Queueing Analysis of Traffic Shaping and Scheduling Applied In The Source ATM End Point

Tien V. Do, László Jereb, György Wolfner
Department of Telecommunications, Technical University of Budapest
Sztoczek 2, 1111 Budapest – Hungary
Email: {do,jereb,wolfner}@hit.bme.hu
Fax : (+36)-1-463-3266 Phone : (+36)-1-463-2096

Abstract

This paper deals with traffic shaping performed at the ATM Adaptation Layer (AAL). First, the application of a shaping scheme based on the Generic Cell Rate Algorithm is discussed. Later on, a mathematical model for the traffic shaping scheme at the AAL using message mode is introduced. Finally, with the use of the mathematical model some results related to the performance of the shaping scheme applied in workstations running a file transfer application are presented.

Keywords

ATM Networks, traffic control, traffic shaping, discrete-time queueing analysis, iterative algorithm

1 INTRODUCTION

In telecommunications networks congestion occurs when the demand exceeds the available resources and as a consequence it causes a performance degradation for services provided by the networks. In fact, traffic congestion is an undesirable side-effect of the interaction between traffic flows of random and dynamic nature due to the activities of the subscribers and the network architecture (topology, link speed, buffer). In ATM networks, the probability of congestion increases, since it is required to carry the traffic of services of a great variety from both LAN and WAN environments (as an example, a multimedia connection may simultaneously require the transfer capability for voice, moving picture and data).

The primary task of traffic control actions is to reduce the risk of congestion in order to protect the network and the users. In general traffic control actions belong either to preventive or reactive categories. Preventive actions try to avoid congestion by taking appropriate actions before occurrence thereof. Reactive congestion control means that the network takes actions only in case congestion occurs.

This paper deals with traffic shaping – a preventive traffic control scheme – in a source ATM end point. The motivations for traffic shaping can be explained as follows. At the source end point the higher-layer protocols or native ATM API (Application Program Interface) can pass data to the AAL for transmission across ATM networks in two ways: message mode and stream mode (ITU-T I.362 1992). In the stream mode, the AAL sends cells as soon as payloads are filled. In the message mode, the AAL accepts an entire packet from a higher-layer protocol and segments it into payloads before forwarding any cells into the networks. In the message mode, large packets used by some applications (e.g. bulk data transfer) may introduce a cell flow of "uncontrolled" large bursts into the network, which may increase the cell loss probability. For an example, the case of a file transfer application can be mentioned where the host of the file server application is connected by a high speed link while the client host is connected by a low speed link to the ATM network. In this case if the submission of cells is not controlled properly severe cell loss can happen.

The second reason behind traffic shaping is the policing function enforced by the ATM network at its entry. The policing function will discard cells if the characteristics of a cell flow are not conforming to the values specified during the connection set-up phase. Until now, only a leaky bucket like Generic Cell Rate Algorithm (GCRA) has been proposed by the standardization bodies (ITU-T, the ATM Forum) for policing the peak and/or sustainable cell rate at the network entrance UNI (User-Network Interface) point.

The use of a GCRA-based traffic shaping in a CPE (Customer Premises Equipment) was proposed (Ajmone *et al.* 1995b) to ensure that the traffic generated by the source is conforming to the Connection Traffic Descriptor and associated parameter values that were negotiated with the public network. Some advantages of a GCRA-based traffic shaping scheme can be emphasized that it increases the efficiency of statistical multiplexing, decreases the cell discarding probability of the submitted cell stream in the network and since only complete data packets are sent, it reduces the need for retransmission (Ajmone *et al.* 1995a, Ajmone *et al.* 1995b). Futhermore, the benefit of the GCRA-based traffic shaping on the throughput of TCP in ATM networks is studied in the paper of (Ajmone *et al.* 1996).

The paper presents a discrete-time queueing model for the GCRA-based traffic shaping scheme at the source end point. The iterative analysis approach is similar to the one presented in (Tran-Gia *et al.* 1988) for a model of a multiplexer in packet switching networks. Recently, after the course of this work the authors have learned that Patel and Biskidian (Patel *et al.* 1996) followed similar approach from technical point of view, but their performance analysis was performed by simulation.

The rest of the paper is organized as follows. In the next two sections, the GCRA-based traffic shaping algorithm and traffic shaper are described. Section 4 presents a discrete-time queueing model for the GCRA-based traffic shaping scheme. Based on a mathematical analysis the performance measures of the GCRA-based traffic shaping algorithm can be derived. Then, we illustrate the application of the shaping algorithm with some performance results of the shaping scheme and finally the paper is concluded.

