Input Rate Regulation and Bandwidth Assignment in ATM Networks: an Integrated Approach

G. Rigolio  Italtel - Central R&D Labs.
Settimo Milanese - Italy

L. Fratta  Politecnico di Milano - Dip. di Elettronica
Milan - Italy

Abstract

In order to better exploit resource utilization in ATM network, statistical multiplexing has been widely proposed; this has the drawback that some flow control must be introduced to provide fairness among users, to protect the network and the honest users from any kind of misemploy. The most common approach is to use some policing functions. Unfortunately almost all the known policing mechanisms so far proposed have shown some intrinsic limitations in the ability to ensure that the agreed call parameters are respected. To overcome these limitations a traffic shaper, which contains both policing and shaping functions, is proposed in the paper. The performance obtained show the effectiveness of this approach which provides a significant statistical multiplexing gain even in the case of sources not respecting the declared connection parameters.

1. Introduction

The Asynchronous Transfer Mode (ATM) has been widely recognized to be a promising technique for implementing an integrated access in high speed transport networks that can easily be shared by end users holding multi rate service calls [1,2,3]. It is in fact accepted that packet switching networks are better suited than circuit switching ones when services with dynamically varying bandwidth requirements must be supported.

The flexibility of packet switching has the drawback of introducing issues like bandwidth allocation and congestion control which must be solved to take full advantage of the statistical multiplexing and to enhance network efficiency. These problems arose since the beginning of the use of the new switching technique and many solutions have been given for conventional packet switched networks [4]. The basic idea common to these existing techniques is to provide virtual circuits to each call thus guaranteeing the required network resources.

In fast packet switching, the high speed operation of the links makes the use of these conventional procedures inappropriate. In fact at any instant it is possible to have thousands of packets in transit across a link and using existing congestion methods an empty buffer must be available in the receiving node for all packets. This leads either to unreasonable buffer requirements or to unacceptable low link utilization.

The new approach in ATM is to transmit packets without preallocated resources and accept a loss of packets when buffers are full. In order to keep the frequency of packet loss small, for each new call the network must select a path that has sufficient bandwidth available to support the connection with an acceptable packet loss rate and to guarantee the required Grade of Service (GoS). A new call is refused if the needed bandwidth is not available.

The bandwidth needed is not only a function of the call parameters, such as peak bit rate, average bit rate and burstiness, but it also depends upon the statistics of the traffic currently multiplexed on the network resources.

Once a traffic model for each class of services has been assumed the bandwidth allocation problem has been solved either by simulation [5,6] or by approximate analytical methods [7].

Many results, available in the literature, show the effectiveness of the statistical multiplexing both among sources of the same class and of different classes.

Unfortunately, the gain in network efficiency does not come for free. In fact once a call has been accepted, the required performance are only statistically guaranteed and some flow control must be operated to provide fairness among the users when congestion situations occur. Such a control may operate at the multiplexing level, but as it is based on the declared call parameters, some actions must be taken at the users interfaces in order to enforce the requested parameters.
Otherwise a user that generates a traffic with parameters value above those declared at the call set up (dishonest user) may force the honest users to suffer a degradation in the GoS. A policing procedure is therefore needed to assure that any change in the user's behaviour does not affect the networks performance as seen by the other users. Furthermore a perfect policing procedure should have no effect on honest users.

The performance analysis of the existing policing mechanisms, such as Leaky Bucket, sliding window and others [5,8] has shown some intrinsic limitations in the ability to ensure that the agreed call parameters are respected.

For instance to police the source average input rate theoretically it would require an infinite time and in practice a trade off must be set between response time and measure accuracy.

It has been observed that in order to detect malicious behaviours the policer parameters must be set to values that cause also rejection at the network access of packets generated by well behaving sources. The limitations observed do not depend on the specific implementation of the policing procedure but are intrinsic in the overall approach.

