Traffic engineering in ATM networks: Current trends and future issues

A Gravey¹, G Hébuterne², R R Mazumdar³ and C Rosenberg⁴

¹ France Telecom, CNET-Lannion A, Route de Trégastel, 22300 Lannion, France

² Département RST, Institut National des Télécommunications, 91011 Evry Cedex, France

³ INRS-Télécommunications, Université du Québec, Ile-des-Soeurs, P.Q. H3E 1H6, Canada

⁴ Département de Génie Electrique, Ecole Polytechnique, Montréal, P.Q. H3C 3A7, Canada

Abstract. In this survey we present an overview of the current state-of-the-art and the future issues to be resolved for the deployment of integrated communications networks based on the Asynchronous Transfer Mode (ATM). Our presentation is from the perspective of traffic and congestion control and management issues which hold the key to the successful achievement of this deployment since it is mainly in these aspects that ATM networks differ from conventional communication networks.

Keywords. ATM; Traffic characterization; control; quality of service.

1. Introduction

In the past the evolution of telephone and data networks has gone ahead independently, the former was performed by the telephone companies while the latter was done by the computing community (the ARPAnet being the case in point). This has basically involved the setting up of a parallel infrastructure or communications backbone suited to the needs of the traffic characteristics of each service. With the need for economic efficiency and a convergence in the demands from both directions to provide a reliable, high bandwidth communications backbone to offer enhanced or new services, there was a consensus in the seventies to develop Integrated Services Digital Networks (ISDN) which would provide a common access for data as well as voice traffic. The ISDN concept was based upon a digital telephone network using common channel signalling to allow basic access rates based upon multiples of 64 kbits/s which is the fundamental building block of ISDN services. However real integration was not performed since the two traditional transfer modes were coexisting just behind the access point.

With the explosion in the communication needs in the eighties driven by the need to provide access to traffic requiring much higher bandwidth (bitrates) than was anticipated the transition towards Broadband ISDN (B-ISDN) networks was made. The idea of B-ISDN is very revolutionary compared to ISDN since the goal is to provide not only a common access to all existing and future communication services but also a common transfer mode to allow an efficient and transparent transport of information while assuring a service dependent Quality of Service (QoS) which is usually specified in terms of delay and/or loss constraints.

Due to the extremely high speeds required for some applications such as real-time video and medical imaging it was quickly realized that the conventional 7-layer ISO architecture on which ISDN is based as well as the error and congestion control procedures existing in layer 2 and 3 would be the limiting bottleneck. This is due to the fact that many of the control procedures associated with traffic management procedures in a X.25-like network such as link-by-link error control would inhibit the speeds and are not necessary thanks to the use of virtually error-free optical fiber transmission links. Besides the ability to react to congestion through feedback would be impossible since by the time the information is fed back the state of the network could be vastly different making the information completely useless. Thus, it was natural for introducing a change in paradigms in the way traffic management and control is to be done.

The shift in paradigms meant that to achieve high speeds it is essential to reduce the overheads by the migration of "intelligence" towards the perimeter or input nodes while providing a "transparent" high-speed and reliable communication link between user and destination. In addition the network architecture should incorporate the flexibility of providing the nice features of circuit-switched networks and the flexibility and efficiency of packet switched networks. These characteristics are to be found in the concept of the Asynchronous Transfer Mode (ATM), originally proposed by CNET in France, which was adopted as the standard by the ITU-T (successor to the CCITT) in 1989 (1.121 1989). The basic concept of ATM is to provide a powerful inter-connection medium with limited functionalities ("intelligence") but offering high performance in terms of bandwidth, QoS and access. This would allow for the means to integrate all existing services and offer high bit-rates to end users through a paring down of the number of layers of the protocols.

In order to appreciate the issues it is worth examining the current services that are envisaged and why the traffic engineering issues are central to the design of ATM networks. Before we do so we state a few provisos: this paper is not an introduction to ATM networks (for an introduction to ATM see CNET (1991) and Minzer (1989)) rather this paper highlights the traffic engineering issues with a presentation of the basic problems, current trends and future issues to be resolved if the goal of providing a fully integrated, versatile and reliable communications network is to be realized. The second point is that we do not survey all the trends but restrict ourselves to the issues in the context of standards that have been agreed upon since if future research is to have any impact it must be within the guidelines of the standards. Finally the issues presented here are not exhaustive and research in this area is being pursued actively the world over.

Let us begin with standard telephone traffic. The analog voice signal is sampled and digitized, resulting in a continuous bit stream: 1 byte every 125 μ s. The main constraints to handle such a service are: no variation in the transport delay of

successive bytes (or at least a very limited variation in case of packetized voice), low delay (say around 20ms to avoid the need for echo cancelling), allowable losses upto ratios around 10^{-3} , low cost (in terms of complexity) of end equipment.

Data transfer is another kind of service, with different characteristics: information arrives in long blocks (from kbytes to Gbytes in the case of file transfers); there is usually no constraint on the delay or on its variation; loss must be kept minimum – less than 10^{-9} ; the complexity of existing end equipment eases implementing elaborate procedures.

Naturally, all intermediate cases could be imagined. Video for professional or entertainment usage; high quality sound; etc. In each case, the service may be characterized in terms of delay or loss constraints and of bandwidth requirements. It is characterized also by the statistical description of the data flow to be carried. Note that most services to come will be *multimedia*, meaning a mix of basic services in a single offer.

The coexistence of these services puts strong requirements on the network. This has been summarized under the concept of *transparency*: the source generates its traffic at its own rate, and ideally the network should carry the traffic at the same rate and without any alteration.

In order to fulfill the above requirements, all services in the B-ISDN are offered above a common medium, the ATM network, that provides basic and powerful interconnection facilities. Information is segmented into fixed-size *cells* at network input nodes. The cell consists of a 48-byte information field and a 5-byte header. Cells are carried along *Virtual Path Connections* (VPC) or *Virtual Circuit Connections* (VCC), identified by the VPI- and VCI-field of the header. The ATM layer performs only routing operations. No error detection/correction takes place in the network (note however that the header is protected by a 1-byte CRC field: cells with corrupted headers are discarded).

The ATM layer of the B-ISDN performs the basic transport function in a simple and efficient way. All its complex procedures are hardware-implemented, yielding a high performance level and the capability to operate at high bandwidths. Specific additional functionalities, such as jitter or error correction, are to be offered on a per-service basis, by an ad-hoc “*ATM Application Layer*” (AAL). Thus we can view the ATM layer as a fixed and short size cell relay network.

Note that the existence of VPC as well as VCC allows for a wide range of semi-permanent and switched accesses.

