CHARGING ISSUES FOR SERVICES IN BROADBAND NETWORKS

Costas Courcoubetis, Vasilios Siris, George D. Stamoulis
Institute of Computer Science, FORTH, and
Dept. of Computer Science, Univ. of Crete,
Heraklion, Greece
and
Yannis Markopoulos
Intrasoft SA
Greece

SUMMARY

Charging and accounting in modern high-speed networks are extremely vital for their successful operation. Tariffs and pricing schemes are needed for the network to recover its costs in a fair way from the diverse population of users, and to effectively allocate network resources. It is widely believed that pricing mechanisms which use feedback can play a part in effectively managing congestion.

Charges will of course be partially determined by the competitive strategies of service providers. However, it is unlikely that a good charging strategy should be wildly out of line with a desire to charge in a manner that reflects user's relative uses of network resources. Although in the case of a traditional circuit-switched telephone network it may be relatively simple to charge for use in a manner that reflects a customer's actual usage, it is not so clear how this can be accomplished in current high-speed networks, where the trend of current developments is towards systems which will allow a number of widely disparate traffic streams to share the same broadband channel. A call, which might be a mixture of voice, video and data, would appear to the network as a stream of cells, and the hope is that calls with a broad range of burstiness characteristics can be efficiently integrated, through statistical multiplexing, to share a common resource. In recent years a number of papers have provided some basis for this hope, by showing that it is possible to associate an effective bandwidth with a source type such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee, expressed in terms of the probability that delay exceeds a threshold or that a cell is lost.

The effective bandwidth depends on characteristics of the source such as its mean and peak rate and there is by now a large literature on methods for policing the peak and mean rates of a call. The underlying idea is that at call admission a contract would be made between user and network specifying in more or less detail the statistical properties of the call, and that policing mechanisms would enforce the contract. Unfortunately policing may limit many of the advantages of a high speed multi-service network, such as the network's inherent flexibility to deal with a call composed of varying and uncertain mixtures of voice, video and data. What is needed is a mechanism that trades off the user's uncertainty about a commencing call against the network's requirement to statistically multiplex calls in an efficient manner.
An effective pricing mechanism should also protect the network from users who seek to gain unfair advantage by `malicious' tactics. One such tactic might be for a user to misrepresent his intended use. Clearly, this is unwelcome, since if users provide accurate estimates of their prospective utilization of the network resources, it should be possible to increase revenue by more efficiently loading the system. Therefore the tariff structure should encourage users to improve the accuracy of their estimates and thereby reduce their cost. Network services will be organized in a layered fashion, with the lowest layer being the bear network services. This hierarchy of services motivates a hierarchical structure for the corresponding tariffs and pricing mechanisms.

If a pricing mechanism is to be feasible and practical it must be based on information that can be obtained from the network. This information should be useful in determining the effective bandwidths of the traffic streams or any other statistical information needed for charging. Moreover, the cost of implementing a pricing mechanism must take into account limitations due to the size of the network, the large number of its users, and the communication costs of retrieving the necessary information. The architecture of the accounting system implementing pricing and accounting is a non-trivial subject. It is anticipated that such a system must interact with the user in order to set the initial contract, and from that point on, monitor the traffic to detect possible contract violations. Besides the added security features that must be deployed, the above functionality imposes certain non trivial real-time requirements to the accounting management software.

We will introduce the basic concepts in network services and network technology, and describe the currently proposed approaches for charging guaranteed quality services and best-effort services in networks such as ATM networks and the current Internet.

1. THE DIFFICULTY OF PRICING NETWORK SERVICES

Charging and accounting in modern high-speed networks are extremely vital for their successful operation. Tariffs and pricing schemes are needed for the network to recover its costs in a competitive and fair way from the diverse population of users, and to effectively allocate network resources. It is widely believed that pricing can play a part in effectively managing congestion. Charges will of course be partially determined by the competitive strategies of service providers. However, it is unlikely that a good charging strategy should be wildly out of line with a desire to charge in a manner that reflects users' relative use of network resources. Economic theory suggests that usage-based pricing will be employed by service providers in the case of perfect competition. This is mainly the type of pricing for network services we consider in this paper.

Quality of service affects pricing
We will demonstrate that modern network technology on one hand simplifies the provision of services by providing the user with a unique interface to a single integrated services network which will handle all the user needs, but on the other hand makes usage-based pricing hard to realize since the same network resources are simultaneously used and shared by a large number of disparate services which makes it hard to decide how to share the network cost among these services.
The first source of complexity arises from the new definitions of network services, where now the service does not depend only on the statistics of the information flow but also on the quality-of-service (QoS) parameters such as maximum delay and information loss probability guaranteed by the network to the connection. Clearly the quality of the network service, should reflect to the price of the service since the network, in order to guarantee it, must reserve a corresponding amount of resources for the above connection.

**Users might give away perfect flexibility**

An added degree of complexity is due to the fact that the standards propose the use of policing and shaping for user traffic entering the network, in order protect the network from excessive usage (abuse). Clearly the more a user agrees to be policed by the network, the more predictable his traffic becomes and hence easier for the network to deal with. Obviously “heavily” policed users should pay less and pricing should take into account the amount of policing a user is accepting. On the other hand heavy policing reduces unnecessarily the flexibility of the user that multiplexing offers. The whole issue becomes even more interesting when the user given the rules of the game (tariffs and network charging mechanisms) selects how to change the statistics of his traffic (traffic shaping) by using his own filter in order to achieve a lower cost under the constraint that his quality-of-service will remain adequate.

**Congestion pricing does not guarantee quality**

Recently economists have been involved in defining economic models for networks where pricing depends on the congestion of the network and results in a socially optimal operation of the network. As we will explain, congestion pricing requires complex mechanisms such as feedback from the network in order to be implemented, which in many cases is not feasible, and can not be used for services requiring a guaranteed quality-of-service; for such services one needs to reserve network resources throughout their duration, and can not use schemes were the allocation of the resources will fluctuate according to demand. On the other hand one could propose an extended pricing mechanism which involves aspects from both congestion and usage pricing that can be successfully used to control congestion of guaranteed quality services at call set-up, and to adequately reflect resource utilization.

