15

Realizing quality of service guarantees in multiservice networks

J. W. Roberts
France Télécom - CNET
38-40, rue du Général Leclerc
92794 Issy Moulineaux Cedex 9
France
(james.roberts@cnet.francetelecom.fr)

Abstract

This paper is motivated by the concern of the multiservice network provider who wishes to offer users quality of service guarantees concerning transparency, accessibility and throughput. We consider how these guarantees can be respected jointly by implementing a simple service model distinguishing stream and elastic traffic, and by providing sufficient network capacity to ensure negligible blocking probabilities for an estimated traffic demand.

Keywords

multiservice network, quality of service, service model, network sizing, charging

1. INTRODUCTION

One of the more significant differences between telephony and computer networking resides in the way the respective networks are planned. Telephone network sizing, a component of teletraffic engineering, has developed over the last century into a fine art, drawing on probability theory, mathematical programming and economics. The planning of most computer networks in comparison is extremely primitive being based at best on empirical rules. There are excellent reasons for this contrast in approaches including the current extreme volatility and unpredictability of computer network traffic. However, as telecommunications moves towards the construction of a universal network handling multiple types of voice, data and video traffic on the same infrastructure, it is becoming urgent for

Performance and Management of Complex Communication Networks
T. Hasegawa, H. Takagi & Y. Takahashi (Eds.)

the traditional operators to extend existing network design practices to allow adequate provisioning for data traffic. Seen from another point of view, as computer networks are increasingly used for real time applications like telephony, they will be obliged to apply some of the engineering methods of the traditional telecommunications operators. The common objective is to provision the network to meet traffic related quality of service objectives given the volume and nature of traffic demand.

In a multiservice network, quality of service depends essentially on two factors: the service model which identifies different service classes and specifies how network resources are shared between their various flows, and sizing rules enabling the necessary amount of capacity to be determined from a demand forecast. Much work on the realization of quality of service guarantees considers each aspect in isolation, either identifying service models which allow an unspecified proportion of users to obtain their required quality of service or extending telephone network dimensioning methods by the convenient assumption that flows can be simply characterized by an "effective bandwidth".

In this paper we aim to cover both aspects of the problem of realizing quality of service guarantees, seeking notably to identify the components of a simple service model capable of meeting user requirements with low complexity network protocols and mechanisms. The objective is not to suggest that this service model be implemented as such but rather to expose the options available and their possible impact on the design and operating principles of a multiservice network.

In the next section we discuss the nature of traffic demand in a multiservice network, identifying two broad traffic categories: "stream", essentially for voice and video communications, and "elastic" for the transfer of digital objects such as data files, texts and pictures. We discuss, in section 3, how the service model can meet the particular quality of service requirements of stream and elastic flows expressed in terms of transparency, throughput and accessibility. Section 4 addresses the problem of sizing the network to meet accessibility requirements for both types of traffic.

2. NATURE OF TRAFFIC AND QOS REQUIREMENTS

It is possible to identify innumerable categories of telecommunications services each having its particular traffic characteristics and performance requirements. Often, however, these services are adaptable and there is no need for a network to offer multiple service classes each tailored to a specific application. In the search for simplicity we prefer to limit present consideration to just two distinct categories which we term "stream" and "elastic".

2.1. Stream 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 where significant delay would constitute an unacceptable degradation. A network service providing time integrity for video signals would also be useful for the transfer of pre-recorded video sequences and, although negligible network delay is not generally a requirement here, we consider this kind of application to be also a generator of stream traffic.

The way the rate of stream flows varies is important for the design of traffic controls. Speech signals are of on/off type with talkspurts interspersed by silences. Video signals generally exhibit more complex rate variations reflecting changing degrees of activity in the filmed sequence. The nature of these variations for stored MPEG open loop coded sequences is discussed, for example, in (Roberts et al, 1996). It is noted, in particular, that moments of the per-frame bit rate and its autocorrelation depend significantly on the nature of the sequence. Importantly for traffic engineering, the bit rate of long video sequences exhibits long range dependence (Garrett and Willinger, 1994) . Roughly speaking, this means that the average rate realized in successive constant length intervals varies significantly even when the length of the interval is large. A plausible explanation for this phenomenon is that the duration of scenes in the sequence has a heavy tailed probability distribution (Frater, 1997) , i.e., the probability a scene lasts longer than x seconds decreases with x less quickly than an exponential distribution. Videoconferencing sequences do not have scene to scene variations. The rate does however depend on factors such as the number of people filmed and their dress making *a priori* prediction difficult. Closed loop coding can be employed to limit the scale of rate variations. In the extreme, a constant rate stream can be generated by a coder at the expense, however, of quality degradation and an additional smoothing delay. Looser closed loop rate control can be used to preserve short time scale rate variations while ensuring that the mean rate conforms to a pre-assigned value (Hamdi et al, 1997) .

