AN INTEGRATED DYNAMIC RESOURCE ALLOCATION SCHEME FOR ATM NETWORKS

R. Bolla, F. Danovaro, F. Davoli, M. Marchese

Department of Communications, Computer and Systems Science (DIST)
University of Genoa
Via Opera Pia, 11/A - 16145 Genova, Italy

ABSTRACT. *Dynamic bandwidth allocation among traffic classes with different performance requirements sharing an ATM link is considered as an integrated control problem with a multi-level structure. At the lower level, call admission control rules are applied that maintain a certain grade of service, in terms of cell loss probability and cell delay, given the buffer space and bandwidth assigned to each class; unlike those used in previous works, these rules are derived on the basis of homogeneous (based on similar quantities) measures of the performance requirements. At the higher level, bandwidth shares are periodically recomputed on-line by an allocation controller, whose goals reflect overall cell loss and refused traffic, as well as overall average delay. These goals are expressed by an optimization problem that is solved by numerical techniques. The whole control system should provide a dynamic feedback controller, capable of reacting in real time to changes in the traffic patterns. Simulation results are presented and discussed, regarding the efficiency of the admission controllers, the performance of the overall scheme, and the capability of reacting to sudden changes in the load of some traffic class.*

1. INTRODUCTION

Even though the Asynchronous Transfer Mode (ATM) has a great flexibility in handling a mix of different traffic types, with different quality of service constraints, and has the advantage of requiring a single switching fabric, nevertheless it raises a host of network management problems that would otherwise be less crucial in other, more "structured", transfer modes. This is due to the statistical nature of resource allocation in ATM, which may not guarantee the satisfaction of certain performance requirements, unless they are specifically enforced by means of real time control.

Congestion control and prevention is particularly relevant in this context, and has received a great deal of attention in the literature (see, for instance, [1] for a survey and [2] for recent papers on this matter). In particular, several works have addressed the issue of admission control ([3 - 13], among others), as a means of guaranteeing quality of service to the connections in progress.

Recently, a hierarchical decomposition of the control task has been proposed [12, 13] that separates the global admission control problem into independent subproblems, one for each traffic class, and periodically coordinates them by means of a capacity reassignment, based on on-line feedback information. We retain this general philosophy also in the present paper. Therefore, we consider an ATM link shared by several traffic and/or service classes, characterized by statistical parameters (like peak and average bandwidth), as well as by performance requirements (in terms of cell loss probability and cell delay). All traffic is supposed to take place on a connection basis. Each specific class is dedicated a call admission controller applying a fixed strategy, which is designed to maintain a certain grade of service, given the buffer space and bandwidth (percentage of cells) assigned to the class. Unlike the previous works [12, 13], where somehow different models and approximations have been used to construct the admission criteria accounting for cell loss probability and cell delay, respectively, the same quantiy will be used here as the basis for both criteria. The bandwidth shares are periodically recomputed on-line by a bandwidth allocation controller, which plays the role of a coordinator in a hierarchical dynamic control scheme, and attempts to minimize a cost function, accounting for overall cell loss and refused traffic. The bandwidth assignments obtained are passed to the call admission controllers, where they are used as parameters affecting the admission rules until the next intervention. The overall control task is thus decomposed between a set of fast "low level" decision makers that base their actions upon a portion of the system's state and a "higher level" agent, which acts periodically, uses centralized information and can take a

