. Based on these results, Elwalid [14] found effective bandwidths for general Markovian sources, which proved to be the maximal real eigenvalue of another matrix, determined by ξ and the Markov fluids, by solving an inverse eigenvalue problem. The additivity of effective bandwidth of multiplexed sources was proved using Kronecker algebra.

Effective bandwidth serves as a useful tool in allocating network resources to satisfy a source's QoS. Since maximum delay at a single node is given by buffer length divided by service rate, a source's loss criterion and single node delay criterion can be jointly satisfied by allocating appropriate buffer space and bandwidth. The elegance of effective bandwidth is that it breaks down a multidimensional problem into 1-D problems which only depend on the characterization of individual traffic sources and the parameter ξ. Furthermore, a buffer-bandwidth trade-off can be obtained by treating effective bandwidth as a function of buffer size while keeping other parameters fixed.

The additivity of effective bandwidth among traffic sources facilitates the negotiation process between a network and its users. If users submit the effective bandwidth of their traffic, the network can then deliver the required QoS by keeping the sum of the effective bandwidths of all the users below its capacity. Policing is necessary, however, to make sure every user transmits no more than what the user claims [15]. Effective bandwidth also provides a good base for pricing since it reflects the amount of bandwidth a service occupies. A pricing policy based on effective bandwidth is suggested in Jiang [36].

Effective bandwidth has a few limitations. First, the existence of effective bandwidth for more general traffic models is unknown. Second, its use is principally limited to the asymptotic regime of large buffers and low loss. Third, effective bandwidth does not allow statistical multiplexing of traffic sources with different QoS. This last limitation can be alleviated by multiplexing services with different QoS into separate groups at the price of some lost multiplexing gain [15].

C. Burstiness Curves

A second recent approach that does not directly model the traffic source random process is the burstiness curve. Cruz [16] studied the performance of deterministic fluids by using two parameters (σ, ρ) : ρ is the service rate and σ is the maximum buffer content if the traffic is fed into an infinite buffer served at rate ρ. By using a buffer of length σ and service rate ρ, no cell loss will occur and the delay jitter will be bounded by σ/ρ. Low [17], [18] extended Cruz's work to define a burstiness curve by treating σ as a function of ρ. The burstiness curve, for a message whose deterministic rate at time r is $m(r)$, is defined by

$$\sigma_m(\rho) = \sup_{0 \leq s \leq T} \int_s^T [m(r) - \rho] dr,$$

the maximal buffer content in a time interval T for the fluid served at rate ρ.

This curve thus gives an explicit trade-off between the buffer and bandwidth required to achieve no cell loss for a given message. For resource allocation purposes, therefore, $m(r)$ need not be known by the network. The user only needs to know an upper bound on $\sigma_m(\rho)$ and request some combination of buffer and bandwidth to satisfy this bound.

This approach has been extended to systems that can tolerate some loss by Wong [19]. The resulting characterization is $L_m(\sigma, \rho)$, the loss suffered by message $m(r)$ when fed into a buffer of length σ served a rate ρ.

Many traffic sources are amenable to such characterizations. However, many stochastic processes such as Poisson processes cannot be deterministically bounded by the two parameters. A compromise between the stochastic and deterministic approaches was proposed by Yaron [20], using an exponential bound for the probability distribution of the traffic. Chang [21] introduced the notion of a minimum envelop rate (MER) for deterministic sources by adding a subadditivity property to Cruz's model. This formulation allowed him to obtain a set of stabililty and multiplexing results. Furthermore, these results are also applicable to stochastic sources if the tails of their distributions can be bounded by decaying exponentials. The notion of MER for such sources was shown to be equivalent to effective bandwidth.

III. Network Characterization

In Section II, user traffic and QoS characterizations were reviewed. In this section, the network's role of determining its capabilities to simultaneously accommodate a mix of different service types and QoS will be reviewed. The setting for this task is initially given by the network's resource management architecture. Within this architecture, the network must first estimate any statistical multiplexing gains to determine its ability to accept new users. The network must then choose a connection access control strategy that maximizes an appropriate performance measure.

A. Resource Management Architecture

The current ATM network architecture for resource management contains two levels: virtual circuit (VC) and virtual path (VP) [22]. A virtual path is a group of connections sharing a common path from source to destination. Virtual circuits are the individual connections within a virtual path. Routing is performed on the virtual path level. When a call arrives, the network first checks whether a virtual path for the source and destination pair exists. If the virtual path exists, then the network verifies whether there is capacity within the VP for a new VC to accommodate this call. Once the call is admitted, no call processing is required at transit nodes and cells will be delivered in order. If the virtual path does not exist or there is no capacity for the VC within the existing VP, the network can either reject the call, request more resources if the virtual path exists, or create a new virtual path if no virtual path exists. By basing network resource control mechanisms (including routing, resource reservation, and congestion control) on the VP level (rather than the VC level), ATM call set up and processing tasks are quick and less costly.

Virtual circuits can be grouped into *circuit bundles* and share bandwidth and buffers based on three variables: path, QoS, and source type. Among the numerous combinations, we explicitly consider the following:

- *Circuit switching:* Each VC is allocated its own bandwidth and buffers.
- *VP/QoS/Type allocation:* All VC's with identical paths, source types, and QoS are statistically multiplexed.
- *VP/QoS allocation:* All VC's with identical paths and QoS are statistically multiplexed.
- *VP allocation:* All VC's with a common path are statistically multiplexed.
- *QoS allocation:* All VC's with identical QoS within a common trunk are statistically multiplexed.
- *Complete sharing:* All cells within a common trunk are statistically multiplexed.

