Internet Traffic, QoS, and Pricing

J. W. ROBERTS

Invited Paper

Based on an analysis of the statistical nature of IP traffic and the way this impacts the performance of voice, video, and data services, we question the appropriateness of commonly proposed quality-of-service mechanisms. This paper presents the main points of this analysis. We also discuss pricing issues and argue that many proposed schemes are overly concerned with congestion control to the detriment of the primary pricing function of return on investment. Finally, we propose an alternative flow-aware networking architecture based on a novel router design called cross-protect. In this architecture, performance requirements are satisfied without explicit service differentiation, creating a particularly simple platform for the converged network.

Keywords—*Admission control, Internet, performance, pricing, quality of service (QoS), traffic.*

I. INTRODUCTION

A recognized goal in networking is to realize the convergence of all communications services, voice, video, and data, onto a common IP platform. It is necessary that this converged network be able to meet the various performance requirements of the range of envisaged applications, implying enhancements to the current "best effort" Internet.

In this paper, we critically examine the prospects for creating this converged network using the mechanisms and protocols of standardized quality-of-service (QoS) architectures like Intserv, Diffserv, and multiprotocol label switching (MPLS). We consider a commercial networking context where the viability of the provider is assured by the sale of transport services. Our point of view is somewhat original in that we consider first the statistical nature of traffic and its impact on the way performance can be controlled. This perspective leads us to question the effectiveness of the usual approaches to realizing QoS and to propose an alternative flow-aware networking architecture.

An overprovisioned best effort network can meet most user requirements and has the advantage of relatively low capital and operational costs. There are, however, several disadvantages that make simple overprovisioning inadequate as a solution for the converged network. It is not possible to provide backup selectively, just for the customers who are prepared to pay for it, so that redundant capacity tends to be provided for either all traffic or none. The network is not able to ensure low latency for packets of interactive real-time services while maintaining sufficiently high throughput for data transfers. QoS depends on the altruistic cooperation of users in implementing end-to-end congestion control. The best effort Internet does not have a satisfactory business model, and few, if any, network providers currently make a profit.

The converged network should have a pricing scheme ensuring return on investment while remaining sufficiently simple and transparent to be acceptable to users. Prices thus need to reflect capital and operational costs, and these need to be kept to a minimum. To this end, it is necessary to perform efficient capacity planning and to implement simple traffic management. Both planning and the design of traffic control mechanisms require a sound understanding of how perceived performance depends on demand and available capacity. It is our analysis of the latter dependence that leads us to question the effectiveness of standard QoS mechanisms.

It turns out that network performance is typically very good as long as demand does not exceed capacity. It is sufficient to give priority to packets of real-time flows to ensure their performance requirements. On the other hand, performance deteriorates rapidly whenever demand exceeds capacity, due to a traffic surge or an equipment failure, for instance. QoS mechanisms thus tend to play the role of overload controls: they preserve the quality of users of premium services in these exceptional situations.

We argue that a better method for dealing with overload is to perform proactive admission control at the level of a user-defined flow. This is the basis of our proposal for a flow-aware networking architecture. A recent development, known as the *cross-protect* router, allows this architecture to meet the distinct performance requirements of real-time and data transfer services without the need to explicitly distinguish traffic classes.

The flow-aware networking proposal is outlined in the penultimate section, Section VI. Before that, we present
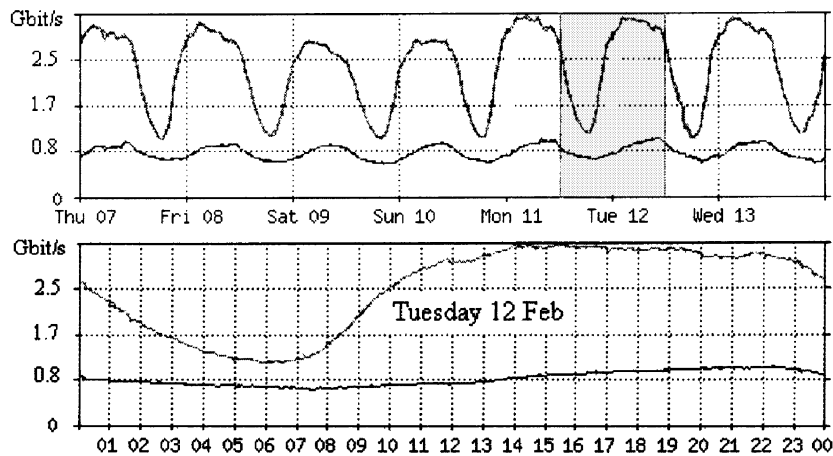
Fig. 1. Weekly and daily demand profiles on an OC192 link.

some key elements of our understanding of IP traffic characteristics (Section II) and of the relationship between performance, demand, and capacity (Section III). In Section IV, in the light of this understanding, we question the usefulness of relying on so-called traffic contracts in standard QoS architectures. In Section V, we argue that congestion pricing is not appropriate in a commercial Internet and that price discrimination cannot be satisfactorily based on QoS differentiation. The proposal for flow-aware networking is somewhat unconventional, and to facilitate understanding, we highlight principal conclusions at the end of each section. General concluding remarks are presented in Section VII.

II. INTERNET TRAFFIC

In the following, we present a brief survey of traffic characteristics having an impact on our ability to realize QoS guarantees.

A. Traffic Variations

Systematic long-term variations are typified by those depicted in Fig. 1. Traffic in bits per second is derived from byte counts sampled at 5-min intervals. There is a clearly recurring busy period, somewhere between 2 and 6 P.M. Traffic in this period attains roughly the same value on successive working days. The network must be provisioned to meet this peak demand while satisfying the quality requirements of users.

To derive the relation between demand, capacity, and performance, we adopt the usual assumption that traffic in the peak period can be modeled as a stationary stochastic process.

Packet-level characteristics of IP traffic are notoriously complex [27]. Arguably, however, this complexity derives from much simpler flow-level characteristics. It proves most convenient to study broad behavior using a flow-based traffic model and to deduce packet level performance, as necessary, in a second phase.

B. Flows and Sessions

By "flow," we mean here the set of packets related to an instance of some user application observed at a given point in the network. A flow is further identified by the fact that these packets arrive closely spaced in time. This is a rather vague definition, but it is sufficient for understanding the nature of IP traffic. A more precise definition is clearly necessary to identify a flow in practice.

Flows generally occur within "sessions." A session observed at a given point in the network consists of a sequence of flows separated by silent periods that we call think times. It is not generally possible to identify a session by simply observing packets in the network. The session relates to some extended activity undertaken by a user such as Web browsing, consulting e-mail, or playing a networked game. An essential defining characteristic is that for all practical purposes, sessions are mutually independent.