2 TRAFFIC SHAPER IN THE SOURCE END POINT

The architecture of the traffic shaper proposed in (Ajmone *et al.* 1995a) is depicted in Figure 1. Messages from higher layer or the applications are directed to the segmentation process in the AAL layer. The cells whose allowed transmission time does not expire are forced to wait in the buffer before submission to the ATM layer. Messages finding no room in the shaping buffer at the arrival instant for all of their cells are lost.

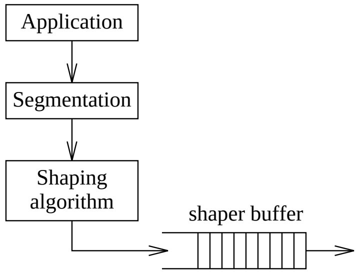

Figure 1 Traffic Shaper

3 TRAFFIC SHAPING ALGORITHM

A GCRA based shaping algorithm depicted in Figure 2 has two key parameters: the first one, T denotes the time increment of the algorithm, while the second one, τ defines the allowed burst length. In the original Generic Cell Rate Algorithm $1/T$ is the peak cell rate, τ denotes CDV (Cell Delay Variation) tolerance.

Cell C_i is tagged with the allowed transmission time Tt_i. The allowed transmission time of a generic cell C_i is the time at which C_i is eligible for transmission and it is calculated given the theoretical arrival time TAT_i and the actual arrival time Ta_i of cell C_i (Ajmone *et al.* 1995a), where TAT_i is defined as the reference point for the algorithm to control the scheduling of cell C_i.

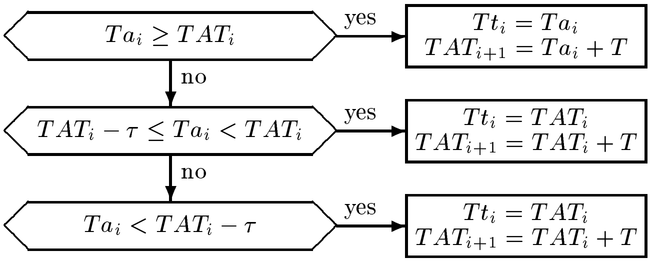

Figure 2 Shaping algorithm

From this description one can easily conclude that the allowed peak cell rate is $\frac{1}{T}$.

4 MATHEMATICAL ANALYSIS

In this Section an algorithm is derived for determining the performance measures of the shaper such as the message loss probability, as well as the cell and message sojourn times in the buffer. The time is discretized into slots of a cell duration, which corresponds with ATM environments. The time unit is exactly the length of cell duration. The parameters of the shaper and message interarrival times are scaled according to the time unit.

4.1 Basic notations

Let us introduce the following notations:

- B: the size of the shaping buffer, in fact it is the buffer found in the network interface cards which is used for buffering cells after the segmentation,
- $Ta^{(n)}$: the arrival time of message n,
- $A_n = Ta^{(n)} - Ta^{(n-1)}$: the interarrival time between the $(n-1)$th and nth message,
- $a_n(k) = \Pr\{A_n = k\}$: the distribution of the discrete-valued random variable A_n. Without loss of generality suppose that $a_n(k) = a(k) \quad \forall n$ which means that the interarrival times A_n between messages are identically distributed,
- X_n: the batch size (number of cells) of message n,
- $x_n(m) = \Pr\{X_n = m\} = x(m) \quad \forall n$: the distribution of the discrete-valued random variable X_n. Similarly to the interarrival time, the message lengths are also supposed to be identically distributed. In this paper only such cases are considered, where the length of messages are upper bounded by the shaping buffer size ($X_n \leq B$),

- $Z(t)$: the difference between the value of TAT and the current time t in the shaper:
$$Z(t) = \max(TAT(t) - t, 0).$$
- Z_n^-: the value of $Z(t)$ in slots just before the arrival instant of message n:
$$Z_n^- = \max(TAT^{(BOM_n)} - Ta^{(n)}, 0),$$
where $TAT^{(BOM_n)}$ is the theoretical arrival time of the BOM (Begin Of Message) cell of message n,
- Z_n^+: the value of $Z(t)$ in slots just after the arrival instant of message n:
$$Z_n^+ = \max(TAT^{(BOM_{n+1})} - Ta^{(n)}, 0),$$
where $TAT^{(BOM_{n+1})}$ is the value of TAT at $Ta^{(n)}$ after performing the shaping algorithm on the EOM (End Of Message) cell of message n, i.e. $TAT^{(BOM_{n+1})}$ is the theoretical arrival time of the BOM cell of the message following message n,
- $C(t)$: the number of cells in the buffer at time t,
- $L(t)$: the number of cells in the buffer, whose allowed transmission time is less than t, or equal to t,
- $W(t)$ the number of cells in the buffer, whose allowed transmission time is greater than t,
- C_n^+, C_n^-, L_n^+, L_n^-, W_n^+, and W_n^-: the values of $C(t)$, $L(t)$, and $W(t)$ just before and after the arrival instant of message n, respectively.