Circuit switching, commonly used in telephone networks, results in a simple connection establishment procedure, but is inefficient if there is much VBR traffic. Complete sharing, commonly used in packet switched networks, is efficient, but lacks mechanisms to insure multiple QoS. VP/QoS/Type allocation represents a minimal amount of statistical multiplexing and could be achieved by grouping all video calls along a route into one VP, all voice calls into another, etc. The current ATM standard specifies that capacity will be allocated to VP's, not VC's, but does not specify how this will be done. The standard thus allows VP allocation or VP/QoS allocation. The VP allocation policy will be more efficient but would require separate mechanisms other than bandwidth and buffer allocation to guarantee the various QoS within a VP. QoS allocation shares capacity among multiple VP's on a common trunk and might be quite efficient; however, it may not satisfy the ATM standard

In addition, within each of these architectures, the resource allocation can be tailored to the traffic characteristics and QoS requirements of a specific service type by choosing appropriate buffer and bandwidth combinations within each group. Choosing the appropriate levels of resource sharing requires that the trade-off between network utilization and resource management cost be carefully weighed.

B. Statistical Multiplexing Gains

Statistical multiplexing allows a combination of traffic sources to share bandwidth and buffer. Savings, or *multiplexing gains*, are achieved when the QoS can be obtained with fewer shared resources than would be required if the sources used separate resources. These savings can be used to accept additional calls, and thus generate additional revenue.

For purposes of discussion, we distinguish between multiplexing gains in bandwidth and in buffer. We also distinguish between gains achieved by statistically multiplexing sources of the *same* type, known as *homogeneous multiplexing gains*, and gains achieved by statistically multiplexing sources of *different* types (but identical paths and/or QoS), known as *heterogeneous multiplexing gains*.

Gains in bandwidth, or instantaneous gains, result from peaks of one traffic source coinciding with the valleys of another source, since the sum of instantaneous rates of statistically multiplexed traffic sources tends to be smaller than the sum of the peak rates of individual traffic sources. Gains in buffer, or temporal gains, result from the buffering of traffic sources. Buffers reduce the burstiness of the traffic by smoothing out the uneven arrrvals of bursts, and thus less bandwidth is required to serve the same amount of traffic after buffering.

The magnitude of these gains is determined by the number of each source type, by the loss criterion, by the differences among source types, and by the shared buffer length. Circuit switching is generally the least efficient policy, and complete sharing is generally the most efficient policy. On the other hand, complete sharing requires complex resource management to insure that each source receives its guaranteed QoS. Queueing theory has developed a set of results concerning homogeneous statistical multiplexing gains. For many common performance measures, gains in service rate (or bandwidth) are proportional to the square root of the number of sources, and gains in buffer are proportional to the number of sources.

Under a QoS constraint such as maximum permissible loss, however, results are not as well-established. Furthermore, the efficiency of intermediate policies is not well-understood. With no buffer, both homogeneous and heterogeneous bandwidth gains appear to be proportional to the square root of the burstiness of each source type (as given by source rate variance divided by source rate mean) [23]. For very long buffers, however, since effective bandwidths add, there is initially no bandwidth gain, but only a buffer gain proportional to the number of sources [14]. This buffer gain can be traded off for bandwidth gain by increasing the buffer length and decreasing the service rate, thereby decreasing a source's effective bandwidth.

These results cover only part of the multiplexing gain combinations. More research is needed to add to our current understanding.

C. Admission Control

When a new call arrives, the network must perform admission control by determining whether to accept this call or not, based on the current network utilization, this call's traffic descriptor, the required QoS of all calls, and the network efficiency. The general view toward admission control is as follows: Given the types of services, and the QoS requirements, the network must first determine the acceptance region, within which all the QoS requirements can be satisfied [23]. However, not all the points within the acceptance region will generate efficient usage of network resources. Network efficiency can be defined by high throughput or high economic efficiency, etc. The next task for the network is to determine what subregions within the acceptance region are the most efficient subregions, reject calls that lead the network to inefficient subregions, and accept calls that bring the network into efficient subregions. However, information for determining the acceptance region is not necessarily available to the network beforehand. In the case where the acceptance region is not known, the network has to make admission decisions on a real-time basis [24].

For admission control in circuit-switching networks, the route of a new call must be determined first. If enough resources can be reserved for the call, the network must then weigh the benefit and the cost of accepting this call. Product-form networks have been studied in [25]. There is a fixed revenue associated with each successful call. The cost of providing the call is equal to the potential revenue generated by other new calls if this call were not accepted and if the circuits occupied by this call were available to those new calls. When the cost exceeds the fixed revenue, this call will be denied even when the circuits are available. Kelly [26] proposed a distributed iterative algorithm to control admission to circuit-switching networks, where revenue is maximized by route optimization and admission control.

In cell-switching networks, the resources needed for each call depend on the burstiness and QoS. However, if an estimate of multiplexing gains produces knowledge of the network's ability to simultaneously accommodate various combinations of sources and QoS, then a similar approach to circuit-switching admission control might be used for connection-oriented services.

IV. Contract Negotiation

Much progress has been made in the first stage of connection establishment, separate user and network characterizations, as discussed in the previous two sections. However, as the theories and concepts of this stage are becoming more mature and unified, the gap between the network and user sides is more apparent. There has been no unified way of bridging across the user's side and the network's side to complete the connection establishment. However, the trend is that connection establishment is viewed as a contract negotiation process involving multiple network and user agents on issues including prices, resource allocation, and QoS.

A distributed contract negotiation process has many attractive properties. First, a network does not need to know every user's information beforehand but can obtain such information in the process when needed. For example, the network does not need to know a user's complete demand curve. The section of a user's demand curve around the current operating point might be sufficient information to the network. Second, each part or each layer of the network only needs to know the information local to them. This lessens the information collection and management tasks for the network and the amount of traffic generated for this purpose. Third, computation of optimal prices and resource allocation can be performed by distributing most of the tasks to local users and lower layers. In all the proposals in this section, users calculate the amount of resources to request for their service by maximizing their consumer surplus, and the network merely collects the results and calculates the best prices.