This organization aims at giving the B-ISDN the higher “flexibility” – ease in introducing new services with specific Quality of Service constraints.

However, a problem arises concerning *traffic and congestion control*. With traditional circuit switched networks, a source cannot exceed the basic rate offered to it (64 kbit/s, in telephone network). Most of the time, the ATM network will offer the source full access to a 150 Mbit/s link. In X.25-like packet networks, link-by-link congestion control and acknowledgments prevent a source from flooding the network with its traffic. The lack of elaborate protocols in the ATM layer prevents the network management from relying on such procedures. In addition the sheer speeds makes the introduction of state dependent control mechanisms of limited value except in the case of congestion notification etc. and at the stage of admitting a connection. Hence, traffic control procedures must rely on “open-loop” measures, i.e., ensuring that traffic inputted into the network behaves “well”. This

basically implies that the traffic control mechanisms have to do the job of verifying the conformance of each traffic stream with the amount of resources allocated to the connection and the state of the network.

The aim of this paper is to point out how traffic and congestion control happens to be the central issue in ATM networks and the current approaches to address these problems.

The organization of this paper is as follows: In §1 we describe the challenge of traffic and congestion control and in particular, we discuss the issue of source characterization and the Quality of Service, source policing, Call Acceptance Control which is the procedure by which a call is accepted and the appropriate bandwidth allocated. In §2 we discuss the present status of traffic control in ATM networks. In Section 3 we discuss the issues which still require further studies like the one related to the choice of a multiplexing scheme to offer appropriate bandwidth which is aimed at increasing network efficiency.

2. Challenge of traffic and congestion control

One of the basic features of ATM networks is that at call set-up a negotiation takes place between the source and network. The source is supposed to indicate its characteristics and requirements in terms of QoS (loss probabilities, delays etc.) and the network decides whether to accept the connection. Based on the QoS and source characteristics the network makes decisions to allocate the appropriate bandwidth and buffers etc. and the network is committed to deliver the QoS if the the sources conforms with respect to its characterization. Of course due to the stochastic variation in the cells emitted by a source there could be a potential gain to statistically multiplex the different sources to gain efficiency of goodput (throughput meeting QoS). For this an efficient characterization of the source is needed.

For a source to declare its complete statistical characterization is completely unreasonable and besides the statistics could be altered (especially in the case of VPC) so one has to seek robust techniques for source characterization with a minimal amount of information to be conveyed. In addition such a minimal characterization can only hold in some approximate sense (i.e., with a closing error) since it is unreasonable to presume that a source will obey the characteristics conveyed in an absolute sense.

Once resources are committed on acceptance of a connection, the network must ensure that the source conforms to the declared characteristics. This control function has been termed source policing and lack of such a control will affect the QoS of background connections when the network tries to maximize efficiency.

According to the ITU-T (I.371 1993), a source submitting a connection request must provide a set of parameters to the network, the Source Traffic Descriptor, whose contents are meant to capture the source intrinsic behaviour. These parameters are part of a traffic contract negotiated between the user and the network. The *Source Traffic Descriptor* (STD) is a subset of the ATM Traffic Descriptor; in addition to traffic parameters, the contract also covers Quality of Service objectives and Cell Delay Variation tolerances. The declared traffic parameters must then be enforced (policed) by the network to protect the grade of service offered to compliant connections.

Hence one of the great challenges is to give an adequate characterization of source

traffic based on a minimal description. The source characterization must also offer maximum flexibility to the source in the way its traffic is generated.

2.1 Characterizing source traffic

There are various ways to describe the cell stream issued by a source. On the one hand, one can classify the sources according to their statistical pattern; on the other hand, “significant” parameters must be defined, to quantify the corresponding traffic process.

Significant parameters should measure the way the given source may influence the network behaviour (i.e., the impact on other already set-up connections), and should allow us to compute the QoS the network may offer the new connection. A typical scenario for the selection of source traffic parameters is in the framework of figure 1. All sources share a common communication resource, and are multiplexed through a finite-size buffer. The STDs must allow us to estimate the resulting QoS, typically loss probability and delay.

Figure 1. The queueing model of multiplexing sources

An important parameter arises naturally, it is the *peak cell rate*. This is the maximum rate at which the source is able to send data and can be viewed as the minimum interval between two consecutive cells. Its formal definition in the ITU-T framework calls for the *Equivalent Terminal* formalism (I.371 1993). Standardization efforts at the ITU-T have focused so far only on issues related to the peak cell rate.

Before we discuss the problem of source characterization beyond the peak cell rate, it is important to note that as a first cut the following categories of sources are naturally identified.

Constant bit rate source (CBR): This is the simplest case; the source generates a cell periodically, at its cell peak rate. The usual telephone traffic belongs to this class (64 kbit/s is 1 cell every 6 ms).

Variable bit rate source (VBR): The simplest video transmission would generate CBR streams. However, video coders efficiently compress the information stream, resulting in variable bit rates. The VBR source is described as having a long-term bandwidth requirement significantly lower than its peak cell rate, with an instantaneous demand varying continuously. Many applications can be viewed as VBR sources (interactive data processing, for instance).

Sporadic source (on/off source): This type of VBR source corresponds to a two-state CBR source. During the “On” state, the source generates a continuous stream.

During the “Off” state, the source remains silent. This definition can be generalized by allowing the process to switch between several CBR levels. Usually, the sojourn in each state has a random duration. One can give numerous examples of this kind of source. File transfer and still picture transmission (medical imaging) are the simplest representatives. For such applications, the peak cell rate of the “On” state is generally quite higher than the mean (it conditions the duration to transmit the picture).

Note that a “good” traffic parameter should be intrinsic, enforceable and usable in the allocation process. Various parameters have been proposed to the purpose of source characterization beyond the peak cell rate to allow statistical multiplexing. The first one is the *mean bit rate*. Roughly, this is the rate at which cells arrive, estimated over a long period of time. Note that only a Poisson arrival process is fully characterized by this parameter. Note also that enforcing such a parameter would, in general, require measures on a very long period.

The *burstiness* is the ratio of the peak to the mean cell rates. It is introduced as a way to quantify the variability of the arrival process.

Other ways can be used to describe the variability, and the correlation of the cell flow. For instance, let $N(t)$ be the number of cells arriving in $(0, t]$. One can compute the mean and variance of $N(t)$, and the “*Index of Dispersion for Counts*” is defined as $I(t) = \text{Var}[N(t)]/E[N(t)]$. One should choose one point of $I(t)$, for instance the maximum value, or the limit as $t \rightarrow \infty$, etc.

These parameters undoubtedly present interesting features; unfortunately, it is not clear how they can be used to allocate resources in a statistical multiplexing context and how they could be efficiently enforced.