In addition to the rate variations of each flow, stream traffic as a random process is also determined by the arrival process and duration of communications. The arrival intensity generally varies according to the time of day. It may be natural to extend current practice for the telephone network by identifying a busy period (e.g., the one hour period with the greatest traffic demand) and modelling arrivals as a stationary stochastic process (e.g., a Poisson process). Traffic demand may then be expressed as the expected combined rate of all active flows: the product of the arrival rate, the mean duration and the mean rate of one flow. Additional variables describing the nature of the rate variations and, notably, the composition of different peak and average rates also have a significant impact on network performance.

2.2. Elastic traffic

The second type of traffic we consider consists of digital objects which must be transferred from one place to another. These objects might be files of alphanumeric data, texts, pictures or video sequences transferred for local storage before viewing. The essential characteristics of elastic traffic are the arrival process of transfer requests and the distribution of object sizes. Observations on Web traffic provide very useful pointers to the nature of these characteristics (e.g., Arlett and Willianson, 1996, and Crovella and Bestavros, 1996).

The average arrival intensity of requests for object transfer varies depending on underlying user activity patterns. As for stream traffic, it should be possible to identify representative busy periods where the arrival process can be considered to be stationary. Measurements on Web sites reported by Arlitt and Williamson (1996) suggest the possibility of modelling the arrivals as a Poisson process. A Poisson process indeed results naturally when members of a very large population of users independently make relatively widely spaced demands.

Statistics on the size of Web documents reveal that they are extremely variable exhibiting a probability distribution with a heavy tail which can be approximated by a Pareto distribution with finite mean but infinite variance. Most objects are very small: measurements on Web document sizes reported by Arlitt and Williamson reveal that some 70% are less than 1 Kbyte and only around 5% exceed 10 Kbytes. The presence of a few extremely long documents has a significant impact on the value of the distribution mean.

It is possible to define a notion of traffic demand for elastic flows in analogy with the definition given above for stream traffic as the product of an average arrival rate in a representative busy period and the average object size. Elastic flows may additionally be characterized by a maximum rate (determined, for example, by the speed of an access line) and a minimum required rate.

2.3. Other types

Some traffic entities are not clearly either of stream or elastic type. This is the case of stored video and audio sequences accessed remotely across the network. The sequences 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. Intermediate network solutions are possible where elasticity is exploited to allow variable network delays with output read at the appropriate (stream) rate from a set top box.

Another category of traffic arises when individual flows and transactions are grouped together in an aggregate traffic stream. This occurs currently, for example, when the flow between remotely located LANs must be treated as a traffic entity by

a wide area network. This is indeed the main expression of demand in existing broadband networks. The characteristics of aggregate traffic flows are generally very difficult to describe succinctly. Observations of LAN traffic (Leland et al, 1994) have revealed its self-similar or long range dependence behaviour. Despite considerable modelling work on self-similar processes and the plausible explanation that the observed characteristics are due to the heavy tailed distribution of the size of transferred items (files, etc.), it remains virtually impossible to adequately quantify this kind of traffic. The advantages of considering an aggregation as a single traffic entity (billing, absence of flow identification, ...) should be weighed against the considerable difficulty of performing traffic management and realizing required quality of service guarantees. The alternative is to recognize individual stream and elastic flows for the purpose of traffic control, as assumed in the present paper.

2.4. Quality of service requirements

Quality of service covers many aspects including transmission quality and reliability. We consider only those aspects of quality of service which are determined by the statistical nature of traffic. We distinguish three such aspects: transparency, accessibility and throughput.

- *Transparency* refers to the time and semantic integrity of transferred data. For stream traffic delay should be negligible while a certain degree of data loss is tolerable. For elastic traffic, semantic integrity is generally required but (per packet) delay is not important. Semantic integrity requires the retransmission of data lost due to congestion (or other reasons) under the control of user and/or network protocols.
- *Accessibility* refers to the probability of admission refusal and the delay for set up in case of blocking. Blocking probability is the key parameter used in dimensioning the telephone network. It is generally considered sufficient to provide network capacity assuring a blocking probability of around 1% in the busy hour.
- *Throughput* is a quality of service measure for elastic traffic defined as the document size divided by the time necessary to transfer the document. User requirements for throughput are quite poorly understood (since networks currently have no target values). A throughput of 1 Mbit/s would ensure the transfer of most Web pages quasi-instantaneously (less than a second).