**The use of prior knowledge about intended resource usage**

It requires some creative thinking to see that pricing can be used in order to immune the network from malicious use or from users which have imprecise knowledge about their traffic. Hence prices can play a role similar to that of policing. It is natural to assume that a more accurate description by the users of their traffic statistics will allow the network to handle more efficiently the anticipated load and hence optimize its operation. Hence through pricing the network should encourage its users to be as precise as possible about their traffic (possibly by learning or keeping statistics), and hence the final charge should reflect both the actual usage and the accuracy of the initial description.
The price of charging
As a last comment pricing should be effectively implementable. By this we mean that the information required by the pricing mechanisms should be easy to obtain and to manipulate. It is well known that a leading cost of the network operation is due to accounting and billing. Hence a prerequisite for a realistic pricing mechanism is its low implementation cost. Modern technology allows sophisticated traffic measurements to be done in hardware, which greatly expands the spectrum of the pricing functions that are realistic to implement.

In this paper we will attempt to illuminate the above issues and propose an approach which encompasses many of the above problems. A key notion in our approach is the notion of an effective bandwidth of a traffic stream; this notion summarizes the use of network resources by the stream, and might depend on a large number of parameters such as traffic statistics, QoS parameters, the multiplexing capability of the network, etc. When an effective bandwidth can be appropriately defined and measured, then the usage-based pricing problem greatly simplifies, and reduces to standard optimization problems. There are very subtle issues with the definition of effective bandwidths as we will explain in the corresponding sections, and there are still many open issues remaining.

The paper is organized as follows: In section 2 we discuss network issues that are related to pricing such as that of user-network contracts and quality-of-service classes. We differentiate between best-effort and guaranteed quality services and discuss how the different characteristics of each affects the pricing scheme. Next, we present the notion of effective bandwidth which in section 3 is used in a simple pricing model for a network supporting both guaranteed quality and best effort services. In section 4 we discuss possible extensions to the simple scheme in relation to social welfare maximization, pricing of available bit rate (ABR) services, and congestion pricing. Finally, section 5 concludes the paper with directions for further research.

2. BROADBAND NETWORK ISSUES RELATED TO PRICING

In this section we review several key broadband network technology concepts relevant for pricing. Our interest lies in the area of Broadband Integrated Services Digital Networks (B-ISDN) based on the Asynchronous Transfer Mode (ATM). Such networks are similar to classical circuit switched telephone networks in that a virtual circuit (virtual connection in ATM terminology) needs to be set up before information is transferred. Information flow in small, constant size (53 byte) packets (called cells). Unlike the telephone network, bandwidth is not reserved for the whole duration of the connection, but rather is used on-demand.

Since there is a large number of such information streams sharing the network resources (link bandwidth and buffer space), an important issue that arises is the effective utilization of the above resources by statistically multiplexing the information flows. In fact, such kind of multiplexing is possible (and can result in significantly higher exploitation of network resources), since most of the traffic sources are bursty; that is, traffic rate may vary considerably with time.
This suggests that usage-based pricing schemes for network services should take into account the economies of scale and scope produced by the multiplexing capabilities of these new generation networks.

2.1 Network Services and QoS

A service provided by the network is the capability for the data stream originating from a particular application to travel through the network by using an end-to-end connection over a virtual circuit and with the network guaranteeing a particular performance (quality-of-service) to the connection. The performance parameters that a connection might specify to the network are cell delay, cell loss probability, cell delay variation (jitter) and minimum throughput. These are lower bounds on the performance requirements of the connection; if these are not met, the user will observe an unwanted deterioration of the quality of the application using the network connection.

Guaranteed and best-effort services

There are two types of services depending on whether these require guarantees on the performance of the network. The best effort services such as email, ftp, web browsing, etc., need no guaranteed quality of service by the network, whereas guaranteed quality services such as video, voice, multimedia, switched circuit emulation, web browsing with minimum performance, etc., need a QoS guarantee. The important issue is that the mechanism through which the network can provide guaranteed quality services is by reserving resources, which is not the case for best effort services.

Another important characteristic of guaranteed services is that these require a VC to be set-up which they will also use whereas best effort services could use the broadcasting mechanism of the network which works in a way similar to the traditional packet-switched networks (this does not preclude virtual circuits for best effort services). Hence another important QoS parameter relevant to guaranteed services is the call set-up time for the connection. Following the same reasoning, since guaranteed services need a particular amount of network resources to be reserved for their own use, it might be the case that such an amount is not available and a call might be blocked or queued until the resource becomes available, which makes call blocking probability another QoS parameter for such services. One would expect that the more a user is willing to pay, the less blocking probability he is willing to suffer.

2.2 User-Network Contracts

An new concept in the B-ISDN networks is the definition of a contract between user and network. From the part of the user, this contract specifies his “obligations” to the network, i.e., the constraints that his traffic should satisfy so that the network in its turn can fulfill its obligation. These constraints are expressed in terms of the “filters” that the user will accept to be placed from the part of the network to his traffic stream. This is the Generic Cell Rate Algorithms (GCRA), [ATMForum94], which are equivalent to leaky buckets. The above policing capabilities that the network could perform to the user traffic is specified by the standards and is restricted to the control
of the sustainable cell rate, maximum mean rate, and maximum burst size. The part of
the information stream of the user that does not conform to the above constraints
might be rejected by the network, or accepted as a type of traffic that is prone to
dropping in the case of congestion. It is intuitively clear that the more constraints the
user accepts, the better the network could plan its operation, and hence the user should
be favorably priced.
From the part of the network, the contract specifies the QoS parameters (a subset
among maximum cell delay, cell delay variation, cell drop probability and minimum
throughput) of the connection. The network by controlling its operation must ensure
that the information cells flowing through the connection will be handled in a way that
the above performance parameters will be satisfied. Of course the network is
committed to the enforcement of the contract provided the user respects its own part
of the contract, i.e., the policing part of the contract. But this is usually enforced by the
policing devices at the periphery of the network.

**How to price a user-network contract**

There are some very important issues related to pricing and the user-network contract.
This could determine an approximate cost for the connection (cost per unit time). The
reason that the price can not be accurately determined is because the contract confines
the user traffic to lie in a specific set determined by the leaky buckets, which serving as
a policing mechanism affect the traffic only if it violates certain conditions. For
example, sending nothing is always a feasible traffic stream under any contract.
On the other hand, one could propose a pricing scheme which assumes the “worst
case” traffic from the part of the user given the particular contract (in terms of
resources necessary to be reserved by the network). There is currently a large amount
of effort and some interesting results related to this approach, i.e., to characterize the
worst case traffic that can be the output of a leaky bucket [HW94, MM94]. Since the
user will not produce in general this worst case input to the network, it should be fair
to the user (and would more adequately reflect actual resource usage) to be charged
according to what he actually sends. This in turn would require sophisticated traffic
measurements, and hence the improved accuracy in pricing should be justified
comparing to the added implementation and management cost. There is also the case
that such a mechanism could not be available in a given network, or that a restricted
set of parameters can be measured. A more sophisticated idea is described next.