Another approach of the source characterization consists in postulating a source process of the type “*n*-stage Markov Modulated Bernoulli Process” (MMBP_n): in each stage, cells are sent according to a pure chance process (Bernoulli), and jumps between states occur according to a Markov process. The parameters are chosen such that the MMBP “looks like” the actual process.

It must be remarked that, for a statistically well-defined class of sources, efficient traffic descriptors can be found. For instance, an On/Off source is perfectly characterized by the length of the emission and silence intervals; the Poisson arrival process is fully characterized by the parameter “ λ ”. Unfortunately, this kind of characterization is of little value since the network would have to identify both the class and the corresponding parameters which could be only done theoretically with an infinite record.

To circumvent these difficulties an approach that has been adopted within the standardization bodies is to search for algorithmic rules i.e., rules such that if the cell count confirms according to some parameters within a prescribed time-interval then the source is said to be conforming. A second and more recent approach is via the use of the notion of “effective bandwidths” which tries to identify a single parameter associated with a source such that if bandwidth greater than the effective bandwidth is allocated the cell stream will meet the required QoS. We briefly discuss these issues and a more recent approach due to the authors based on quantile measures.

Algorithmic descriptors: Recently an algorithmic procedure called the Generic Cell Rate (GCRA) orum. This is very close to the (σ, ρ) characterization introduced by

Cruz (1991). These methods are in fact a particular instance of a broader class of algorithmic characterizations as shown by Berger (1994). The idea is that an arrival stream is said to be (σ, ρ) constrained if in the interval $[t_0, t]$

$$N[t_0, t] \leq \rho(t - t_0) + \sigma$$

Suppose that the traffic stream is (σ, ρ) constrained for all t , then it is easy to see that from the above form the mean rate is bounded by ρ and σ is a measure of tolerance (a measure of variability) of the stream. Of course it is difficult a priori for a source to give such a characterization but such a characterization is directly enforceable by a simple *leaky bucket* (see discussion below in the context of control). These descriptors have gained much acceptance due to their simple characterization and the means to enforce the parameters. However such a characterization does not yield tight performance bounds since it is essentially a deterministic measure which gives “worst-case” bounds. Much work needs to be done on obtaining tight bounds from which we can study the multiplexing gains when several streams are multiplexed on a common line.

Effective bandwidths: The notion of an effective bandwidth of a source was first introduced by Hui (1988). The idea is to specify a single parameter called the effective bandwidth such that if the effective bandwidth is less than the multiplex rate (bandwidth allocated) the QoS constraints are met. This idea has been further refined by Kelly (1991) for shared resource systems and in the context of buffered resources by Elwalid & Mitra (1993). More recently a detailed exposition of this idea and its relation to the so-called rate functions in large deviations theory has been presented by Whitt (1995). By now the idea is fairly well understood for buffered and unbuffered systems when viewed as systems in isolation with a given mix of traffic. We briefly discuss this idea below.

Suppose there are N types of ON-OFF sources whose ON distributions are specified by $F_i(x)$, $i = 1, \dots, N$. Suppose the sources wish to access a common multiplex of bandwidth C and a given source requires one unit of bandwidth when ON. Suppose we wish to allocate bandwidth such that the loss probability is less than $e^{-\gamma}$. Then it has been shown by Hui that there exists a parameter $\alpha_i(F_i)$ which depends on C and γ such that if

$$\sum_{i=1}^N \alpha_i(F_i) < C$$

then the sources will suffer blocking less than $e^{-\gamma}$. Kelly (1991) showed this also holds for delay constraints. The parameter α_i is called the *effective bandwidth* and represents the bandwidth requirements of the source to meet its QoS constraint. Note $\alpha_i(F_i) \leq 1$.

One of the important issues to resolve is given that there is a new request for a connection can one compute the effective bandwidth of the new connection without recalculating the effective bandwidths of all existing connections and the effects of the scheduling policies? Recent work on the notion of “decoupling bandwidths” by Di Vecchia *et al* (1994) is a step towards this direction but appears overly conservative. Moreover in the buffered resources context, the effective bandwidth is asymptotic as the buffer size increases and thus reduced utility since large buffers

will contribute to longer delays. Another issue concerns the enforcement of such a parameter. Nevertheless this does not mean that such a concept is not useful.

Time-quantile measures: While characterizing source behavior with one or a few parameters is a worthy goal it is of interest to ask the question what statistic associated with the arrival counting process (other than the obvious true interarrival distributions) characterizes the process from the point of view of performance and control objectives? In recent work Rosenberg *et al* (1993) showed that the so-called time- ε -quantile process yields such a useful criterion. The time-quantile process associated with a point process is the function, denoted $M_\varepsilon(T)$ which corresponds to the $(1 - \varepsilon)$ quantile for the number of counts in an interval of length T . If ε corresponds to the loss level required such a function determines the maximum number of arrivals in an interval of length T with prob $(1 - \varepsilon)$. It can be shown that statistics such as the indices of dispersion, mean and burst rate can be extracted from such a measure which can be determined empirically if the source is stationary in the long term [?]. Moreover this measure is directly enforceable. If one defines the function $r(T)$ corresponding to the maximum number of cells accepted by a leaky bucket in any interval of length T then a cell loss rate $< \varepsilon$ is assured if $M_\varepsilon(T) < r(T)$ for every T (Rosenberg & Lague 1994)

In spite of the above developments traffic characterization remains an important issue since the gains of statistical multiplexing can only be achieved by a tight characterization which may imply complex procedures at the user end for determining it and heavy signalling to pass the information to the network. Moreover, in some remarkable empirical measurements on the aggregated LAN traffic reported by Leland *et al* (1994) it has been suggested that such traffic has fractal characteristics. This renders much of the issues such as effective bandwidths moot since the long range dependence makes the distributions heavy tailed and hence large deviations approaches might fail. Recently, Norros (1994) studied a model based on a simplification that the fractal traffic is modeled as a shifted fractional Brownian motion. Even based on this approximation he showed that the loss characteristics are crucially dependent on the *Hurst parameter* (which governs the rate of decay of the covariance and is a measure of the fractal dimension) and furthermore that having long buffers does not improve the situation and thus this type of traffic must be included into the categories of traffic models.

For a comprehensive survey of source characterization, the reader is referred to Roberts (1991) and the recent proceedings of the International Teletraffic Congress (Labetoulle & Roberts 1994).