According to the definitions stated above, $Z(t)$ decreases by one in each slot between two arrival instants of two arbitrary messages until it reaches zero. Consequently, the following relation can be easily set up between Z_{n+1}^- and Z_n^+:

$$Z_{n+1}^- = \max(Z_n^+ - A_{n+1}, 0). \tag{1}$$

On the basis of the definition of $C(t)$, $L(t)$, and $W(t)$, the following equality can be also obtained:

$$C(t) = W(t) + L(t). \tag{2}$$

If there are some cells waiting in the buffer at time t, the number of slots to the TAT of the EOM cell of the latest message accepted before t can be expressed as $Z(t) - T$ and the number of slots to the allowed transmission time is $Z(t) - T - \tau$. Provided that at slot t there are $W(t) = k$ cells waiting in the buffer whose allowed transmission time has not expired, the number of slots to the allowed transmission time of the first cell of them in the buffer can be derived as $Z(t) - k \cdot T - \tau$ and the following inequality holds:

$$0 \le Z(t) - k \cdot T - \tau < T, \tag{3}$$

and therefore $W(t)$ can be expressed as:

$$W(t) = \left\lfloor \frac{\max(Z(t) - \tau, 0)}{T} \right\rfloor \tag{4}$$

where $\lfloor r \rfloor$ denotes the largest integer number which is not larger than r.

In case $\tau < T$, the allowed transmission time of maximum one cell can expire at any slot, thus it can be forwarded immediately and $\forall t: \ L(t) = 0$. In case $\tau \geq T$ there may be more than one cell waiting in the buffer. It may be caused either (i) by the messages arrived later than $TAT - \tau$ or (ii) by those cells whose allowed transmission time has expired but they are not allowed to enter the network because of the cells standing ahead due to (i) or (ii). In the sequel cases $\tau < T$ and $\tau \geq T$ will be distinguished.

4.2 Case A: $\tau < T$

(a) Basic equations

In this case the number of cells waiting in the buffer at time t can be directly obtained from $Z(t)$, thus, the state evolution of the shaper at the arrival instants can be described by the distribution of $Z(t)$ just before and after the arrival instant of message n. Let us introduce the following notations:

$$z_n^-(i) = \Pr\{Z_n^- = i\},$$
$$z_n^+(j) = \Pr\{Z_n^+ = j\}.$$

Using Equation (3) the evolution from Z_n^- to Z_n^+ can be expressed as follows:

$$Z_n^+ = \begin{cases} Z_n^- + X_n \cdot T & \text{if } Z_n^- + X_n \cdot T < (B+1) \cdot T + \tau, \\ Z_n^- & \text{if } Z_n^- + X_n \cdot T \geq (B+1) \cdot T + \tau. \end{cases} \tag{5}$$

Applying the rule of total probability, one can obtain $z_n^+(k)$ in the following form:

$$z_n^+(k) = \sum_{j=1}^{\lfloor \frac{k}{T} \rfloor} z_n^-(k - jT) \cdot x_n(j) + z_n^-(k) \sum_{j=b(k)+1}^{B} x_n(j), \tag{6}$$

where $k = 0, \cdots, (B+1) \cdot T + \tau - 1$, and

$$b(k) = \left\lfloor \frac{(B+1) \cdot T + \tau - 1 - k}{T} \right\rfloor$$

describes the number of free slots for cells in the buffer.

From (1) $z^-_{n+1}(k)$ can be expressed as:

$$z^-_{n+1}(k) = \pi_0(z^+_n(k) * a_{n+1}(-k)), \tag{7}$$

where $\pi_0(f(k))$ denotes the following operation on the distribution $f(k)$:

$$\pi_0(f(k)) = \begin{cases} 0 & k < 0, \\ \sum\limits_{i=-\infty}^{0} f(i) & k = 0, \\ f(k) & k > 0 \end{cases}$$

and * denotes the discrete convolution operation.

The equilibrium state distributions $z^-(k)$ and $z^+(k)$ can be obtained by using iteratively equations (6) and (7):

$$z^-(k) = \lim_{n\to\infty} z^-_n(k), \ and\, z^+(l) = \lim_{n\to\infty} z^+_n(l). \tag{8}$$

(b) Performance measures

The performance measures can be obtained by using the equilibrium distribution $z^-(k)$.

