Quality of Service by Flow Aware Networking

By J.W. Roberts and S. Oueslati-Boulahia

France Telecom-CNET

38-40 rue du Général Leclerc, 92794 Issy-les-Moulineaux Cedex 9, France

The paper addresses the issue of providing quality of service guarantees in the Internet. After a brief discussion of Internet traffic characteristics, we consider the possibility of performing multiplexing with predictable performance for stream and elastic traffic using open loop and closed loop control, respectively. Quality of service depends essentially on providing sufficient capacity to handle expected demand. We argue that flow awareness is additionally necessary to ensure that traffic is directed over routes with available capacity and to avoid congestion collapse in case of overload. Proposed flow aware controls allow simple volume based charging and the development of an economic model similar to that of the telephone network.

Keywords: Internet quality of service; traffic characteristics; service differentiation; admission control; adaptive routing.

1. Introduction

The issue of providing quality of service guarantees in the Internet has far-reaching implications which go well beyond the definition of protocols and traffic control mechanisms. Quality of service is closely related to pricing and, through that, to the underlying economic model on which the network as a business is based. In advocating flow aware networking, we are inspired to a large extent by the highly successful model on which the public telephone network is based.

Quality of service in the telephone network is ensured by overprovisioning. Demand at historically fixed price levels is estimated based on past records and capacity is provided to handle that demand with a very low probability that a new call must be blocked. Adaptive routing algorithms and call admission controls are employed to maximise network efficiency and provide protection against the effects of overloads and failures. Optimally, in a competitive environment, flat rate and per call charges are set such that overall revenue covers the cost of network provision and operation. The network provider has the necessary economic incentive to expand capacity as demand grows to ensure that quality of service is preserved.

The Internet community tends to eschew the telephone model for a variety of reasons. It is considered, notably, that Internet traffic, which is generated by a wide variety of applications each with its own characteristics, is much less predictable than telephone traffic. Moreover, different applications typically have quite different quality of service requirements necessitating the definition of distinct service classes. After experimention with the *Intserv* model (White & Crowcroft 1997), based on explicit resource reservation for individually signalled transactions, the current consensus is that quality of service in the core of the network must be assured more simply by applying controls at the level of broadly defined aggregates of packets. The *Diffserv* model (Blake *et al.* 1998) allows users to identify their packets as belonging to one of a certain number of classes, each class being handled differently in the network nodes.

There remains considerable uncertainty about how a network provider can apply the Intserv and Diffserv service models to meet user quality of service requirements and how it can price the different service classes to generate sufficient revenue to cover costs. Consideration of the statistical relationship between demand, capacity and performance leads us, indeed, to doubt that there is a satisfactory answer to the above questions. We argue in the present paper that it is necessary rather to define a new service model allowing traffic controls to be applied at flow level, where, by flow, we mean the succession of packets relating to a given instance of some application such as a voice signal or the transfer of a Web page.

2. The nature of Internet traffic

To be able to make quality of service guarantees depends on a sound understanding of the statistical nature of network traffic. In this section we discuss traffic characterisation at packet, flow and aggregate levels and suggest that Internet traffic can be modelled as a stationary stochastic process.

(a) *Self-similarity at packet level*

It is now well known that the arrival process of IP packets is extremely irregular with intensity variations occurring at multiple time scales. Data traffic is asymptotically self-similar and even exhibits multi-fractal behaviour at very small time scales (Feldmann *et al.* 1999). For such traffic it proves extremely difficult to define a parsimonious characterisation capable of capturing its impact on network performance. It is clear, in particular, that parameters currently used to define the traffic offered to a wide area network by individual customers, namely the parameters of a "token bucket", are woefully inadequate†.

The main cause of self-similarity in IP traffic is extreme variability in the size of the documents transferred. In particular, the size of Web documents is known to have a distribution with an infinite variance (Bestavros & Crovella 1996). It proves more natural to describe Internet traffic in terms of "flows" rather than packets.