When the user population is large, and each user contributes a small proportion of the overall traffic, independence naturally leads to a Poisson session arrival process. Empirical evidence suggests this property is one of the rare Internet traffic invariants [19]. This fact allows relatively simple mathematical modeling, as discussed below, despite the complexity of the arrival processes of individual flows and packets.

C. Streaming and Elastic Flows

With respect to QoS requirements, we distinguish two kinds of flow termed streaming and elastic.

Streaming flows transmit an audio or video signal for real-time play-out. Correct reconstitution of the signal requires low packet loss and delay. The quality of a streaming application also clearly depends on the signal bit rate. Flows generally have variable bit rate due to the use of compression coding.

Elastic flows transfer digital documents corresponding to an e-mail, a Web page, or an MP3 track, for instance. The rate of elastic flows can be varied without significant detriment to perceived performance, which depends on the overall transfer time. The QoS requirement here relates to the response time or, equivalently, to the average throughput over the entire transfer.

It is possible to distinguish different classes of streaming or elastic applications according to their precise performance
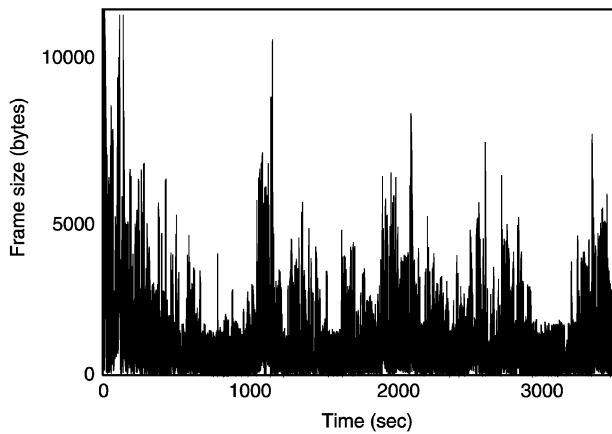
Fig. 2. MPEG4 video trace of *Silence of the Lambs*.

requirements. However, these requirements are rarely absolute, and applications can generally adapt to the quality that is technologically and economically feasible for the network to offer.

D. Characterizing Variable-Rate Traffic Streams

Fig. 2 depicts the rate of an MPEG-4 coded video sequence in bytes per frame. This reproduces a sequence analyzed in [18] and downloaded from the trace library indicated in that paper.[1] The rate varies over multiple time scales exhibiting so-called self-similar behavior. A practical consequence of such variability is that it is very difficult to succinctly characterize a video flow in a way that is useful for traffic control purposes.

In particular, the leaky bucket (or token bucket) is not a useful descriptor for such traffic. Experiments reported in [31] led to the conclusion that the burst parameter needs to be excessively large even when the rate parameter is significantly greater than the actual flow mean rate (see also [37]). This observation has important consequences for QoS architectures that rely on an *a priori* traffic specification, as discussed later.

While a single elastic flow, according to our definition, can be characterized simply by the size of the transferred document, the composite traffic stream corresponding to an aggregate of flows (all the traffic from one LAN to another, for instance) is typically as variable as the video trace depicted in Fig. 2. Aggregate traffic has properties of long-range dependence and self-similarity, as noted by Leland *et al.* [27] and confirmed many times since. It is again extremely difficult to describe such traffic succinctly in a way that is useful for traffic control.

E. Conclusions on IP Traffic Characteristics

The following are the most significant observations.

- IP traffic in the busy period can be characterized as a Poisson arrival process of user sessions, each session comprising an alternating sequence of flows and think times.
- Flows can be classified as either streaming or elastic according to whether their performance requirements

relate to packet loss and delay or overall response time, respectively.
- Streaming flows and aggregates of elastic flows typically exhibit self-similar rate variations that are very difficult to describe succinctly.

III. THE TRAFFIC–PERFORMANCE RELATION

Understanding the traffic–performance relation between demand, capacity, and performance is the key to realizing controlled QoS in a cost-effective way. In the next sections, we discuss the nature of this relation for streaming and elastic traffic.

A. Streaming Traffic Performance

We first consider streaming traffic performance under the assumption that flows have access to a dedicated link.

1) Constant Rate Flows: If streaming flows all have constant rate, performance at flow level is like that in a multi-service circuit switched network (see [33, Ch. 18]). Quality is guaranteed by applying admission control to ensure the overall rate of flows in progress is within link capacity.

A useful traffic management device is to apply a common admission control condition to all flows independently of their particular peak rate. The condition is such that all flows are blocked whenever a flow with the maximum peak rate would necessarily be blocked. This is useful notably in situations of overload when, otherwise, only the flows having the smallest rate would be admitted.

The economies of scale of circuit switching are well known: allowable link utilization compatible with a given blocking probability target (1%, say) increases with the ratio of link capacity to maximum supported flow rate. Networks are most efficient when they federate a large number of demands each having a small bandwidth requirement. It costs more to provision for a given blocking probability as the flow rate increases. A network operator, therefore, has an economic incentive to limit the maximum rate for which blocking is guaranteed to be negligible.

In a packet-switched network, jitter is an important issue, even when all flows have nominally constant rates. When flows are multiplexed in router queues, packets suffer variable delays so that initially periodic flows become jittered. Jitter may increase as flows are repeatedly multiplexed along their path. Our research on the formation of jitter suggests this phenomenon can be controlled, however, simply by ensuring the sum of rates of flows in progress is not more than a certain (high) proportion of link capacity (90%, say) [11].

2) Variable-Rate Flows: Most streaming flows are variable rate, possibly with extreme self-similar behavior, as discussed in Section II-D. In this section, we illustrate some significant results on traffic performance and admission control for variable-rate flows.

It is convenient to assume flows are like fluids with a well defined instantaneous rate. With this fluid-flow model, there is a clear distinction between buffered and bufferless multiplexing. Buffered multiplexing aims to smooth an arrival rate excess with respect to link capacity C by momentarily
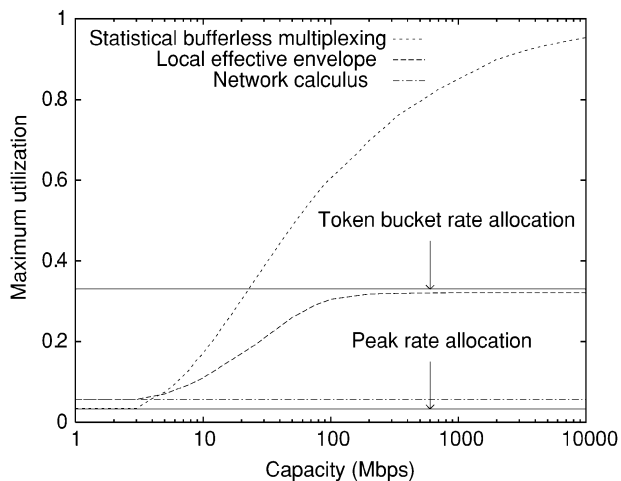
Fig. 3. Achievable utilization depending on assumed multiplexing criterion.

storing the excess in a buffer of size B. Bufferless multiplexing dispenses with the buffer and relies on the overall arrival rate staying less than C.