0743-166X/93 $03.00 © 1993 IEEE

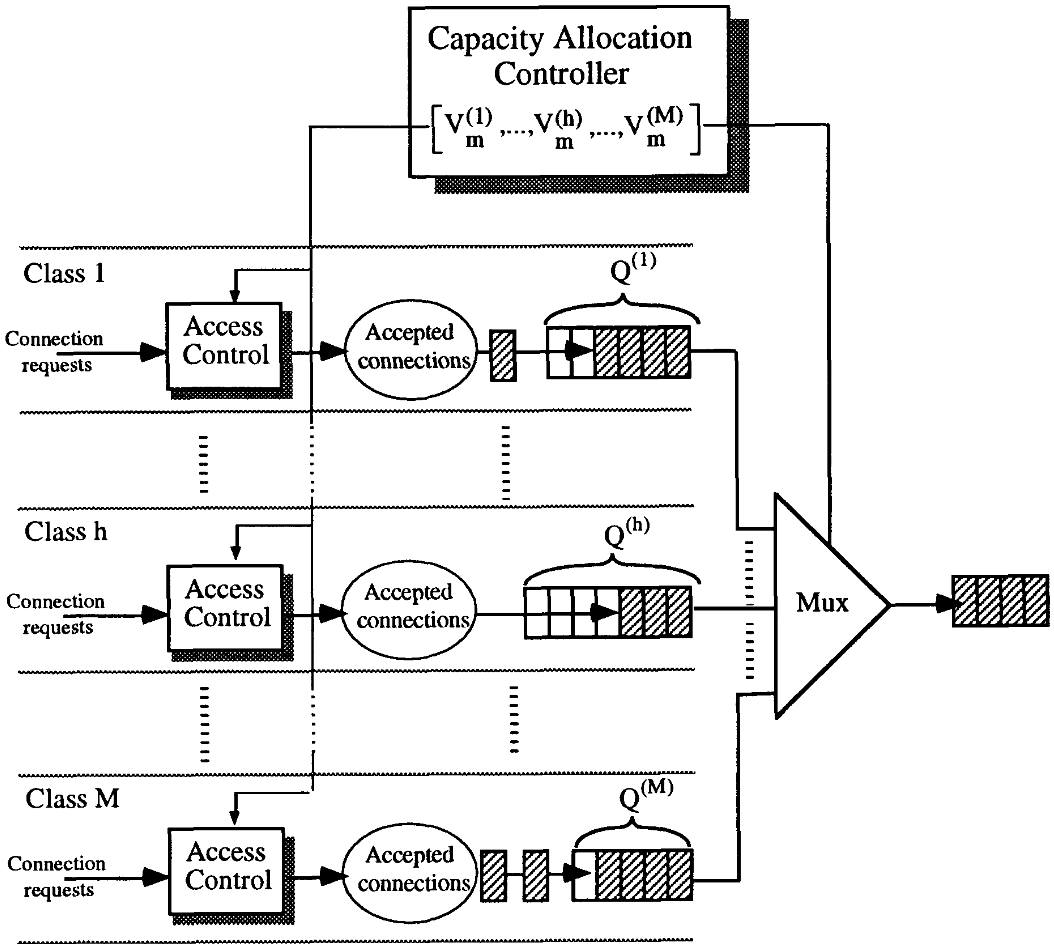

Fig. 1. Structure of the overall control system

longer computing time.

The paper is organized as follows. In the next Section, we describe the control architecture, the admission control rules, and the cost function to be minimized by the bandwidth allocation controller. A simulative analysis of the effectiveness of the admission control rules is given in Section 3. In Section 4 we report and discuss several simulation results that show the average as well as the dynamic behaviour of the overall control architecture. Section 5 contains the conclusions.

2. STRUCTURE OF THE MODEL AND OF THE CONTROL SYSTEM

Our model essentially regards an ATM link outgoing from an access node in the network, and its structure is represented in Fig. 1. The ATM channel carries fixed length cells, whose duration corresponds to that of a time slot and represents the discrete time unit (in the following, k denotes the discrete time variable). Upstream of the ATM channel are an access multiplexer, M admission controllers (one for each user traffic class) and a bandwidth allocation controller. The multiplexer is assumed to operate synchronously, i.e., input and output times are synchronous. Each admission controller h implements a decision rule for the acceptance of incoming calls that depends on the current number of accepted connections for the specific class, on the statistical and performance characteristics of the connections, and on the bandwidth (which will be termed "virtual capacity" in the following) it has been assigned for the current period by the allocation controller. The latter divides the total capacity C_T of the ATM channel (in Mbits/s) into virtual capacities $V_m^{(h)}$, h=1,...,M, where m=0, K, 2K,... represent the instants of intervention, and K is the length of the intervention period (in slots); the m-th assignment holds constant for the time slots k=m, m+1, ..., m+K-1. The assignment is made on the basis of the dynamic variations in the traffic flows, with the goal of providing a fair sharing to the admission controller. To this aim, a suitable cost function, that takes into account the expected number of lost cells pertaining to the whole offered traffic over the following K slots, is minimized at each instant m, where a new K-slot period begins.