(b) *Flow level characterisation*

Flows may be broadly divided into two categories (Roberts 1999): stream flows, generally corresponding to audio and video applications, having an intrinsic rate which must be preserved by the network; and elastic flows, corresponding to the transfer of digital documents, whose rate adapts to available capacity. A stream flow is characterised by its duration and how its rate varies. An elastic flow can be characterised more simply through the size of the document to be transferred. For

† Traffic conforms to a token bucket of parameters r bits/sec and b bits if the volume of bits $A(s,t)$ emitted in an interval (s,t), satisfies $A(s,t) \leq r(t-s)+b$.

the sake of simplicity we do not specifically consider the case of adaptive real time applications.

To complete the traffic description it is necessary to specify how flows arrive. The well-established practice of modelling telephone traffic in the busiest period as a stationary Poisson arrival process of calls of independent duration can be carried over to some forms of stream traffic. In describing the arrival process of elastic flows, it may be necessary to pay more attention to the way in which individual users behave in a Web session, for instance. However, there is some empirical evidence to support the hypothesis that elastic flow arrivals in a backbone link may also be assimilated to a Poisson process (Nabe *et al.* 1998).

(*c*) *Traffic demand*

Traces depicting traffic intensity on backbone links typically reveal quite predictable behaviour (Thomson *et al.* 1997, Roberts & Oueslati-Boulahia 1999). Intensity in working days consistently attains the same level and remains roughly constant over an afternoon busy period lasting several hours suggesting the possibility of modelling traffic as a stationary stochastic process. The average rate attained may then be interpreted as an expression of demand given by the product of the flow arrival rate and the average volume in bits of each flow (size of elastic flows, duration $\times$ average rate of stream flows).

Network performance depends essentially on whether demand is greater than or less than available capacity. Quality of service clearly depends both on the capacity being in place and on the traffic being able to access that capacity by appropriate routing. Additional controls are necessary to preserve performance in case of overload. Though we argue in the following that routing and overload controls should be performed at flow level, currently proposed evolutions to the Internet service model aim rather to perform traffic management on the basis of traffic aggregates.

(*d*) *Characterising traffic aggregates*

Through aggregation, quality of service requirements are satisfied in a two step process: the network guarantees that an aggregate has access to a given bandwidth; this bandwidth is then shared by the flows constituting the aggregate using mechanisms capable of meeting their individual QoS requirements. The situation would be clear if the guarantee provided by the network were for a fixed constant bandwidth. In practice, because traffic in an aggregation is generally extremely variable, a constant rate is not usually a good match to user requirements. In frame relay and ATM networks, current practice is to considerably overbook capacity (the sum of guaranteed rates may be several times greater than available capacity), counting on the fact that users do not all require their guaranteed bandwidth at the same time. In addition, the aggregate traffic is generally allowed to exceed the nominally guaranteed bandwidth. Excess packets are marked and considered to be expendable in case of congestion.

Undeniably, the combination of overbooking and admitting excess traffic leads to a commercial offer that is attractive to many customers, especially in comparison to the cost of a leased line of equivalent "guaranteed" capacity. It does, however, lead to an imprecision in the nature of the offered service and in the basis of charging which we believe will prove unacceptable as the multiservice networking market gains maturity.

3. Predicting performance

The feasibility of quality of service guarantees depends on being able to predict the performance of implemented traffic controls. We distinguish open loop control suitable for stream traffic and closed loop control suitable for elastic flows.

(a) *Open loop control*

Strict delay bounds can be assured for a flow (or flow aggregate) which is controlled at the network input by a token bucket filter (Cruz 1991). The usefulness of such bounds is somewhat limited in practice for many reasons: the token bucket is not a useful traffic descriptor; delay bounds are only attained in unrealistic "worst case" scenarios; most applications do not require absolute delay guarantees. In addition, to realize the bounds requires complex scheduling. We believe it is preferable to perform controlled statistical multiplexing allowing delays to be "guaranteed" with a high probability.

Consider an isolated link and assume for the sake of simplicity that packet flows can be assimilated to fluids with clearly defined instantaneous rates. We can then distinguish statistical multiplexing schemes according to whether or not they rely on a buffer to absorb momentary input rate overloads. When buffering is used, it proves very difficult to predict performance without knowing very precise details concerning the way the input rate varies over time (Roberts *et al.* 1996). Performance is much more easily predictable with bufferless multiplexing.