A. Objectives and Negotiation

The network's objective has often been expressed in terms of average throughput, delay, and cell loss. However, these criteria do not distinguish between different users of service types. Economic efficiency in terms of revenue [25], [26] or total user benefit [27], [28], when used as a network objective, can help differentiate between the value various services place on access and performance. Most recent contract negotiation proposals attempt to maximize total user benefit. Properly used, this approach also prevents congestion and degradation of service qualities and supports a variety of service classes. Maximizing user benefit might be unreasonable if the network is owned privately and if profit maximization is the objective. However, if the network is owned by the government, whose objective is to best serve the users of the network, this measurement of efficiency can be justified.

Incentives [28], especially monetary incentives (prices), are essential for the successful implementation of an efficient network. Prices play an important role in congestion control and in resource allocation among all the users and different network levels. As a result, pricing is an inseparable part of the contract negotiation process. Design of an incentive compatible pricing mechanism is a challenging task, which requires solving a combination of technical and economic problems.

The user's objective for using network services has often been considered as getting the best QoS possible. However, if prices are charged to users, then they have to weigh the benefit of service against the cost charged. The benefit may be some monetary measure of how much a user values the service. If the benefit exceeds the cost, the users will most likely use the service. We borrow from economics to define the difference between benefit and cost as *consumer surplus*. Most recent contract negotiation proposals assume that a user's objective for using the network is to maximize that user's consumer surplus.

At the beginning of the distributed negotiation process, both sides know their objectives. A user agent has information about his traffic stream (Section II) and valuation of the service, but is unaware of the network's available capacity and the total market demand for network services. On the other hand, a network agent knows the available capacity (Section III) but is unaware of user's valuations of services and their desired QoS. The dominant mechanism in negotiations is as follows. By setting prices, the network agent signals to the user agents the available capacity and the market demand. A user agent chooses the amount of resources needed for his desired application. At market-clearing prices, the network and user agents agree on the prices, amount of resources for each connection, and QoS. In the following sections, we briefly review a number of recent contract negotiation proposals. Most reflect the dynamics between users and the network described above and strive to achieve economic efficiency through distributed approaches. These proposals present considerably different pricing schemes, and some do not have the two distinctive connection set-up stages or the negotiation order described above. However, each of these proposals has its own contribution, and their work lends insight into pricing of high-speed network services.

B. Smart Market Pricing for the Internet

Mackie-Mason [29] introduced a smart market-pricing scheme, for each packet transmitted on the Internet, to control congestion and to improve network efficiency. This paper was one of the first papers to show the importance of the joint effort of engineers and economists to design technically feasible and economically efficient pricing policies for network services. Each user submits a bid for each packet to transmit. The network transmits all packets whose bids exceed the cut-off price. The network sets the cut-off price equal to the equilibrium price, where demand meets capacity, or to the marginal cost of transmitting one more packet, whichever is applicable. The marginal cost consists of a noncongestion cost for the network to transmit the packet plus a congestion cost. The congestion cost accounts for the burden that transmission of a packet imposes on other users due to added delay and loss of packets experienced when the network is congested. The prices thus fluctuate dynamically with the utilization level of the network and with users' demands.

With smart market pricing, users have incentives to reveal the true values of their packets since nobody can manipulate the price to their own advantage by lying to the network about the true value. Furthermore, only one round of negotiation is needed to reach the market-clearing price, and thus it is computationally simple and can be applied in real time. In practice, it is very hard for a user to measure willingness-to-pay (benefit) for a single packet since this value often depends on the other packets in the same information stream. It is easier to measure the value of an entire application. Furthermore, if real-time applications are deployed on the Internet, per-packet pricing might not accurately reflect the bursty nature of a traffic source, since there is little incentive for users to generate less bursty traffic sources.

In addition to the above per-packet pricing policy, Parris [30] studied the effect of per-packet pricing, set-up pricing[1] and peak load pricing[2] on network performance, as measured by revenue, call-blocking probability, link utilization, and cost recovery. Simulation results show that network revenue first increases and then decreases while blocking probability only decreases as the per-packet price increases. Set-up pricing causes blocking probability to decrease compared to per-packet pricing for the same per-packet prices. Peak load pricing raises network revenue and decreases blocking probability by smoothing out network usage.

C. Priority Pricing for Computer Networks

Cocchi [28] proposed a priority pricing policy for multiple service disciplines[3] in computer networks, borrowing the concept of Nash implementation from economics and game theory. A *Nash equilibrium* is an equilibrium point in a game where the strategy chosen by each player is the best strategy, given the choices of all other players. No player has incentive to deviate from the equilibrium point unilaterally, but such an equilibrium does not necessarily generate the best overall performance of all players. A *Nash implementaton* of a socially optimal policy exists if the socially optimal operating point is also a unique Nash equilibrium point. Four types of user requests (e-mail, FTP, telnet, and voice) are considered as simple examples in this paper, and the benefit functions of the services are computed using the parameters total delay, loss probability,[4] average throughput, and round-trip time. Priorities are chosen by each application based on the transmission prices of low- and high-priority packets. A user's objective is to maximize consumer surplus. The network's objective is to maximize total user benefit, by setting the right prices, so that users will act in both individually and socially optimal manners by self-selecting the packet priorities for their applications. The problem of how to choose the right prices for the network and the right priorities for the users is treated as a game problem by Cocchi. The decision reached by both the network and users is a Nash equilibrium of this game. From simulation runs with various network configurations, Cocchi found that there is a price range for each priority level that will result in a Nash implementation.

Cocchi's work laid a firm ground for future research in this area by showing that it is possible to design a priority pricing policy that leads to efficient usage of computer networks. However, more research is needed to determine how the optimal pricing policies can be obtained without time-consuming simulations and how prices relate to the underlying costs of the services. Mackie-Mason's and Cocchi's approaches are very similar in that both try to allocate network resources among users based on how much a user values a particular service. The priority pricing scheme is practical and effective when the number of service disciplines is very few, but it becomes less so when the number becomes large or even unknown.