In the model that was used in [12, 13] to derive the access control rules and the cost function, we supposed the slot assignment to be such that class h can see an average service capacity $V_m^{(h)}$ independently of the actual utilization of the channel. This would imply that available channel slots be distributed among the traffic classes in proportions that are in accordance with the) $V_m^{(h)}$'s, even if a certain class has temporarily not enough flow to fill its assigned slots and some waste may be created.

On the other hand, in the actual implementation of the scheme that we used in the simulations to be described further on, the assigned capacity values are treated as priorities, allowing a cell of a lower priority class to seize a slot if no higher priority cell is available, as is done in [14]; in this case, our calculations are based on a conservative assumption, and should result in upper bounds on the delays and cell loss probabilities.

Obviously, we must have that

$$\sum_{h=1}^{M} V_m^{(h)} = C_T \qquad m=0, K, 2K, \ldots \qquad (1)$$

A similar partitioning scheme is also adopted for the buffer pool of the access multiplexer to the ATM channel, as shown in Fig. 1. The multiplexer is made up by M buffers of length $Q^{(h)}$ [cells], one for each traffic flow. However, to keep our derivation analytically tractable, the quantities $Q^{(h)}$ are not assigned dynamically; they are determined a priori (off-line), on the basis of the performance requirements and the declared traffic intensity of the corresponding class. A possible procedure is outlined in [12, 13]; however, its use is closely related to the admission control criterion adopted there. In the present case, we fix the dimension of the buffer arbitrarily. A scheme for determinig this quantity optimally will be the subject of further study.

As regards the nature of class h traffic, we suppose it to be made up by bursty connections with statistical traffic characteristics that are identical and independent of each other. Each bursty connection can be represented by means of a two-state model (active and idle, respectively). The transitions between these two states form a two-state Markov chain, and we denote by $\alpha^{(h)}$ and $\beta^{(h)}$ the probabilities of transition from the idle to the active state, and from the active to the idle state, respectively. For each traffic class, these probabilities can be easily derived as in [5], if we suppose the statistical characteristics of the traffic flow to be known. Moreover, in order to take sources at different speed into account, we assume that, during a slot interval, an active connection may generate a cell with probability $\Gamma^{(h)}$, where $\Gamma^{(h)}$ is equal to the ratio between the peak bit rate at which an active call generates bits and the speed of the channel, $P^{(h)}/C_T$. When a connection is idle it does not generate cells.

Let $N^{(h)}$ be a given number of multiplexed connections; as they are independent of each other, the steady state probability $v_{n^{(h)},N^{(h)}}$ of having only $n^{(h)}$ active connections out of $N^{(h)}$ network connections is

$$v_{n^{(h)},N^{(h)}} = \binom{N^{(h)}}{n^{(h)}} (w_a^{(h)})^{n^{(h)}} (w_i^{(h)})^{N^{(h)}-n^{(h)}} \qquad (2)$$

where $w_i^{(h)}$ and $w_a^{(h)}$ are the steady-state probabilities of a connection being idle and of a connection being active, respectively, and are given by

$$w_i^{(h)} = \frac{\beta^{(h)}}{\alpha^{(h)} + \beta^{(h)}} \quad ; \quad w_a^{(h)} = \frac{\alpha^{(h)}}{\alpha^{(h)} + \beta^{(h)}} \qquad (3)$$

Since, due to our previous assumptions, each traffic class effectively "sees" a virtual multiplexer with buffer length $Q^{(h)}$ and channel capacity $V_m^{(h)}$, we can derive the cell loss probability for each single multiplexer, independently of the others. This was also done in [12, 13], and we will use the corresponding expressions in the following, where needed.

a) *Admission control*

Each admission controller acts independently of the others. The acceptance algorithm that we use is based on two controls, satisfying loss and delay requirements, respectively.