To meet transparency requirements for both stream and elastic traffic the network must implement an appropriately designed service model. The accessibility requirements must then be satisfied by network sizing taking into account the random nature of user demand. Throughput for elastic flows is determined both by how much capacity is provided and how the service model shares this capacity between different flows.

3. TOWARDS A SIMPLE SERVICE MODEL

As the Internet seeks to extend the current best effort service model to allow quality of service guarantees and network operators adopting ATM are faced with a complex choice among the panoply of standardized service classes, it is interesting to speculate on what could be the components of the simplest service model. In the paragraphs below we distinguish the need for open loop and closed loop control for stream and elastic traffic flows, respectively, and briefly discuss the impact of the adopted charging scheme.

3.1. Open loop control for stream traffic

We pretend that stream flows can be handled most efficiently by means of open loop or preventive traffic controls. Reactive controls, relying on users adjusting their rate in response to network traffic conditions, may lead to more efficient sharing of a limited resource. However, if we assume the network operator has the ambition to provide sufficient capacity to ensure good quality of service, the investment required does not depend significantly on the type of control. Open loop control can be employed to perform statistical multiplexing with predictable performance as described below.

In a fluid analogy where the instantaneous rate of a traffic stream can be defined unambiguously, statistical multiplexing schemes can be distinguished according to whether or not they rely on buffering. With bufferless multiplexing, data loss is avoided by maintaining the overall arrival rate less than the service rate. We maintain that this is the preferred multiplexing scheme for stream traffic ensuring that traffic suffers minimal delay and that the characteristics of the stream remain unaltered throughout the network. Unlike buffered multiplexing, the data loss rate depends only on the stationary rate distribution, and not on more complex traffic characteristics such as long range dependence.

In practice, to account for the non-fluid nature of real traffic, a small multiplexer buffer is required and it is necessary to precisely define the notion of rate taking account of jitter. For an ATM network, a practical realization of "bufferless" multiplexing is known as rate envelope multiplexing (Roberts et al, 1996; ITUa, 1997). Buffer dimensioning and jitter control for rate envelope multiplexing are greatly facilitated when the peak rate is defined with "negligible CDV" obtained by spacing traffic to this rate at the network ingress (Brichet et al,1997).

We do not believe it to be either necessary or useful to characterize stream flows beyond their peak rate, by means of a traffic filter like the leaky bucket, for example. The parameters of the leaky bucket either define a bounding traffic envelope which is too loose to be useful for resource allocation or simply imply that the flow must be shaped to a quasi-constant stream in order to be compliant. The additional information provided by the leaky bucket parameters is unnecessary

since bufferless multiplexing can be performed conservatively with a guaranteed loss rate simply by using known peak rates and the actual activity of connections derived by measurement, as proposed for example by Gibbens et al (1995). Furthermore, in a network where bandwidth is shared between stream and elastic flows, the precision of admission control for the former is clearly less critical than in a dedicated network.

Bufferless multiplexing is efficient when the peak rate of multiplexed connections is a small fraction of the multiplexer service rate (Roberts et al, 1996). In this case, variations of the overall rate about its mean value are of relatively small amplitude and the probability of overload is low even at high loads. If the peak rate is high, on the other hand, momentary overload and data loss can occur with high probability when just a small number of connections emit peak rate bursts simultaneously and this severely limits attainable utilization. This is a manifestation of the scale economies effect in any statistical shared resource: sharing is only really efficient when each user individually has a relatively small requirement.

We would argue that the limit on peak rate is not a serious limitation in any moderately large network where a link would serve several tens or even hundreds of simultaneous stream flows. LAN interconnection or other services involving aggregations of individual traffic flows do generate high peak rates which would need to be handled with controlled delay if the aggregate included stream flows. The difficulty of doing so, given the particular characteristics of traffic aggregation, was one reason for suggesting in section 2 that this type of service should be avoided in favour of a service model recognizing individual flows.

3.2. Closed loop control for elastic traffic

We reject the use of open loop controls for elastic traffic. A peak rate limitation allowing efficient rate envelope multiplexing would be counterproductive since elastic transfers should be able to use as much bandwidth as possible to optimize throughput. Reliance on significant network buffering as an alternative to rate envelope multiplexing is fraught with the difficulties of performance prediction when offered traffic must be defined using inadequate exogenous traffic descriptors such as the parameters of a leaky bucket.