**Pricing a contract under uncertain user information**

An important proposal due to Kelly [Kel94b] is that the pricing of a connection should
be a function of the initial contract of the prediction of the user about his traffic and of
the actual traffic sent. The reason is that the network would like to use the information
from the contract and from the initial user prediction in order to assess if enough
resources are available to satisfy its part of the contract to the new connection (while
satisfying the other already existing contracts with the other on-going connections;
note that contracts can not be re-negotiated when a call is accepted). Hence a user
should pay in the best case the exact cost of resource usage (if his initial prediction
were accurate), and pay in the rest of the cases more; this extra amount should be used
by the network in order to compensate for its loss of revenue because of being overly
conservative in accepting new calls.
One could ask the following question. Why does a user not choose an extreme amount of policing so that he will be subjected to cheaper tariffs, or why is the user not “pre-shaping” his traffic, so that the traffic that enters the network becomes of a “cheaper” type. The answer is that there is a lower bound on how much the user can shape his traffic or be subjected to policing, since it is the application using the above traffic stream that imposes the requirements on the worst case QoS parameters for the connection. Hence, for example, decreasing the peak rate (traffic shaping) by using buffering at the user end would increase the delay of the cells, which in turn might be not acceptable to the application.

**Choosing the best contract**

Summarizing, the above discussion implies that pricing is related to the user-network contract and that choosing the best contract from the part of the user should be determined partly by the way contracts are priced. Hence in a usage-based pricing scenario, the network defines the tariffs according to the previous discussion and the user, knowing his traffic needs, chooses the cheapest contract which does not violate his minimum performance requirements from the part of the network. Such a choice is by no means simple, see [deVeciana94, CWWS96]. One could extend this line of argument one step further, and combine the above procedure with the possibility by the user to perform a limited amount of traffic shaping.

### 2.3 A Classification of Network Services

The Quality-of-Service of a connection is expressed in terms of delay, variance of delay and cell loss. Current standards propose four classes of network services and hence of user-network contracts depending on whether the information flow is sensitive to delay and information loss. These are the following [ATMForum95].

1. **Constant Bit Rate (CBR):** This service class is intended for real-time applications (those requiring tightly constrained delay and delay variation) with a constant cell rate. It will be used for telephone, video-conferencing, leased line service, etc.

2. **Variable Bit Rate (VBR):** This service class is intended for variable bit rate applications with time constraints. This class will support statistical multiplexing, and it is motivated by the bursty nature of certain types of traffic. Recent specifications differentiate between real-time VBR (VBRrt) which has tightly constrained delay and delay variation, and non-real-time VBR (VBRnrt) which has bounds on mean transfer delay. It will be used for variable bit rate video, audio, and other multimedia applications.

3. **Available Bit Rate (ABR):** This service class supports non-real time traffic with no specific delay requirements. The network supports feedback congestion control in response to changing network characteristics. A minimum cell rate (MCR) can be specified. It will be used for file transfers, web browsing with performance guarantees, and other data applications.
4. Unspecified Bit Rate (UBR): This service class supports best-effort traffic with no delay requirements or traffic service guarantees such as minimum throughput and cell loss. It can be used for traditional computer communication applications, email, etc.

From their definition, CBR and VBR should be used for guaranteed services, whereas UBR for best effort services. ABR fits in between since it provides a limited guarantee (in terms of cell loss).

Depending on the type of service a user needs from the network, the contract with the appropriate parameters will be set. It is interesting that for the same information flow requirements, more than one of the above service types could be chosen. For example, if one would like to browse the WWW with a certain guarantee on the delay it would take to download large picture files or view video clips, then he could choose the ABR type with a guaranteed minimum throughput, or choose the VBR type by selecting the policing parameters in order to reflect the idiosyncrasy of the traffic to be generated (peak rate, mean rate, burst size), and require the appropriate QoS parameters from the network. The above observation shows that the user’s choice will depend on the way the various contracts are priced, and it implies a consistency requirement to the pricing schemes. Of course, using a particular contract type it might be more expensive since this would provide the user with a larger set of capabilities, although he might not be able to use them.

2.4 Network Control for Congestion

In ATM networks congestion is due to instantaneous high loads that can appear to the switches of the network, which are due to the burstiness of the input traffic and the multiplexing that takes place. The effects of congestion are the large delays that the cells will experience due to the building of large queues of cells at the switches in the congested area, and the dropping of cells when the buffers become full. Since the important new feature of the above networks is the capability of guaranteeing a certain performance level to the on-going connections (according to the corresponding contracts), the capability of effectively dealing with congestion is a key issue.

Due to the high speed of the links (155 Mbps, 622 Mbps, 2.5 Gbps and beyond) and to the proportionally large round-trip delays (due to the speed of light), it turns out that traditional congestion control using feedback will fail. This is the case because from the time congestion builds up until the network can find out about it and slow-down the inputs that contribute to the congestion, an extremely large amount of data will enter the network and get probably lost. Also, as we have already mentioned, once a guaranteed service connection is accepted by the network (and the corresponding contract is “signed”) the network has no right to alter any of its responsibilities with respect to this connection. Hence the only way to make guaranteed services immune to congestion is to be able to accurately predict the probability of congestion, and to sign a new contract only if the resulting probability of congestion due to this added user remains within the pre-specified levels. This “preventive” approach for congestion control is implemented in the call acceptance mechanism of the network. What essentially call acceptance does is to check if there are enough network resources to be
reserved for the new service, in which case it accepts the call (in the opposite case it blocks the call).

The above discussion suggests that the amount of network resources “used” by a connection while being multiplexed with other connections becomes a key quantity for accepting calls and for charging according to usage. We elaborate further in this direction since it is key to our approach.

2.4.1 Handling Guaranteed Service Traffic

**Effective bandwidths and call acceptance**

A call, which might be a mixture of voice, video and data, would appear to the network as a stream of cells, and the hope is that calls with a broad range of burstiness characteristics can be efficiently integrated, through statistical multiplexing, to share a common resource. In recent years a number of papers [dVW93,EM93,GH91,Kel91a], have provided some basis for this hope, by showing that it is possible to associate an effective bandwidth with a source type such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee, expressed in terms of the probability that delay exceeds a threshold or that a cell is lost.