Taking into account the condition in (5), the message blocking probability expressed as follows:

$$\begin{aligned} P_{BM} = & \sum_{k=0}^{(B+1)\cdot T+\tau-1} z^-(k) \cdot \Pr\{\text{message blocked } |Z^- = k\} \\ = & \sum_{k=0}^{(B+1)\cdot T+\tau-1} z^-(k) \sum_{j=b(k)+1}^{B} x(j), \end{aligned} \tag{9}$$

The delay of the ith cell of an arriving message finding $Z(t) = k$ just before the arrival instant can be derived in the following way:

$$cd(i,k) = max(k + (i-1)\cdot T - \tau, 0). \tag{10}$$

Using this result the mean cell delay can be calculated as:

$$MCDT = \frac{\sum\limits_{k=0}^{(B+1)\cdot T+\tau-1} z^-(k) \sum\limits_{j=1}^{b(k)} x(j) \cdot \sum\limits_{i=1}^{j} cd(i,k)}{\sum\limits_{k=0}^{(B+1)\cdot T+\tau-1} z^-(k) \sum\limits_{j=1}^{b(k)} x(j)\cdot j}. \tag{11}$$

The mean message delay is the mean delay of the EOM cells of messages:

$$MMDT = \frac{\sum_{k=0}^{(B+1)\cdot T+\tau-1} z^{-}(k) \sum_{j=1}^{b(k)} x(j)\cdot cd(j,k)}{\sum_{k=0}^{(B+1)\cdot T+\tau-1} z^{-}(k) \sum_{j=1}^{b(k)} x(j)}, \tag{12}$$

Finally, the distribution of the EOM cell delay is written as:

$$\Pr\{EOM_d = h\} = \frac{\sum_{k=0}^{(B+1)\cdot T+\tau-1} z^{-}(k) \cdot \sum_{j=1}^{b(k)} x(j)\cdot \delta(cd(j,k)-h)}{\sum_{k=0}^{(B+1)\cdot T+\tau-1} z^{-}(k) \sum_{j=1}^{b(k)} x(j)}, \tag{13}$$

where

$$\delta(l) = \begin{cases} 1 & l = 0, \\ 0 & \text{otherwise}, \end{cases}$$

4.3 Case B: $\tau \geq T$

(a) Basic equations

As stated in 4.1, in this case the distribution of $Z(t)$ is not sufficient to describe the state of the shaper as there may be some cells with expiring time stamp waiting in the buffer. In order to take these cells into account the following two-dimensional random variable and distribution are introduced:

- $S(t) = (Z(t), C(t))$: a two-dimensional random variable,
- S_n^+ and S_n^-: the value of $S(t)$ just before and after the arrival instant of message n, respectively,
- $s_n^+(i,j)$ and $s_n^-(i,j)$: the distribution of S_n^+ and S_n^-:

$$s_n^+(i,j) = \Pr\{S_n^+ = (i,j)\},$$

$$s_n^-(i,j) = \Pr\{S_n^- = (i,j)\}.$$

Note that either (Z(t),C(t)) or (Z(t),L(t)) may describe the state of the shaper with buffer. The analysis is shown with the first state variable pair.

The relation between S_n^+ and S_n^- can be described as follows:

$$S_n^+ = \begin{cases} S_n^- + (X_n \cdot T, X_n) & \text{if } C_n^- + X_n \leq B, \\ S_n^- & \text{if } C_n^- + X_n > B, \end{cases} \tag{14}$$

and then the distribution $s_n^+(k,l)$ is given by

$$s_n^+(k,l) = \sum_{j=B-l+1}^{B} s_n^-(k,l) \cdot x(j) + \sum_{j=1}^{l} s_n^-(k - T \cdot j, l - j) \cdot x(j). \tag{15}$$

Concerning the number of cells waiting in the buffer with expiring time stamp the following expression can be derived:

$$L_{n+1}^- = max(L_n^+ - A_{n+1} + W_n^+ - W_{n+1}^-, 0), \tag{16}$$

and (16) follows

$$C_{n+1}^- = max(C_n^+ - A_{n+1}, W_{n+1}^-). \tag{17}$$

Based on (17), and (1) S_{n+1}^- is expressed as:

$$S_{n+1}^- = (max(Z_n^+ - A_{n+1}, 0), max(C_n^+ - A_{n+1}, W_{n+1}^-)). \tag{18}$$