In the absence of buffering, data loss must be limited by ensuring a sufficiently low probability that the input rate exceeds the link rate. This probability depends only on the stationary distribution of the individual flow rates and is consequently insensitive to correlation in the rate process. Bufferless multiplexing is compatible with reasonably high utilisation (60 %, say) if the peak rate of multiplexed flows is small (no more than 1% of link rate, say).

(b) *Closed loop control*

Elastic traffic is, by definition, suited to the use of closed loop control whereby the rate of flows is adjusted to make maximal use of available bandwidth. In the interests of developing insight into the performance of closed loop control, we discuss below a simple performance model assuming that bandwidth is shared perfectly fairly.

Consider a single bottleneck link of capacity C dedicated to handling elastic flows. We assume flows arrive according to a Poisson process of rate λ and that when n flows are in progress each is served at rate C/n. Flow sizes are assumed to be independently drawn from a general distribution of mean θ. With these assumptions the considered system can be recognised as an M/G/1 processor sharing queue (Kleinrock 1976). Let $\rho = \lambda\theta/C$ be the link utilisation and assume $\rho < 1$. It is well known that the number of flows in progress then has a geometric distribution, $Pr[n\,\text{flows}] = \rho^n(1-\rho)$, and that the expected response time of a flow of size p is $E[\text{response time}] = p/[C(1-\rho)]$.

These results demonstrate that the performance of a link shared using closed loop control is satisfactory even though the packet arrival process is self- similar (due to an infinite variance flow size distribution). Clearly, however, if the offered load ρ is greater than one, the considered model is unstable: the number of flows in progress increases indefinitely as more and more new flows arrive while the throughput achieved by any one flow tends to zero.

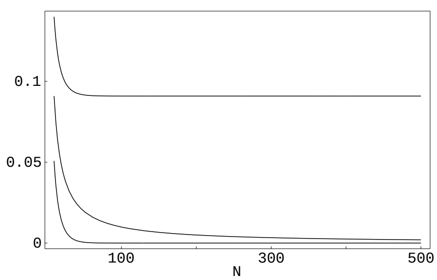

Figure 1. Blocking Prob., $\rho = 0.9, 1.0, 1.1$

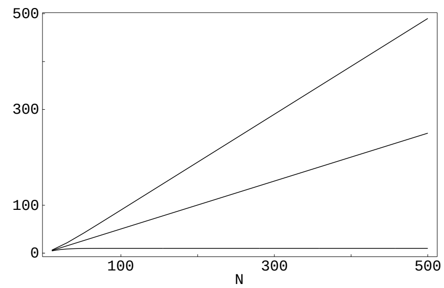

Figure 2. ENRT, $\rho = 0.9, 1.0, 1.1$

Admission control is a means to preserve useful throughput in the event of overload (Massoulié & Roberts 1999). The objective here is not so much to preserve a minimum acceptable throughput for admitted flows (a few tens of kilobits/sec, say) as to preserve network efficiency in overload. Suppose new flows are blocked when the number sharing the isolated link considered above attains N. The probability of blocking is then $B = (1-\rho)\rho^N/(1-\rho^{N+1})$.

Figure 1 shows how B depends on N. In underload, the blocking probability is negligible as soon as N is greater than 50. In overload, on the other hand, B tends rapidly to the fluid limit $(\rho - 1)/\rho$ and is independent of N. Define the *expected normalised response time* (ENRT) as the response time of a document of size p divided by p and multiplied by the link rate C. This measures the response time in multiples of the time it would take to transfer the document if it had exclusive use of the link. Figure 2 plots this quantity as a function of N. There is again a clear distinction between performance in underload (response time remains very small) and overload (response time increases with N). To limit response time in overload while avoiding unnecessary blocking in underload, a reasonable choice for N would be 100, although any value between 50 and 200 would also be acceptable.

4. Service differentiation

It is clear that stream and elastic flows have distinct quality of service requirements. It is also true that the same responsiveness is not required for all elastic flows (Web consultation, file transfers, e-mail...). In this section we discuss possible means of realising service differentiation.