The effective bandwidth depends on characteristics of the source such as its mean and peak rate, its amount of burstiness, of the QoS required by the connection, and is in general less than its peak rate (and of course above its mean rate). It is beyond the scope of the present paper to derive explicit formulas of effective bandwidths. Sophisticated mathematical tools [CW95,CFW94] show that for large systems (large in the amount of network resources available, i.e., bandwidth and buffer at the switches, and hence capable of a large amount of multiplexing), effective bandwidths can be effectively computed for a large variety of traffic streams, where the QoS criterion is the probability that a is lost or that the time a cell spends in the network exceeds a large amount of delay.

Summarizing the previous ideas, the effective bandwidth approach is an important mathematically justifiable abstraction of the complex actual problem since:

*It allows to summarize the details of the network resources by defining an abstract networking resource whose quantity, the effective network capacity, depends on the context definition, i.e., types of contracts to be fulfilled, actual networking resources (in term of buffer and bandwidth), etc. Then, to each traffic stream one can associate a scalar, its effective bandwidth in the above context, which is dimensioned consistently with the abstract networking resource. The maximum resource utilization subject to the constraints of the contracts is achieved when the sum of the effective bandwidths of the connections that share the resource is equal to the amount of the effective network capacity.*
Note that the definitions of the total resource amount and of the effective bandwidths are context-dependent! These critically depend on what is important in terms of performance guarantees, and not only on the statistics of the cell flows.

The important consequence of the above framework is that the percentage of the networking resource used by a stream is exactly the ratio of its effective bandwidth divided by the amount of the effective network capacity. This consists the base of our usage-based pricing scheme we propose for guaranteed services (as we will explain, it does not make sense to define in a similar way effective bandwidths for best-effort services).

**A general mathematical formula**

Reviewing some of the mathematical results from [CW95, KWC96], if $X(t)$ is a stationary point process denoting the arrivals of cells in continuous time $t$, and $A[0, a]$ denotes the number of cells that $X(t)$ produced in the interval $[0, a]$, the effective bandwidth $eb$ of $X(t)$ can be defined as ($\mathbb{E}$ denotes expectation):

$$eb(s, a) = 1/sa \log \{ \mathbb{E}(\exp\{sA[0, a]\}) \},$$

for appropriate choices of the parameters $s, a$ (these depend among others on the QoS requirements, the duration of the burst of the process relative to the available buffer, etc.). Then if the amount of the resource is large relatively to the needs of the individual connections (i.e., allowing a large amount of multiplexing), the condition that a quantity $C$ of the actual resource suffices to provide the above QoS becomes (“nearly”) equivalent with the condition that the sum of the effective bandwidths of the individual connections be less than $C'$, the effective network bandwidth, where $C'$ depends on $C$ and on some other parameters of the system (QoS, etc.). It is beyond the scope of this paper to provide more details on this subject.

What is important is that in many cases one can successfully come up with reasonable values for $s, a, C'$. There are also ways to estimate on-line by performing direct traffic measurements the value of the quantity $\mathbb{E}(\exp\{sA[0, a]\})$, which is needed for the calculation of the effective bandwidth. This method is promising because it does not require any a priori knowledge about the traffic statistics, and special techniques allow in many cases the estimation to take short time, much less than one minute. Of course there are also many cases where such techniques are not effective, in which case one should use heuristic techniques.

2.4.2 Handling best-effort traffic

Consider first UBR traffic. Since this type of traffic does not require any guarantee from the network, there is no contract to be “signed”, and the network accepts this traffic at its own risk. If cells are lost, it is the responsibility of the connection to detect it and retransmit them. A natural way to carry this traffic is by assigning to it lower priority than the guaranteed services traffic so that it uses the remaining resources available, without interfering with the higher priority traffic (which uses the network essentially for it self).
An intermediate scheme can be used for ABR. Resource reservation is needed only for the minimum amount of throughput required by the particular stream, which must also be served at a lower priority than the streams of guaranteed services (but higher than UBR).

**Can we define effective bandwidths for best-effort traffic?**
The notion of effective bandwidth is directly related to the notion of service guarantee. The effective bandwidth is the minimum requirement in terms of networking resource quantity, needed by the stream in order to perform adequately. Since in best effort service (forget for the moment the minimum throughput requirement of ABR) there is no such guarantee, there can not be any sensible effective bandwidth definition.

On the other hand, best-effort traffic uses network resources as well. Which is the appropriate measure for resource usage? We will argue that for UBR and simple ABR (with zero minimum throughput guarantee), this is precisely the actual number of cells carried, or, in terms of usage rate, the average rate of cells (number of cells carried over the duration of the call).

**The use of flow control for best effort services**
It turns out that flow control can play an important role in controlling congestion for best-effort services. As we already mentioned, if the propagation delays are relatively small as in the case of Local Area Networks (LANs), congestion information could be used as a feedback to the inputs of the network and indicate that these should reduce their rates. There are currently many such proposals some of which are incorporated to standards. There is an important similarity between the handling of best effort services in ATM networks and the way the Internet functions (the TCP/IP stack of protocols). This remark motivates our approach for pricing such services.

**2.5 The Resource Allocation Problem for Network Services**

Based on the above discussion we formulate the following simple abstract network resource allocation problem which captures many of the key concepts. We assume that:

1. The network consists of a single link of bandwidth $C$ and with equivalent network bandwidth $C'$,

2. at time $t$ there are $N$ guaranteed connections $g_1$, ..., $g_N$ and $M$ best-effort (UBR and simple ABR) connections $b_1$, ..., $b_M$, which use the network,

3. the effective bandwidth of connection $g_i$ is $EB_{g_i}$, $i = 1, ..., N$,

4. the actual rates of the connections at time $t$ are $R_{g_i}(t)$, $R_{b_j}(t)$, $i = 1, ..., N$, $j = 1, ..., M$.

Then the condition that the QoS requirements are met by the network, and the fact that best-effort receives lower priority than guaranteed services implies that the following condition must be met
\[ \sum_i E B_g i \leq C', \quad i = 1, \ldots, N \]  
(c1)

Moreover, since the total actual rate cannot exceed the bandwidth of the link, we need that

\[ \sum_i R_g i (t) + \sum_j R_b j (t) \leq C, \quad i = 1, \ldots, N, \quad j = 1, \ldots, M. \]  
(c2)

The above conditions are key for justifying the pricing approach of the next section.