In high-speed networks, applications with different characterizations of QoS may present challenges to pricing methods designed for traditional computer networks. The rest of this section will review recent pricing proposals designed specifically for ATM networks. These proposals suggest issues related to pricing concerning traffic characterization, the effect of statistical multiplexing, and resource management architecture.

D. Resource-Based Pricing

Low [27] extended his work on burstiness curves (reviewed in Section II) to connection establishment, using a distributed iterative negotiation process between the network and users. The network's objective is to maximize total user benefit, and a user's objective is to maximize his consumer surplus by requesting the best combination of buffer and bandwidth while maintaining no cell loss. The network posts prices for bandwidth and buffer on each link, and the user then decides how much of each resource to request for all the links on his route. The iteration process continues as the network posts new prices based on the demand and utilization level of the resources and as users update their requests for resources based on the newly posted prices. For the aggregated demand functions specified in the paper, this iterative process will converge to the optimal solution where welfare is maximized.

The advantage of this pricing policy is that the network does not need to know detailed individual user traffic characterizations, and thus there is no need for traffic policing. The network provides the resources, and it is up to the users to package the resources into the desired services. Murphy [31] demonstrates that the separation between resources and services enables the network to provide a large and even unknown number of service classes. However, statistical multiplexing among user-information streams cannot be incorporated in the current model. The main contribution of this model is that each user's demand elasticity for bandwidth versus buffer is used to improve the network usage, as defined by an economic viewpoint instead of the traditional performance viewpoint.

A similar approach was taken by Parris [32], who extended Ferrari's work [33] on real-time connection establishment in packet-switching networks to propose a resource-based pricing policy. Prices are the weighted sum of reserved network resources, including buffer, bandwidth, CPU time, and delay. This policy is very similar to Low's pricing policy in that the

[1] Set-up pricing here means that users are charged a fixed one-time admission fee and for the number of packets sent.

[2] Peak load pricing means that users are charged at higher rates during peak network usage periods and are charged at lower rates during low network usage periods.

[3] A service discipline is defined as a function from users' requests to assigned network services.

[4] Loss probability here means the percentage of packets not delivered within a certain time frame.

composite prices are the weighted sum of a set of parameters that characterize network resources; both policies are resource reservation types of connection establishment. The two policies differ in that Parris's method allows some sharing among users and thus policing is necessary.

E. Effective Bandwidth Pricing

Effective bandwidth seems to be promising in user-traffic characterization and admission control. Usually policing is necessary to monitor that a user's traffic source does not exceed the effective bandwidth he declared. Kelly [34], however, devised a pricing structure to encourage users to share with the network their true traffic parameters, in place of policing, for on-off Markov fluid sources. The network represents a user's effective bandwidth as a function, $B(Z)$, of a certain parameter such as the user's mean rate. The user tells the network his estimate, z, of the parameter Z. Users are then charged the amount $a_z + b_z Z$, where Z is the measured quantity of the parameter and where a_z and b_z are chosen so that $a_z + b_z Z$ is tangent to $B(Z)$ at z. Kelly proved that a user faces the minimal expected charge if and only if $z = E[Z]$, assuming $B(Z)$ is concave and that the minimal expected charge is equal to the user's effective bandwidth.

This scheme guarantees that users will tell the truth about their estimate of their traffic parameters, and thus avoids the need for traffic policing. Furthermore, it might be possible to combine this approach with other pricing approaches to achieve optimal resource allocation without policing.

F. Marginal Cost Pricing for ATM Networks

As described in Section III, the architecture of ATM networks consists of virtual paths and virtual circuits. Murphy [35] suggested a distributed pricing policy to allocate bandwidth on these levels to achieve maximal network efficiency, as defined by the difference between total user benefit and cost. Each user has a benefit function in bandwidth, describing the user's valuation of the service. The cost function is convex in the virtual path's utilization, and can be considered to include some congestion cost. By setting the price equal to the marginal cost, the network maximizes its objective, and the user chooses the right amount of bandwidth based on the posted service price to maximize his consumer surplus.

In this paper, an individual user's demand elasticity in bandwidth is considered and resource allocation using pricing among different network levels is addressed. In the subsequent paper [31], statistical multiplexing was introduced in simulations, where the network posts prices according to the buffer content and users choose bandwidth based on the prices. Some interesting results from these experiments are that network-utilization-based pricing can control traffic admission and consequently network utilization to a large extent; and that by choosing the appropriate price, the network is able to assure a basic service quality such as no cell loss and very small delay without traffic policing or enforcement. Users then decide the desired service quality levels and the corresponding bandwidth needed, tailored to their own applications and the cost.

V. Linking User and Network Characterizations

Most recent developments in connection establishment address a limited set of connection establishment problems. For example, Low focused on the trade-off between buffer and bandwidth, and Murphy [35] emphasized resource allocation at user, circuit bundle and virtual path levels. However, their proposals do not include statistical multiplexing. Effective bandwidth is promising as a traffic characterization but has not been adopted to maximize network's economic efficiency. To form a complete connection establishment process, we must combine a traffic model, QoS metrics, service discipline, resource management architecture, statistical multiplexing, admission control, user and network objectives, and resource allocation in a manner such that the individual parts interrelate.

This section will explore ways to integrate various proposals from previous sections into a complete distributed connection establishment procedure which encourages network efficiency through optimal resource allocation among virtual circuits, circuit bundles, and virtual paths. As we will see in the rest of this section, the notion of effective bandwidth might provide the linkage among user traffic and QoS characterization, admission control, pricing base and multiplexing gain. For a more concrete example, where an analytical model is established and some interesting results have been obtained, readers may refer to Jiang [36]. By exploring linkages between separate connection establishment pieces, we hope to suggest problems which must be solved to form a complete connection establishment process.

