Quality of service guarantees and charging in multiservice networks *

James W. ROBERTS
France Télécom - CNET

May 1998

Abstract

Quality of service requirements are satisfied conjointly by the service model, which determines how resources are shared and by network engineering, which determines how much capacity is provided. In this paper we consider the impact of the adopted charging scheme on the feasibility of fulfilling QoS requirements. We identify three categories of charging scheme based respectively on flat rate pricing, congestion pricing and transaction pricing.

1 Introduction

There are two main candidates for a future universal multiservice network: the Internet and the ATM based B-ISDN. Some would say that the Internet has already won the contest since it is already providing multimedia services to some tens of millions of users throughout the world. However, the network suffers chronically from congestion and this is severely limittting the development of new applications. The need to introduce quality of service guarantees is thus now widely recognized, although there is still no consensus on how this should be accomplished.

The B-ISDN has been designed with quality of service as a prime concern with multiple service classes standardized by the ITU and the ATM Forum with particular usage categories in mind. However, although ATM is increasingly used in the backbone network, the prospects of deploying an end to end ATM network still appear remote.

Quality of service depends on two factors: the network service model, which determines how resources are shared, and the network provisioning strategy,

*published in the Japanese journal, IEICE Transactions on Communications, Vol. E81-B, No.5, May 1998

which determines how much capacity is available. Both factors depend significantly on the way users are charged for network services. Fundamental differences between Internet and B-ISDN service models can often be traced to the fact that the former is based on flat rate pricing while the latter is conceived as a generalization of the telephone network with per call charging. Adequate provisioning clearly depends on the existence of an economic incentive: a network provider will only install enough capacity if it is economically viable to do so given the charging scheme.

The economy of the Internet is the subject of active research. The recent book "Internet Economics" [15] gives an excellent introduction to the subject. Web pages maintained by MacKie-Mason [12] and Varian [21] and the bibliography established by Klopfenstein [10] point to further useful references. The survey of "frequently asked questions", "firmly expressed opinions" and "partially baked ideas" compiled by MacKie-Mason and Varian is a particularly useful overview [14].

Telephone network economics has also received a large amount of attention recently but from a largely different community. It is here mainly a question of regulating state or private monopolies and providing the necessary economic conditions for the introduction of fair competition. We have by no means made a comprehensive survey of relevant literature in this area, the author's understanding of the telephone pricing issue being derived mainly from the books by Curien and Gensolen [5] and Baumol and Sidak [1].

The present paper is addressed mainly to readers who, like the author, have a background in network engineering and the design of traffic controls. It is intended to stress the importance of including the economic point of view in these activities. We hope that economists will excuse an imperfect knowledge and understanding of their science and accept the following discussion as a contribution to a necessary multi-disciplinary approach to multiservice network design.

We discuss the nature of quality of service requirements for a broad categorization of traffic types. We then consider possible charging schemes, distinguishing flat rate pricing, congestion pricing and transaction pricing, before considering their impact on the choice of service model.

2 Quality of service requirements

In discussing quality of service requirements it is useful to distinguish two broad categories of traffic:

- *stream* traffic entities are flows having an intrinsic duration and rate (which may be variable) whose time integrity must be (more or less) preserved by the network; such traffic is generated by applications like the telephone and interactive video services such as videoconferencing;

- *elastic* traffic entities are digital objects which must be transferred from one place to another; these objects might be files of alphanumeric data, texts or pictures, for example.

Stored video and audio sequences accessed remotely across the network can be considered as stream traffic if they are emitted at their natural playback rate or as elastic traffic if the entire sequence is transferred for storage at the destination prior to playback beginning.

Quality of service depends on the statistical nature of traffic through three main phenomena:

- *transparency,* referring to the time and semantic integrity of transferred data;
- *throughput*, a quality of service measure for elastic traffic defined as the document size divided by the transfer time;
- *accessibility*, the probability of admission refusal and the delay for set up in case of blocking.

Stream traffic requires time integrity while a certain degree of data loss is tolerable. Elastic traffic on the other hand can by definition tolerate delays occasioned by network queueing or flow control but transport should preserve semantic integrity.