Packet loss and delay targets must be realized by performing admission control. To illustrate possible admission control options and their performance, we consider a case study with data drawn from [13]. A number of statistically identical flows share a link of capacity C. They have a peak rate of $p = 1.5$ Mb/s with variations bounded by a leaky bucket with burst parameter $b = 95$ kb and rate parameter $r = 150$ kb/s.

In [13], the characteristics of the sources are not further specified. To proceed with our comparisons, we make the assumption that the leaky bucket parameters are chosen to ensure a low nonconformance probability of 10^{-6}. A particular traffic source meeting this requirement for the given leaky bucket has on–off rate variations with exponentially distributed on and off periods. The mean rate is $m = 50$ kb/s, and the mean activity period is of 3 ms (4500 b).

Note the $3:1$ difference between the mean rate m and the leak rate r. This is a typical order of magnitude though somewhat arbitrary, since we could have chosen other traffic characteristics satisfying the assumed conformance probability. The difference would be much greater for a flow with self-similar variations as in Fig. 2. Our choice of source characteristics is motivated by reasons of analytical tractability.

Fig. 3 compares achievable utilization against link bandwidth for four possible admission control approaches when the maximum delay is 50 ms:

1) peak rate allocation;
2) applying deterministic network calculus using the leaky bucket and peak rate parameters [26];
3) applying the stochastic relaxation of network calculus described in [13], i.e., traffic sources are worst case but independent and the objective is a 10^{-6} probability of exceeding the delay target;
4) applying admission control for bufferless multiplexing with a 10^{-6} rate overload probability, assuming the underlying mean rate m is known.

The comparison between peak rate allocation, network calculus, and statistical network calculus is discussed in [13]. The advantage of (realistically) assuming source independence is considerable, especially when the link capacity is large (more than 100 times the peak rate, say). The difference between the third and fourth approaches is also significant. It illustrates our observation in Section II-D that relying on declared leaky bucket traffic parameters is typically very conservative. The relative gain in achievable utilization on a high capacity link is r/m.

Of course, it is not usually possible to know the value of m in advance. The gain from using bufferless multiplexing depends on a further assumption that admission control can be efficiently based on measurement. We claim that this is the case, especially when the number of multiplexed sources is large. Possible approaches are discussed in [20] and [22].

As seen in Fig. 3, bufferless multiplexing is efficient when the flow peak rate represents a small fraction of link capacity. This observation corresponds to a scale economies phenomenon similar to that previously discussed in the context of blocking for constant rate flows.

An additional advantage of bufferless multiplexing is the fact that flow traffic characteristics are broadly the same on leaving the queue as on entering. The same measurement-based admission control (MBAC) can thus be applied throughout the network. Packet delay is low and can be controlled by exploiting the fact that jitter remains "negligible" as discussed in Section III-A1 [11].

A further gain in achievable utilization could be achieved by performing statistical buffered multiplexing with a buffer size compatible with the allowed delay budget (50 ms in the above case study). Unfortunately, this gain relies on being able to account for the detailed traffic characteristics such as burst length statistics and correlation that impact queue behavior. These characteristics are generally unknown *a priori*, and we are unaware of any satisfactory MBAC solution.

B. Elastic Traffic Performance

In this section, we present a summary of results on the performance of elastic traffic, assuming that flows share bandwidth fairly. Fairness is often cited as a objective [7], but is only ever realized approximately in practice. The traffic models are not, therefore, proposed as a precise method for evaluating network performance. They are rather a means for gaining insight into the factors influencing elastic traffic performance and into the scope for realizing quality guarantees.

Under the reasonable assumption discussed in Section II-B that user sessions arrive as a Poisson process, performance of a fairly shared bottleneck is excellent as long as demand is somewhat less than capacity [5]. In the absence of any external rate limits, expected flow throughput is equal to the mean residual capacity $C - A$, where C is the link capacity and A is the demand (flow arrival rate $\times$ mean flow size). Throughput on a network path is mainly determined by the link with the smallest residual capacity [10].

Usually, flows are subject to an external rate limit considerably smaller than the residual capacity. In other words, network links are rarely bottlenecks. The external limit is due,

for example, to the user's access rate or to the load of the server delivering the document. Assuming that a given flow cannot exceed a peak rate p representing the external limit, its throughput observed on a certain link is given approximately by the minimum of p and $C - A$.

The above results are very robust under the assumption of fairness. They apply to flows of any size. They do not depend on precise traffic characteristics such as the flow size distribution of the structure of sessions [5].

Potential for unfairness, due to the dependence of TCP throughput on the round trip time, for example, is generally severely limited by the flow peak rate. Furthermore, even when the link is a true bottleneck, the bias in realized mean throughputs is considerably less than the bias in instantaneous rates due to the fact that the population of active flows is continually changing. Even flows with a low relative share acquire high throughput when the number of concurrent flows is small [9].

While network links can thus appear transparent to the flows of most users in normal load conditions, the impact of congestion can be severely felt when demand exceeds capacity. The impact of overload was studied in [12]. Flow throughput tends to decrease rapidly until some applications are no longer sustainable and flows are abandoned. Note that when demand A is greater than capacity C, at least $A - C$ b/s must disappear due to abandoned or postponed transfers.

Rather than relying on impatience to stabilize a congested link, we maintain that it is preferable to proactively limit demand by performing admission control. The criterion for rejecting a new flow should be such that, in normal load ($A < 0.9C$, say), the probability of blocking is negligible, while in overload ($A > C$), the control rejects excess traffic and ensures admitted flows experience acceptable throughput. It turns out that such a criterion is to reject a new flow if its admission would otherwise make the instantaneous throughput of ongoing bottlenecked flows less than around 1% of link capacity [3].