(*a*) *Discriminating between stream and elastic traffic*

There are considerable advantages in allowing stream and elastic flows to share the same links. Giving priority to the service of stream flow packets ensures maximal responsiveness for the underlying audio and video applications without penalising the throughput of elastic flows. The stream flows would be admitted in the conditions of bufferless multiplexing at a rather low load (assuming the majority of traffic is elastic), facilitating measurement-based admission control. The loss rate of stream flows would be very low and there is no obvious advantage to be gained by differentiating service with respect to this criterion. Elastic flows would naturally exploit all the bandwidth left by the stream traffic, gaining in average throughput compared to a system of equivalent capacity divided into dedicated parts.

(b) *Impact of load on elastic flow throughput*

One possibility for creating differentiated elastic flow classes is to dedicate bandwidth to each and to operate that bandwidth at different loads. Indeed, in the processor sharing model, expected throughput $C(1-\rho)$ depends linearly on ρ and it would appear easy to create different quality of service classes. In practice, however, flow rates are generally limited elsewhere in the network, notably by the speed of access lines. Figure 3 shows how the expected normalised response time depends on load when the maximum flow rate r is equal to one tenth of the link capacity. It is clear that the load has little impact on response times until it gets close to $1-r/C$. There is thus very little room between very good quality when $\rho < 1-r/C$, and bad quality when $\rho > 1$. By adequate provisioning, it is relatively easy to offer excellent service but virtually impossible to target any intermediate quality level.

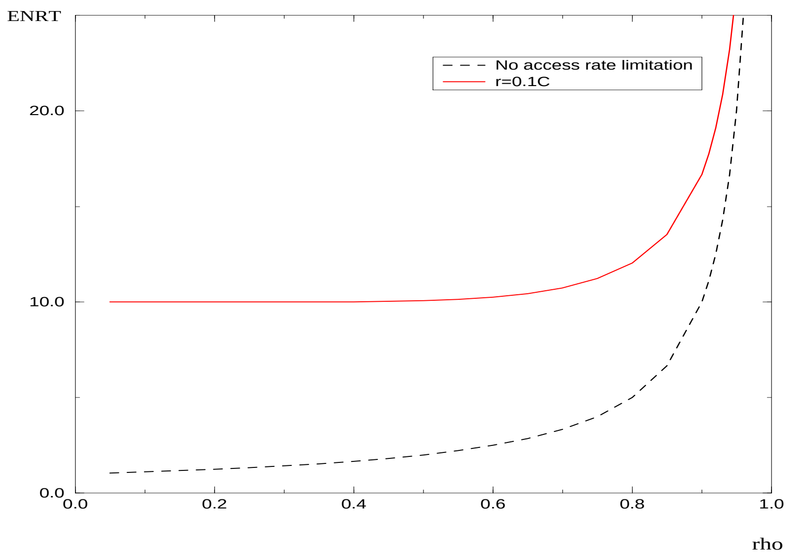

Figure 3. ENRT as a function of ρ

(c) *Discriminatory bandwidth sharing*

Another possibility for service differentiation is to deliberately share bandwidth unequally. We have used simulation to investigate weighted sharing assuming the document size has a Pareto distribution. Figure 4 shows how the expected normalised response time depends on document size for two classes of flow sharing

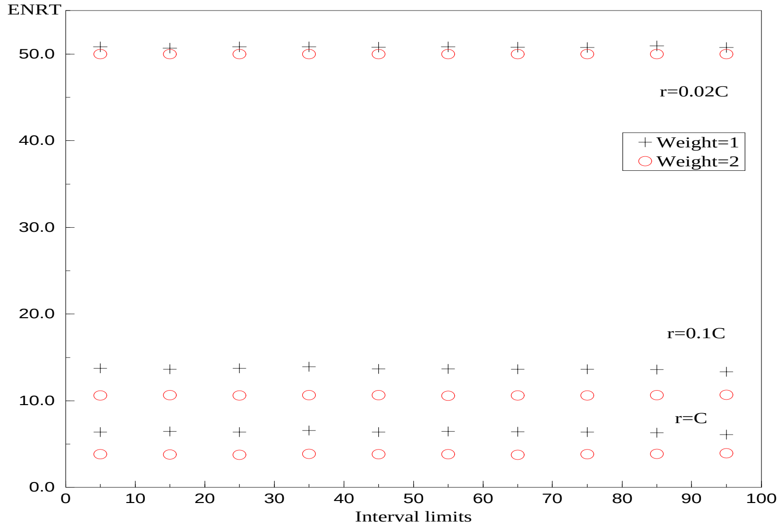

Figure 4. ENRT with unequal bandwidth sharing, $\rho = 0.8$, $r = .02C, .1C, C$ Mbit/s