Throughput requirements for elastic traffic entities are not well known. It is useful to distinguish traffic emitted for immediate attention, like web pages, and traffic which is essentially deferable such as e-mail. For the latter throughput in the sense defined above is not really relevant since transfers can be delayed for minutes or hours without significant inconvenience. For the great majority of web documents, throughput of some tens or hundreds of kilobits/sec would be sufficient for quasi-immediate local display. The transfer of larger documents (e.g., entire data bases) may require much higher rates.

Accessibility is only relevant to quality of service in a network employing admission control. The probability of blocking depends in general on the rate required by a given transaction and a network provider will only aim to meet a target value for demands up to a certain maximum rate. A blocking probability of around 1% in the busy hour is a typical target in the telephone network.

Quality of service requirements are satisfied jointly by the network service model, which determines how resources are shared, and by network engineering procedures which determine how much capacity is provided. The service model alone allows transparency and throughput guarantees for some users; accessibility and throughput for a given population of users or an entire class of traffic additionally requires adequate network provisioning.

Depending on the definition of the service model, the network may or may not see stream and elastic traffic flows as individual transactions. It is usual

presently, for example, to offer network services for LAN interconnection where the traffic entity is in fact an aggregation of flows. The administrative advantages of this approach (billing, absence of flow identification) must be weighed against the difficulty of providing quality of service guarantees for such traffic. These difficulties stem in large part from the particularly complex characteristics of traffic aggregations [11].

3 Charging options

The feasibility of satisfying quality of service requirements depends significantly on the charging scheme employed. We distinguish three broad categories.

3.1 Flat rate pricing

With flat rate pricing, users pay a fixed charge, every month say, independently of the volume of traffic they produce. The price generally depends on the capacity of their network access line. It should be calculated to cover all the network provider's costs including any settlement charges incurred to provide interconnection with another provider's network. This is how the large majority of Internet users are charged today.

The major advantage of flat rate charging is its simplicity leading to lower network operating costs. A weakness is its inherent unfairness, a light user having to pay as much as a heavy user. The level of the flat rate charge excludes potential users having a lower, though positive, evaluation of the value of using the network. Flat rate schemes are also open to abuse by resale of access capacity. A more immediate problem is the absence of restraint inherent in this charging scheme which may be said to contribute to the present state of congestion of the Internet.

3.2 Congestion pricing

Charging can be used to modulate network traffic dynamically by exploiting the elasticity of demand with respect to price. Congestion pricing allows users to pay for the level of quality of service they judge necessary. The payment is not intended to cover the cost of the network but only the cost of congestion represented by the inconvenience caused to other users who are denied quality of service. Network costs must still be recovered by a flat rate, as in the previous scheme. Congestion pricing relies explicitly on a service model where users can express the worth they attach to their traffic.

A well known congestion pricing scheme is the so-called "smart market" [13]. 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 tariffed at a rate determined by the highest bid among the rejected packets. The

cost of carrying each packet is thus related to the marginal value (represented by the bid) of the traffic which is squeezed out.

If the revenue gained by the network provider exceeds the cost of adding extra capacity, he will have an incentive to expand bottleneck links in order to admit more traffic and thus gain even more. The network will cease to expand when the marginal cost of capacity is equal to the marginal cost of congestion, leading to a micro-economic optimum.

The optimal charging paradigm has a number of disadvantages which have been pointed out by Shenker and co-authors [19]. A major problem is that of identifying the value of a unit of traffic such as a packet when the true worth of a communication is associated with the higher level transaction (e.g., a document transfer or a telephone conversation). The complexity of the scheme and the implied billing system may severely limit the practicality of its implementation.

It is argued in [19] that it is necessary to sacrifice economic optimality in order to take account of structural aspects of the underlying service model. The authors suggest that it is sufficient to offer differentially priced service classes with charges increasing with the guaranteed level of quality of service. Users regulate their charge by choosing or not to use a higher quality of service class in times of congestion. The simplest service model fulfilling this objective has just two differently charged service classes.

A simple two-tier charging scheme would result from an evolution of the Internet in which users are required to pay usage charges for reservations (depending on reservation parameters and duration) but only the flat rate for best effort traffic. In case of congestion, users requiring a certain minimum throughput would be obliged to reserve capacity and thus pay more for their communication.