### 2.6 Standardization Requirements on Pricing

The standardization bodies which do relevant work are ITU (International Telecommunications Union), ATM Forum, ISO (International Organization for Standardization) and ETSI (European Telecommunications Standards Institute). Their recommendations so far that influence pricing are the ones defining traffic classes and contract types along the lines we have mentioned already. Besides this indirect influence, there is current on-going work with the goal to standardize the pricing parameters and how pricing will be incorporated in the negotiation of the contracts between users and the network. The immediate concern of the standardization work is on the formatting and structure (the syntax) of the pricing information and on the interface requirements between users and network, and not on the semantics. Of course, although particular pricing algorithms are beyond the scope of such activities, one needs correct paradigms in order to make sure that the suggestions will not be unnecessarily restrictive. Important such issues are the availability of feedback from the network to the user about his current charging status, the possibility of negotiating different QoS parameters in conjunction with pricing, the traffic parameters influencing the charge calculation, the special handling of the accounting information (security, privacy), etc.

#### Network management and pricing

Accounting Management is one of the five network management functional areas defined by ISO and is widely accepted by most organizations and vendors of networking equipment. Its main function is to measure the use of network and service entities by users, and use this information to compute the corresponding charge. The standards suggest the information flow about pricing to be based on a Call Detail Record which is generated at the network nodes and is sent periodically to higher level management nodes with the updates on the connection’s resource usage profile (volume of communication, status of QoS provision, etc.). The charging for a connection should be computed according to the pricing policy followed by the above accounting applications. These standards remain vague about what needs to be measured, and do not address the issues of the contract negotiations enriched by pricing information, or the needs for dynamic access by the users to the charging information. There is a danger that by overlooking the complexity of the problem, the proposed “syntax” for pricing will be not expressive enough.
3. PRICING NETWORK SERVICES

There are many important issues one should consider when designing the pricing structures and tariffs for an integrated services network. As we have already made clear in the previous section, there are many important technological details one should take into account in order to capture the essence of what comprises really the usage cost in such a network. It is important to remember that the amount of load of the network can not be determined unless one specifies with respect to which QoS criterion this load is to be defined. (luckily enough, the QoS parameters are interrelated, and this simplifies the matter). Hence one needs an approach that unifies the seemingly disparate requirements of the various service types into a consistent framework.

We start with defining a simple economic model which justifies our choice of effective bandwidths as a basis for pricing. We then discuss issues related to the implementation of the model.

3.1 The Economic Model

Usage-based pricing and demand for services
Tariffs and pricing schemes are needed for the network to recover its costs in a competitive and fair way from the diverse population of users, and to effectively allocate network resources. It is widely believed that pricing can play a part in effectively managing congestion. Charges will of course be partially determined by the competitive strategies of service providers. However, it is unlikely that a good charging strategy should be wildly out of line with a desire to charge in a manner that reflects users’ relative use of network resources. Economic theory suggests that usage-based pricing will be employed by service providers in the case of perfect competition. Clearly such a policy depends on the demand functions for the various services which in turn depend on the price elasticity parameters and the way services are priced.

A simple pricing model
Consider the case where the demand for guaranteed services is known in terms of the total effective bandwidth required by such services, and there is an arbitrary demand for best effort services. In this case we can argue that the following policy has some very desirable properties.

- Obtain from the “wholesale” market capacity $C$ such that its effective value $C'$ is equal to the total effective bandwidth of the guaranteed services,
- if the price of $C$ is $\gamma$ ECU per unit time, then charge the guaranteed services with a charge equal to $\gamma/C'$ ECU per unit of effective bandwidth and per unit time,
- admit best-effort services for free!

The reason for purchasing the above resource capacity is that if one gets a smaller value, then because of multiplexing reasons (needs some algebra) the effective capacity will be proportionally smaller, the market price at least the same (since one anticipates lower prices for larger purchases), and hence the price per unit of effective bandwidth
will be higher, which is bad for competition. Similarly, if one buys more capacity, he must either charge for this extra capacity the best-effort customers or charge the guaranteed services at a higher rate.

Note that the assumption that the effective capacity will be always fully utilized implies that under this pricing scheme the service provider will recover his cost. Furthermore, the “paradox” of how can we ever serve best-effort traffic while being “full” with guaranteed service traffic is easily explained since, it follows from condition (c2), that the average rate (in the long run) at which the network serves best-effort is $C$ minus the sum of the average rates of the guaranteed services traffic. The latter is in general considerably less than $C$, even if the equivalent network capacity $C'$ is equal to the sum of the effective bandwidths.

**Important remark**

Our model has an important social implication, which is certainly very desirable: *Under our usage-based scheme, if the demand for guaranteed services is high enough, we could recover the cost of the network while accepting best-effort services for free!* The implication of the above is that the bill for the network will be paid by the “expensive and rich” users (video, etc.) whereas network access to individuals for simple use (WWW access, email, etc.) can be for free.

### 3.2 Implementation

We now discuss important issues about what should be measured from the user traffic in order to charge in a non-trivial way. Then we describe the implementation of the simple model of the previous section.

#### 3.2.1 Some interesting issues

*Measurements and pricing*

*The inability to measure certain traffic parameters can severely threaten the ability to price guaranteed services in a usage-based way.* The theory of effective bandwidths indicates that in order to determine effective resource usage one needs to consider traffic parameters which measure burstiness. Indeed, burstiness and unpredictability are key factors leading to congestion. Since a specific amount of capacity accommodating some bursty connections can provide the same QoS to a larger number of less bursty connections with the same average rate, it is obvious that pricing according to average rate can not reflect actual usage. Similarly, call acceptance and pricing according to peak rate is extremely conservative and unfair. But can we actually price according to the effective bandwidth of the stream, which definitely reflects actual resource usage?

The answer to the above question is directly related to the information the accounting mechanism can get from the traffic. Of course, if the users could know and supply to the network the effective bandwidth of their data connections, this would simplify the problem. This could be achieved by some sophisticated users by performing their own measurements and keeping statistical information about the data communications needs of their applications. In any case, the network could never rely to the user declarations.
since the users might be malicious, or might very well not know what traffic statistics their connections will generate.

Modern hardware and network management technology allows the simultaneous real-time calculation of many statistical parameters for large numbers of connections in parallel, and their further processing in order to compute the necessary charging functions. This is one of the important research tasks of the project CASHMAN [CASHMAN96].

An important fact is that the network must use some prior knowledge (or be able to make a reasonable guess) about the resource requirements of a new incoming connection, in order to determine if it can be accepted given the existing load. The more accurate this information is, the less chance will be for the network to reject unnecessarily new calls, or overload the system and disappoint its already existing customers. Hence there is an incentive to be given to the users for providing initially information accurate as much as possible. And the only way to provide incentives is through pricing.

**Policing and pricing**

Of course, there is always a way out from the above problem, by the network assuming a worst case behavior from the user.