2.2 Open-loop control of a source

As already mentioned, in the ATM context feedback or reactive control of sources is not feasible due to the high delay-bandwidth product. This is due to the fact that the delays in feeding back the status of the links are of comparable order to the transmission times so that by the time status information is received the state could be drastically different making the information useless. Hence the way the control function is envisaged is via open-loop or pro-active means by which desirable behaviour is enforced at the input point. In the ATM context this is referred to as

source policing. The requirements of such a control mechanism are that the function must be easy to implement and must react fast to non conforming behavior while being transparent to conforming streams.

The problem could be mathematically posed as follows: Let $\varphi(\cdot)$ denote the policing control which acts on the input stream $A(0, t)$ yielding an output stream $D(0, t)$ i.e.,

$$\varphi(A(0, t)) = D(0, t)$$

Then the requirement is that $D(0, t)$ which is the policed input to the network must be such that it conforms with the negotiated contract to meet the QoS and $\frac{D(0, t)}{t}$ should be maximal i.e., the amount of throughput should be maximized.

It turns out that the map $\varphi(\cdot)$ which is optimal is the so-called *leaky bucket* (LB for short) mechanism which we describe below: consider a *virtual queue* in which service tokens are generated at constant rate R and an arriving cell consumes a service token. There are a maximum of N tokens available and thus if on arrival there are no tokens available then the corresponding cell is either dropped or “tagged” depending on the philosophy of the network (see below). Note when service tokens are available the cells transit through the LB without incurring any delay.

Let $r(T)$ denote the maximum number of cells accepted in an interval of length T by the mechanism ($r(T)$ is a deterministic upper-bound on the number of cells allowed by the LB to enter the network over any time interval of length T). Then it is easy to see that $r(T) \leq N + RT$ or it is (N, R) constrained at the output. The interpretation is that N is the maximum instantaneous burst allowed and R the bound on the long-term average.

This is a particularly simple mechanism and was proposed by Turner (1986) in his seminal paper. The optimality of the mechanism in the sense that it is the mechanism which maximizes the throughput subject to the (N, R) constraint on the output stream was recently shown by Anantharam & Konstantopoulos (1994).

ITU-T has defined an algorithm which generalizes the LB: it is the *Virtual Scheduling Algorithm* (VSA) – or Continuous State Leaky Bucket (I.371 1993). It can be seen as a LB where the parameters N and $T = 1/R$ would not be restricted to integer numbers.

In spite of the simple form of the control mechanism the problem of controlling a source is far from trivial which we discuss below.

Ideally, a Source Traffic Descriptor (STD) should be controlled at the interface point between the source and the network. Actually, the control process may not be performed at such a point, most of the time: as illustrated on figure 2, either the source is connected by means of a customer equipment (e.g., a local network), or the control is performed at an Inter-Carrier Interface (ICI). As a consequence, the control device only measures the source process “corrupted” by one or more exogenous processes. Some parameters have to be set in the traffic contract, in order to take into account the modification of the STDs in the customer equipment.

Of primary concern here is the perturbation on the cell arrival process of the source to be controlled. Let us assume, for simplicity, that a CBR source is observed. Cells ought to be detected at epochs of the form $T_0 + kT$ (T is the emission period). Most of the time, cells encounter some *Cell Delay Variation* (also called *jitter*) – that is they arrive sooner or later as compared with theoretical epochs. Possibly, one may even observe *bursts* of cells (i.e., groups of back to back cells). Since the phenomenon is inherent to an asynchronous transfer mode, this is the task of the control process

Figure 2. The general configuration: the Source is not connected directly at the UNI

to take it into account. The influence of the jitter must be upperbounded and the tolerance of the control algorithm must be tuned accordingly.

It has been recognized that cell jitter is composed of two basic complementary phenomena, namely cell clumping and cell dispersion (Roberts & Guillemin 1992). To illustrate the cell clumping effect, consider an initially periodic cell stream with period T passing through an ATM network. When this cell stream is observed at some point along the connection it may happen that the distance between two consecutive cells is less than the initial period T because of random delays. This is precisely what is called cell clumping. Its impact on the ATM layer performance is obvious since by creating bursts, it may affect the Cell Loss Ratios and thus must be controlled for the ATM layer sake (this has been realized a long time ago and cell clumping is taken care of in I.371 (see below)). It may also happen that two consecutive cells experience increasing delays so that their inter-arrival time at the observation point is greater than the initial period T , giving rise to a cell gap in the cell stream. This is the phenomenon of cell dispersion, which is critical for connections supporting a circuit emulation (e.g., ATM Adaptation Layer, AAL, of type 1) and/or when resequencing (e.g., two-layer video) has to be performed in the receiving terminal.

Corresponding to each control device, one can define the *worst-case* traffic. This is the “worst” source that is observed as conforming to the negotiated contract. By “worst source” is meant here the source inducing the higher loss ratio for its companion flows in any queue, like the one of figure 1. The safest approach is then to allocate resources based on the worst case traffic, the general belief concerning the worst case traffic for a LB is that the worst case is given by an On/Off periodic source filling the bucket with a burst at its peak rate and then remaining silent during a time just long enough to regenerate all the credits. For instance, assume one intends to control the peak rate of a CBR source, with a Leaky-Bucket, of parameters (R, N) . The worst-case source corresponding to the (R, N) -LB is an On/Off process, with On periods generating $N/(1 - R)$ cells and silent periods of duration N/R . The network accepting the CBR source must be ready to carry in fact such an On/Off process. See Doshi (1993) for more considerations about the meaning to be given to the concept of “worst case source”.

Numerous studies have been done (see Rathgeb 1991) concerning the capabilities and relative performances of these mechanisms. Is it possible for instance to control the mean rate – say m – of a connection? Let us try a LB with R slightly higher than m . The value of N has to be chosen such as to compensate for statistical fluctuations in the cell generation process which implies in general the use of a large N which results in poor response time and protection against bursts. Similarly, controlling the peak rate D would require R around D , and theoretically $N = 1$

would be sufficient if there was no jitter but N should compensate for jitter. It is seen that in order to control by a LB the mean rate (resp. the peak rate) of a source that has been modified by an exogeneous process, the leak rate R should be chosen as an upper-bound of the mean rate (resp. peak rate) and N should be chosen to account for the allowed “modification”, i.e., the jitter.

All studies (see Rathgeb 1991 for instance) show clearly the impossibility to control any *statistical* parameter, as for instance a mean rate, or an IDI.

Two reasons can explain this: first of all, a statistical descriptor makes sense only for a source known to belong to a well-defined “class” (this is the case if the source is a Poisson process; its mean rate is described by λ ; but the Poisson nature of the source has to be checked at the same time. Otherwise, there exists an infinite number of ways for a process to achieve a given “mean rate”); secondly, a statistical parameter requires often lengthy measurement procedures – since they are defined as ergodic limits and thus of limited value due to the intrinsic non-stationary nature of real traffic.