A second possibility is the two-tier best effort Internet service advocated by Clark where users identify their packets as being "in" or "out" with respect to an "expected capacity profile" [3],[4]. Only "in" packets are charged above the flat rate . The profile may be defined as a long term constraint on expected use as part of the data defining a subscription. Alternatively, a user could define an expected profile for a shorter term session or choose to mark packets as in or out according to his own priority criteria. The definition of the expected capacity profile is left open although the token bucket filter is quoted as an example.

The proposition of Songhurst and Kelly [20] for pricing ABR connections in an ATM network is also based on two-tier charging. In this case, the cost of an ABR connection depends on the minimum cell rate parameter MCR. Users pay more per bit transferred as the value of MCR increases allowing the user to pay to maintain a larger share of throughput in case of congestion.

Congestion pricing has the considerable advantage of allowing users the possibility of expressing the value of their traffic and gaining corresponding priority in access to network resources. If the network provider responds appropriately by investing congestion revenue in extra capacity, the scheme should also lead to an optimally dimensioned network with low congestion.

Likely user perception of congestion pricing is unclear, however, since the cost of a given transaction depends on invisible factors: how can users tell if the network provider isn't deliberately causing congestion? why should they pay more to an inefficient provider? Note that congestion pricing is not generally employed in other service industries subject to demand overloads such as electricity supply or public transportation or, indeed, the telephone network.

3.3 Transaction pricing

In the telephone network, switches and links are generally sized to ensure that demand congestion occurs only exceptionally. This operating model relies on being able to reliably predict demand, given announced charges, and to size the network to avoid congestion when that demand prevails. Since demand fluctuates over time, the representative traffic volume used for provisioning is generally that of an appropriately defined "busy hour". Enough capacity is provided to handle the busy hour traffic with good quality of service (less than 1% of calls blocked, say).

The price must be set at a value allowing the network operator to recover the cost of investment and is determined on a long term basis, ideally at an optimal level such that the revenue from an additional unit of demand, stimulated by further lowering the price, would be equal to the extra cost of carrying that unit. Differential pricing on a time of day basis is used to smooth out the demand profile to some extent but this is not generally viewed as a congestion control mechanism.

The question of charging for the telephone service has been the subject of much study over the last few years as the market has been opened to competition. It is generally recognized that prices for different services (e.g., local, long distance) and market segments (e.g., residential, business) must align more closely with costs. It remains difficult, however, to evaluate these costs taking into account resource sharing and the realization of economies of scale and scope. Certainly, the marginal cost of handling an additional transaction is not sufficient to cover the major part of the costs of a network provider. A more satisfactory basis is to relate charges to an "average incremental cost" per unit of demand, as discussed by Baumol and Sidak [1], taking account of all scale and scope economies attributable to resource sharing. The average incremental cost includes the fixed capital and operational costs of these shared resources.

The average incremental cost of a transaction in a multiservice network depends on the amount of resources necessary to handle the flow concerned. This amount generally depends on many parameters including the traffic characteristics and performance requirements of the transaction and the bandwidth and buffer capacity of the network links it uses. Songhurst and Kelly have discussed a rational transaction pricing scheme for multiservice networks based on the formula:

$$\text{price} = a(x) \times \text{duration} + b(x) \times \text{volume} + c(x)$$

where a, b and c depend in a rather complex way on the traffic and quality of service characteristics of the transaction x [20]. The complexity of this type of scheme may be reduced in the case of a large network where individual transactions only use a small fraction of available resources.

By the scale economies effect, the efficiency of resource sharing increases as network capacity grows. Indeed, networking is generally only economically advantageous when a large number of small users share a large resource. In this case, it may be shown that required bandwidth for a stream flow tends to its mean rate for whatever loss rate while the throughput of an elastic flow remains satisfactory even as links tend to saturation [17].

In this large network limit, it may be natural to replace the above formula by:

$$\text{price} = b \times \text{volume} + c(x)$$

i.e., a flat rate per byte charge b for all transactions plus a set up charge c which may depend of the nature of the connection.