Generally, 1% of link capacity is much greater than what a user would consider acceptable (for most flows, we have $p \ll 0.01C$). However, it is important to note that there is nothing to be gained by relaxing the admission threshold: link capacity is used (almost) to full in overload, and the blocking probability is approximately $(A - C)/A$, *for any admission threshold*; it is, therefore, preferable to choose a relatively high threshold, as this allows flows without a rate limit to complete their transfer more quickly.

C. Integrating Streaming and Elastic Traffic

Streaming and elastic flows can share the same transmission capacity as long as packets of streaming flows are given priority in multiplexer queues. If bufferless multiplexing conditions are ensured for streaming flows, packet delay, loss, and jitter are low and controllable, as if the flows had access to dedicated capacity. Possible delays behind a long data packet do not change the "negligible jitter" property [11].

Integration is beneficial for both types of traffic: streaming flows see a link with a low effective load leading to very small probabilities of loss and low delays; elastic flows can profit from all the link capacity not currently used by streaming traffic and consequently gain greater throughput.

As discussed in the last section, admission control is necessary to protect performance in situations of demand overload. In an integrated system with a majority of elastic traffic, admission control is facilitated. All flows, streaming and elastic, would be rejected whenever the bandwidth available to a new elastic flow is less than a threshold of around 1% of link capacity, as in Section III-B. Applying the same admission condition to all flows equalizes blocking probabilities. It also facilitates control, since it is unnecessary to signal the flow peak rate. If streaming traffic is not the minority, more complex measurement-based admission criteria, as discussed in Section III-A, would also be applied.

Note finally that for an overprovisioned link whose capacity is very much larger than the peak rate of all flows, it is unnecessary to give priority to streaming traffic. Packet level performance is that of a bufferless multiplexer. This is the current situation of most IP backbones and explains why their performance is sufficient for a voice-over-IP (VoIP) service, for example. Unacceptable delay can occur in case of demand overload or when some sources of elastic traffic do have a peak rate greater than the average residual capacity.

D. Conclusions on IP Traffic Performance

We summarize below our observation in each or the preceding subsections.
Traffic and performance for streaming flows:
- Bufferless multiplexing ensures negligible packet delay and is efficient when flow peak rates are a small fraction of the link rate (less than 1%, say).
- Buffered multiplexing is not controllable or leads to exaggerated overprovisioning if admission control relies on *a priori* traffic descriptors.
- Packet loss in bufferless multiplexing can be controlled using MBAC without the need for *a priori* traffic characterization beyond the flow peak rate.
- There is a maximum peak rate that it is economically efficient for a network to support (i.e., with small blocking probability and sufficiently high link utilization).
- Blocking all flows whenever a maximum peak rate flow would be blocked avoids service bias in overload, when only low rate flows would otherwise be accepted, and facilitates control.

Traffic and performance for elastic traffic:
- Fairness is a useful objective for statistical bandwidth sharing leading to performance that is largely insensitive to detailed traffic characteristics.
- Throughput in normal load conditions is unconstrained by most network links, being mainly determined by an external "peak rate" limitation.
- Dependence of performance on demand tends to be binary in nature with virtual transparency until demand attains capacity and a sudden deterioration in overload.
- Admission control is necessary to preserve flow throughput in case of overload.

- MBAC is relatively easy for elastic traffic.
- A reasonable admission condition is that the current bottleneck fair rate is not less than 1% of the link rate, whatever the latter might be.

Integrating streaming and elastic traffic:
- Integration, performed by giving priority to streaming flow packets and ensuring fair sharing of residual bandwidth by elastic flows, improves bandwidth efficiency and facilitates control.
- MBAC can be employed to ensure negligible packet delay and loss for streaming flows and adequate throughput for elastic flows and is particularly simple when the majority of traffic is elastic.
- Applying the same admission conditions to all flows simplifies control and removes bias which would otherwise favor the least demanding traffic flows.

IV. QoS Guarantees

In this section, we consider the notion of QoS guarantees in the light of the previous discussion on traffic characteristics and the traffic–performance relation. By "QoS guarantee," we mean the type of guarantee envisaged in service level agreements where performance targets are assured for traffic of given *a priori* characteristics. Such "traffic contracts" are proposed in one form or another in the QoS architectures Intserv and Diffserv, as well as in certain applications of MPLS.

A. Traffic Contract

The notion of "traffic contract" implies the following three actions.

1) Users specify their traffic, in terms of characteristics and location, and declare their performance requirements.
2) On the basis of this *a priori* specification, the network performs admission control, only accepting the new contract if all performance requirements, of this and previously admitted contracts, can be satisfied.
3) To ensure that the user respects its side of the contract, either its traffic is policed (or conditioned) at the network ingress, or its allocation of resources is controlled by per-flow schedulers managing router queues.

The contracts may variously apply to individual connections for microflows or aggregates, or to more vaguely defined sets of connections.

B. Contracts for Microflows

In Intserv, traffic contracts can be established for individual microflows. This possibility is frequently dismissed as impractical due to questions of scalability. In this paper, we concentrate on some other difficulties related to the previous discussion on traffic and performance.

The first difficulty resides in the choice of traffic descriptor. The first asynchronous transfer mode (ATM) standards [23] recognized that a traffic descriptor should have three properties:

1) user understandability;
2) usefulness for resource allocation;
3) verifiability by the network.

Unfortunately, these properties seem to be mutually exclusive. For example, the parameters of a leaky bucket are verifiable by design, but they are hard for a user to assign (for a flow like that in Fig. 2, say) and are hardly useful for resource allocation, as discussed in Section III-A2.

It is also rather difficult to meet precise performance targets with respect to packet loss and delay unless these correspond to what can be achieved with bufferless multiplexing. In the latter case, and particularly when the packets of microflows to which the contract applies are given priority in router queues, loss and delay are very small. There is no gain for a provider in relaxing the performance requirements realizable with bufferless multiplexing, even if the applications are more tolerant.

Deterministic guarantees, envisaged in Intserv Guaranteed Service, are somewhat meaningless when network calculus leads to exaggerated overprovisioning, as in the results of the case study illustrated in Fig. 3. One may ask why a provider would advertise a delay bound of 50 ms when the actual delay is always less than 1 ms or so.

Note finally that to require users to previously declare their traffic parameters is a significant constraint to impose on customers and is arguably unnecessary with the use of bufferless multiplexing with MBAC. The only required traffic parameter then is the flow peak rate.

C. Contracts for Tunnels

Instead of a microflow, contracts are frequently established for aggregates of flows. Such contracts might be used to create a permanent MPLS tunnel as part of a virtual private network (VPN), for example, with the aggregate corresponding to the traffic between two sites.