However, from a practical standpoint only sample-path based schemes are of importance: no end-user would be able to predict the value to be allocated to the “mean rate” a connection is to carry, or to bound efficiently the maximum value of its IDI. Thus the most promising mechanisms are based on monitoring sample-paths which are essentially deterministic in nature.

Moreover, all studies show clearly that a single algorithm (i.e., one (R, N) -LB) can by itself control satisfactorily *only one characteristic* of the source. Also, they show the weakness of these mechanisms, where N is chosen according to the cell jitter, which is to some extent external to the source definition.

As a consequence, no attempt is made today to control anything but *operationally defined* parameters. Peak rate and jitter are defined by means of an algorithm (VSA) which gives an instantaneous answer: a cell arriving at time t is conforming or not

depending on whether or not it passes successfully the control process. This approach has gained such considerable success that the “Sustainable Cell Rate” has been introduced within the ATM Forum as an upper bound for the mean rate and the burst size by means of the same algorithmic procedure.

2.3 Quality of service

In the discussion so far the Quality of Service (QoS) has been defined in terms of delay or overflow probabilities. In almost all studies so far the preference has been on long-term or stationary overflow probabilities as a QoS measure. However, the true QoS is more correctly reflected by the rate and amount of cells that are lost. This is basically related to transient characteristics of excursions of the underlying processes. In particular in addition to the stationary overflow probability, which we denote by $P(W > x)$ where W denotes the queue state or congestion process, the following measures associated with the overflow process need to be considered:

1. The transient overflow probability $P(\sup_{t \leq s \leq t+T} W_s > i/W_t < x)$ which corresponds to the probability of an overflow in the next T time units given that the system is not in congestion.
2. The length of the congestion period and the number of cells or bursts arriving during a congestion period since these measures determine the amount of information lost and hence have direct implications for the AAL layer.

The above measures are related to transients and difficult to obtain analytically. These issues have been addressed recently in the context of buffered systems (modeled as $G/GI/1$ queues) (Guillemin & Mazumdar 1995a) and unbuffered systems like multiplexers (modeled as $M/M/\infty$ queues) in Guillemin & Mazumdar (1995b) and Iscoe *et al* (1993). One remarkable fact is that in the region of interest for QoS in ATM systems the ratio of the transient overflow probability to the stationary overflow probability can be several orders of magnitude higher. Since any overflow results in cell loss or tagging and depending on the source the durations of excursions could be large, the transient characteristics are better indicators of QoS and thus cannot be ignored. Much work needs to be done on this issue for more relevant multi-rate models which will allow the determination of the true QoS offered to a connection as seen from the AAL.

3. ATM Networks: the current situation

3.1 Standards and their implications

ITU-T I.371 (1993) specifies what should be the traffic parameters of the source, and how to control them. Note that traffic contracts are naturally operator dependent, and as such outside the scope of Standards Organizations. However, the need to handle connections supported by more than a single network forces operators to agree on a minimum set of requirements, concerning traffic contracts and conformance with them.

One has to distinguish between *conformance testing* and *source policing*. The former refers to measurement procedures aiming at verifying if the source conforms to its negotiated traffic contract (and is clearly under the responsibility of Standards), while the latter encompasses all actions to be taken when non conforming cells are detected (and is operator specific). The network's commitment concerning QoS is thus directly related to cell conformance (Gravey & Hébuterne 1994).

ITU-T I.371 (1993) has defined the basic traffic contract at the peak cell rate of the connection. The source process is characterized by two parameters: the intrinsic Peak Rate of the connection, and the Cell Delay Variation (CDV) clumping tolerance. Specifically, for a given initial period T intrinsic to the connection (T is usually referred to as the peak emission interval of the connection (I.371 1993), cell clumping is quantified by using a reference algorithm, namely the Virtual Scheduling Algorithm (VSA) as defined in ITU-T Recommendations (I.356 1992; I.371 1993, Annex 1).

The VSA, also known as Generic Cell Rate Algorithm (GCRA) within the ATM Forum (1993), depends on two parameters, namely the increment I and the limit L . The increment I is usually equal to some reference period (e.g., the peak emission interval of a connection) and the limit L is some tolerance around this period, accounting for random cell transfer delays. In other words, when a cell stream is observed with a $VSA(I, L)$ to check conformance with respect to the cell rate $1/I$, a certain number of cells, up to some limit related to L , are allowed to arrive with an instantaneous cell rate larger than $1/I$. Specifically, a group of cells forms a cell clump when the $VSA(T, \tau)$ is ahead of the current time and the parameter τ quantifies the cell clumping effect on a connection in the strict sense (resp. the wide sense with a precision ε) if all cells (resp. a fraction of at least $(1 - \varepsilon)$ of cells) pass the $VSA(T, \tau)$. The impact of cell clumping on the peak cell rate control

function located in the Usage/Network Parameter Control (UPC/NPC) and then on resource allocation has been investigated in Roberts & Guillemin (1992) and Boyer *et al* (1992) respectively.

The first public ATM networks offer subscribers only peak cell rate traffic contracts. Note that it amounts to CBR source characterization only. However, the presence of jitter forces a negotiation based on the couple (T, τ) at the call set up.

The specification of cell dispersion within the standardization bodies is not as complete as that of cell clumping, especially because its close relationship with AAL issues.

Concerning policing actions, I.371 describes two possibilities: either the non-conforming cells are immediately discarded, or they are marked as non-conforming (*tagging*). A tagged cell is allowed to enter the network, but when congestion is detected within the network, tagged cells are destroyed first. Tagging uses the *Cell Loss Priority* bit in the header. This option implies that efficient selective discard mechanisms are implemented in the network (see below and e.g. Gravey & Hébuterne (1991), Gravey *et al* (1991), Kroner *et al* (1991) for more details).

The CLP bit allows another option, namely to offer two levels of QoS, more precisely two levels of Cell Loss Rate (CLR). The corresponding traffic contract implies to control both the CLP=0 flow and the aggregated CLP=0+1 flow. Serious issues have already been raised concerning the control of such hybrid contracts (Gravey *et al* 1991; Gravey 1994; Guillemin *et al* 1995). For instance, the control on the aggregated 0+1 flow could “destroy” (or mark as non-conforming) CLP=0 cells already accepted by the specific control.