Note that the latter simplified transaction pricing scheme gives no incentive to users to identify elastic traffic flows as deferable rather than intended for immediate delivery. By definition, this type of traffic can use any spare network capacity over and above that required to handle traffic with quality of service guarantees. However, this does not imply that a deferable traffic service class would have zero average incremental cost. This cost must be evaluated in common with that of other elastic traffic with which it would be indistinguishable in the absence of congestion.

Transaction pricing works well for the telephone network because there is only one service whose use can be forecast very accurately. On the other hand, the traffic in any computer network is extremely varied and indeed changes all the time. It is simply not possible to predict future trends based on current usage and, indeed, current usage is itself extremely difficult to characterize and measure. Furthermore, Internet traffic overall is growing at an exceptional rate of more than 100% per annum such that any capacity expansion is rapidly saturated and congestion can hardly be avoided. The following two questions must be addressed:

- is it possible to forecast multiservice traffic sufficiently reliably to be able to provision the network to make demand congestion an exceptional event?
- will transaction pricing perform satisfactorily when demand congestion does occur?

The unpredictability of Internet traffic is undeniable and largely explains and justifies the absence of any established network traffic engineering practice in the Internet. We would argue, however, that this situation is indicative of an initial transitory phase which will disappear as soon as the network has grown

sufficiently to absorb latent demand from the population at large. In a mature network serving a large population of "ordinary" users, it is not unreasonable to suppose that demand for stream and elastic traffic can be characterized on a statistical basis with the idiosyncrasies of individual flows being insignificant with respect to a stable ensemble statistical behaviour.

The second question is particularly relevant in the transitory network growth phase when it may be difficult for a network to keep up with demand. The impact of overload depends on whether or not admission control is employed. Indeed, the absence of admission control appears incompatible with transaction pricing: overload would lead to unacceptable transparency and throughput performance degradation for a paid transaction. Use of admission control is intended to ensure that performance is acceptable for admitted transactions with overload manifested by increased blocking probabilities.

3.4 Other charging issues

As pointed out in [19], effective cost recovery is not the only issue to be resolved in designing the charging scheme. The following questions are particularly relevant:

- who pays for any data exchange: sender, receiver or some third party?
- who receives the payment, the network provider to which the user is directly connected or all network providers used by the flow?
- should users be charged for goodput or for all emitted data, particularly when congestion results in the need for multiple retransmissions?
- what basis should be used to charge for multipoint communications?

4 Choosing a service model

In this section we consider how the adopted charging scheme influences the definition of the service model. We examine successively three major components of the service model: the definition of service classes, the use of admission control and requirements for network mechanisms to allow quality of service guarantees.

4.1 Service classes

If the network provider offers flat rate pricing there is no case for introducing service classes with different degrees of the same quality of service measure (e.g., expected packet delay) since users have no incentive to choose other than the best. However, it remains useful to distinguish between service classes specialized for stream and elastic flows, respectively, since their quality requirements are different: stream transactions require time integrity for flows with a bounded

peak rate; elastic flows require a minimum throughput but both users and network gain by providing a much higher rate whenever possible. Users have a quality of service, rather than financial, incentive to choose the service class corresponding to their application.

Congestion pricing where the individual packets have explicit priority determined by their value may be seen as an enhancement to best effort service which *avoids* the need to introduce distinct service classes. The simplicity of this solution is clearly attractive for the current Internet. To enable quality of service through resource reservation, on the other hand, leads to an indirect definition of service classes through the identity of a flow and the conformity of its packets with respect to the negotiated traffic specification.

Transaction pricing explicitly relies on the definition of service classes and on the characterization of a flow within a service class by a set of traffic and performance parameters. The problem of defining statistical traffic parameters which are relevant for resource allocation and yet can be policed by the network has still not received a satisfactory solution.

If price were essentially determined by the traffic volume realized during the transaction and not by pre-announced traffic parameters, a simple distinction between two service classes could be sufficient: one class for stream flows and another for elastic flows. Stream flows would be characterized by their peak rate (for admission control purposes) and elastic streams would be assigned a common minimum throughput guarantee, no further differentiation being necessary or useful under this charging scheme.