an isolated bottleneck link of capacity 100 Mbit/s. Class 1 flows receive twice as much bandwidth as flows of class 2 when they are not limited by the access rate. We show results for three access rates, expressed as a fraction of the link rate, $r = .02C$, $r = .1C$ and $r = C$. Link utilisation is 0.8.

To calculate the points in this figure, we first class the simulated documents in increasing size order. We then define 10 contiguous size intervals such that each contains the same number of documents. The symboles give the ENRT computed as an average over all documents in the corresponding interval.

Clearly, the access rate r has a significant impact. For r equal to $.02C$, discrimination is ineffective at the considered load. Discrimination is apparent when the link is the only bottleneck ($r = C$). Note, however, that throughput is then excellent, even for the underprivileged class. In overload, class 1 flows do always attain a throughput twice that of class 2 flows but throughput then tends to zero for both classes, in the absence of admission control.

(*d*) *Differentiated blocking probabilities*

The above results show that there is limited scope for service differentiation in underload since all flows then receive good quality of service. In overload, admission control is necessary to preserve performance. It can also be used to perform effective differentiation.

Consider a link of capacity C simultaneously used by elastic flows belonging to two classes producing loads ρ_1 and ρ_2, respectively. The classes differ through their access priority. Flows of class i are blocked when the number of flows in progress of either class is greater than or equal to N_i. We assume $N_1 > N_2$ so that flows of class 1 receive priority service.

Numerical results confirm that effective differentiation is obtained for a wide range of N_1 and N_2 (Roberts & Oueslati-Boulahia 1999). The particular values $N_1 = 100$ and $N_2 = 50$ constitute a reasonable choice. For $\rho_1 + \rho_2 < 1$, both classes see negligible blocking. When $\rho_1 < 1$ and $\rho_1 + \rho_2 > 1$, class 1 flows are not blocked while the fluid limit $(\rho_1+\rho_2-1)/\rho_2$ applies to class 2. If $\rho_1 > 1$, class 1 sees blocking $(\rho_1 - 1)/\rho_1$ while virtually all class 2 flows are blocked.

5. Flow aware networking

The above models demonstrate that quality of service is generally excellent if capacity is overprovisioned, whether flows are identified as such or not. Flow awareness is necessary to ensure that flows are routed over paths which are indeed overprovisioned and to allow admission control when necessary to prevent congestion collapse.

(a) *Flow identification*

Admission control can be performed simply if it is possible to identify the start of a new flow. Rejecting the first packets of a new flow is generally sufficient signal to a source that the network is congested. In the particular case where the majority of flows correspond to a TCP connection, new flows can be identified on recognising the SYN or SYN/ACK packets of the three-way set up handshake. This technique has already been successfully employed to considerably improve the effective throughput of a congested Internet access link (Kumar *et al.* 1999).

This approach may be sufficient in a network with fixed routing. However, to perform flow aware routing it is additionally necessary to ensure that all packets of a given flow follow the same path. This requires the creation of per-flow state explicitly identifying the flows in progress and indicating their route. Minimal flow state would include an identifier derived from the packet header† and the epoch of the last observed packet.

In an imagined implementation using Multiprotocol Label Switching (Callon *et al.* 1999), flow state would be associated with the edge router incoming interface on which the flow arrives. It would be stored in lists corresponding to the forwarding equivalence classes (FEC) of that router. Every packet has a unique FEC (defined principally by its destination address) associated, in our implementation, with a set of label switched paths (LSP). On a packet arrival, in addition to the regular address look-up required to identify the FEC, it is necessary to verify whether or not the packet belongs to an existing flow by comparing its identifier with those of the list. If so, the packet is routed over the LSP indicated in the list and the last packet epoch is updated. If not, it is necessary to perform flow routing by choosing an appropriate LSP. If all available routes are congested the packet is discarded. The congestion status of an LSP would be determined by its available bandwidth as described next.