Connection Acceptance Control procedures (CAC) have to take into account the jittered stream, in order to allocate resources to accepted connections. This is done by assuming that all sources behave according to the worst-case the control tolerates, for safety. The network utilization is thus the consequence of the following trade-off:

- either the connections request a low value for the CDV tolerance (the jitter the flow is allowed to be affected by) – say, less than 1 ms; this clearly limits the ATM gains, and increases the cost of subscribers equipments (need to shape outgoing traffic);
- or the connection may have a relatively high CDV tolerance (i.e., several ms) – i.e., the network accepts high jitter, the control device having accordingly unfavourable worst-case sources. This results in quite poor network utilization.

It is however possible to overcome this dilemma, by performing *active* control actions: Instead of using the negotiated parameters (T, τ) for conformance testing only, it has been proposed to use them to reconstruct the initial stream, as far as possible. This is a “reshaping” action: arriving cells are buffered and sent with a spacing at least equal to T . A control action is performed at the same time by limiting the buffer size, so as to discard very long bursts (possibly due to excessive jitter or to malevolent sources).

The “Spacer-Controller” (Boyer *et al* 1992) has a great advantage as compared with a “passive” control, like Leaky Buckets. Instead of limiting the size of bursts entering the network, it suppresses them, the outgoing flow is strictly conforming to its peak rate (actually, a residual jitter may be observed, due to multiplexing inside the Spacer Controller).

As a consequence, CAC procedures in a network which make use of the Spacer-Controller may be implemented without any reference to worst-case traffic. Note however that the choice to use such a shaping mechanism is operator specific.

The ATM-Forum (1993) extends ITU-T Recommendations, especially by defining the *Sustainable Cell Rate* Traffic Descriptor. It has been introduced by the ATM Forum to characterize *within the ATM Layer* and by means of an algorithmic rule, the statistical activity on an ATM connection. It is composed of two traffic parameters, namely the Sustainable Cell Rate $1/T_{SCR}$ and the Intrinsic Burst Tolerance (IBT) η_{BT} . Roughly speaking, the SCR is an upper bound on the source mean cell rate, while the IBT sets a limit on the size of bursts at peak cell rate. The definition of the SCR traffic descriptor relies on an intuitive approach aimed at defining an envelope for the traffic generated by an ATM source. The algorithmic rule used to define the SCR traffic descriptor is the GCRA, which has been adopted within the standardization bodies as *the* reference algorithm for cell rate monitoring (I.371 1993) and cell conformance testing (I.356 1992; Guillemin *et al* 1994a).

As for "0,0+1" contracts, interworking of Peak Rate and SCR controls on a given connection have to be carefully defined. In recent work Guillemin *et al* (1994b) showed that the specification of the SCR traffic descriptor, as stated by the ATM Forum, is incomplete. As it is defined now, the SCR characterization leaves the source with an infinite number of (T_{SCR}, η_{BT}) couples to choose from. This ambiguity is likely to be a potential source of problems. While admissible couples could be considered equally suitable from the source perspective, such is not the case from a networking point of view. The parameters of the SCR determine the worst case traffic pattern allowed in the network, which is generally recognized as the basis for resource allocation. A poor choice of SCR might thus involve lower network utilization (Rosenberg and Hébuterne (1994)); furthermore, the consequences depend on the multiplexing scheme and other network specific factors. In all cases, further information should be provided in order to insure that values for $1/T_{SCR}$ and η_{BT} are selected properly (Guillemin *et al* 1994b). It has been shown that in order to obtain a useful notion of the SCR from these different points of view it is necessary to define a "cost functional" and the SCR which in some sense optimizes this functional for "worst-case" traffic can lead to a useful measure.

The ITU-T Recommendations as well as the ATM-Forum (1993) mention other traffic control mechanisms which are discussed below in the section dealing with the "future" of ATM networks. One of them is the *Explicit Forward Congestion Indication* (EFCI), which is viewed as a "Traffic Control Function" in the ATM-Forum, while it is a "Congestion Control Function" within ITU-T (I.371 1993). This mechanism can be considered a high-level feedback or reactive mechanism which is only used in the case when congestion is present. Another one is termed with the generic name of *Fast Reservation Mode*.

Before dealing with the "future" of ATM networks (in terms of what remains to be done to provide an efficient broadband integrated service network), we describe below some current European experimental ATM networks intended for testing on a large scale a certain number of ideas and mechanisms.

3.2 Examples: the ATM-Pilot, the Brehat Project

European Public Network Operators have devised a public ATM network within the ATM-Pilot project. The first "public" ATM connection ought to be opened by the

end of 1994 (ATM-Pilot 1994).

The options taken during the specification phase of the network reflect the caution regarding the possibility of elaborate controls. Connections are offered on the basis of peak cell rate traffic contracts. No attempt is made to make use of the CLP bit to carry two levels of loss-sensitive cells. Concerning the control mechanisms and their actions, there are naturally operator specific, and will be defined soon.

However, the network will serve as a first step towards more efficient multiplexing schemes. The goal is to gain insight with ATM traffic and multiplexing mechanisms through a few benchmark services and experimentations.

"Brehat" is another experimental ATM network, led by France-Telecom (see Onno *et al* 1993). The first step of the project (December 1993) offers peak rate connections and control by means of Leaky Buckets. The second phase, expected for the end of 1994, will make use of the Space Controller (Boyer *et al* 1992) for the control of ingoing traffic.

4. ATM Networks: the future

The previous Section has presented the current situation on public ATM networks (at least in the European context). No attempt is made in this first step to offer services beyond the peak rate connections.

Evidently, this is to be seen as a first step, necessary to get acquainted with traffic phenomena, and with source and control schemes. The next step will consist in increasing network efficiency by offering some kind of multiplexing. The ultimate step would be to introduce multiplexing schemes allowing the network to be operated at utilization rates around 80% – and sources multiplexed according to their mean rates. This is not the goal of this paper to justify this evolution or to develop the related technical points. Only, for such multiplexing schemes to be implemented safely, the corresponding traffic control mechanisms have to be devised.

At this point, it is worth emphasizing a few remarks.

First of all, the question arises of the need for such improvements *within the ATM network*. That is, for all additional functions beyond the basic ones, the decisions to be taken are: should the function take place in the ATM network; or should it be integrated inside one (or several) AAL; or should a specific overlay network be built over the AAL, dedicated to specific service, and implementing the function. This is typically true for error correction, but could also be discussed for multiplexing. In the case where multiplexing schemes would be in use in AALs or in dedicated networks, the need of traffic control functions would not vanish: they would shift to these levels automatically.

Note however that different solutions may coexist at different levels, users and network managers being able to take advantage of the best one according to the application. For instance, reactive mechanisms (see below) behave quite differently if operated in a large public network (where round trip delays may reach tens of milliseconds) and in a Local Area Network (LAN) of limited geographic extension.