4.2 Admission control

The need for admission control can be avoided with flat rate and congestion pricing at the expense, however, of any real guarantees with respect to transparency or throughput. Transaction pricing on the other hand, as noted previously, relies on admission control. As a general statement, admission control allows transparency and throughput guarantees at the cost of variable accessibility quality of service. The use or not of admission control is one of the most significant differences between Internet and B-ISDN service models.

Reliance on admission control adds considerable complexity to the network, necessitating connection oriented operation with resource reservation. Admission control algorithms can also be very complex, depending on the nature and number of quality of service criteria which must be satisfied for different flows.

For stream flows, admission control is easiest and best understood in the case of so-called "bufferless" multiplexing: delay and loss performance is guaranteed (statistically) by ensuring that the combined input rate of flows multiplexed on a given link is less than the link bandwidth with very high probability. The most promising procedures, like that proposed by Gibbens et al [9], are based on measurements of real traffic combined with sure knowledge of flow peak rates. A stream transaction would not be accepted if, given current traffic

levels, a constant rate flow at the stream peak rate would lead to a non-negligible probability of rate overload.

Bufferless multiplexing is a mathematical model based on the interpretation of stream traffic as a superposition of variable rate fluid flows. In practice it is necessary to provide a non-zero buffer to account for the asynchronism of packet arrivals from different flows and for the imprecision of rate definition in the presence of jitter. The ATM multiplexing scheme known as Rate Envelope Multiplexing allows controlled performance if the peak rate of flows can be guaranteed by spacing cells at network ingress [18],[2].

Required precision of admission control can be relaxed when link capacity is shared with a non-negligible amount of elastic traffic. The latter can easily adjust to the rapid rate variations of the combined stream traffic. Admission control for elastic flows could be performed simply by comparing the sum of minimum required throughputs to available link bandwidth, after subtraction of the current estimated requirement for stream flows. Admission of elastic flows would of course also need to take account of memory requirements, as determined by the implemented queue scheduling mechanism.

Note finally that admission control allows a form of congestion pricing to be applied at transaction level. A given transaction might be admitted or rejected depending on whether or not its value, as declared by the user, were greater than an estimated "shadow price" equal to the expected value of hypothetical subsequent transactions which would be rejected if the considered transaction were accepted.

4.3 Mechanisms for service differentiation

Different service models impose widely varying requirements on network queueing and flow control mechanisms. In fact, requirements on queue scheduling and protocols for flow control are complementary. Simple FIFO queueing may be adequate to ensure quality of service if end to end protocols closely adjust flow rates to avoid excessive overflow. On the other hand, individual flows may have to be protected by complex scheduling mechanisms, like weighted fair queueing, if the network ensures no rate enforcement.

FIFO works well in the Internet when the rate of elastic flows is adjusted according to the congestion control algorithms of TCP and the volume of stream traffic, generally transported using the non-responsive protocol UDP, remains well within the capacity of network links. However, performance does rely on the cooperative behaviour of users which is less and less sure as the Internet expands. The importance of using queue mechanisms to reduce the impact of flows which are "not TCP friendly" is stressed by Floyd and Fall [6].

The introduction of congestion pricing would require mechanisms of varying complexity, ranging from a simple selective rejection device for the two-class service differentiation scheme of Clark [4] to some unspecified means to implement the "auction" in the case of the smart market.

Head of line priority queueing can provide quality of service differentiation. It is possible to closely control the quality of service of stream traffic in the highest priority service class if admission control is also employed. It may not, however, be feasible to realize guarantees for lower priorities without implementing supplementary flow controls like those of TCP or ABR. The same high priority queue can be used for different service classes if the realized quality of service is sufficient for the most exigent class. To realize different loss rates for different classes, for example, would require a more sophisticated class-based queueing discipline.

In a network where users pay for quality of service, in the case of transaction pricing for example, it appears unreasonable to rely on users executing an end to end protocol like TCP in preference to implementing network level flow control. The ABR service class is a particularly complex flow control protocol devised for ATM [7]. A simpler protocol may be sufficient if rate control is applied to individual elastic transactions rather than to bursty aggregates. The protocol is simpler still if there is no requirement to differentiate between flows because they all have the same per byte charge.