A. User Characterization

The user characterization must link together a description of a user's desires, a traffic model, and a definition of QoS. We begin with the linkage between the user's desires and QoS. Assume for simplicity that all network services can be categorized as either real-time or nonreal-time. Assume that the QoS for real-time services can be described by cell loss probability (subject to a maximal delay), and that the QoS for nonreal-time services can be described by completion time (subject to a maximal cell loss). Expanding these limited QoS descriptions is a challenge to the research community. It is reasonable to assume that a user's satisfaction with a network service can be measured by a benefit function depending on the received QoS, such as pictured in Fig. 1(a) and (b). Benefit functions for nonreal-time services must then change with elapsed time and remaining file size. Fig. 2(a) shows the function $\text{ben}(\text{completion time})$ at the beginning of the transmission of a file of size f. Fig. 2(b) shows the benefit function at a later time $t = t'$ with the remaining file of size f'.

This benefit-QoS link must also be combined with a traffic model. Suppose that a user transmits a stream whose characteristics depend on the source and on the desired QoS. A *real-time* user's stream might then be described by effective bandwidth as a function of desired pretransmission cell loss (Fig. 3(a)), given the network channel[5] chosen by the user.

[5] The QoS of a network channel is defined by cell loss probability and maximal delay jitter.

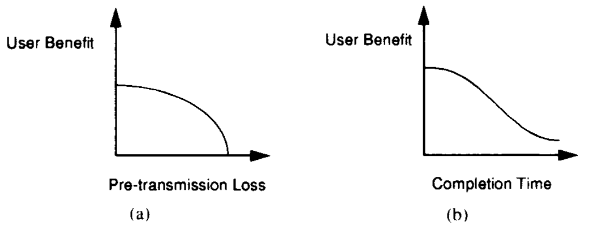

Fig. 1. User benefit functions. (a) Benefit function for real-time service. (b) Benefit function for nonreal-time services.

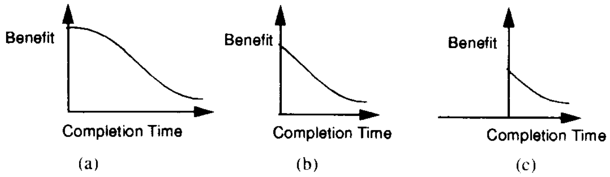

Fig. 2. Change of benefit functions in time for nonreal-time service. (a) $t = 0$ and file size $= f$. (b) $t = t'$ and file size $= f'$. (c) $t = t''$ and file size $= f''$.

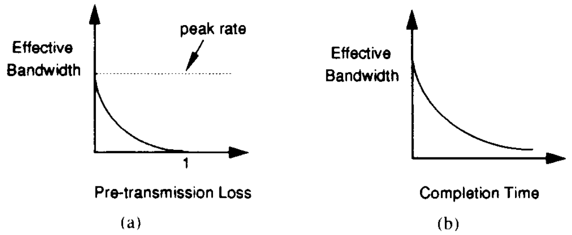

Fig. 3. Effective bandwidth versus QoS. (a) Effective bandwidth for real-time services. (b) Effective bandwidth for nonreal-time services.

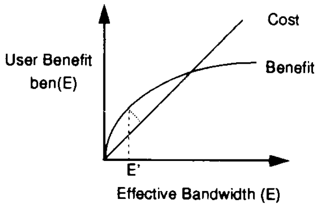

Fig. 4. User benefit of a real-time service versus effective bandwidth.

Similarly, a *nonreal-time* user's stream might be defined by effective bandwidth as a function of desired completion time (Fig. 3(b)). Although the mapping for real-time streams is fairly straightforward, the mapping for nonreal-time streams requires the design of a separate mechanism.

A direct mapping between user benefit and effective bandwidth such as pictured in Fig. 4 could be obtained by combining Fig. 1 and Fig. 3. This accomplishes a linkage among benefit, QoS, and traffic model for all service types.

Alternative methods of linkage are surely possible. In particular, it would be desirable to incorporate buffer usage directly into the traffic model. Bandwidth can almost always substitute for buffer[6] while buffer can only substitute for bandwidth to a certain extent. A user's traffic stream could instead be characterized by a "burstiness curve" detailing acceptable combinations of bandwidth and buffer, as in Low's work. However, this approach severely restricts sharing of buffer and bandwidth. It might also be possible to model the tradeoff between effective bandwidth and buffer to characterize a traffic stream. This would allow a more complex traffic model; however, when multiple users share the same buffer and bandwidth, it is difficult to separate individual users' usage of network resources.

B. Network Characterization

The network characterization must link together the user's traffic model and QoS definition with a service discipline, a resource management architecture, and an admission control policy. We begin with the linkage between the users's traffic model/QoS and the service discipline. If effective bandwidth is adopted as a traffic model, and if the source or combination of sources will be served at a constant rate, then the resulting loss directly follows. Other traffic models may require a separate analysis to determine the QoS that results using a particular service discipline.

We next consider linking this characterization with the resource management architecture and finding the resulting network capacity including any effect from statistical multiplexing. As mentioned above, a network's capabilities to offer a particular service mix is characterized according to the partition of resources at three levels: virtual circuit, circuit bundle, and virtual path. A circuit bundle groups VC's with common characteristics as given by the resource management architecture. For instance, under VP/QoS/Type allocation, a circuit bundle contains VC's with common paths, source types, and QoS. Under QoS allocation, circuit bundles are defined on each trunk by QoS only. Greater sharing should generally result in greater efficiency, but we must be able to predict which service mixes the network can accommodate with guaranteed performance.

Given the bandwidth allocated to a circuit bundle, the effective bandwidth of each source can be determined. Since the effective bandwidth of statistically multiplexed sources with equal QoS is simply the sum of their effective bandwidths, the acceptance region is a simplex. This choice of traffic model thus simplifies the calculation of the network capacity, but this simplicity comes at a cost. Because of the limitation that effective bandwidth is additive only among the traffic sources with the same ζ (as defined in Section II), circuit switching, VP/QoS and VP/QoS/Type allocations are the only feasible choices if effective bandwidth is used as the traffic model. In particular, VP allocation is not possible, even though it might result in greater network efficiency, because we are unable to predict the resulting acceptance region. Other traffic models could be used if they can be similarly linked with a resource management architecture.