Reactive control by means of a flow control protocol like ABR/ATM or TCP/IP is preferable for elastic traffic, allowing transmission rates to adjust to the maximum allowed by current traffic levels and the capacity of end terminals. Both protocols aim to fully exploit available network bandwidth while achieving fair shares between contending flows. A significant difference between ABR and TCP is the degree of control over the bandwidth sharing which is afforded to the network. In our alternative simple service model we would retain the capability included in ABR for the network to enforce necessary rate changes and not rely solely on

cooperative user behaviour. The precise ABR algorithm would not need to be as complicated as that of the present standard, however, particularly if it were optimized like TCP for the transfer of individual digital objects.

Insight into the throughput performance of an elastic service class can be derived from the following simple model. Consider a single bottleneck link of capacity c offered traffic in the form of digital objects arriving according to a Poisson process of rate λ. The size of each object is drawn independently from a general distribution with mean s. When n objects are being transferred on the link, the closed loop control realizes a perfect fair share immediately such that each transfer takes place at rate c/n. These assumptions define the classical processor sharing queue for which some interesting performance results are known (Kleinrock, 1975).

Firstly, the distribution of the number of active transfers and their expected throughput is insensitive to the distribution of object size, i.e., they depend only on the mean of the distribution. Let the load of a link of capacity c offered the above traffic be $\rho=\lambda s/c$. The number of transfers in progress N_t is geometrically distributed: $\Pr\{N_t=n\}= \rho^n(1-\rho)$, and the average throughput of any flow is equal to $c(1-\rho)$.

These results demonstrate that even if the object size distribution is heavy tailed (as is indeed the case), the system is stable for a load ρ less than one. This is in marked contrast with the corresponding result for a FIFO M/G/1 queue where a heavy tailed object size distribution with infinite variance leads to infinite expected delay for any positive load. Heyman *et al* (1997) have studied a similar model for elastic bandwidth sharing.

The control loop may be designed to offer differential service rates to different users. The corresponding generalization of the above model is discriminatory processor sharing as considered, for example, by Fayolle *et al* (1980). Assuming m distinct user classes, bandwidth is shared in proportion to a parameter g_i associated with each class i such that, when the number of transfers in progress from class i is equal to n_i, a user of class j receives a service rate $\phi_j = g_j / (n_1 g_1 + \ldots + n_m g_m)$. Expected throughput depends on object size and is no longer insensitive to the object size distribution.

Figure 1 shows results obtained for a particular configuration* . A link is shared by two user classes with parameters $g_1=1$ and $g_2=2$. The figure shows the reciprocal of the expected throughput (i.e., the expected response time divided by the object size) for each class as a function of object size for three different object size distributions of unit mean (the same distribution for both classes): Erlang with

* these results were derived by Anas Lalou during his internship at the CNET.

squared coefficient of variation cv^2=0.067; exponential (cv^2=1); hyper-exponential with cv^2=250. Link load is 0.67 with each class contributing half. From the figure we derive the following observations:

- the sharing parameters ensure effective discrimination for the transfer time of short objects;
- discrimination increases with the variance of object size;
- as object size increases, throughput for both classes tends to the limit $c(1-\rho)$;
- for very short objects, throughput is roughly independent of the object size distribution.

Additional results show that discrimination increases with increasing load.

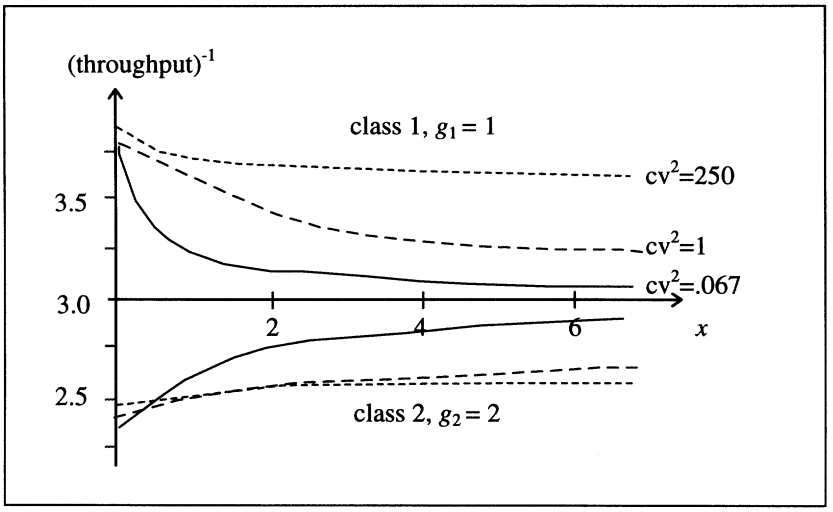

Figure 1. Normalized response time for an object of size x

The limiting large object throughput is explained by the fact that, whatever its sharing parameter g_i, a very long transfer utilizes all the bandwidth except that required by other users, equal on average to $c\rho$. The relative independence of the throughput of very small objects with respect to the object size distribution reflects an approximate insensitivity of the distribution of the number of jobs of both classes present in the system at an arbitrary instant.