Instead of solving the bandwidth assignment problem on declared source parameters and then enforce these parameters, the new approach discussed in the paper consists on guaranteeing traffic parameters through a shaper device and then optimizing the bandwidth assignment for a required GoS. This integrated approach seems to be more significant as the traffic characteristics are related to the physical parameters of the shaper that can be set by the user at the call set up. The grade of service is obtained through a suitable bandwidth assignment based on the parameters guaranteed by the shaper, rather than on that enforced by the policer, affected, as exposed above, by a great inaccuracy.

In order to analyze and compute the performance, in section 2 the traffic model usually proposed in the literature is reviewed. In section 3 the integrated approach is described together with the shaper function definition and performance analysis. In section 4 the numerical results on the bandwidth assignment obtained by approximate analysis are presented and compared to the case of no traffic shaping. Conclusions and directions for further works are given in section 5.

2. Source model

A source model, widely used in the literature, which is able to account for a large variety of services is the bursty model. A source is said to be bursty if it interleaves active periods, in which it transmits at the peak bit rate, with idle periods, in which it remains silent. A bursty source can be characterized by the following parameters:

- **Bp**: peak bit rate, i.e. the maximum transmission bit rate of the source.
- **Bm**: mean bit rate, i.e. the mean transmission bit rate of the source averaged over the duration of a call.
- **T**: mean peak duration, i.e. the mean duration of the time interval during which the traffic source transmits at the peak rate.

The burstiness is defined as the peak to mean ratio (i.e. \( b = \frac{B_p}{B_m} \)).

If \( B_p = B_m \), \( T \) is equal to the call duration and the source is of stream type.

For most of the new broadband services as packet voice, coded with silence suppression, imaging services, document retrieval and data on demand, sources with a high level of burstiness must be considered.

Both active and silence periods are assumed to be exponentially distributed with average \( T \) and \( T(b-1) \), respectively.

As standard ATM cells have a fixed size \( n_{\text{cell}} = 424 \) bits (48 data bytes + 5 header bytes) the burst average length (L) measured in cells is given by \( L = \frac{T}{n_{\text{cell}}/B_p} \).

All the results presented in the paper have been obtained assuming source traffic with \( B_p = 10 \) Mbit/s, \( L = 100 \) cells and \( b \) equal to 5 or 10.

3. Shaper performance

The shaper considered in our analysis is mainly composed of a server operating at the bit rate \( B_s \) (shaping device) and a "sliding window" mechanism that allows the transmission of at most \( m \) cells on a time interval of \( D \) slots (policing device) (see Fig. 1).

A bursty source can be characterized by the following parameters:

- **Bp**: peak bit rate, i.e. the maximum transmission bit rate of the source.
- **Bm**: mean bit rate, i.e. the mean transmission bit rate of the source averaged over the duration of a call.
- **T**: mean peak duration, i.e. the mean duration of the time interval during which the traffic source transmits at the peak rate.

The burstiness is defined as the peak to mean ratio (i.e. \( b = \frac{B_p}{B_m} \)).

If \( B_p = B_m \), \( T \) is equal to the call duration and the source is of stream type.

For most of the new broadband services as packet voice, coded with silence suppression, imaging services, document retrieval and data on demand, sources with a high level of burstiness must be considered.

Both active and silence periods are assumed to be exponentially distributed with average \( T \) and \( T(b-1) \), respectively.

As standard ATM cells have a fixed size \( n_{\text{cell}} = 424 \) bits (48 data bytes + 5 header bytes) the burst average length (L) measured in cells is given by \( L = \frac{T}{n_{\text{cell}}/B_p} \).

All the results presented in the paper have been obtained assuming source traffic with \( B_p = 10 \) Mbit/s, \( L = 100 \) cells and \( b \) equal to 5 or 10.

3. Shaper performance

The shaper considered in our analysis is mainly composed of a server operating at the bit rate \( B_s \) (shaping device) and a "sliding window" mechanism that allows the transmission of at most \( m \) cells on a time interval of \( D \) slots (policing device) (see Fig. 1).

Cells generated by the source enter the shaper through the sliding window and join the shaping device, i.e. a queue able to contain up to \( h \) cells and a transmitter with the bit rate \( B_s \).