As far as the first control is concerned, we impose an upper limit $\varepsilon^{(h)}$ on the long-term time-averaged value of cell loss rate, namely

$$\sum_{n=0}^{N^{(h)}} P_{loss}^{(h)}(n)\, v_{n,N^{(h)}} \leq \varepsilon^{(h)} \tag{4}$$

where $P_{loss}^{(h)}(n)$ represents the steady-state value of the "instantaneous" (in the sense of [5]) cell loss probability; its expression in our case is given in [12, 13]. From (4) we can obtain the maximum number of connections $N_{max,L}^{(h)}(m)$ that $V_m^{(h)}$ and $Q^{(h)}$ can support as regards the cell loss requirement. Note that we are using steady-state values, due to the fact that the decision period K is supposed to be sufficiently long with respect to the cell as well as the connection activity dynamics.

As far as the delay requirements are concerned, the use of a $Q^{(h)}$-cell buffer ensures that a cell of the h-th class experiences a maximum delay (in slots) $Q^{(h)}/(V_m^{(h)}/C_T)$. It is worth noting here that, in general, this maximum delay may vary during the connection holding time due to the variations of the amount of capacity allocated by the controller.

$D^{(h)}$ being the value of the user delay requirement (in slots) for the h-th class, we can indicate with

$$\hat{Q}_m^{(h)} = \left\lfloor \frac{V_m^{(h)} D^{(h)}}{C_T} \right\rfloor \tag{5}$$

the queue length above which a cell experiences a delay exceeding the requirement. Then the following two cases may occur at the beginning of each K interval:

1) $\hat{Q}_m^{(h)} \geq Q^{(h)}$

in this case no further admission control is necessary to guarantee the delay requirement. Therefore, a new connection of the h-th class arriving at time slot k can be accepted in the network, provided that the cell loss requirement is satisfied, and we take the maximum number of connections $N_{max,D}^{(h)}(m)$ that $V_m^{(h)}$ and $Q^{(h)}$ can support as regards the delay requirement to be equal to $N_{max,L}^{(h)}(m)$;

2) $\hat{Q}_m^{(h)} < Q^{(h)}$

in this case an admission control that takes into account the delay requirement specifically is to be implemented.

We impose the delay requirement by requiring that the probability of the cell delay exceeding the value $D^{(h)}$ be lower than a given threshold $\delta^{(h)}$, that is

$$\Pr\{\text{ delay} > D^{(h)}\} < \delta^{(h)} \tag{6}$$

This is the same criterion proposed in [12, 13], but now we use a different approach in the calculation of $\Pr\{\text{delay} > D^{(h)}\}$. Moreover, we compute it directly instead of transforming the delay constraint into a bound on the maximum utilization. Thus, (6) becomes

$$\sum_{n=0}^{N^{(h)}} P_{delay}^{(h)}(n)\, v_{n,N^{(h)}} \leq \delta^{(h)} \tag{7}$$

where $P_{delay}^{(h)}(n)$ is computed in a similar way as $P_{loss}^{(h)}(n)$, that is

$$P_{delay}^{(h)}(n) = \sum_{i=0}^{Q^{(h)}} \pi_i^{(h)} \frac{\sum_{j=0}^{n} \left[\tilde{z}_{ij}^{(h)}\, \Omega^{(h)} + z_{ij}^{(h)} (1-\Omega^{(h)})\right] f_j^{(h)}(n)}{\sum_{j=0}^{n} j\, f_j^{(h)}(n)} \tag{8}$$

where

$$f_i^{(h)}(n) = \begin{cases} 0 & i < 0 \\ \binom{n}{i} (\Gamma^{(h)})^i (1-\Gamma^{(h)})^{n-i} & 0 \leq i \leq n \end{cases} \tag{9}$$

$$z_{ij}^{(h)} = \max[i+j-\hat{Q}_m^{(h)} - \max(i+j-Q^{(h)}, 0), 0] \tag{10}$$

$$\tilde{z}_{ij}^{(h)} = \max[i+j-1-\hat{Q}_m^{(h)} - \max(i+j-1-Q^{(h)}, 0), 0] \tag{11}$$

and

$$\Omega_m^{(h)} = \frac{V_m^{(h)}}{C_T} \tag{12}$$

is the probability of a class h cell being served in a slot. Here, $\tilde{z}_{ij}^{(h)}$ and $z_{ij}^{(h)}$ represent the number of cells exceeding the delay requirements (excluding the lost ones), when i cells are in the buffer and j cells are incoming, in the case that class h is served or not, respectively. Finally, $\pi_i^{(h)}$ is the steady state probability of having i class h cells queued in the buffer.