Using equation (18) the distribution of the evolution from S_n^+ to S_{n+1}^- is expressed as:

$$s_{n+1,1}^-(i,j) = \pi_{0,0} \left(\sum_{k=-\infty}^{\infty} s_n^+(i+k, j+k) \cdot a(k) \right), \tag{19}$$

where $\pi_{0,0}$ denote the following operation on the distribution function $f(i,j)$ of a two-dimensional random variable

$$\pi_{0,0}(f(i,j)) = \begin{cases} 0 & i < 0 \; or \; j < 0 \\ \sum_{k=-\infty}^{0} \sum_{l=-\infty}^{0} f(k,l) & i = 0 \; \& \; j = 0 \\ \sum_{k=-\infty}^{0} f(k,j) & i = 0 \; \& \; j > 0, \\ \sum_{l=-\infty}^{w(i)} f(i,l) & i > 0 \; \& \; j = w(i), \\ 0 & i > 0 \; \& \; j < w(i) \\ f(i,j) & i > 0 \; \& \; j > w(i) \end{cases} \tag{20}$$

and

$$w(i) = \left\lfloor \frac{max(i-\tau,0)}{T} \right\rfloor .$$

Equation (19) can be rewritten in the form of the convolution of two two-dimensional discrete-valued random variables in order to use Fast Fourier Transform:

$$s^-_{n+1,1}(i,j) = \pi_{0,0}(s^+_n(i,j) * \mathbf{a}(-i,-j)), \tag{21}$$

where

$$\mathbf{a}(i,j) = \begin{cases} a(i) & \text{if } \; i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly to case $\tau < T$ the equilibrium distributions $s^+(i,j)$ and $s^-(i,j)$ can be obtained by using iteratively equations (15) and (19 or 21):

$$s^+(i,j) = \lim_{n\to\infty} s^+_n(i,j) \;\; and \;\; s^-(i,j) = \lim_{n\to\infty} s^-_n(i,j). \tag{22}$$

Furthermore, the buffer content distribution can be written as:

$$c^-_n(k) = \Pr(C^-_n = k) = \sum_{i=0}^{\infty} s^-_n(i,k), \text{ and } c^-(k) = \lim_{n\to\infty} c^-_n(k). \tag{23}$$

(b) Performance measures

The message blocking probability can be obtained as follows:

$$P_{BM} = \sum_{i=1}^{B} c^-(i) \sum_{j=B-i+1}^{B} x(j). \tag{24}$$

The delay of the lth cell of a message finding $S(t) = (i,j)$ at arrival can be given by

$$cd(i,j,l) = max(i + (l-1)T - \tau, 0) + max(j + l - max(i + (l-1)T - \tau, 0), 0)).$$

Using this result the mean cell delay can be written in the following form:

$$MCDT = \frac{\sum_{i=0}^{(B+1)\cdot T+\tau-1} \sum_{j=0}^{B-1} s^-(i,j) \sum_{k=1}^{B-j} x(k) \sum_{l=1}^{k} cd(i,j,l)}{\sum_{i=0}^{(B+1)\cdot T+\tau-1} \sum_{j=0}^{B-1} s^-(i,j) \sum_{k=1}^{B-j} x(k) \cdot k}, \tag{25}$$

and the mean sojourn time of messages is computed by

$$MMDT = \frac{\sum_{i=0}^{(B+1)\cdot T+\tau-1} \sum_{j=0}^{B-1} s^-(i,j) \sum_{k=1}^{B-j} x(k) \cdot cd(i,j,k)}{\sum_{j=0}^{B-1} \sum_{i=0}^{(B+1)\cdot T+\tau-1} s^-(i,j) \sum_{k=1}^{B-j} x(k)} \tag{26}$$

Moreover, the distribution of the EOM cell delay can be also obtained as

$$\Pr\{EOM_d = h\} = \frac{\sum_{i=0}^{(B+1)\cdot T+\tau-1} \sum_{j=0}^{B-1} s^-(i,j) \sum_{k=1}^{B-j} x(k) \cdot \delta(cd(i,j,k) - h)}{\sum_{i=0}^{(B+1)\cdot T+\tau-1} \sum_{j=0}^{B-1} s^-(i,j) \sum_{k=1}^{B-j} x(k)}. \tag{27}$$

5 APPLICATION SCENARIO

In this section we consider the configuration of a file transfer application running on workstations connected to the ATM network. The host of the file server application is connected by a high speed link of 150 Mbit/s while the host of the client is connected by a 25 Mbit/s link to the ATM network. Note that the assumption of the 25 Mbit/s client link can be relaxed, and for that case we can suppose that the server and client hosts are connected by the virtual channel (or path) connection of 25 Mbit/s through the ATM network.

In order to avoid the congestion in the 25 Mbit/s pipe traffic shaping is applied as a form of traffic control. Therefore, the allowed peak cell rate from the server is 25 Mbit/s, which follows that T is of 6 cell durations in a link of 150 Mbit/s.