The output traffic characteristics depend on both input traffic and shaper parameters. More specifically, after the sliding window the mean bit rate is equal to \( \min(B_p \cdot m/D : B_m) \) and the maximum burst length is limited to the value \( m \), while after the shaping device the peak bit rate is equal to \( \min(B_p,B_s) \).
In order to obtain a traffic shaping with no cell loss some constraints exist on the choice of the shaper parameters. In particular \( m/D \) must be chosen so that \( B_p \cdot m/D \geq B_m \). This guarantees that all the offered traffic can be transmitted. Moreover the queue in the shaper, when \( B_s \) is smaller than \( B_p \), may grow too much and cells could be discarded due to the finite buffer size. In order to investigate the performance of the shaper a simulation campaign has been run assuming \( h = 900 \) cells, i.e. equal to 9 times the mean burst length of the offered traffic. The shaping device mean queue length as function of \( B_m/B_s \) is shown in Fig. 2, for offered traffics of burstiness 5 and 10.

![Fig. 2 - Mean queue length for different value of burstiness vs. \( B_m/B_s \)](image)

In order to keep the cell delay relatively low, which also guarantees a negligible buffer overflow probability, the ratio \( B_m/B_s \) must be smaller than 0.6. In fact, in this range of values, it has been observed a buffer overflow probability of two orders of magnitude smaller than the cell loss probability introduced by the sliding window mechanism.

The role of the sliding window mechanism is that to police the traffic in order to limit the non well-behaving sources. Moreover this function must be obtained without affecting the well behaving sources. It is therefore important to set the parameters \( m \) and \( D \) so that the cell loss probability suffered by well behaving sources is negligible. The trend of cell loss probability when \( m/D \) changes has been investigated by simulation. The results obtained are therefore limited to values (order of \( 10^{-5} \)) that can be measured in a reasonable simulation time.

As an example the behaviour of cell loss probability versus \( (m/D) \cdot B_p/B_m \) is shown in Fig. 3 for a window size \( D = 20000 \) slots. Note that \( (m/D) \cdot B_p \) represents the capacity (maximum throughput) of the shaper. To guarantee the required cell loss probability to well-behaving sources this capacity must be at least twice the average offered traffic. This means, unfortunately, that the increased offered traffic, generated by malicious sources, may force into the network a traffic with an average value which is twice the declared one. As it is the worst case traffic that can enter the network, it must be considered in the bandwidth assignment to guarantee that the well behaving sources performance are not degraded. Moreover the combination of shaping and policing functions limits the peak bit rate to \( B_s \) and the maximum burst length to \( m \).

![Fig. 3 - Cell loss probability vs. sliding window parameters](image)

It is worthy noting that a malicious user may theoretically force the input traffic up to twice its declared average value, but it is obtained at the cost of an high rate of cell loss, whose values are given by the curve in Fig. 3. For instance if a source doubles its average offered traffic without changing the shaper parameters, it observes a 50% cell loss which will vanish any attempt to sneak in extra information and reduce the grade of service to an unacceptable level.

This direct "automatic" feedback to the source has a discouraging effect against cheating in the traffic parameter declaration. A further function of the shaper is that of providing the user a measure of his traffic parameters. More specifically, the user, by measuring cell loss and delay, may choose the appropriate \( B_s \) that satisfies his requirements and characterizes his traffic at the network interface.

To realize the shaper two parts must be implemented, the queue with the variable server...
and the sliding window. The first one is obtained through a buffer of size \( h \), read with frequency \( B_s/n_{cell} \) according to an usual FIFO policy. For the implementation of the second part a shift register of \( D \) bits can be used. The register is shifted of one position at each cell time; the current bit is turned "on" if there is a new cell, otherwise the bit is forced "off". A counter stores the number of the "on" bits of the shift register (note that the update procedure of the shaper simply adds the new bit and subtract the overflowing one at each cell time); the cell is then actually transmitted if the current value of the counter is lower than \( m \), otherwise it is rejected.

4. Bandwidth assignment

The goal of this section is twofold. First we evaluate the gain obtained by multiplexing shaped sources with respect to the case of no shaping. Second we compute the bandwidth assignment needed to guarantee a required GoS when the shaper described in the previous section is used and malicious sources are assumed (worst case).