Summing up, the access control rule satisfying both requirements is simply the following: a new connection of the h-th class arriving at time slot k, m≤k≤m+K-1, can be accepted in the network if

$$N_c^{(h)}(k) + 1 \leq \min\{ N_{\max,D}^{(h)}(m), N_{\max,L}^{(h)}(m) \} \qquad (13)$$

being $N_c^{(h)}(k)$ the number of connections of the h-th class in progress at slot k.

b) *Bandwidth reassignment*

As regards the higher control layer, at each decision instant, the virtual capacities $V_m^{(h)}$ are dynamically reassigned by the allocation controller by means of a process that minimizes a suitable cost function J_P. The function to be minimized is chosen in such a way as to take into account the expected number of lost cells up to the next decision instant.

To this aim, by assuming quasi-stationarity of the connection request processes over the K slot decision interval, the structure of the cost function has been taken as

$$J_P = \sum_{h=1}^{M} \sigma^{(h)} \left[\sum_{n=0}^{N_c^{(h)}(m)} P_{loss}^{(h)}(n)\, v_{n,N_c^{(h)}(m)} + \right.$$

$$\left. + \xi \left(\sum_{n=0}^{N^{(h)}(m)} P_{loss}^{(h)}(n) v_{n,N^{(h)}(m)} - \sum_{n=0}^{N_c^{(h)}(m)} P_{loss}^{(h)}(n) v_{n,N_c^{(h)}(m)} \right) \right] \qquad (14)$$

where the constants $\sigma^{(h)}$, h=1,...,M, are weighting coefficients (due to the possibly largely different scales of loss probabilities), and ξ is a tradeoff coefficient. $N^{(h)}(m)$ is constant throughout decision interval m, and we choose for it a value that represents an estimate of the traffic activity for the h-th class in this interval, on the basis of the activity measured during the previous one. As a consequence, we set the value of $N^{(h)}(m)$ equal to the sum of the number $N_c^{(h)}(m)$ of connections in progress at the instant of decision and the number of blocked requests $N_b^{(h)}(m)$ in the preceding decision interval, i.e.,

$$N^{(h)}(m) = N_c^{(h)}(m) + N_b^{(h)}(m) \qquad (15)$$

Then, the first sum in brackets in (14) represents the long-term time-averaged value of cell loss rate for the h-th class (in the sense explained in [5]), due to the connections in progress, whereas the difference in parentheses represents the additional loss that would have been incurred if all calls presented in the previous interval had been accepted.

Thus, the tradeoff coefficient ξ can be used to increase the importance of call refusals that are not explicitly accounted for otherwise; as regards the choice of the weighting coefficients $\sigma^{(h)}$, it can be made in order to reflect the relative importance attributed to the various traffic classes by the network manager. Appropriate ranges of these parameters in the above sense may be determined through extensive simulation studies.

As we mentioned, at each decision instant m, every admission controller communicates the values $N_c^{(h)}(m)$ and $N_b^{(h)}(m)$ to the allocation controller, which minimizes the above defined cost function with respect to $V_m^{(h)}$, h=1,...,M. The latter quantities are computed by the allocation controller and communicated to the respective admission controllers (and, obviously, to the cell scheduling process), whose decisions in the next interval depend on the assigned capacity value.