In order for this approach to pricing to be realistic and not to be overly conservative and unfair towards the user, the user should choose the policing part of the contract in a way that closely reflects his actual traffic. This is definitely not a desired approach since it reduces the flexibility of the user and the multiplexing gain ATM is offering. Also in most cases the worst case traffic remains “far” from the actual traffic (in terms of effective usage) even if one chooses the optimal policing parameters.

### 3.2.2 Implementing the simple pricing model

An important concept in the approach we present in this paper is the separation of the pricing problem in two complementary parts. The first part was the formulation of the simple pricing model based on the abstract resource allocation problem under constraints (c1) and (c2), where many details of the underlying network and service technology are hidden. The second part is the mapping problem of the actual network and service requirement situation to the abstract model.

A requirement for the above approach to work is that abstract problem is “expressive” enough in order to capture most of the key issues relevant for determining the price of all the types of services described in the previous sections. In other words, the important issue is the possibility of a uniform treatment of the complicated details of the contracts for guaranteed services due to the large diversity of QoS parameters. We can not claim that this problem is solved completely, but there are some generic cases where the approach works well, and we have indications that the remaining issues can be successfully addressed as well.
**Important implementation assumption:** The available network technology provides the capability of performing sophisticated traffic measurements needed to compute for the on-going connections the effective bandwidths from equation (*).

Our theory works even in the case of restricted user traffic information by taking a more conservative approach in terms of required traffic measurements.

**The mapping problem**
The first step is the definition of the actual networking resource quantity $C$ and its price $\gamma$. For simple networks (single link approximation) $C$ is the total available bandwidth under price $\gamma$. The general discussion on how to produce this simple scalar quantity in the case of networks and switches with complex buffers reservation possibilities is beyond the scope of this paper and is in most cases still open. The approach for pricing guaranteed services is summarized in Figure 1.

The information from the QoS requirements and the dimension of the network is used to calculate the parameters $s$ and $C'$. The statistics of the cell flow are used for deriving the value of $a$ and for calculating the main term of the equation (the log-moment generating function of process). This can be done by actual measurements in conjunction with some prior user information (there could be traffic profiles that the network uses based on prior history information about its users).

**Discussion**
We summarize some of the issues and assumptions
1. we do usage and not congestion pricing,
2. we assume that given any QoS contract, and in a given networking environment there is a “sensible” effective bandwidth formulation for the resource allocation problem, i.e., there is a resource of amount $C$ and the effective bandwidth has a form similar to $(\ast)$ for which we can compute the necessary parameters; our theory so far allows the multiplexing of connections with identical QoS requirements,
3. the pricing scheme for best-effort can only be used for UBR services and ABR with no minimum throughput guarantee,
4. the average rate of a best-effort connection refers to the actual rate of delivered cells which is the same as the input rate if losses are not substantial (see next remark).
5. the QoS parameter for cell loss required by guaranteed services is typically so small that cell loss does not affect the effective bandwidth of such connections.

4. EXTENSIONS

4.1 A Two-Layer Network Service Model

We know that in actual networks, there will be a hierarchical provision of services, where service providers will buy simple services and repackage them with some added value to be sold to their customers. Our basic model is consistent with the above notion and can be easily used to formulate a two-layer service model (and possibly to a multi-layer model) as follows.

There are two layers of services. The top layer is the “retail” layer, where there are many service providers, each of which sells a subset of guaranteed and best-effort services along the lines of the previous discussion. In order to do so he multiplexes an amount $C$ of resource (bandwidth) which he buys for an amount $\gamma$ from the “wholesale” market of networking resources, which is the bottom layer of the model. In this bottom layer, the prices are determined by sheer competition, by the usual supply-demand rules. Hence the equilibrium of the bottom layer economy will produce the whole sale resource prices, and the amount $C$ a service provider should buy at the market price will depend on his capability to attract a sufficient number of customers to pay for the cost. For a general formulation and solution of the “wholesale” pricing problem the reader is referred to [LV93,MM94].

This formulation of the multi-layer pricing model natural fits in the ATM layered service framework where the bottom layer sales Virtual Path bandwidth (bearer service provision), and the upper layers buy the above bandwidth to construct a Virtual Path based Private Network, used to provide the traditional services (video, voice, data, etc.). It is interesting to note that buying large chunks of capacity provides multiplexing advantages which in turn reduce the usage-based pricing costs, and hence provide competitive advantage. Of course, the market for services should provide the necessary demand in order to ensure high resource utilization. For a formulation of this problem see [CWW96].

4.2 Optimizing Social Welfare

One can consider a different economic model where there is a demand for services specified by price elasticity parameters, and the goal is the maximization of a standard social welfare function being the sum of the user surplus and the payment to the network. The technology of the network and services remain the same, and hence the
feasible region of the resource allocation problem remains the same as described by equation (c1) and (c2).

The results in [dVB95] suggest that at the optimal pricing point, guaranteed services are priced proportionally to their effective bandwidths, and best-effort services proportionally to their mean. It is also the case that given two services with the same average bit-rate, the first being of the guaranteed type and the second being best-effort, will always be priced in a way that the guaranteed one incurs a higher price (per unit time).

This important result confirms the important role effective bandwidths play in pricing. It also suggests that networks might be charging best-effort services as well, hence requiring the measurement of mean rates. Note that this is also the case if in our model we have only customers that are best-effort, in which case we should also charge according to the mean.

### 4.3 Pricing ABR Services

The pricing scheme for best-effort presented previously is essentially for UBR traffic. The case of ABR is substantially more complex and there are many difficult technological issues that must be resolved for an ATM network to be able to support that service (one might need an intermediate priority level and specialized scheduling and buffer management at the switches) To price ABR, the minimum guaranteed bandwidth will be priced according to the effective bandwidth of the corresponding CBR service, and the “rest” of the traffic as UBR. Defining what “rest” of the traffic means, is a non-trivial matter. Assuming that losses will be small, one could measure during small intervals the input rate and compute the difference between this rate and the guaranteed rate. Then the average of these positive differences over the duration of the call would consist the average rate of the UBR component of the connection.

If losses are substantial, then one should compute the charge on the traffic that actually went through, which is a non-trivial matter requiring more measurements at multiple points in the network, and a substantial cost for moving and processing data.

### 4.4 Pricing Reveals Information About User Behavior

We have already discussed the fact that the network can greatly benefit form the user providing an accurate initial declaration about his traffic resource usage requirements, and what is needed is a pricing-based mechanism that trades off the user's uncertainty about a commencing call against the network's requirement to statistically multiplex calls in an efficient manner (which requires pre-call traffic information).