The above results lead us to question the usefulness of "fairness" as a criterion in the definition of the closed loop control. The primary objective is to assure the necessary throughput quality of service requirement and this depends as much on the current traffic level as the way bandwidth is shared. Throughput of large objects is not affected by the rate assigned to the transfer of short objects which start and

finish within the transfer time of the former. Overall throughput can therefore be improved by giving priority to short objects with limited detrimental effect on the longer transfers. The sharing weight assigned to an object transfer has a quite unpredictable impact on quality of service. Throughput gain occurs mainly for small objects whose transfer time is in any case very small. For large objects, whose transfer time is significant, a user would gain little by choosing (i.e., buying) a higher relative proportion of link rate.

The processor sharing model illustrates how performance can deteriorate suddenly as offered load ρ increases through 1: if link bandwidth c is high, thoughput performance is good even when ρ is close to 1. For heavier loads, throughput is zero and the number of transfers in progress increases indefinitely (of course, the model then ceases to be accurate, since many real users will abandon transfers as soon as they begin to notice such congestion). More graceful performance degradation would be obtained if the service model incorporated admission control. Limiting the number of active transfers on a given link preserves transparency and throughput quality of service at the cost of restricting accessibility.

In the interests of simplicity and in view of the above remarks, the present service model will guarantee the same minimum throughput to all elastic flows. This allows simple admission control by comparing available bandwidth to the number of admitted sources multiplied by the guaranteed throughput, and also reduces the complexity of the flow control algorithm and necessary bandwidth sharing mechanisms.

3.3. Charging

Many differences between the service models of B-ISDN and the Internet owe their origin to the different charging principles envisaged in the two networks. In (Roberts, 1998) we discuss possible charging schemes and their impact on the realization of quality of service guarantees. One conclusion is that so-called "transaction pricing", where users pay in relation to the resources used for each stream or elastic flow, necessarily implies the systematic use of admission control to ensure transparency and throughput performance guarantees. Conversely, if admission control is not possible, the network must rely on flat rate pricing or, at best, some form of congestion pricing (users pay a premium to receive priority access to scarce network resources) to recover network infrastructure and operation costs. In the present service model we opt for transaction pricing and the use of admission control for both stream and elastic traffic. Transaction pricing allows prices to be closely related to costs which we believe to be necessary to ensure network provider solvency in a competitive environment. Of course, the charging scheme also includes a fixed fee to cover the cost of user dedicated resources.

The price level per transaction may depend on the detailed traffic characteristics of the flow in question, including whether it is stream or elastic. However, we argue

in [Rob98] that, in a large network provisioned for good accessibility quality of service, the amount of resources (average bandwidth × duration) effectively allocated to a flow by admission control is nearly equal to the volume of data transferred. This suggests the possibility of a flat rate per byte charge for all carried traffic.

With this simple charging principle, we have a two-class service model where users would have a quality of service rather than financial incentive to choose the appropriate class for stream and elastic traffic, respectively: stream flows would be guaranteed negligible delay, elastic flows maximal throughput.

Blocking probability depends on the required peak rate of stream flows and the minimum throughput guarantee for elastic flows. It is envisaged that the network would be dimensioned to ensure a low blocking probability only for peak rates and throughputs up to a certain value. To allow the realization of scale economies, this value would need to be set to small fraction of the rate of network links.

3.4. Queuing mechanism

The choice of service model has a direct impact on the complexity of the queueing mechanisms necessary to share memory and bandwidth between flows of the different service classes. A major advantage of simplifying the service model is the reduced complexity necessary for these mechanisms.

From the above discussion, the present service model is equipped with two service classes, employs admission control to ensure a common minimum throughput for elastic flows, and implements an ABR-like flow control to prevent any elastic flow from seizing more than its fair share of available bandwidth. For this service model, it is possible to rely on a simple two-priority queueing mechanism. Giving highest priority to stream flows, coupled with "bufferless multiplexing" admission control, guarantees negligible delay. The second priority queue, long enough to absorb the latency of the flow control algorithm, can be served in FIFO order since delay in this queue and realized throughput are bounded due to the limited number of admitted flows.