The critical parameters of this application are the maximum size of messages the application is allowed to send to the AAL and the burstiness of the message arrival process. Concerning the maximum message size called Maximum Transfer Unit (MTU) two cases are taken into consideration.

In the first scenario Ethernet-based network parts are connected to the ATM network, and in order to avoid the fragmentation of IP datagrams the MTU size is set to be the MTU of Ethernet-based networks (1500 byte), therefore the maximum allowed message size is of 32 cells.

The second case is motivated by the default MTU value of 9180 bytes for IP over ATM environments (Atkinson 1994), and thus the MTU size of the second cases is of 192 cells*.

To model the message interarrival process a negative binomial distribution is used

$$Pr\{A_n = k\} = \left(\begin{array}{c} y+k-1 \\ k \end{array} \right) p^y (1-p)^k, \forall n \quad 0 \leq p < 1, y \text{ real} \tag{28}$$

where $p = \frac{1}{E(A) \cdot CoV^2}$, $y = \frac{E(A)}{E(A) \cdot CoV^2 - 1}$, $E(A) \cdot CoV^2 > 1$, for which the different choice of the Coefficient of Variation (CoV) results in various arrival processes and as a consequence, various burstiness degree of the message arrival processes. Note that the choice of $CoV = 1.0$ leads to the Bernoulli message interarrival process.

In the case of file transfer applications from the operation of the higher layer protocol (TCP/IP) it is quite reasonable to assume that the messages are sent to the AAL with the maximum allowed size.

*Note, that for example in workstations where Unix is used as an operating system and TCP/IP is for the network layer communication protocol the MTU - Maximum Transfer Unit- can be easily modified by `ifconfig` command.

6 NUMERICAL RESULTS

All the numerical results reported in this section are obtained with the choice of the average load of 10 Mbit/s, The choice of load is reflected by the fact that the typical LANs can generate average traffic volume in this range.

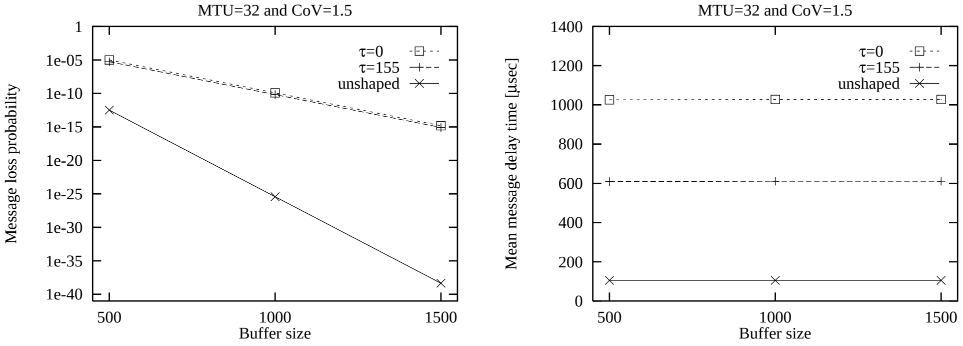

Figure 3 Message loss and mean message delay versus the shaping buffer size, MTU=32 cells

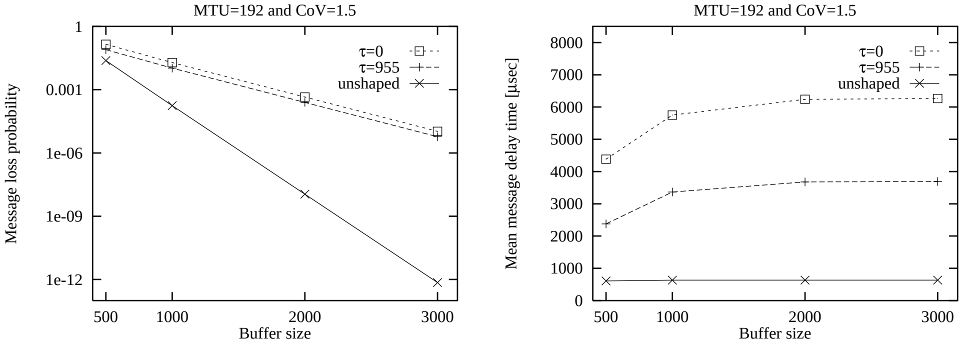

Figure 4 Message loss and mean message delay versus the shaping buffer size, MTU=192 cells