Here again, the user has considerable difficulty in specifying traffic in terms of a descriptor like the leaky bucket. Aggregate traffic displays extreme (self-similar) random rate fluctuations and is notoriously difficult to describe.

In frame relay and ATM networks where such contracts are used, it turns out that users tend to significantly overestimate their demand. A common admission control method is then to overbook available capacity: contractual committed rates are added up and divided by a "fudge factor"[2] to determine the necessary amount of capacity.

Notice that this is a very imprecise form of MBAC, the fudge factor being determined from long-term observations of overall traffic. The difference between actual usage and declared traffic parameters varies widely from user to user, however, making this kind of empirical approach particularly imprecise.

D. Nonlocalized Traffic Contracts

Instead of setting up a set of point-to-point traffic contracts, VPN customers prefer to only specify overall traffic aggregates entering or leaving the nodes of their network. In

---

[2]The exact value is proprietary, but might be as high as five or even ten.

this case, there is no indication of the amount of traffic to be offered to individual links in the provider network.

There are some proposals for dynamically adjusting tunnel bandwidths [17] or designing the network to handle any load compatible with the aggregate traffic descriptors [4]. However, the most practical solution for providers is again to employ a form of MBAC, as for point-to-point tunnels. In this case, the declared traffic parameters are even more irrelevant for resource provision. The provider simply practices over-provisioning based on monitored traffic levels.

QoS for VPN traffic is better than in the best effort network, since the potential for overload is smaller and backup capacity can be used selectively for this class of traffic in the event of failures. This is a valuable assurance of availability, but not a QoS guarantee in the usual sense.

E. Service Differentiation

Nonlocalized traffic contracts are also envisaged in the Diffserv model. Users declare traffic descriptors for the traffic they emit (or receive) in distinct traffic classes. The network provider performs admission control (or equivalently allocates resources for the different classes) in order to respect class-based QoS guarantees.

The most stringent guarantees are for a traffic class based on the expedited forwarding (EF) per-hop behavior. There is, to our understanding, no satisfactory solution to meeting deterministic QoS guarantees without precise information on the paths used by this traffic (see [16] to understand the nature of the problems posed). On the other hand, statistical performance guarantees can be met by giving priority to EF packets and ensuring sufficient capacity is available to perform bufferless multiplexing. The declared user traffic descriptors are then hardly useful, since provisioning must again be based on measurements of the traffic using given links.

The AF group of per-hop behaviors is intended to allow differentiated QoS, corresponding to gold, silver, and bronze service classes, say. Exactly how distinguishable levels of QoS can be achieved remains unclear. Some propose to apply different overbooking fudge factors to the traffic of distinct classes (e.g., overbook bandwidth reserved for gold service by a factor of two, for silver by a factor of four, etc.). This is, to say the least, not based on any analysis of the traffic–performance relation, and the outcome is difficult to predict.

F. Conclusions on QoS Guarantees

A significant conclusion is that the traffic contract does not constitute a satisfactory basis for QoS guarantees.

- QoS guarantees for a microflow cannot reasonably be based on an *a priori* traffic descriptor that is generally a very poor characterization of actual traffic.
- The fact that users systematically overestimate the traffic parameters for tunnels obliges providers to overbook resources, negating the very notion of QoS guarantee in any real sense.
- When the path to be used in a traffic contract is not specified, as in Diffserv, for example, traffic descriptors

are of practically no use for resource allocation and admission control must be measurement based.
- Giving priority to premium traffic[3] protects the QoS of users of the privileged classes as long as their overall demand remains less than capacity.

V. PRICING AND QoS

Pricing is, of course, a vital networking issue and has given rise to a great amount of research in recent years. The subject is complex and a comprehensive discussion is largely beyond present scope (see [32] for a recent survey). We concentrate on the relationship between pricing and QoS. This relationship is often obscured by the dual role of pricing: to assure return on investment and to control congestion.

A. Return on Investment

Return on investment is the prime objective of the network provider. It is necessary that levied charges cover all the capital and operational costs of running the network. Prices of different items (e.g., connection charge, modem rental, usage charges) should be somewhat related to the costs incurred, but there remains considerable flexibility in the way they are attributed.

The cost of an individual IP flow is difficult to evaluate. It is not appropriate to use the marginal cost of handling its packets, since this is arguably negligible. It is more a question of devising a means for sharing overall network costs in an appropriate way. Considerable work on exactly how this should be done has been performed in the context of telephone network interconnection charges (e.g., see [2]). Long run average incremental costs are frequently used to determine interconnection charges.

A similar formalism is not necessary for the unregulated Internet, but the way interconnection charges are evaluated does show that even traffic handled by otherwise idle resources still incurs a cost and is susceptible to charging. A reasonable assumption is that the cost of a flow is proportional to the volume of data transmitted. The cost may also depend on the burstiness of the flow or on whether it is streaming or elastic. However, these considerations are of secondary importance, as they arguably have a negligible impact on provisioning (see Section III).

B. Price Discrimination

Cost is not the only factor determining price. In particular, price discrimination is economically efficient when there exist distinct market segments with different willingness to pay for basically the same service. Many different devices can be employed as a key to discrimination. The airline industry is a useful reference. Business-class comfort justifies a price difference with tourist class that largely exceeds the difference in cost. Further price discrimination is practiced in tourist class by the weekend stay-over clause, which

---

[3]Priority can be realized by many different mechanisms, including priority queuing, weighted fair queueing (WFQ), or weighted random early detection (WRED), but the result is basically the same.

allows leisure travelers to pay less than business travelers for exactly the same QoS.

The need for price discrimination in the Internet is often identified with a need to offer distinct QoS classes. Unfortunately, our understanding of the way QoS depends on traffic volume and characteristics (see Section III) suggests it is not easy to create a networking equivalent to business class and tourist class.

QoS guarantees through traffic contracts (for individual flows or traffic aggregates) are only advantageous in situations of overload. This, of course, may be a useful distinction if overloads are frequent or have very serious consequences when they do occur. However, there is no means to ensure that a premium service is manifestly and consistently better than best effort in the same way that business class is better than tourist class.

Fixing the price of a traffic contract is problematic. The cost of a flow depends ultimately on the volume of data emitted, and not on the traffic parameters declared in the traffic contract. One must, therefore, question the long-term sustainability of charging based on a contractual traffic descriptor.

In Section III, we insisted on the distinction between streaming and elastic traffic. This distinction might constitute a key to price discrimination. However, willingness to pay is not systematically greater for real-time audio and video flows than for data transfers. The per-byte transport cost of both types of traffic is roughly the same.