4. NETWORK SIZING

In this section, we consider how the network can be sized to ensure that quality of service is adequate for an assumed traffic demand. We assume the network uses the above simple service model and therefore relies on admission control to ensure the transparency and throughput requirements of accepted flows. Sizing then has the objective of meeting the accessibility quality of service requirements. We interpret the latter to imply a busy hour blocking probability less than ε_s for stream flows of peak rate p_s and a blocking probability less than ε_e for elastic flows with required minimum throughput t_e, for appropriately chosen parameter values.

4.1. Provisioning for stream traffic

To determine the network capacity required to meet a target blocking probability, it is necessary to make assumptions about the arrival process of new demands, their rate and their duration. For illustration purposes, we consider a simple traffic model consisting of one link receiving calls from a very large population of users.

First assume that it is possible to identify m distinct homogeneous call classes, each class having a common rate distribution. Calls from class i arrive according to a Poisson process of intensity λ_i (calls per second) and have an expected call duration of $1/\mu_i$ seconds. Their peak rate is p_i. For a fixed (fairly large) link capacity c, the impact of a call of class i on the probability of cell loss can be summarized in a single figure, the effective bandwidth (cf. ITUb, 1997; Roberts et al, 1996): the effective bandwidth e_i is such that that the probability of data loss is negligible (less than a target value) as long as $\sum n_i e_i \leq c$, where n_i is the number of class i calls in progress.

Although measurement based admission control does not rely on the identification of the different call classes (a new request is denied if its peak rate is less than an estimate of available bandwidth), for dimensioning purposes we can assume a call of class j will be blocked if $\sum n_i e_i \leq c - p_j$. With this blocking condition and the assumption of Poisson arrivals, the distribution of the n_i has a well known product form (e.g., Roberts et al, 1996) enabling computation of the blocking probability. Note that the call blocking probability and the probability of data loss for a given set of admitted calls are insensitive to the distribution of call duration.

A reasonable approximation for the blocking probability of a flow with peak rate p_s when c is large with respect to the e_i is given by:

$$B_s \approx \frac{p_s}{\delta} E\left(\frac{a}{\delta}, \frac{c}{\delta}\right) \qquad (1)$$

where $a = \sum e_i \frac{\lambda_i}{\mu_i}$, $\delta = \sum (e_i^2 \frac{\lambda_i}{\mu_i}) / a$ and $E(a,n) = \frac{a^n / n!}{\sum_{i \leq n} a^i / i!}$ is Erlang's formula.

Formula (1) is a simplification of the formulae given by Lindberger (1994). It is less accurate but more clearly demonstrates the structural relationship between performance and traffic characteristics.

It is well known that application of Erlang's formula leads to scale economies: to achieve a low blocking probability and high utilisation (a/c), it is necessary to have a large capacity c. For multirate traffic with blocking probabilities given by (1), the same requirement implies a high value of c/δ. The line labelled "stream" in Figure 2 shows how the achievable utilization a/c in a simple Erlang loss system varies with c for a target blocking probability of 0.01.

In dimensioning the bandwidth of a link for given traffic, it is necessary to take account of the fact that equivalent bandwidths e_i, and consequently blocking probabilities, depend on the capacity c. An iterative procedure may thus be necessary using a first estimate of δ to calculate the required capacity c and then repeatedly recalculating δ and a new value of c until convergence (ITUb, 1997).

The blocking probability in a network can be derived approximately on assuming links are statistically independent and using fixed-point algorithms, as outlined in (Roberts et al, 1996), for example.

4.2. Provisioning for elastic traffic

With the simple service model of section 3, minimum throughput of elastic flows, t_e, is guaranteed by admission control. We assume here that the criterion for link sizing is simply that the probability of blocking should be less than ε_e.

Consider first an isolated link handling only elastic flows. Assuming Poisson arrivals, a common minimum rate t_e, exact fair shares (i.e., processor sharing service) and a link bandwidth of $c=n \times t_e$, the probability of blocking is equal to the saturation probability in an M/G/1 processor sharing queue of capacity n:

$$B_e = \rho^n (1-\rho) / (1-\rho^{n+1}) \tag{2}$$

where ρ is the link load defined in section 3.2.