If one can calculate an acceptance region, within which the indicated mix of service types can be accommodated with guaranteed QoS, then it remains to link an admission control policy. In general, one may wish to block some calls that

[6] Except that cell level congestion cannot be eliminated without a short buffer, even when the service rate is greater than the peak rate.

might be accommodated, in order to maximize some system objective. A common admission control policy, however, is to accept all calls that can be served within QoS constraints. With effective bandwidth as a traffic model, this policy would simply be to accept a new call request iff the sum of effective bandwidths of all users within its chosen bundle is no more than the bundle's capacity.

C. Negotiation

The negotiation phase must link together the user's desires and the network's desires, under the constraints given by the user and network characterizations. This linkage has been provided by previous researchers using prices. Pricing can also play an important role in distributing the optimization tasks into various network levels and geographically local network areas since it provides user incentives, acts as a media through which user and network agents communicate, and signals the optimality of network resource allocation. Proper choices of what to price and how to charge users results in network efficiency; improper choices create arbitrage opportunities [37].

The most common usage-based pricing schemes are mean-rate pricing[7] and peak-rate pricing.[8] However, under the presence of statistical multiplexing gain, the bandwidth a bursty traffic source requires to guarantee QoS is between its mean rate and peak rate. Effective bandwidth can be considered as composed of two parts: mean rate + burstiness. If users are charged an amount equal to *effective bandwidth* × *price*, then this charge includes a fixed price per packet plus an amount based on burstiness.

To link user and network desires, a negotiation might be accomplished through a periodic exchange of price and demand. A user could first choose a circuit bundle that can satisfy his QoS. A circuit bundle for real-time applications might have a short buffer to satisfy the delay requirement whereas a circuit bundle for data might have a relatively long buffer for the low cell loss requirement. In a short periodic automated cycle, a circuit bundle and its users could negotiate by exchanging price per effective bandwidth unit and corresponding demand. Because of the short cycle, burst level admission control could be accommodated for traffic sources with long and unpredictable bursts. In a long cycle, bandwidth allocations among circuit bundles within a virtual path could be updated through negotiation between the circuit bundles and the virtual path. In a even longer cycle, virtual paths that share common physical trunks could compete for bandwidth on the physical trunks along each route. Note that resource allocation would thus be performed at three levels, and each level would communicate only with its adjacent levels. There would be prices at each network level, as results of the negotiation between adjacent network levels, which signal the optimality of resource allocation. Computations of optimal resource allocations would be distributed among network levels and among local network areas. For instance, a user would only negotiate with its circuit bundle level, and the negotiation would be transparent to its virtual path level or to any other circuit bundles.

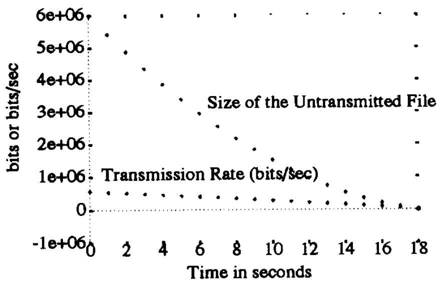

Fig. 5. Transmission rates for each negotiation cycle.

To link this negotiation with the user characterization, assume that a user's objective is to maximize his consumer surplus, defined in Section IV, by choosing the appropriate QoS. Since a mapping is available between QoS and effective bandwidth, the user could choose an amount of effective bandwidth to purchase depending on its current price. For example, a real-time user might adjust the compression level for his video transmission according to the current price for bandwidth on the network channel defined by the required maximum delay. Similarly, a nonreal-time user might spread out the transmission of data or transmit the data at once depending on the posted price.

The user could maximize his consumer surplus by choosing the optimal effective bandwidth E' depending on the posted price per effective bandwidth unit ρ as shown in Fig. 4

$$\max_{E} \mathrm{ben}(E) - pE, \qquad \text{subject to } E \geq 0.$$

As mentioned above, this approach requires a model for how nonreal-time users choose instantaneous bandwidth in reaction to its current price. We are unaware of much research on this topic. To investigate what such a model might include, for simplicity assume a user reevaluates this choice periodically, and thus transmits at a piecewise constant rate. His effective bandwidth is thus equal to this constant rate during the cycle. A primitive model could have the user choose his transmission rate during the next cycle in order to maximize his incremental consumer surplus. A simplified example illustrates how this method works. Assume that price is fixed at 2×10^{-7} per effective bandwidth unit throughout the transmission, that a user has a benefit function $\mathrm{ben}(T_c) = 1+1/(1+T_c/45)$, where T_c is the completion time, that the file size to be transmitted is 6×10^6 bits, and that the negotiation cycle is 1 s. The transmission rate resulting from this scheme is computed and shown in Fig. 5. It remains to be demonstrated whether such a scheme has desirable dynamics.

To link this negotiation with the network characterization, assume that a network's objective is to maximize total user benefit of all network users. The prices should then be set to the intersection point where supply and demand meet at the optimal solution. To find the market-clearing prices, an iterative negotiation process could be involved. The higher network level could post its price for bandwidth, and lower network levels could respond to the price by submitting

[7] Users are charged based on the mean rate of their traffic sources.

[8] Users are charged based on the peak rate of their traffic sources.

requests for effective bandwidth. The higher network level would then sum up the demand of all the users. If the demand is higher than its capacity (or supply), the higher network level would raise its price; otherwise, it would lowers its price, until eventually demand meets supply where the market-clearing price is found.

A policy for such price updates must be designed, and the dynamics of the resulting process must be analyzed. In addition, if any sharing of network resources is allowed, e.g., under VP/QoS/Type or VP/QoS architectures, the process should be studied to investigate the effect of multiplexing gain on prices at each level. At the optimal solutions, marginal total user benefit in bandwidth should be equal among all competing circuit bundles and virtual paths.