4.1 Statistical multiplexing schemes

The effectiveness of ATM networks as efficient carriers of information is going to depend very heavily on the efficiency of multiplexing. The ideal multiplexing mechanism – at least for the users' viewpoint, consists in sources sending their traffic

blindly, once the connection is set-up. This is usually referred to as *straightforward statistical multiplexing*. However, simple multiplexing where all the users have access to the available bandwidth will not work efficiently enough and more elaborate schemes based on bandwidth reservation schemes will be needed. Various mechanisms can be envisaged based on movable windows, thresholds or partial bandwidth sharing schemes. One of the most promising schemes in the ATM context is the so-called *generalized processor sharing* (GPS) or weighted fair queueing scheme (WFQ).

Weighted Fair Queueing (WFQ) is a service discipline designed to provide both delay and throughput performance guarantees, under the assumption that connections are constrained by LBs. This scheme was introduced in Parekh & Gallagher (1993, 1994) as a passive means of controlling congestion in high speed integrated services networks. It is, in essence, a generalization of round robin service discipline, where each session is guaranteed a minimal throughput and any remaining server capacity is split proportionally among connections.

All the performance bounds obtained for such schemes are essentially deterministic based on the assumption that the traffic is (σ, ρ) constrained. What is needed are analyses based on statistical assumptions. These models fall under the class of the so-called multi-rate loss models and these models both under straight forward multiplexing and bandwidth reservation have been addressed by Kelly (1991) and Theberge & Mazumdar (1994, 1995).

Another promising scheme is based on the "*Fast Reservation Mode*" (FRM) which covers all protocols by which the resources allocated to a connection can be modified in real time during the duration of the call.

A proposal for such a mechanism is made by Tranchier *et al* (1992). The connection is set-up based upon the usual peak rate contract. The protocol is operated by a command unit in the UPC/NPC. Upon request of the source, the protocol send a Request for Modification along the path of the connection. At each node, the Request is processed, and either accepted or rejected. If it is accepted, the resource allocated to the connection are updated. After completing its travel, the Request returns the command unit. If the bandwidth increase is accepted, the command updates the parameters (T, τ) of the traffic control device (LB, VSA) in the UPC.

The advantage of this multiplexing mechanism is to permit the network to accept long bursts, without long buffers. Note that the protocol may be operated by the network or by the user (within a VP, e.g.). See Tranchier *et al* (1992) for additional discussion.

For such schemes to guarantee the negotiated QoS, the acceptance control must be carefully tuned.

Assume the peak rates are already controlled -- and preferably enforced. For simplicity let us assume identical sources, with peak rate D and mean rate m . Let C be the link capacity. With peak rate allocation, the number of simultaneous connections is around C/D (assuming 100% efficiency). Statistical multiplexing allows a number of connections between C/D and C/m .

Statistical multiplexing requires us to define traffic parameters characterizing the statistical behaviour of the source (i.e., characterization beyond the peak cell rate). For this purpose, the ATM-Forum (1993) has defined the Sustainable Cell Rate. This descriptor limits the size of the bursts the source is allowed to send at the peak rate of the connection, upperbounding the actual mean rate. It is defined with respect to a VSA.

From the viewpoint of traffic control, FRM can be seen as statistical multiplexing using CBR sources – i.e., the only traffic control operates on the usual parameters (T, τ) of the peak rate management – with the improvement that these parameters are dynamically controlled.

4.2 *Explicit notification*

The principle of reactive, explicit congestion notification is well known. It has been adopted for instance in the Frame Relay networks. See e.g. Makrucki (1991) and Newman (1993) for its use in the ATM context. Two possibilities may be envisaged:

- Forward Congestion Notification (FCN). Once a node inside the network detects a congestion, it informs the receiver; the receiving end has then to take any flow control action to slow down the source.
- Backward Congestion Notification (BCN). In this case, the network element informs directly the sender (in the “backward” direction, implying a bidirectional connection).

The use of the FCN mechanism is mentioned, both in ITU-T (I.371 1993) and in the ATM-Forum (1993). Since an ATM connection is bidirectional, backward notification is always possible, *via* FCN. However, since responding to FCN/BCN is optional for end users, such mechanisms cannot be used alone.

FCN/BCN could offer interesting traffic management opportunities. This is especially true in local area network. Consider the typical LAN, with round trip delays around $50 - 100 \mu\text{s}$. If the duration of congestions is much larger than this time, the source may usefully react. Note that $100 \mu\text{s}$ is the time needed to send around 35 cells on a 150 Mbit/s link. In Wide Area Networks, or in large public networks, round trip delays may reach tens of milliseconds. In this case, the source could have sent up to 10^4 cells before receiving any notification. Note too that the LAN environment may be considered as more “cooperative” as compared with a public network. See Newman (1993) for a more complete discussion.

Finally, FCN/BCN mechanisms may help efficiently to regulate traffic and recover from congestion, especially in a LAN environment. As traffic control mechanisms, it remains to integrate them inside a complete set (including enforcing mechanisms); it remains also to determine which traffics would benefit from such reactive mechanisms.

5. Conclusions

In this paper, we have surveyed a number of topics related with the control of traffic in ATM networks. The need of efficient control schemes comes from the flexibility ATM is to offer to its users, in order to achieve a true multiservices network.

It was soon understood that it was both useless and impossible to control statistical characteristics of traffic sources. So the studies on traffic control have focussed on defining operational parameters related with traffic phenomena. These parameters are negotiated at call set-up, and then monitored during the connection duration through deterministic and real-time algorithms. Recall, that for a traffic parameter to be called “operational”, a measurement algorithm (or a conformance test) has to be defined with it.

Until now, Peak Cell Rate is defined this way (in duality with Cell Delay Variation), and the first implementations of public ATM networks should offer only peak rate based connections.

However, the need of more efficient multiplexing schemes spurs traffic specialists to devise new control mechanisms. Any multiplexing scheme is based on some traffic characteristics of the source. This must correspond to an operational parameter, and to Connection Acceptance Control procedures.

Statistical multiplexing (based on the Sustainable Cell Rate), Fast Reservation schemes, operating at the end user data units timescale, are declared candidates for these extensions. However, it is unclear if the ATM network is the preferred layer for elaborate schemes, or if elaborate multiplexing has to be offered on a per service basis, that is inside the ATM Application Layer.

In that case, the new traffic control algorithms would have to migrate towards upper layers (AALs or above). Nevertheless, traffic control will remain an important task of B-ISDN traffic management.