An effective pricing mechanism should also protect the network from users who seek to gain unfair advantage by “malicious” tactics. One such tactic might be for a user to misrepresent his intended use. Clearly, this is unwelcome, since if users provide accurate estimates of their prospective utilization of the network resources, it should be possible to increase revenue by more efficiently loading the system. Therefore the pricing structure should encourage users to improve the accuracy of their estimates and thereby reduce their cost.
An approach offering a solution to the above problem is proposed by Kelly[]. It extends nicely the general approach presented so far and can be summarized as follows.

Besides the traditional parameters in the contract negotiated during call set up, the user is presented by the network a set of tariffs from which he must choose one. A tariff consists of the parameters of the charging function that the network will use to charge the traffic of the user, and this function uses as arguments certain statistical parameters of the traffic.

The user will choose the tariff which minimizes his expected cost given his prior knowledge about the values of his traffic parameters, and hence his choice reveals to the network important information about his (anticipated) traffic profile. The way these tariffs are constructed guarantees that

1. if the prior knowledge of the user is accurate, then the resulting charge will be proportional to his effective bandwidth, hence the method extends our basic approach,
2. if the user is not accurate, then his charge will be higher than his actual effective bandwidth,
3. the more accurate the user gets the less his expected charge will be.

This method is based and requires the measurement of traffic parameters necessary for determining the effective bandwidth of the connection (*).

A nice feature is that the method still gives interesting results if one requires and measures only very simple parameters such as peak and mean rate. In this case, the tariffs are computed by assuming a worst-case scenario (i.e., by assuming the worst input traffic with the above parameters) and the resulting charge is conservative and exceeds the actual usage. The important property (3) remains valid in this case.

4.5 Congestion Pricing

Congestion pricing is an important mechanism used for controlling access to congestible resources such as freeways, network servers and communication links. The key idea in congestion pricing is that the pricing of a service should reflect the ongoing congestion, and hence discourage users which do not really need it for accessing the resource and contribute to the congestion. The optimal congestion toll is the resource access cost imposed by congestion pricing, and has the property that it achieves socially optimal resource utilization. This cost is the incremental social cost (sum of additional cost that each of the remaining users will incur due to the new user) at the given time (depends on the state of the system). Then a user will use the service only if his incremental benefit for using the service is higher than the sum of the congestion toll plus his private usage cost at the time of the congestion. Hence each user will voluntarily make the social optimal decision.
There are many benefits in using such a pricing scheme, such as allocating excess capacity only when really needed, and avoid the bad effects of unnecessary congestion. Congestion pricing is currently proposed for the Internet, where fixed pricing (namely, pricing of the access line solely) has led to congestion. For related research on pricing for the Internet see [PIEW95].

In order to use the above concepts in the case of networks, one must solve the following issues.

1. Define what cost is. This must be defined in terms of QoS parameters such as cell delay, cell loss probability, jitter, call blocking probability, total transaction time, etc.,
2. be able to compute the incremental social cost on-line and in real-time,
3. be able to communicate this information to the access point of the network.

In many cases congestion pricing is not realistic because computing the incremental social cost on-line and in real-time can be hard for the following reasons.

1. Defining what cost is can be very complex. For a particular user, the cost must be defined in terms of his particular QoS requirements, and hence there is no “universal” monetary cost measure,
2. even in the case of a network consisting of a single link, computing the incremental social cost is hard since the existing connections could have different QoS requirements and hence different cost definitions, and hence computing the incremental change of the above sum of the costs by the addition of the new connection might be extremely complex,
3. in the case of networks one should repeat the above procedure for each link to be potentially used by the connection,
4. this information should be consistent and available at the access points of the network which implies high computation and communication costs.

Nevertheless, we believe that a restricted use of congestion pricing can be used in ATM networks for best-effort services with no QoS guarantee as follows.

**Pricing simple best-effort services**

An important feature available in ATM which is not available in the Internet technology, is the existence of low-level flow control mechanisms which are build-in into the network. In other words, (if such a mechanism is available) we get for free explicit congestion indications at the entry points of the network. The current standards assume that the sources will react to the congestion indications in a sensible way but provide no way to ensure it, hence the need for policing.

The mechanisms for flow control are of two basic types. One is “credit-based”, and a source can not put a cell in the network unless there are credits available, in which case it reduces the number of tokens by one; the network sends credits back to the source to reflect the availability of networking resources. Hence a credits corresponds to a resource unit (for example buffer space) needed by a cell.
The other mechanism is “rate-based”. When the buffer occupancy at a switch exceeds a particular level, the network makes sure that a congestion notification arrives soon after at the entry points of the network to the connections that contribute to the congestion phenomenon. These connections must use the information available in the notification in order to reduce their input rate in a way which will ensure that congestion will not occur. Periodically the sources compete for extra bandwidth, and they get it if no congestion notification arrives.

One can easily see that above congestion control mechanisms can be easily extended to incorporate pricing. Following the congestion-pricing approach from [], the information fed back to the users should include also the congestion cost (calculated by the network). The users will respond to the congestion notification by checking the current access price posted in the notification. Similarly, in the case of tokens, each token should be priced accordingly, and a user will use it only if his utility justifies the cost. There are some important details missing in order to make the approach rigorous, but these are beyond the scope of this paper.

We should remind the reader that congestion pricing provides no service guarantee, and hence it can only be used in ATM for best-effort services which do not require any such guarantees (UBR, simple ABR with no cell loss requirements).

**Managing blocking of best-effort services with congestion pricing**

An important performance parameter for guaranteed service connection is the probability of being blocked because of insufficient resources at call set-up. Hence if the price of the unit of the effective bandwidth charged by the network is constant and does not reflect demand for bandwidth, we have the situation that the bandwidth allocation will not reflect actual user needs. Hence we should use congestion pricing to remedy this situation. There are many possible scenarios one could propose, all of them using prices determined by effective bandwidths. An interesting conjecture is that the price of the unit of effective bandwidth should be priced according to congestion independently of the size of the connection (its total effective bandwidth). An important issue is the calculation of the blocking cost that is caused by a connection to the rest of the system.

A simpler model proposed by Stahl [S95b] uses a queue where blocked connections wait for their turn. There is important work to be done in order to define the right economic model which extends the congestion-pricing model in [GSW95a,b] for best-effort services to include guaranteed services with the above type of congestion.

5. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

We have provided an overview of the networking technology for the B-ISDN based on ATM, highlighting the issues relevant for pricing. We have explained the key new concepts such as the support of guarantee and best-effort services, the contracts between the users and the network at call set-up which define the responsibilities and the requirements of the two sides, and the need for exploiting the large multiplexing capabilities. All these are directly related to the way network connections should be charged.
We have provided strong motivation for using the notion of effective bandwidth in order to determine the usage of network resources by the traffic streams by proposing a simple economic model where under usage-based pricing, the guaranteed services should be charged according to their effective bandwidth, and best-effort services according to their mean rate (or carried for free depending on the system utilization by guaranteed service traffic). The use of effective bandwidths for charging produces a coherent framework and simplifies greatly the problem. We addressed the issues related on the on-line calculation of effective bandwidths by direct traffic measurements.

It was beyond the scope of this paper to provide the technical details on the exact definition of effective bandwidths, and it is true that many important details were hidden “under the rug”! The basic thesis is that given a traffic stream and a service requirement, one can define its effective bandwidth in the above context. We have to admit that there is still research needed in order to validate this thesis. So far we know that for the most basic requirements such as small loss probabilities and large delays the existing theory provides the necessary results and tools for the case of large systems, which is the case of the target networks (large in the sense that these will carry many simultaneous connections). For these systems we know that the effective bandwidths are reasonable approximations of the actual resource usage by the various streams. Unfortunately, the standards made things more (and perhaps unnecessarily) complex by introducing more “exotic” QoS parameters (such as jitter) and allowing a contract to contain any possible combination of such parameters. For such “exotic” contracts we do not have an immediate answer, but we believe that the situation can be greatly simplified, since one should find in a contract the dominant QoS parameters (in most realistic contracts one parameter is the bottleneck, or there can be some other parameter which can be used as an indicator for the actual one), and price according to this parameter.

Important issues for further research are the construction of general hierarchical service models by extending the two-layer model proposed in this paper, where one could investigate issues such as tariff consistency and refinement. We mentioned in the previous section some ideas about extending congestion pricing in order to control blocking of best-effort services.

Some more questions of general interest include architectural issues related to the availability and abstraction of the charging information to be used by intelligent networks and smart agents.

**Some more implementation issues**
Besides the issue of “what” to measure, another important issue “where” to measure. The existing technology provides constraints of where certain measurements can be performed, since parts of the network are either already available, or since the standards did not predict such a need for sophisticated measurements, the products already being sold in the market do not support some of the required features.

Hence in order to implement the methodology suggested in this paper, we need to use specialized hardware which should be place in a transparent way in the existing
networks at the places that the existing infrastructure allows. This might be a serious complication in the application of our methodology.

Another important issue is the question of how sensitive is our approach in the case of simpler measurements, i.e., in the case that a subset of the statistical parameters of the traffic can be estimated in the network. There could be many reasons for this simplification, one being the case of the hardware not being able to perform in real time all the necessary calculations needed in order to compute the terms in (*). There is ongoing research with promising results towards classes of simple and inexpensive to implement algorithms, which provide approximations to (*), see [CKW96].

An important remark is that theory suggests a methodology to price using simpler parameters (for example the mean and the peak rate) in a way that the pricing will always be conservative, and hence will still immune the network and reflect usage. Of course, users will be priced an a “worst case” fashion, which might not always be fair. One needs to investigate the quantitative aspects of this approach, namely the decrease of measurement and computation overhead, and the amount of the excess charge to the user that will occur in most realistic situations.

A most important issue we conclude with, is the cost of charging itself. From traditional telephony we know that billing contributes the largest part of the cost of a call. Clearly, the approach in this paper suggest a sophisticated mechanism, which could be expensive to implement. Hence an important issue is the “optimal” implementation of such an approach by (a) following a “traditional” approach and extend existing architectures in order to meet the new requirements (by standards, real-time constraints, HW and SW computation requirements, topology), or by following some innovative approach similar to the new billing methodologies proposed for the Internet. Clearly one should justify the added value of the extra sophistication our approach requires, since there some important arguments for using much simpler pricing schemes with trivial implementation cost [AS95].

Acknowledgments: We would like to thank Frank Kelly and Richard Weber for many helpful suggestions, and George Manos for his interesting comments.

REFERENCES
[CASHMAN95]Charging and Accounting Schemes in Multi-Service ATM Networks. ACTS Project CASHMAN AC-039 of the EC. Project Web Page URL: http://www.isoft.intranet.gr/cashman/


Costas Courcoubetis was born in Athens, Greece. He received his Diploma (1977) from the National Technical University of Athens, Greece, in Electrical and Mechanical Engineering, his M.S.(1980) and Ph.D (1982) from the University of California, Berkeley, in Electrical Engineering and Computer Science. He was a Research Assistant in ERL (Electronic Research Lab) in U.C. Berkeley, an AMTS (Associate Member of Technical Staff), Summer 1981, Bell Laboratories, an Adjunct Assistant Professor in the EE department, Columbia University, Visiting Lecturer at the EECS department, U. C. Berkeley, 1986, from 1982 until 1990 Member of the Technical Staff (MTS) in the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ, and since 1992 Associate Professor at the Computer Science department, University of Crete, Heraklion, Greece, and researcher at the Institute of Computer Science, FORTH. His interests are performance analysis of computer systems and high-speed (ATM) networks, network management, formal methods for the specification and verification of concurrent and real-time systems.

Vasilios Siris was born in Rhodes, Greece. He received the B.S. degree in Physics from the University of Athens, Greece, in September 1990, and the M.S. degree in Computer Science from Northeastern University, Boston, MA, in June 1992. Since 1993 he has been working toward a doctoral degree in Computer Science at the University of Crete and is a graduate assistant at the Institute of Computer Science, FO.R.T.H., Heraklion, Greece. His research interests include architecture, performance and pricing of ATM networks, architecture of an integrated services Internet, and Internet economics.

George Stamoulis received the Diploma in electrical engineering (1987, with highest honors) from the National Technical University of Athens, Greece, and the M.Sc. (1988) and Ph.D. (1991) degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. From 1991 to 1993 he was a research associate with the communication networks group of the Department of Electrical and Computer Engineering at the National Technical University of Athens, participating in RACE projects. At the same time, he served in the Hellenic Navy, as a lecturer in the Hellenic Naval Academy. Until 1995 he was with the Development Programs Department of INTRACOM, as a coordinator of RACE projects. He has been appointed (1994) Assistant Professor at the Department of Computer Science, University of Crete, while he is also affiliated with ICS-FORTH (1995). His research interests are in routing and performance evaluation of communication networks and multiprocessing systems, in IBC network management, and in queueing theory. He is a member of the Technical Chamber of Greece and Sigma Xi.