Alternative keys to price discrimination are probably more acceptable than necessarily vague QoS guarantees. For example, the speed of a DSL modem is a significant price factor in current networks. There is also considerable scope for service bundling and the design of specific pricing packages. These alternatives can effectively segment the market in the same way that the weekend stopover clause segments the market for tourist class air travel.

C. Congestion Pricing

Most research on network pricing is concerned with congestion control and not return on investment. The best known example of congestion pricing is the "smart market" proposed by MacKie-Mason and Varian [28]. In the smart market, users include a bid in each packet. In case of congestion, the users offering the lowest bids are discarded first and accepted packets are priced at a rate determined by the highest bid among the rejected packets.

From this example, it is clear that congestion pricing is not concerned with return on investment. When the network is not congested, there is no charge, so that a well-provisioned network gains no revenue from the smart market. The objective is rather to optimally share a scarce resource by inciting users to reveal their utility and attributing the resource those who gain the most.

The smart market was proposed more as an illustration of the principle of congestion pricing than as a practical system. A more pragmatic approach was advanced by Shenker *et al.* [34]. These authors suggest that it is sufficient to offer differentially priced service classes with charges increasing with the guaranteed level of QoS. Users regulate their charge by choosing or not to use a higher QoS class in times of congestion. A proposal along the same lines using Diffserv classes of service was recently advanced by Shu and Varaiya [35].

Kelly has proposed an alternative congestion pricing framework [24]. His "self-managed networking" scheme is based on a reactive congestion control like that of TCP where explicit congestion notification (ECN) marks are issued to signal imminent congestion. Each mark received by the user implies a unit charge. In the event of congestion, users with high utility continue their transmissions. They receive more marks and pay a surcharge but successfully complete their transaction. Users with low utility will refrain from transmitting until the congestion ceases.

Despite the popularity of the above schemes in the networking research community, there are serious reservations on the use of congestion pricing by a commercial network operator. In the first place, network resources are generally not scarce. The provider can easily upgrade capacity and will do so before congestion occurs if return on investment is assured. Congestion may then be interpreted by users as a sign of bad management. Since other charges must already cover network cost, users might find it unreasonable to pay extra when bad planning or bad maintenance results in congestion.

It is difficult to find examples of other service industries where congestion pricing is successfully employed. Most, like the telephone network, use pricing to share overall costs as discussed in Section V-A. They ensure by provisioning that congestion occurs rarely.

Congestion in the telephone network is manifested by blocking. The use of admission control ensures admitted traffic is completed in good conditions. This appears as a natural condition for the application of simple usage-based charging: the network sells a service that is always of adequate quality; if demand temporarily exceeds supply, not all customers can be satisfied; however, only satisfied customers have to pay.

Experience in the commercial Internet and similar service industries shows that customers have a very strong preference for simplicity and risk avoidance [30], [1]. It is unlikely on these grounds alone that they would ever accept the unpredictability of congestion pricing. Complexity is also an issue for the provider whose operating costs are significantly lower with a simple volume-based charging scheme.

D. Conclusions on Pricing and QoS

Our conclusions on pricing are as follows.

- Return on investment must be assured by appropriately sharing network capital and operating costs between users.
- Price discrimination is economically efficient but should be based on criteria other than pretended QoS guarantees.
- Congestion pricing, used to efficiently share a scarce resource, is not a satisfactory charging basis for a commercial network operator.

- User preference for simplicity and transparency can be satisfied by a simple volume-based charging scheme in a network equipped with admission control.

VI. Flow-Aware Networking

The preceding considerations on the nature of IP traffic, its impact on realized performance, the feasibility of QoS guarantees, and the acceptability of complex pricing schemes in a commercial network lead us to question the appropriateness of commonly proposed architectures for the converged Internet. We believe it is necessary to implement an alternative architecture that we have called flow-aware networking. In the following paragraphs, we briefly outline this alternative vision.

A. Flow Identification

The flow constitutes the appropriate level of granularity for traffic control. It is the closest identifiable object which can be assimilated to a transport service provided by the network. Applying admission control at this level allows the network to protect the QoS of ongoing flows.

In Section II, we gave rather loose definitions of flow and session. These were sufficient for traffic modeling, but to implement flow-level admission control, it is necessary to be more precise. It is necessary to strike a balance between the requirement to identify an entity for which an admission decision makes sense for the user and the need to realize a simple "on the fly" recognition of a new flow.

One possibility with considerable flexibility would be for the user to set a flow ID field in the IP header (as envisaged in IPv6) with the flow being identified by the association of this field with the source address, the destination address, or both. Two bits of the flow ID could be used to specify which of the IP addresses are relevant, as appropriate for a given application. This would allow all the elements of a Web page to be identified as a single flow, for example, by associating the same flow label with the destination address.

Alternative flow identifiers could be used, including the usual five-tuple of IP addresses, protocol, and port numbers, although these do not allow the same flexibility for a user to choose the entity to which traffic controls should apply. In this case, a blocked flow might result in the partial download of a Web page, for example. Observing that this frequently happens in the current Internet when some in-line images fail to display, the consequences would not necessarily be serious. Recall that the alternative to per-flow admission control is indiscriminate packet discard with consequent reduction in throughput for all flows using the congested link (see Section III-B).

B. Flow-Level Implicit Admission Control

Consider a link handling streaming and elastic traffic using priority queuing, as envisaged in Section III-C and assume users identify their flows as either streaming or elastic. Admissibility conditions must be such that performance would be preserved if a new flow of either type were admitted. The admission condition is that current priority load is less than one threshold, determined as in [20] to ensure low packet
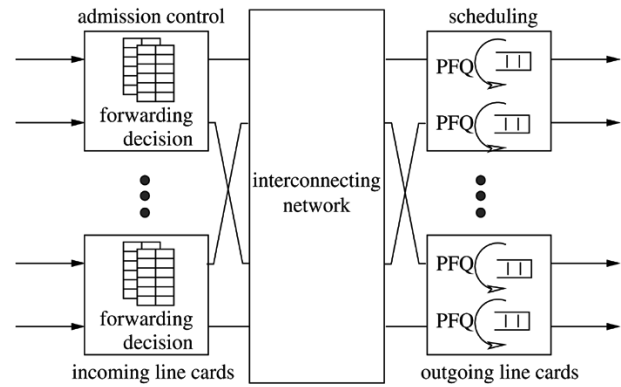

Fig. 4. *Cross-protect* router.

delay and loss for streaming flows, and the available bandwidth[4] for an elastic flow is greater than another threshold, determined as in [3].

To propose per-flow admission control immediately raises concerns of complexity and scalability. The implementation we have in mind limits such problems by avoiding the need for signaling and requiring minimal per-flow state [3].