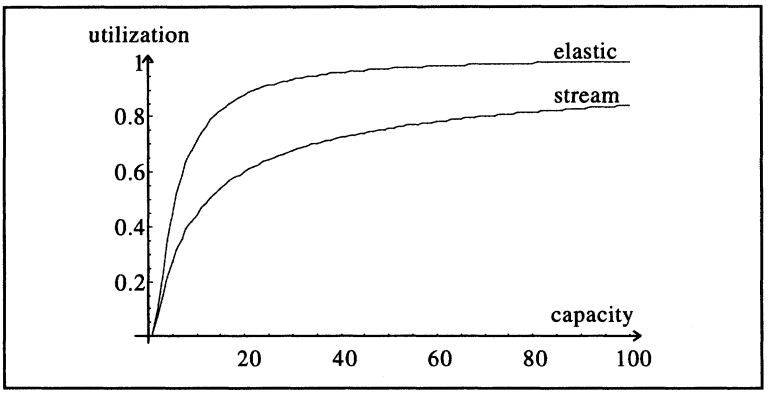

Figure 2. Achievable utilization ρ_{lim} against link capacity n

Since elastic flows use bandwidth more efficiently, blocking probability (2) can be considerably less than the corresponding probability for stream traffic requiring

constant rate t_e, as given by Erlang's formula $E(n\rho,n)$. The line labelled "elastic" in Figure 2 shows achievable utilization ρ for elastic traffic such that B_e, given by (2) is equal to 0.01. These results illustrate clearly the scale economies effect and the greater efficiency of elastic sharing. This advantage may, however, be somewhat mitigated in a network where flows cannot generally attain a full share of available link bandwidth rate because of congestion on another link of their path, including the access link.

Throughput in a network depends on the utilization on all links of the communication path. For example, if the closed loop control protocol realizes min-max fairness (e.g., as in Charny et al, 1996), the realized throughput of any transfer is a complicated function of all transfers in progress throughout the network. In the example of Figure 3, the flow on link A cannot expand its service rate to the full capacity because it is constrained on link B. Throughput in a network thus tends to decrease as the number of links used by a flow increases. Flows over short paths may, on the contrary, gain in throughput since they benefit from the constraints restricting other flows routed over long paths.

A flow which cannot use its full allocation on a given link lasts longer and thus leads to an increased probability of blocking. The link blocking probability is, however, always less than that predicted by the Erlang formula which could be used as a conservative estimate for dimensioning.

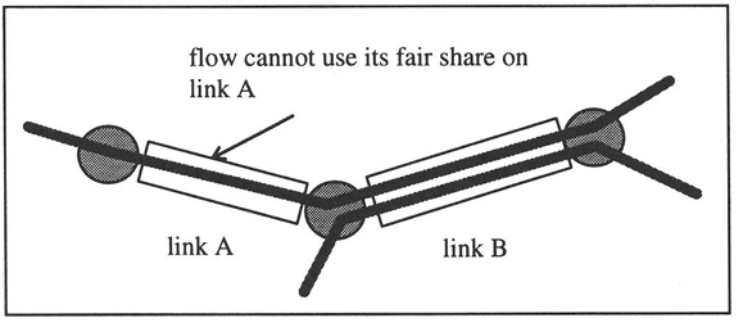

Figure 3. Min-max fair shares on a three-node network

Note finally that we have evoked the problem of sizing for stream and elastic traffic considered separately. In practice, of course, both types of flow share the same bandwidth and it is necessary to derive traffic engineering methods which allow the network to satisfy blocking probability objectives for both types of traffic simultaneously. The optimal design of a multiservice network realizing target blocking probabilities for a specified volume of stream and elastic traffic is an open problem.

5. CONCLUSION

In this paper we have stressed that the realisation of quality of service guarantees in a multiservice network depends jointly on the definition of the service model which specifies how resources are shared and on the network engineering procedures which determine how much capacity must be provided.

We have argued that the service model should distinguish two broad categories of traffic: stream traffic, characterized by intrinsic rate and duration, and elastic traffic, whose rate can be modulated to fully exploit currently available bandwidth. It is suggested that the service model should be designed principally for the transfer of individual flows (e.g., one videoconference, one file transfer, ...) and not for the transport of aggregations of flows (e.g., LAN interconnection) whose statistical characterization is notoriously difficult.

We have sought to identify minimal requirements for a simple service model capable of satisfying the respective transparency and throughput requirements of the two types of traffic. The choice depends on the adopted charging scheme. We assume here that user charges are determined by pricing each transaction rather than relying uniquely on a flat rate, with or without the adjunction of some form of congestion pricing. This charging scheme imposes the use of admission control to ensure that transparency and throughput requirements are satisfied.