(b) *Measurement-based admission control*

To account for rate limitations affecting elastic flows outside the considered domain and to account for the variability of stream flow rates, it is necessary to estimate available bandwidth by measurement. The idea is that the edge router continually estimates the bandwidth that would be available to a new elastic flow routed between the end points of every path of which it is the origin. The precise

† We ignore possible complications due to IPSEC encryption.

mechanism by which this estimation could be made is the subject of on-going research. One possibility is to create a "phantom" TCP connection between the path end points, as proposed by Afek *et al.* (1996), and simply measure its realized short term throughput.

A flow would be accepted or rejected on a given route according to whether the available bandwidth were greater than or less than a certain threshold. Applying the same threshold to stream and elastic flows ensures that the maximum stream peak rate is equal to the minimum elastic throughput. Based on the discussion of Section 3, this threshold could be set to around 1% of link capacity. Since the objective is to avoid the negative effects of demand overload and not to ensure a contractual minimum rate for elastic flows, the estimation of available bandwidth does not need to be particularly accurate.

(*c*) *Flow aware routing*

With present Internet routing protocols, in the absence of topology changes, the path available to any given flow is fixed. The flow is routed over that path even if it is currently congested and a more lightly loaded alternative may be available.

The careful planning necessary to ensure that traffic offered to all links is within their capacity in a network with fixed routing is made particularly difficult by the uncertain characterisation of aggregate traffic alluded to in Section 2. Telephone networks generally employ adaptive routing where the path of each call is chosen on its arrival depending on the current congestion status of the paths available. Adaptive routing leads to more efficient use of installed capacity and considerably improves the resilience of the network with respect to planning uncertainty and equipment failures. Adaptive routing in the Internet could be applied effectively if the network were flow aware. To route flows, rather than packets or aggregates of flows, allows the application of techniques already perfected in the telephone network and appears as the more stable and controllable alternative.

We have evaluated a number of possible per flow routing algorithms by means of simulation (Oueslati-Boulahia & Oubagha 1999; Oueslati-Boulahia & Roberts 1999). Considered algorithms make routing decisions based on the value of two path metrics: the number of hops and the available bandwidth. A path is feasible if its available bandwidth is greater than the admission threshold fixed for the network (1% of the minimal link rate, say). The well-known "widest-shortest" algorithm consists in choosing the feasible path with the largest available bandwidth among those with the smallest number of hops. We have shown that the performance of this algorithm can be improved by employing a form of "trunk reservation" whereby the available bandwidth admission threshold increases with the number of hops (Oueslati-Boulahia & Roberts 1999). This device prevents the choice of long paths in heavy traffic, leading to lower blocking probabilities and higher throughput since link bandwidth is then used more efficiently.

(*d*) *Flow unaware pricing*

Flow aware networking would make it possible to perform flow *unaware* pricing based simply on counting the number of bytes transmitted across a particular interface. We pretend that volume pricing is appropriate because, by the use of admission control, all packets except the discarded first packet of rejected flows are effective and all flows receive adequate quality of service.

The purpose of such pricing is to recover the cost of investment in network infrastructure, with users paying in relation to their utilisation of this infrastructure. The price level should ideally be such that revenue covers costs, the network provider then having the incentive to expand capacity as necessary to stay ahead of congestion. It is neither necessary nor useful to price stream and elastic flows differently. Users requiring negligible delay will naturally declare their flows as stream while those seeking high throughput will choose the elastic class.

Clearly, different pricing packages are possible within this scheme, as for the current telephone service, including flat rate charges for users not exceeding some utilisation threshold. Note that this model is indeed closer to that of the telephone network than many proposed Internet pricing schemes where users pay in relation to perceived or expected quality of service levels.

6. Conclusions

Internet traffic can be characterised most easily at flow level making the significant distinction between stream and elastic flows. Traffic in the busiest period can reasonably be modelled as a stationary process whose intensity is the product of the flow arrival rate and the expected volume of data in any flow.