We have in mind a window-based protocol like TCP where, however, the need for congestion avoidance is greatly alleviated by the use of admission control. A fixed window size can provide a reasonably fair share of spare bandwidth over and above a minimum throughput guaranteed by limiting the number of admitted flows. The required queueing mechanism would be a simple, but large, FIFO queue in second priority behind the queue reserved for stream flows.

In section 2, we identified the category of deferable elastic traffic (e-mail, etc.) which requires no throughput guarantees. This could be handled by a third and last priority queue served only in the absence of traffic from the stream and immediate delivery elastic flows. To clearly distinguish two elastic service classes, access to this third priority could be restricted to specialized (mail) servers: users pay to deposit their documents in a local mail server; the mail service provider pays the network provider for transporting the document to the mail server corresponding to one or more destination users.

Priority queues have limited applicability if it is necessary to identify more than two or three service classes or if it is necessary to satisfy individually specified flow quality of service requirements. Much more complex mechanisms, including "weighted fair queueing" [16] and "earliest deadline first" scheduling [8], have been extensively studied in the last few years. There are two main motivations:

- to satisfy deterministic delay guarantees;
- to protect individual flows from the traffic of other users.

In this paper we have not considered the need for absolute delay guarantees, as provided for in the guaranteed service category defined by the IETF.

If such delays are necessary, it is clear that simple priority queueing, as considered above, is insufficient. However, we pretend that the stream traffic category only requires that the probability of delay exceeding a given limit be negligibly small. We also believe that this objective can be achieved simply by employing bufferless multiplexing with measurement based admission control, as previously mentioned.

Per flow scheduling as a means to ensure that a flow receives a "fair share" of bandwidth, independently of the traffic on other flows, may be seen as a requirement in a network having no control on input rates. The requirement is less obvious if admission control is employed and the network itself ensures that stream flows respect their traffic contract and that elastic flows are prevented from exceeding their current rate allocation.

5 Conclusion

Quality of service requirements in a multiservice network handling a mixture of stream and elastic traffic concern transparency, throughput and accessibility. The feasibility of fulfilling these requirements depends both on the service model, which defines how resources are shared, and on how much capacity is made available. Charging has a central role with respect to both factors.

We have identified three broad charging schemes: flat rate pricing, congestion pricing and transaction pricing. From the point of view of the business model, however, the options for a network provider reduce to two since the revenue raised by congestion pricing is not intended to cover network infrastructure and operation costs.

The business model of current Internet service providers is generally based on flat rate pricing: users pay only an access capacity dependent connection fee. This fee must be sufficient to cover the cost of both user-dedicated and shared resources. The network provider has limited incentive to invest in additional shared resources to avoid congestion since this would cost more but produce no additional revenue (except by attracting customers from a competitor).

The present Internet best effort service model, with the absence of admission control and reliance on end to end flow control, can provide virtually no quality of service guarantees. Proposed enhancements include the introduction of resource reservation and service differentiation. We interpret the use of these new facilities as a kind of congestion pricing since users only have an incentive to use the premium services in case of congestion. The appearance of non-TCP friendly transport protocols is leading to an additional requirement for more sophisticated queue management mechanisms.

The business model of the telephone network (with some notable exceptions) is based on transaction pricing. The price of transactions is fixed, demand at that price is forecast and the network is provisioned to handle that demand with high quality of service. This model may be generalized to a multiservice

network on condition that admission control is systematically employed. It is then feasible to charge for each transaction since required transparency and throughput are guaranteed.

We have suggested a simple service model based on transaction pricing with just two service classes, one for stream traffic and one for elastic traffic. Stream traffic flows, characterized by their peak rate, would be handled using "bufferless" multiplexing with measurement based admission control ensuring negligible probability of rate overload. Admission control would also be employed to ensure minimum guaranteed throughput for elastic flows. A simple network layer flow control, possibly based on a fixed sized window, could be sufficient to ensure efficient and fair sharing of available link bandwidth. Resource sharing would be assured by a simple queue with head of line priority for stream flows and FIFO service in each class. In a large network, the cost related transaction price could be determined, for both stream and elastic flows, by just the volume of data transmitted. It remains to more fully explore the feasibility of this simple model which should for the time being be classified among the "partially baked ideas" of [14].