A newly arriving flow can be recognized as such "on the fly" without explicit signaling. The flow ID of every packet would be compared to a list of flows in progress. If the flow exists, the packet is forwarded; if not, the admission test is applied. If the flow can be admitted, its ID is added to the list; if not, the packet is simply discarded. The loss of this first packet would be interpreted by the user's application as flow rejection, rather like the loss of a probe in an endpoint MBAC [14] or the loss of the SYN or SYN–ACK packet of a TCP connection [29].

Maintenance of the list of flows in progress appears as the most complex task. Consultation of the table is necessary for every packet and must be performed as rapidly as a route lookup. However, this appears to be perfectly feasible using purpose-built application-specific ICs (ASICs), even at line rates of 10 Gb/s [15].

In case of flow rejection, most users or their applications will make retrials. These consist in packet reemissions and are no more troublesome to network performance than reemissions of lost packets by TCP. However, an interesting possibility would be to allow the retransmitted packets to test alternative network paths, thus realizing a kind of adaptive routing. This could be achieved by performing load balancing using a hash function applied to fields of the packet header to choose between alternative outputs. The fields currently used for load balancing are the IP addresses. The proposition is to additionally include the user-defined flow label (or port number). If this label is changed on each attempt, with high probability, the user will test the availability of alternative output links on successive retrials.

C. The Cross-Protect Router

The requirement to explicitly distinguish streaming and elastic flows is a significant constraint. It is notably necessary to police the peak rate of streaming flows and (in the

---

[4]The bandwidth a new flow would attain assuming fair sharing.

absence of signaling) to fix a maximum value for this to be assumed in performing admission control. The fairness of capacity sharing by elastic flows still relies on user cooperation in implementing end-to-end congestion control. These disadvantages are removed with a recent development to the flow-aware networking architecture called *cross-protect* [25].

A *cross-protect* router associates admission control on the incoming line cards and a novel kind of fair queuing on the outgoing line cards, as illustrated in Fig. 4 [25].

The priority fair queuing (PFQ) scheduler realizes unweighted start-time fair queuing [21] with the following modification: the packets of any flow emitting at a rate less than the current max–min fair rate are given head-of-line priority. In this way, streaming flows of peak rate less than the fair rate are multiplexed in the conditions of bufferless multiplexing and, consequently, experience very low packet delay and loss. The value of the fair rate is maintained sufficiently high for envisaged audio and video applications by means of admission control. Any streaming flow whose peak rate exceeds the fair rate will lose packets and be obliged to adapt.

The architecture is called *cross-protect* because admission control and fair queuing are mutually beneficial: admission control limits the number of flows to be scheduled[5] ensuring scalability while the scheduler readily provides the load measurements necessary for admission control. *Cross-protect* is also an appropriate description for the service protection provided to individual streaming and elastic flows whose quality is unaffected by user misbehavior.

A major advantage of *cross-protect* is the absence of any *explicit* class of service distinction. This considerably simplifies network operations, which are essentially limited to those of the current best effort network. In periods of light traffic, users experience low packet delay even for flows with a peak rate greater than the limit assumed in defining admission conditions.

D. Pricing and Flow-Aware Networking

It is not necessary that pricing be flow aware. Indeed, pricing considerations are much simpler than for classical QoS architectures. Usage-based charging can be based on simple byte counting, since all packets, except discarded probes, correspond to flows with assured quality. There is no reason to differentiate streaming and elastic flows for either traffic control or charging.

In Section V, we noted that the main advantage of premium classes of service over best effort was an assurance of quality in case of exceptional demand overload. This distinction is a useful key for price discrimination since, some users are willing to pay more for strict "five–nines" availability guarantees. This kind of discrimination can easily be realized in flow-aware networking by applying differentiated admission conditions. If best effort flows begin to be rejected when the

available rate is below 2% of link capacity, say, and premium flows are rejected only if the available rate drops below 0.5%, the latter are virtually never blocked unless their own traffic exceeds capacity [6].

VII. Conclusion

Analysis of IP traffic characteristics and how these impact network performance leads us to question the appropriateness of currently proposed QoS architectures as a basis for the converged network. We also doubt that pricing can be used to control congestion in a commercial setting where the main role of charging is return on investment.

This analysis leads us to propose an alternative architecture called flow-aware networking. Flow-aware networking meets performance requirements for individual user-designated flows without the need for explicit class-of-service differentiation or the negotiation of traffic contracts. This is achieved by performing implicit MBAC and implementing a novel per-flow scheduler called PFQ. Derived mutual benefits from the conjunction of admission control and scheduling, together with the fact that per-flow performance is protected against malicious use, lead us to call the proposed router mechanisms *cross-protect*.

Flow-aware networking can be introduced incrementally by progressively equipping individual routers. *Cross-protect* can also be used in parallel with the existing mechanisms of Diffserv and MPLS to provide additional quality assurances for best effort traffic. Ultimately, flow-aware networking will allow the development of a business model rather like that of the telephone network. Users would pay in relation to the volume of traffic they generate, with admission control ensuring that this traffic is effective and, therefore, susceptible to charging. A range of price packages, including flat-rate charging, can be used to discriminate between different market segments more effectively than reliance on QoS differences that are practically uncontrollable.

Implementation of flow-aware networking brings a number of interesting technical challenges that we have not fully discussed here. These include the definition of an efficient MBAC and the realization of priority fair queueing in combined input–output queue routers. Users must be provided the means to flexibly define what should be considered by the network as a flow. Applications need to be designed to respond efficiently to the implicit signalling constituted by probe packet discard. However, the most difficult challenge we face is to make people aware of the fact that current proposals for the converged network are unsatisfactory, from both technical and economical points of view, and that flow-aware networking may well be the only viable alternative.

---

[5]This number is measured in hundreds and is independent of the line rate [25].

REFERENCES

[1] J. Altmann and K. Chu, "A proposal for a flexible service plan that is attractive to users and Internet providers," in *Proc. INFOCOM*, 2001, pp. 953–958.

[2] W. J. Baumol and J. G. Sidak, *Toward Competition in Local Telephony*. Cambridge, MA: MIT Press, 1994.

[3] N. Benameur, S. Ben Fredj, S. Oueslati-Boulahia, and J. Roberts, "Quality of service and flow-aware admission control in the Internet," *Comput. Netw.*, vol. 40, pp. 57–71, 2002.

[4] W. Ben-Ameur and H. Kérivin, "New economical virtual private networks," *Commun. ACM*, vol. 46, no. 6, pp. 69–73, June 2003.