In the minimization of the cost function J_P, account must be taken of the equality constraint (1), as well as of the inequality constraints

$$V_m^{(h)} \geq V_{\min}^{(h)}(m) \qquad h = 1,\ldots,M \qquad (16)$$

that serve the purpose of ensuring service quality to the $N_c^{(h)}(m)$ connections already in progress. Actually, $V_{\min}^{(h)}(m)$ is the minimum capacity that is necessary to cope with these connections, and can be computed so as to satisfy both cell loss and delay requirements. First, note that setting $N^{(h)} = N_c^{(h)}(m)$, the value $V_m^{(h)}$ that satisfies (4) with equality represents the minimum capacity $V_{\min,L}^{(h)}(m)$ necessary to satisfy the loss requirement.

On the other hand, the delay requirement is guaranteed by imposing that

$$V_{\min}^{(h)}(m) \geq \min\{\tilde{V}^{(h)}, V_{\min,D}^{(h)}(m)\} \equiv \tilde{V}_{\min,D}^{(h)}(m) \qquad (17)$$

where $V^{(h)}_{\min,D}(m)$ is obtained using (7) with equality and setting $N^{(h)} = N^{(h)}_c(m)$ and

$$\tilde{V}^{(h)} = \left\lfloor \frac{Q^{(h)}_m D^{(h)}}{C_T} \right\rfloor \qquad (18)$$

Thus, $V^{(h)}_{\min}(m)$ can be taken altogether as

$$V^{(h)}_{\min}(m) = \max\{\tilde{V}^{(h)}_{\min,D}(m), V^{(h)}_{\min,L}(m)\} \qquad (19)$$

The minimization of (14) under constraints (1) and (16) is a mathematical programming problem that can be performed by means of a gradient projection method [12, 13].

3. PERFORMANCE EVALUATION OF THE ADMISSION CONTROL RULES BY SIMULATION

In this Section, we show some results of analytic computations and simulations that we use to test the correctness and the performance of the proposed access control rule and to compare it with the access rule we proposed in [13]. The following data have been used:

$C_T = 100$ Mbit/s; $M = 2$; $K = 3 \cdot 10^5$ cells
T_s = slot duration = $4.24 \cdot 10^{-6}$ s (53 bytes/cell)
$P^{(1)} = 384$ kbit/s; $P^{(2)} = 1$ Mbit/s
$r^{(1)} = 0.5$; $r^{(2)} = 0.2$
(corresponding to burstiness 2 and 5, respectively)
$B^{(1)} = 10$; $B^{(2)} = 100$ (average burst length)
$1/\mu^{(1)} = 1.5$ s; $1/\mu^{(2)} = 0.8$ s
(average connection duration)
$\varepsilon^{(1)} = 1 \cdot 10^{-4}$; $\varepsilon^{(2)} = 1 \cdot 10^{-4}$
$\delta^{(1)} = \delta^{(2)} = 1 \cdot 10^{-3}$;
$D^{(1)} = 10$ slots; $D^{(2)} = 100$ slots
$N^{(1)}_a = 200$; $N^{(2)}_a = 100$ (average traffic intensity)
$Q^{(1)} = 11$ cells; $Q^{(2)} = 12$ cells

Some of the above values are far from representing a real situation; they were chosen especially to limit the length of the simulation runs necessary to obtain a significant number of events. This is true, in particular, for the average duration of the connections. However, one of the main purposes of the scheme is that of coping, to a certain extent, with dynamic variations in the call processes; in our case of relatively short connections, this is achieved by keeping the reallocation interval K also relatively short. In case of longer duration of the connections (with the same values of traffic intensity), the situation would be substantially unchanged by correspondingly enlarging the reallocation interval. We may note, in passing, that a larger value of K renders the computing time for the execution of the reallocation algorithm less critical.

Fig. 3 shows, for different capacity allocation, the maximum number of class 1 calls that the access rule allows to accept and the maximum number of calls, evaluated by simulations, that could be accepted, without violating the quality of service constrains (maximum delay and loss probability). We can note that the simulation points are always above the points computed by the access control rule, which means that the rule always respects the constraints. Moreover, the maximum number of calls acceptable using the rule is in general close to that obtained by simulations, and very close when the capacity is above 80 Mbits/s. This indicates a good efficiency of the system, especially for high capacities.

It can be interesting to note that the small jumps in the curve representing the maximum number of calls acceptable by the access rule are due to the presence of the integer part in (5).