In this paper, we have developed the idea that the adopted charging scheme has a significant impact on the nature of the service model. An additional "firmly held opinion" derives from the converse of this statement: a service model without systematic use of admission control precludes the use of transaction pricing and may consequently jeopardize the solvability of the network provider.

References

[1] W. J. Baumol, J. G. Sidak, "Toward Competition in Local Telephony", The MIT Press, Cambridge, 1994.

[2] F. Brichet, L. Massoulié, J. Roberts, "Stochastic ordering and the notion of negligible CDV", Proc. ITC15 (V. Ramaswami, P.E. Wirth (Eds), "Teletraffic Contributions for the Information Age"), Elsevier, 1997.

[3] D. D. Clark, "A model for cost allocation and pricing in the Internet", in L. W. McKnight, J. P. Bailey (Eds), "Internet Economics" [15], 1997.

[4] D. D. Clark, W. Fang, "Explicit allocation of best effort packet delivery service", Laboratory for Computer Science, MIT, Preprint, available via URL http://diffserv.lcs.mit.edu/ Papers/exp-alloc-ddc-wf.pdf, 1997.

[5] N. Curien, M. Gensollen, "Economie des Télécommunications - Ouverture et Réglementation" (in French), ENSPTT/Economica, Paris, 1992.

[6] S. Floyd, K. Fall. "Router mechanisms to support end-to-end congestion control", Lawrence Berkeley National Laboratory, Preprint, 1997.

[7] E. J. Hernandez-Valencia, L. Benmohamed, S. Chong, R. Nagarajan, "Rate control algorithms for the ATM ABR service", Europ. Trans. Telecom. Vol 8, No 1, Jan-Feb, 1997.

[8] L. Georgiadis, R. Guérin, V. Peris, R. Rajan, "Efficient support of delay and rate guarantees in an Internet". Proc. ACM SIGCOMM'96, 1996.

[9] R. Gibbens, F. Kelly, P. Key, "A decision theoretic approach to call admission control in ATM networks", IEEE JSAC, Vol 13, No 6, August, 1995.

[10] B. C. Klopfenstein, "Internet economics: an annotated bibliography", to be published in Journal of Media Economics, available via URL http://www.bgsu.edu/departments/tcom/annota.html, 1997.

[11] W. Leland, M. Taqqu, W. Willinger, D. Wilson, "On the self-similar nature of Ethernet traffic", IEEE/ACM Trans. Networking, Vol 2, No 1, Feb, 1994.

[12] J. K. MacKie-Mason, "Intenet economics", Web pages via URL http://www.spp.umich.edu/telecom/net-economics.html, 1997.

[13] J. MacKie-Mason, H. Varian, "Pricing the Internet", in B. Kahin, J. Keller (eds), "Public Access to the Internet", Prentice Hall, 1995.

[14] J. K. MacKie-Mason, H. R. Varian, "Economic FAQs about the Internet", in L. W. McKnight, J. P. Bailey (Eds), "Internet Economics" [15], 1997.

[15] L. W. McKnight, J. P. Bailey (Eds), "Internet Economics", The MIT Press, Cambridge, (earlier versions of published papers available in the Jounal of Electronic Publishing via URL http://www.press.umich.edu/jep/econTOC.html), 1997.

[16] A. K. Parekh, R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case", IEEE/ACM Trans. Networking, Vol 2, No 2, April 1993.

[17] J. Roberts, "Realizing quality of service guarantees in multiservice networks", to appear in Proceedings of IFIP Conference PMCCN'97, Chapman and Hall, 1998.

[18] J. Roberts, U. Mocci, J. Virtamo (Eds), "Broadband Network Teletraffic (Final Report of COST 242)", LNCS 1155, Springer Verlag, 1996.

[19] S. Shenker, D. Clark, D. Estrin, S. Herzog, "Pricing in computer networks: Reshaping the research agenda", Preprint, 1995.

[20] D.J. Songhurst, F.P. Kelly, "Charging schemes for multiservice networks", in Proc. ITC15 (V. Ramaswami, P.E. Wirth (Eds). "Teletraffic Contributions for the Information Age"), Elsevier, 1997.

[21] H. R. Varian, "Network economics", Web pages via URL http://www.sims.berkeley.edu/ resources/infoecon/Networks.html, 1997.