In Fig. 3 and 4 the performance parameters versus the buffer size are plotted. The selection of $\tau = 155$ and $\tau = 955$ (with $T = 6$) corresponds to the situations of the maximum allowed burst length is of 32 and 192 cells, respectively. Moreover, two other choices of the traffic shaping parameters form the bounds of the traffic shaper behaviours: the choice of $\tau = 0$ means the strictly spaced cells leaving the traffic shaper with the intercell time of 6 cells, while the "unshaped" curves ($T = 1$) represents the case when no traffic shaper is applied. The curves indicate the fact that the tail distribution of the buffer content is asymptotically exponential. The results show that if the buffer size is in the order of several thousand cells, the message loss probability due to delaying cells in the shaping buffer is very low (in the order of 10^{-15}). It can be observed that the maximum allowed message size has a strong impact on the necessary buffer size in order to sustain the low message loss probability. Therefore, in the sequel we will show the numerical results with the buffer size of 500 cells for the case of MTU=32 cells, and 2000 cells for the case of MTU=192 cells.

The impact of the message interarrival process on the shaping performance is demonstrated in Fig 5 and 6, where the performance parameters are plotted versus CoV. It can be observed that the higher burstiness of the arrival process (larger value of CoV) significantly increases the probability of the arriving messages finding no place in the shaping buffer. The performance degradation expressed by the message loss probability and average message delay can be in the range of several orders of magnitude.

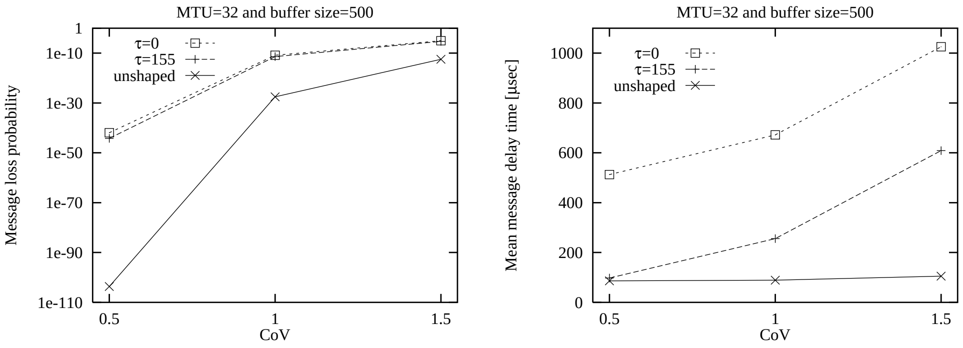

Figure 5 Message loss and mean message delay versus CoV, MTU=32

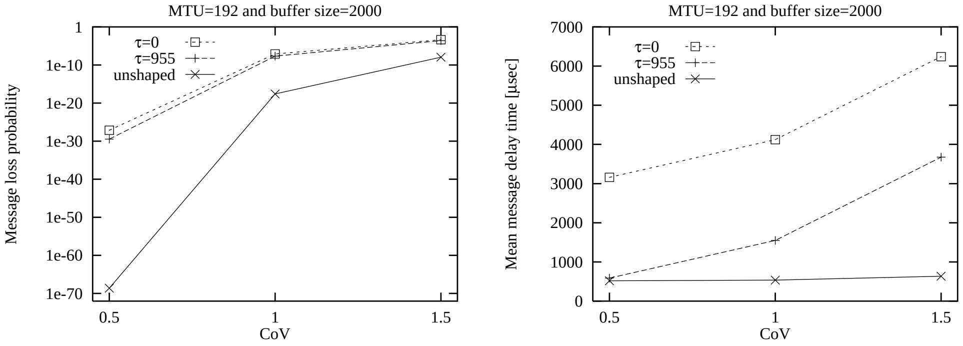

Figure 6 Message loss and mean message delay versus CoV, MTU=192

It can be also experienced from the figures that the modification of τ results in only small changes in the message loss probability but drastic changes in the average message delay.