[5] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," in *Proc. ACM SIGCOMM*, 2001, pp. 111–122.

[6] S. Ben Fredj, S. Oueslati, and J. Roberts, "Measurement-based admission control for elastic traffic," in *Proc. ITC 17: Teletraffic Engineering in the Internet Era*, J. Moreira, *et al.*, Ed., 2001, pp. 161–172.

[7] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[8] T. Bonald, S. Oueslati-Boulahia, and J. W. Roberts, "IP traffic and QoS control: The need for a flow-aware architecture," presented at the World Telecommunications Congr., Paris, France, 2002.

[9] T. Bonald and L. Massoulié, "Impact of fairness on Internet performance," in *Proc. SIGMETRICS 2001/Performance 2001*, pp. 82–91.

[10] T. Bonald and A. Proutière, "On performance bounds for balanced fairness," *Perform. Eval.*, vol. 55, pp. 25–50, 2004.

[11] T. Bonald, A. Proutière, and J. W. Roberts, "Statistical performance guarantees for streaming flows using expedited forwarding," in *Proc. IEEE INFOCOM*, 2001, pp. 1104–1112.

[12] T. Bonald and J. Roberts, "Congestion at flow level and the impact of user behavior," *Comput. Netw.*, vol. 42, pp. 521–536, 2003.

[13] R. R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Satistical service assurance for traffic scheduling algorithms," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 2651–2664, Dec. 2000.

[14] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang, "Endpoint admission control: Architectural issues and performance," in *Proc. ACM SIGCOMM*, 2000, pp. 57–69.

[15] Flow-based routing: Rationale and benefits (white paper) (2003). [Online]. Available: http://www.caspiannetworks.com/files/Apeiro_Flow_State.pdf

[16] A. Charny and J.-Y. Le Boudec, "Delay bounds in a network with aggregate scheduling," in *Lecture Notes in Computer Science, Quality of Future Internet Services*. Heidelberg, Germany: Springer-Verlag, 2000, vol. 1922, pp. 1–13.

[17] N. G. Duffield, P. Goyal, A. G. Greenberg, P. P. Mishra, K. K. Ramakrishnan, and J. E. van der Merwe, "Resource management with hoses: Point-to-cloud services for virtual private networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 679–692, Oct. 2002.

[18] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation," *IEEE Network*, vol. 15, pp. 40–54, Nov.–Dec. 2001.

[19] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Trans. Networking*, vol. 9, pp. 392–403, Aug. 2001.

[20] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1101–1114, Aug. 1995.

[21] P. Goyal, H. Vin, and H. Cheng, "Start-time fair queueing: A scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Trans. Networking*, vol. 5, pp. 690–704, Oct. 1997.

[22] M. Grossglauser and D. Tse, "A time-scale decomposition approach to measurement-based admission control," *IEEE/ACM Trans. Networking*, vol. 11, pp. 550–563, Aug. 2003.

[23] JT-I371: Traffic control and congestion control in B-ISDN (2000). [Online]. Available: http://www.ttc.or.jp/e/summary/b-isdn/pdf/i371.pdf

[24] F. P. Kelly, "Models for a self-managed Internet," *Philos. Trans. R. Soc. London*, vol. A358, pp. 2335–2348, 2000.

[25] A. Kortebi, S. Oueslati, and J. Roberts, "Cross-protect: Implicit service differentiation and admission control," in *2004 Workshop High Performance Switching and Routing*, 2004, pp. 56–60.

[26] J. Y. Le Boudec and P. Thiran, *Lecture Notes in Computer Science, Network Calculus*. Heidelberg, Germany: Springer-Verlag, 2001, vol. 2050.

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

[28] J. K. MacKie-Mason and H. Varian, *Pricing the Internet: Chapter Public Access to the Internet*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[29] R. Mortier, I. Pratt, C. Clark, and S. Crosby, "Implicit admission control," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 2629–2639, Dec. 2000.

[30] A. M. Odlyzko, "Internet pricing and the history of communications," *Comput. Netw.*, vol. 36, pp. 493–517, 2001.

[31] A. R. Reibman and A. W. Berger, "Traffic descriptors for VBR video teleconferencing over ATM networks," *IEEE/ACM Trans. Networking*, vol. 3, pp. 329–339, June 1995.

[32] P. Reichl, D. Hausheer, and B. Stiller, "The cumulus pricing model as an adaptive framework for feasible, efficient, and user-friendly tariffing of Internet services," *Comput. Netw.*, vol. 43, pp. 3–24, 2003.

[33] J. Roberts, U. Mocci, and J. Virtamo, Eds., *Lecture Notes in Computer Science, Broadband Network Teletraffic*. Heidelberg, Germany: Springer-Verlag, 1996, vol. 1155.

[34] S. Shenker, D. D. Clark, D. Estrin, and S. Herzog, "Pricing in computer networks: Reshaping the research agenda," *ACM Comput. Commun. Rev.*, vol. 26, pp. 19–43, Apr. 1996.

[35] J. Shu and P. Varaiya, "Pricing network services," in *Proc. IEEE INFOCOM*, 2003, pp. 1221–1230.

[36] B. Suter, T. V. Lakshman, D. Stiliadis, and A. K. Choudhury, "Buffer management schemes for supporting TCP in gigabit routers with per-flow queueing," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1159–1170, Aug. 1999.

[37] D. E. Wrege, E. W. Knightly, H. Zhang, and J. Liebeherr, "Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical tradeoffs," *IEEE/ACM Trans. Networking*, vol. 4, pp. 352–362, June 1996.

J. W. Roberts received the B.Sc. degree in mathematics from the University of Surrey, Guildford, U.K., in 1970 and the Ph.D. degree from the University of Paris, Paris, France, in 1987.

He has been with France Telecom R&D, Issy-Moulineaux Cedex, France, since 1978. His research has been mainly in the field of performance evaluation and design of traffic controls for multiservice networks. He was chairman of three successive European COST projects on the performance of multiservice networks, this activity culminating in the publication of *Broadband Network Teletraffic* (Heidelberg, Germany: Springer-Verlag, 1996). He is currently co-organizing joint research activity in the European Network of Excellence, Euro-NGI. He has published extensively and has been a Member of several journal editorial boards, including *Computer Networks*.

Dr. Roberts received the France Telecom R&D Prize for research in 2004. He has been a Member of the Editorial Board of the IEEE/ACM TRANSACTIONS ON NETWORKING. He was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Internet Quality of Service in 2000. He is a Member of many conference program committees in the networking field, including INFOCOM and SIGCOMM. He was TPC Cochair for IEEE INFOCOM 2003.