The suggested simple service model has two service classes: one, for stream traffic, operated with "bufferless multiplexing" admission control, the second designed to provide a common guaranteed minimum throughput to all elastic flows. We have not worked out the precise details of the flow control algorithm necessary for sharing bandwidth above the minimum guaranteed rate but anticipate a network level protocol, considerably simpler than that of the ABR service class in ATM.

Accessibility quality of service requirements must be satisfied by network engineering. Sizing procedures for stream traffic using the notion of effective bandwidth have already been widely studied. The throughput and blocking performance of elastic traffic, on the other hand, is much less well understood. We have proposed a simple processor sharing model for a single bottleneck link which illustrates some interesting properties. Many challenging problems remain open, however, including the evaluation of expected throughput in a network and the derivation of sizing procedures accounting for stream and elastic traffic integration.

The suggested simple service model is not intended as a practical proposition for the Internet or the B-ISDN. It is rather intended to clarify the issues involved in providing quality of service guarantees in a multiservice network. In the search for simplicity we have neglected many possible additional requirements. The present study can help in evaluating the possible cost in added complexity of taking these requirements into account.

6. REFERENCES

M.F. Arlitt, C. Williamson (1996) Web server workload characterisation: The search for invariants. *ACM Sigmetrics 96.*

F. Brichet, L. Massoulié, J. Roberts (1997) Stochastic ordering and the notion of negligible CDV. *Proceedings of ITC15* (V. Ramaswami, P.E. Wirth (Eds). *Teletraffic contributions for the information age*), Elsevier.

M. Crovella, A. Bestavros (1996) Self-similarity in World Wide Web traffic: Evidence and possible causes. *ACM Sigmetrics 96.*

A. Charny, K.K. Ramakrishnan, A. Lauck (1996) Time scale analysis and scalability issues for explicit rate allocation in ATM networks. *IEEE/ACM Trans. Networking*, Vol.4, N°.4.

G. Fayolle, I. Mitrani, R. Iasnogorodski (1980) Sharing a processor among many jobs. *Journal of the ACM*, Vol 27, N°3, pp519-532.

M. Frater (1997) Origins of long range dependence in variable bit rate video traffic. *Proceedings of ITC15* (V. Ramaswami, P.E. Wirth (Eds). *Teletraffic contributions for the information age*), Elsevier.

R. Gibbens, F. Kelly, P. Key (1995) A decision theoretic approach to call admission control in ATM networks. *IEEE JSAC*, Vol 13, N°6.

M. Garrett, W. Willinger (1994) Analysis, modeling and generation of self-similar VBR video traffic. *Proceedings of SIGCOM 94.*

D. Heyman, T. Lakshman, A. Neidhart (1997) A new method for analysing feedback-based protocols with application to engineering Web traffic over the Internet. *ACM Sigmetrics '97.*

M. Hamdi, P. Rolin, J. Roberts (1997) Rate control for VBR video coders in broadband networks. *IEEE JSAC*, Vol 15, N° 6.

ITUa (1997) Recommendation E.736, Methods for cell level traffic control in B-ISDN, ITU, Geneva.

ITUb (1997) Recommendation E.737, Dimensioning methods for B-ISDN, ITU, Geneva.

L. Kleinrock (1975) *Queueing Systems*, Vol 2, J. Wiley & Sons.

K. Lindberger (1994) Dimensioning and design methods for integrated ATM networks. *Proceedings of ITC 14* (J. Labetoulle, J. Roberts (eds). *The fundamental role of teletraffic in the evolution of telecommunications networks*) Elsevier

W. Leland, M. Taqqu, W. Willinger, D. Wilson (1994) On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking*, Vol 2, N°1.

J. Roberts, U. Mocci, J. Virtamo (1996) *Broadband Network Teletraffic* (Final Report of COST 242), LNCS 1155, Springer Verlag.

J. Roberts (1998) Quality of service guarantees and charging in multiservices networks, To appear in *IEICE Trans Commun.* Special issue on ATM traffic control and performance evaluation, 1998.

7. BIOGRAPHY

Jim Roberts was awarded a Doctorate by the University Pierre et Marie Curie, Paris, France in 1987. He began working in the field of teletraffic engineering and performance evaluation with the British Post Office in 1971 and has been with France Télécom since 1978. His current research is mainly concerned with the performance evaluation and design of traffic controls for multiservice networks. He is an associate rapporteur for ITU Study Group 2 activities on B-ISDN traffic control and dimensioning methods. He is a member of a several journal editorial boards and conference programme committees in the networking field.