The ability to predict performance when a certain volume of traffic is offered to a link of given capacity depends on the type of multiplexing employed. We advocate the use of open loop "bufferless" multiplexing for stream flows and closed loop control for elastic flows. Assuming the latter shares bandwidth equally, we have shown using simple models that perceived performance depends essentially on whether demand is less than or greater than capacity. In the latter case, it is important to apply admission control to prevent a form of congestion collapse manifested by an increasing number of flows in progress, each taking a longer and longer time to complete.

Integration of stream and elastic flows on the same links, with queuing priority given to packets from stream flows, simplifies traffic control. When a significant proportion of traffic is elastic, admission control applied to both stream and elastic flows can be performed by comparing the bandwidth currently available for a new elastic flow to a certain threshold. By applying different admission thresholds, it is staightforward to offer a form of service differentiation allowing privileged access to certain categories of flow in case of congestion.

To realise flow aware networking in a high speed network is clearly not easy. We envisage the identification of flows "on the fly" at the edge of a label switched domain. The availability of bandwidth on the paths available from a given edge router would be monitored continually allowing a form of measurement-based routing and admission control. The choice of path for a new flow should take account both of the path length in hops and its current available bandwidth.

Since all flows which are admitted have guaranteed quality of service, it is natural to apply a volume-based pricing scheme with the same per byte charge for both elastic and stream flows. The price levels would be fixed, as in the telephone model, to cover the cost of the network infrastructure, the latter being provided in sufficient quantity to avoid congestion in all but exceptional circumstances.

References

Afek, Y., Mansour, Y. & Ostfeld, Z. 1996 Phantom: A simple and effective flow control scheme. *Computer Communication Rev. (SIGCOMM)* **26**, no. 4.

Bestavros, A. & Crovella, M. 1996 Self-similarity in World Wide Web traffic: Evidence and possible causes. In *Proc. ACM Sigmetrics 1996*.

Blake et al. 1998 An Architecture for Differentiated Services. RFC 2475.

Callon, R. et al. 1999 A framework for Multiprotocol Label Switching. Internet draft. Work in progress.

Cruz, R.L. 1991 A calculus of network delay. Part I: network elements in isolation. *IEEE Trans. on Information Theory* **37**, 114-131.

Feldmann, A., Huang, P., Gilbert, A.C. & Willinger, W. 1999 Dynamics of IP traffic: A study of the role of variability and the impact of control. *Computer Communication Rev. (SIGCOMM)* **29**, no. 4.

Kleinrock, L. 1976 *Queuing Systems* **2**. Wiley-Interscience.

Kumar, A. et al. 1999 NETMASTER: Experiences in using non-intrusive TCP connection admission control for bandwidth management of an Internet access link. Submitted.

Roberts, J.W., Mocci, U. & Virtamo J. (eds) 1996 *Broadband Network Teletraffic (Final Report of Action COST 242)*, Springer Lecture Notes in Computer Science, vol. 1155.

Roberts, J.W. 1999 Engineering for Quality of Service. In *Self-similarity in network traffic* (eds K. Park & W. Willinger). Wiley. To appear.

Roberts, J.W. & Oueslati-Boulahia, S. 1999 Flow aware networking. Submitted.

Massoulié, L. & Roberts, J.W. 1999 Arguments in favour of admission control for TCP flows. In *Proc Int. Teletraffic Congress - ITC16, Edinburgh, UK, June 1999*, pp. 33-44. Elsevier.

Nabe, M., Murata, M. & Miyahara, H. 1998 Analysis and modeling of World Wide Web traffic for capacity dimensioning of Internet access lines. *Performance Evaluation* **34**, 249-271.

Oueslati-Boulahia, S. & Oubagha, E. 1999 An approach to routing elastic flows. In *Proc Int. Teletraffic Congress - ITC16, Edinburgh, UK, June 1999*, pp. 1311-1320. Elsevier.

Oueslati-Boulahia, S. & Roberts, J.W. 1999 Impact of trunk reservation on elastic flow routing. Submitted.

Thomson, K., Miller, G. J. & Wilder, R. 1997 Wide-area Internet traffic patterns and characteristics. *IEEE Network*, November/December 1997, 10-23.

White, P. & Crowcroft, J. 1997 Integrated services in the Internet : State of the art. In *proc. IEEE* **85**, no. 12.