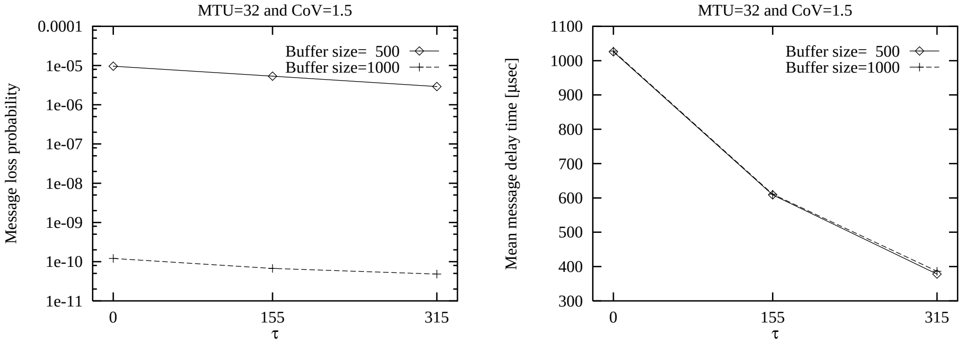

Figure 7 Message loss and mean message delay for different τ, MTU=32

The cumulative distribution function of the message delay is presented in Figure 8 which shows the dynamic of the message delay process. It can be observed that the presence of the traffic shaper decreases the probability of the arriving messages finding the buffer empty from 0.75 to 0.39. Moreover, 95% of messages departs from the shaper within 2.5 ms in the case of MTU=32 cells and 15 ms in the case of MTU=192 cells, respectively.

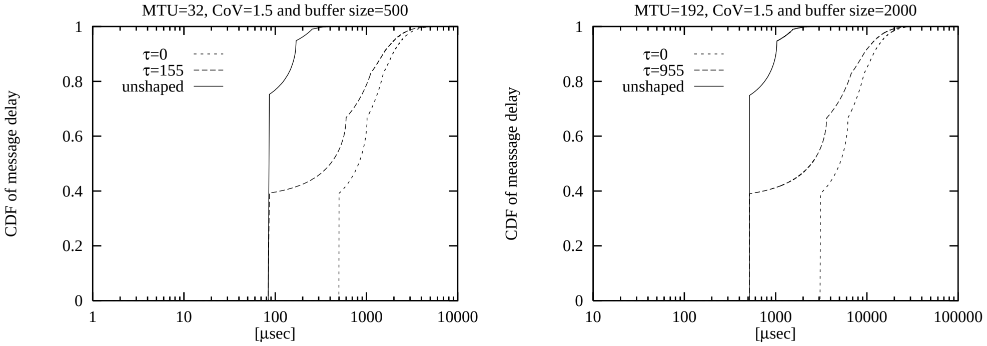

Figure 8 CDF of message delay

7 CONCLUSION

This paper considers the GCRA-based shaping algorithm in the source end point. The algorithm guarantees that the traffic submitted by the user is conforming to the negotiated traffic contract. In addition, a small modification of the network interface card is needed in order to implement the proposed shaping algorithm.

A discrete-time queueing model with general arrival processes is also presented for the GCRA-based shaping algorithm. The numerical results show the strong impact of the burstiness of the arrival process on the performance of the traffic shaper. Moreover, in the investigated case the additional delay introduced by shaping is in the order of several msec, which is quite low comparing with the delay recommendation ITU-T G.114 (For example, it states that less than 150ms is acceptable for most user applications).

In addition, the determination of the lower bound of cell/message queueing delay in the ATM end-point can be mentioned as a further contribution of this paper, which results can be obtained by applying the discrete-time queueing analysis in the case of $T = 1$ and $\tau = 0$.

ACKNOWLEDGEMENT

This work was performed with the support of the European Community through the COPERNICUS ATMIN 1463 project.

REFERENCES

M. Ajmone-Marsan, A. Bianco, R. Lo Cigno, and M. Munafó. (1996) Some simulation results about TCP connections in ATM networks. In D. D. Kouvatsos, editor, *Performance Modelling and Evaluation of ATM Networks.* Chapman and Hall, London (IFIP), 1996.

M. Ajmone-Marsan, A. Bianco, T. V. Do, L. Jereb, R. Lo Cigno, and M. Munafó. (1995a) ATM simulation with CLASS. in *Performance Evaluation*, 1995.

M. Ajmone-Marsan, T. V. Do, L. Jereb, R. Lo Cigno, R. Pasquali, and A. Tonietti. (1995b) Some simulation results on the performance of traffic shaping algorithms in ATM networks. In D. D. Kouvatsos, editor, *Performance Modelling and Evaluation of ATM Networks.* Chapman and Hall, London (IFIP), 1995.

R. Atkinson (1994): Default IP MTU for use over ATM AAL5, Naval Research Laboratory, Internet RFC 1626

ITU-T Recommendation I.362. *B-ISDN ATM Adaptation Layer (AAL) Functional Description*, 1992. Geneve, Switzerland.

B. Patel and C. C. Biskidian. (1996) End-Station Performance Under Leaky Bucket Traffic Shaping. *IEEE Network Magazine*, September-October 1996.

P. Tran-Gia and H. Ahmadi. Analysis of a discrete-time $G^{[x]}/D/1-S$ queueing system with application in packet-switching systems. In *IEEE INFOCOM'88,pp. 861-870*, New Orleans LA., 1988.