For both purposes the bandwidth needed to transmit with a required GoS the traffic obtained by multiplexing \( N \) sources must be evaluated. This computation is done by using the Uniform Arrival and Services (UAS), also referred as Fluid Flow model, for finite buffer size, first described in [9] and extensively applied in [7]. The method computes the cell loss probability for a given bandwidth allocation and a given number of identical bursty sources. This calculation must be iterated in an algorithm searching for the appropriate value of the bandwidth using a logarithmic interpolation method. The advantage of the analytic method on the simulation approach is that very low probabilities can be computed so that the usually required values for ATM (\( 10^{-9} \), \( 10^{-10} \)) can be considered.

The results obtained for cell loss probability equal to \( 10^{-10} \) and source traffic burstiness 5 and 10 are shown in Fig. 4 and 5, respectively. In this case all sources are assumed to well behave, i.e. to generate their traffic according to the declared parameters. The behaviour for different values of \( B_p/B_s \) are shown and compared with the case of no shaping (dotted line). The gain in bandwidth increases with \( B_p/B_s \), but one should not forget that it is obtained at the cost of accepting a delay which also increases as shown in section 3.

The performance shown in Fig.4 and 5 correspond also to those obtained if an ideal policer, which constrains source traffic to nominal values, is supposed. These represent the upper bounds to the performance of the shaper approach.

The lower bounds are obtained by solving the bandwidth assignment in the worst traffic case which is obtained assuming that each of the \( N \) multiplexed sources generate a continuous traffic at the peak bit rate.

This offered traffic after being shaped as in Fig. 1 becomes of bursty type with average \( (m/D) \cdot B_s \), constant burst length equal to \( m \) and burstiness \( b = D/m \).

The bandwidth needed to guarantee the required GoS versus the number of such sources is given in Fig. 6 and 7 for different values of \( B_p/B_s \) and for \( D \) equal to 20000 and \( m \) equal to 8000 and 4000, respectively. This corresponds to values of burstiness after the shaper/policer of 2.5 and 5 respectively, which are half the nominal values declared by the sources of the cases analyzed.

To summarize and compare the results previously presented, the bounds to bandwidth assignment in the case with and without traffic shaping for \( B_p/B_s = 2 \) and a source of traffic with burstiness 10 are shown in Fig. 8.

Curve a) corresponds to the case of traffic sources which well behave either by nature or because an
ideal policer is assumed. Still assuming well behaving sources the use of the shaping function allows the bandwidth assignment represented by curve b). The gain with respect to the case of no shaping function (curve a) is remarkable as the number of sources that can be multiplexed is almost doubled.

To guarantee a required GoS when sources are as malicious as possible, the peak bit rate assignment, represented by curve c), must be chosen if no control exists. The use of a conventional policer, such as the sliding window mechanism, allows a very little saving in bandwidth as represented in curve d), thus showing the ineffectiveness of this approach.

On the contrary, the combination of shaping and sliding window policer, as proposed in the shaper described in section 3, provides the remarkable improvement represented by curve e).

Note that curves d) and e) represents an upper bound to bandwidth assignment as they guarantee the performance also in the very unfortunate conditions when all sources at the same time force the traffic to the maximum value.

The gaps between the bounds in Fig. 8 are quite large and justify the use of any policing mechanism more effective than that considered here in order to reach performance near to the lower bounds. However, the most significant improvement is obtained through the shaping function which should be carefully considered for the introduction in the ATM environment.

5. Conclusions

Differently from the existing approaches to exploit statistical multiplexing in ATM, the one proposed in the paper adds shaping functions on the traffic sources. The traffic shaper at the cost of cell loss and delay, that are design parameters for each service, guarantees traffic parameters on which statistical bandwidth allocation can be done. The analysis carried out in the paper has shown that, no matter to the behaviour of the sources that may generate traffic not according to the declared parameters, a significant advantage is still provided by statistically multiplexing with a guaranteed grade of service.

The bounds of the performance show that further improvements could be obtained by using more effective policing functions, and that the shaping function significantly increases the overall network efficiency.

References