VI. Challenges

High-speed networks such as ATM provide great flexibility to users to package resources into services, since the networks are principally designed to provide resources and basic QoS instead of particular services. This flexibility can result in a wide range of services if we can better link user desires with network availability. In Sections II and III, we reviewed recent approaches to characterization of traffic sources, QoS, and network capacity. In Section IV, we reviewed recent proposals for contract negotiation. Finally, in Section V, we discussed possible ways to combine these methods to form a complete connection establishment process. The attempt to link traffic models, QoS, service disciplines, resource management architectures, admission control, and user and network objectives exposed a number of outstanding problems. In this section, we expand on these problems and present them as challenges to the research community.

A. User Characterization

1) Richer set of QoS descriptions: A richer set of QoS metrics from a user's perspective needs to be developed. Pretransmission loss and completion time are not sufficient to describe complex services, including services that are neither real-time nor nonreal-time, such as telnet. For instance, Cocchi [28] suggests that the QoS description should include total delay, loss probability,[9] average throughput and round trip time. Any such additional QoS metrics need to be linked to traffic models, service disciplines, and user benefit.

2) Real-time and nonreal-time service integration: Connection establishment requires a mapping between user benefit and network resources. This link is likely to be provided by QoS, but different types of service, e.g., real-time and nonreal-time, are likely to be described using different QoS metrics. For each service type, it is necessary to model user demand for network resources and the QoS resulting from the network service discipline, so that multiple service types can be effectively integrated. A simple set of service types and QoS metrics was chosen in Section V to demonstrate the required linkages, but a richer set of service types requires more complex mappings. Even with this simple set, the mapping for nonreal-time service bandwidth demand given a user's benefit function and network price is nontrivial. As prices fluctuate from one negotiation interval to another, the primitive mechanism suggested above may not maximize a user's consumer surplus over the entire transmission period.

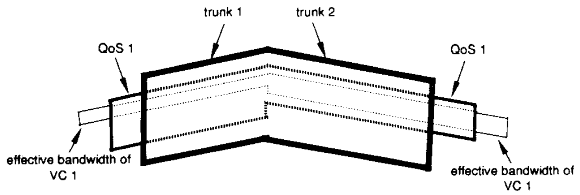

Fig. 6. Effective bandwidth for a virtual circuit under QoS allocation

3) Demand elasticity: Differences in users perceptions of benefit can result in increased economic efficiency if used wisely. Consider a simple example. A network contains two resources: A and B. User 1 is indifferent to resource type A or B for his service while user 2 requires resource type A for her service. If user 1 chooses the type A resource, then only one user can use the network. If user 1 chooses the type B resource, however, then both 1 and 2 can use the network.

The negotiation process suggested above takes advantage of different user's varying benefits from bandwidth to increase economic efficiency. However, other differences between users are not similarly utilized. In particular, it would be advantageous to also measure each user's demand elasticity in buffer usage. In addition, knowledge of cross demand elasticity between circuit bundles could result in users adjusting demand in reaction to bundles' relative prices instead of prechoosing a bundle.

B. Network Characterization

1) Increased sharing: Under VP/QoS/Type or VP/QoS allocation, effective bandwidths and multiplexing gains are constant along a virtual path. For QoS allocation, however, a virtual circuit may have different effective bandwidths on each link along its route if the capacity of its circuit bundle varies from trunk to trunk, as shown in Fig. 6. As a result, admission control is more complex. QoS allocation is among the most efficient resource management architectures. Unfortunately, the current understanding of traffic characterization, admission control and multiplexing gain is not mature enough to accommodate this.

Another mechanism for increasing sharing is to take advantage of the smoothing effect of buffers. Buffers often smooth out some burstiness of a traffic stream so that when the stream is fed into the next link on a virtual path, it needs less effective bandwidth. Therefore, the effective bandwidth needed along a virtual path could be decreasing going downstream. This extra savings in capacity should be considered in the process of connection establishment to improve network efficiency.

2) Buffer allocation: The discussion above centered around pricing and allocating bandwidth. Other network resources were either assumed to be plentiful, e.g., buffers, or ignored. If we had a better analytical understanding of how to separate individual user's usage of a buffer shared by multiple users,

[9] Loss probability here means the percentage of packets not delvered within a certain time frame.

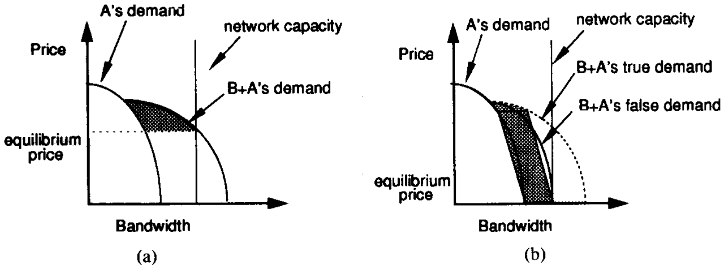

Fig. 7. An example of user B's incentive to lie about his demand curve.

then buffer could also be priced. The network could then allocate both bandwidth and buffer to maximize network efficiency, by allowing users to maximize their consumer surplus by choosing the optimal balance between bandwidth and buffer. Such an approach would also likely result in the ability to remove the circuit bundle level, by allowing a virtual path to contain all virtual circuits with identical paths, and still guarantee performance.

C. Negotiation

1) Convergence: In many pricing schemes, distributed and iterative negotiation processes are involved in finding the optimal prices and resource allocation. It is crucial that these processes converge to a near-optimal solution in a limited amount of time. Mackie-Mason's negotiation process converges after one iteration. Low's process also converges, under the assumption of continuous and decreasing convex aggregated demand curves. It is not clear whether Murphy's scheme will converge under this assumption. We have suggested that a similar iterative process might be used in a complete connection establishment process, but mechanisms to dynamically post prices on each level must be designed, and their convergence measured. It is unclear how often each level should reallocate its resources, and how long the negotiation at each level must last.