References

- Anantharam V, Konstantopoulos T 1994 Reflection mappings and flow control. Preprint, University of Texas, Austin (presented at Workshop on Stochastic Models, Ulm, Germany)
- ATM Forum 1993 The ATM Forum: UNI Specification. Version 3.0, Boston
- ATM-Pilot 1994 Info-Booklet of the ATM-Pilot Network
- Berger A E 1994 Desirable properties of traffic descriptors for ATM connections in broadband ISDN. *Proceedings ITC 14* (eds) J Labetoulle, J W Roberts (New York: Elsevier Science)
- Boyer P, Guillemin F, Servel M, Coudreuse J P 1992 Spacing cells protects and enhances utilization of ATM network links. *IEEE Networks*
- CNET 1991 Special issue on ATM. *Echo des Recherches* (English issue) France Telecom CNET
- Cruz R L 1991a A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Info. Theory* 35: 114-131
- Cruz R L 1991b A calculus for network delay, part II: network analysis. *IEEE Trans. Info. Theory* 35: 131-141
- Di Vecinia G, Courcoubetis C, Walrand J 1994 Decoupling bandwidths in queues. Preprint, University of California, Berkeley
- Doshi B T 1993 Deterministic rule based traffic descriptors for a broadband ISDN: worst case behavior and connection acceptance control. *Proceedings of Globecom '93* Houston, paper 48.A.1
- Elwalid A, Mitra D 1993 Effective bandwidths of general Markovian sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* 1: 329-343

- Gravey A 1994 Cell conformance and quality of service guarantees in ATM networks. *14th International Teletraffic Congress* Antibes (France)
- Gravey A, Hébuterne G 1991 Mixing time and loss priorities in a single server queue. *13th International Teletraffic Congress* Stockholm
- Gravey A, Hébuterne G 1994 Cell loss ratio commitments in ATM networks (to be presented in ICC'94)
- Gravey A, Boyer P, Hébuterne G 1991 Tagging versus strict rate enforcement in ATM networks. *Globecom'91* Phoenix
- Guillemin F, Mazumdar R 1995 On excursions of the workload process in $G|GI|1$ queues. *Stochastic processes and their applications* (to appear)
- Guillemin F, Levert C, Rosenberg C 1994a Cell conformance testing with respect to the peak cell rate in ATM networks. *Comput. Networks ISDN Syst.* (to appear)
- Guillemin F, Rosenberg C, Mignault J 1994b Guidelines for sustainable cell rate specification in ATM networks. *Comput. Networks ISDN Syst.* (submitted)
- Guillemin F, Mazumdar R, Simonian A 1995 On heavy traffic approximations for transient characteristics of $M|M|\infty$ queues. *J. Appl. Probab.* (to appear)
- Hui J Y 1988 Resource allocation in broadband networks. *IEEE Selected Areas Commun.* JSAC-6: 1598-1608
- I.121 1989 ITU-T Recommendation I.121: *Broadband aspects of ISDN*. Blue Book, vol. III.7, Geneva
- I.356 1992 ITU-T Recommendation I.356: *B-ISDN ATM layer cell transfer performance*, Geneva
- I.371 1993 ITU-T Draft Recommendation I.371: *Traffic control and congestion control in B-ISDN*. Geneva
- Iscoe I, McDonald D, Qian K 1993 Capacity of ATM switches. *Ann. Appl. Probab.* 3: 277-295
- Kelly F P 1991a Loss networks. *Ann. Appl. Probab.* 1: 319-378
- Kelly F P 1991b Effective bandwidths at multi-class queues. *Queueing Syst.* 9: 5-15
- Kroner H, Boyer P, Gravey A, Hébuterne G 1991 Priority management in ATM switching. *IEEE J. Selected Areas Commun.* : 418-427
- Labetoulle J, Roberts J W (eds) 1994 *Proceedings of the International Teletraffic Congress, ITC 14* (New York: Elsevier Science)
- Leland W E, Taqqu M S, Willinger W, Wilson W S 1994 On the self-similar nature of ethernet traffic. *IEEE/ACM Trans. Networking*

- Makrucki B A 1991 On the performance of submitting excess traffic to ATM networks. *IEEE Globecom'91* Phoenix
- Minzer S E 1989 Broadband ISDN and asynchronous transfer mode (ATM). *IEEE Commun. Mag.*
- Newman P 1991 Backward explicit congestion notification for ATM local area networks. *IEEE Globecom'93* Houston
- Norros I 1993 Studies on a model for connectionless traffic based on fractional Brownian motion. *Presented at the Conf. on Appl. Probab.* Paris (to appear in *Queueing Syst.*)
- Onno G, Lemonier M, Kerbérenes G, Gonzales J 1993 Services and applications in the Brehat Project. *ISSLS'93* Vancouver
- Parekh A, Gallager R 1993 A generalized processor sharing approach to flow control intergrated services networks – the single node case. In *Proc. Infocom'92* Florence. Also 1993 in *IEEE/ACM Trans. Networking* 1
- Parekh A, Gallager R 1994 A generalized processor sharing approach to flow control intergrated services networks – the multiple node case. *IEEE/ACM Trans. Networking* 2: 137-150
- Rathgeb E 1991 Modeling and performance analysis of policing mechanisms for ATM networks. *IEEE J. Selected Areas Commun.* 9:
- Roberts J W (ed.) 1991 COST 224, Performance evaluation and design of multi-service networks. Final report, Commission of the European Communities
- Roberts J W, Guillemin F 1992 Jitter in ATM networks and its impact on peak cell rate control. *Perform. Eval.* Special Issue on Modelling of High Speed Telecommunications Systems
- Rosenberg C, Hébuterne G 1994 Dimensioning traffic control devices in an ATM network. *Proceedings Conference on Telecommunication Systems* Nashville
- Rosenberg C, Lague B 1994 A heuristic framework for source policing in ATM networks. *IEEE/ACM Trans. Networking* (to appear)
- Rosenberg C, Guillemin F, Mazumdar R 1993 On quantile measures for traffic characterization in B-ISDN. *Proc. ITC Specialists Seminar* Bangalore
- Theberge F, Mazumdar R 1994 Approximations for loss probabilities in large multi-rate systems. Preprint INRS-Telecommunications, (also submitted for publication)
- Theberge F, Mazumdar R 1995 Multi-rate loss systems with bandwidth reservation: Algorithms and analysis. *IEEE Selected Areas Commun.* Special Issue on Foundations of Networking (to appear)
- Tranchier D, Boyer P, Rouaud Y, Mazeas J Y 1992 Fast bandwidth allocation in ATM networks. *ISS'92* Yokohama City

Turner J 1986 New directions in communications (or which way in the information age ?). *Proceedings of the Zurich Seminar on Digital Communications* Zurich, pp 25-32

Whitt W 1995 Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommun. Syst.* (to appear)