2) Truth telling: As it was pointed out in earlier sections, prices are used as incentives to encourage users to act in a socially optimal way. However, if users know that the information they provide will affect prices, users may lie about their information to manipulate the prices to their own advantage. This will only occur if users have both the ability and incentive to manipulate prices. A simple example, shown in Fig. 7, demonstrates a user's potential ability and incentive. Suppose a network intends to set the price to the equilibrium point shown in Fig. 7(a) where capacity meets demand, or to set the price equal to zero if demand is less than capacity, through the iterative negotiation process described in the previous section. Suppose, however, that user B gave a false demand curve to move the equilibrium price to zero. The shaded regions indicate user B's consumer surplus. Even though the bandwidth obtained by user B is reduced by doing so, user B does not pay any price for the bandwidth, and therefore user B's consumer surplus in Fig. 7(b) may be greater than that in Fig. 7(a).

Prices can be based on measures so that truth-telling is desirable to the user. Such mechanisms were introduced in Mackie-Mason [29] and Kelly [34]. Alternatively, the negotiation itself can be designed to encourage truth-telling. Vickrey [38] suggested charging a consumer the price determined by supply and total demand of all users except this consumer. He showed that the consumer will not benefit from lying to the supplier about his demand if he cannot manipulate the price under this mechanism. It is not clear if these mechanisms can be applied to complete and distributed connection establishment processes.

REFERENCES

[1] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 932–989, Sept. 1991.

[2] S. Q. Li, "Study of information loss in packet voice systems," *IEEE Trans. Commun.*, vol. 37, pp. 1192–1202, Nov. 1989.

[3] D. P. Heyman, A. Tabarabai, and V. Lakshman, "Statistical analysis and simulation study of video teleconference traffic in ATM networks," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 2, pp. 49–58, Mar. 1992.

[4] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 1, pp. 446–458, Aug. 1993.

[5] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.

[6] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1598–1608, Dec. 1988.

[7] ——, "Switching and traffic theory for integrated broadband networks," Norwell, MA: Kluwer, 1990.

[8] F. P. Kelly. "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, pp. 5–15, Sept. 1991.

[9] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, May 1991.

[10] R. Guerin, 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, Sept. 1991.

[11] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.

[12] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 9, pp. 165–170, 1991.

[13] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Probab.*, vol. 20, pp. 646–676, 1988.

[14] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.

[15] G. Kesidis and J. Walrand, "Traffic policing and enforcement of effective bandwidth constraints in ATM Networks," preprint.

[16] R. L. Cruz, "A calculus for network delay, pt. I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–141, Jan. 1991.

[17] S. Low and P. Varaiya, "Burstiness bounds for some burst reducing servers," in *Proc. Infocom'93*, Mar. 1993, pp. 2–9.

[18] ——, "A simple theory of traffic and resource allocation in ATM," in *Proc. Globecom'91*, Dec. 1991, pp. 1633-1637.

[19] M. Wong and P. Varaiya, "A deterministic fluid model for cell loss in ATM networks," in *Proc. Infocom'93*, Mar. 1993, pp. 395–400.

[20] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Networking*, vol. 1, pp. 372–385, June 1993.

[21] C. S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 913–931, May 1994.

[22] J. Burgin and D. Dorman, "Broadband ISDN resource management: The role of virtual paths," *IEEE Commun. Mag.*, vol. 29, pp. 44–48, Sept. 1993.

[23] I. Sidhu and S. Jordan, "Multiplexing gains in bit stream multiplexors," *IEEE/ACM Trans. Networking*, to be published.

[24] H. Saito, "Call admission control in an ATM network using upper bound of cell loss probability," *IEEE Trans. Commun.*, vol. 40, pp. 1512–1521, Sept. 1992.

[25] S. Jordan and P. P. Varaiya, "Throughput in multiple service, multiple resource communication networks," *IEEE Trans. Commun.*, vol. 39, pp. 1216–1222, Aug. 1991.

[26] F. P. Kelly, "Routing in circuit-switched networks: Optimization, shadow prices and decentralization," *Advanced Appl. Probab.*, vol. 20, pp. 112–144, 1988.

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

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

[29] J. K. Mackie-Mason and H. R. Varian, "Pricing the Internet," in *Proc. 2nd Int. Conf. Telecommun. Syst. Modelling, Anal.*, Nashville, TN, Mar. 24–27, 1994, pp. 378–393.

[30] C. Parris, S. Keshav, and D. Ferrari, "A framework for the study of pricing in integrated networks," Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. TR-92-016, Mar. 1992.

[31] J. Murphy and L. Murphy, "Bandwidth allocation by pricing in ATM networks," preprint.

[32] C. Parris and D. Ferrari, "A resource based pricing policy for real-time channels in a packet-switching network," preprint.

[33] D. Ferrari and D. C. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 368-379, Apr. 1990.

[34] F. P. Kelly, "On tariffs, policing and admission control for multiservice networks," *Oper. Res. Lett.*, vol. 15, 1994.

[35] J. Murphy, L. Murphy, and E. C. Posner, "Distributed pricing for embedded ATM networks," in *Proc. Int. Teletraffic Congr. ITC-14*, Antibes, France, June 1994.

[36] H. Jiang and S. Jordan, "The role of price in the connection establishment process," *European Trans. Telecommunications and Related Technologies*, to be published.

[37] P. Srinagesh, "Economic issues in the pricing of broadband services," preprint.

[38] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *J. Finance*, vol. XVII, pp. 8–37, May 1961.

Scott Jordan (S'83–M'90) received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, in 1990.

He has been an Assistant Professor at Northwestern University, Evanston, IL, since 1990. His research and teaching interests are in the modeling and analysis of behavior, control, pricing in computer/telecommunication networks, production, queueing, and other stochastic systems.

Hong Jiang (S'95) received the M.S. degree from Northwestern University, Evanston, IL, in 1993 in electrical engineering. She is working toward the Ph.D degree in the field of networking.

Her research interests are in pricing and resource allocation for high-speed networks and network performance modeling.