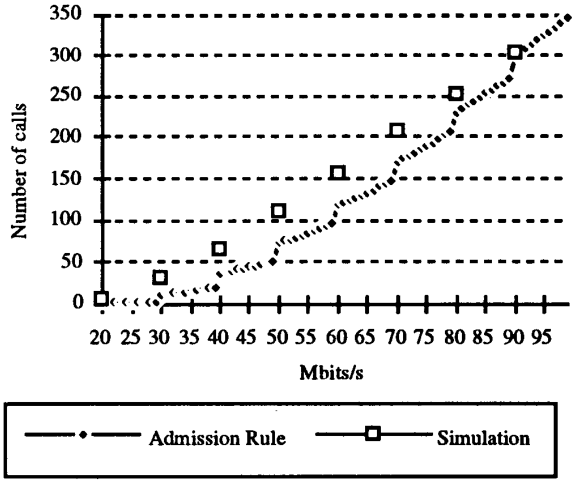

Fig. 3. Maximum number of acceptable calls computed by simulation and by the actual access control rule, respectively, versus the allocated capacity.

Fig. 4 compares the results seen in Fig. 3 with those obtained using the access rule, termed "previous access rule" in the figure, which we proposed in [13] . The comparison is done only for capacity values above 50

Mbits/s, because the previous rule is unable to enforce the constraints under a certain minimum capacity threshold. In the present case, it can be seen from the figure that this threshold is between 60 and 65 Mbits/s, where the points computed by the rule start being located under the simulation points. The new access rule clearly appears to follow better the trend of the simulation results; moreover it works well for every capacity values and it is more efficient than the previous one for capacities above the 80 Mbits/s.

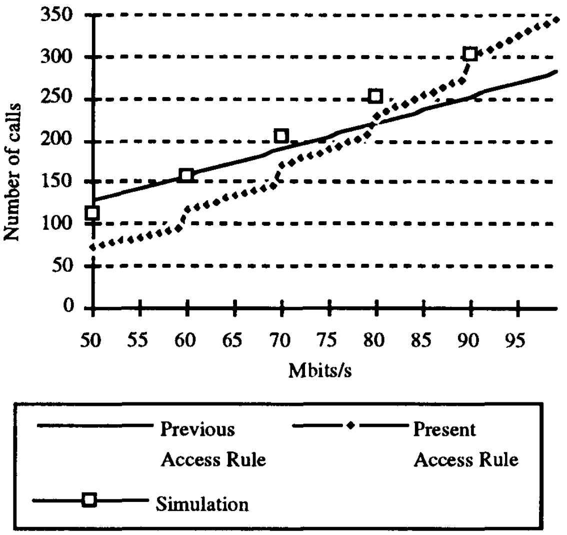

Fig. 4. Maximum number of acceptable calls computed by simulation, by the previous and by the actual access control rule, respectively, versus the allocated capacity.

We must note that for low values of capacity, the system appears to be able to support a very small number of calls; at 20 Mbits/s, for example, only 4 calls can be accepted (see Fig. 3). This number is far from the one that we could obtain by assigning the peak bandwidth to each call. This seemingly paradoxical situation is due to the very restrictive constraint on the delay. In fact, the maximum delay is fixed to ten slots expressed in terms of the maximum transmission capacity, which is 100 Mbits/s. So, for example, if we allocated, say, less than 10 Mbits/s to class 1, the scheduler would give less than one cell out of ten to this class. Thus, with less than 10 Mbits/s it is impossible to meet the delay constraint, and the maximum number of acceptable calls must be zero. We use such a restrictive constraint to test our system in the worst possible conditions.

Actually, the delay constraint is the more evident factor that influences the efficiency of the access rule, for low values of the capacity, but not the only one.

In Figs. 5, 6, and 7, we show the feasible load regions (in terms of the maximum number of acceptable calls for both classes), obtained by using the peak capacity assignment, the mean assignment, and the access rule, respectively. Once fixed the value of connections for one class, the corresponding value for the other is obtained by first finding the minimum capacity necessary to support the fixed connections under the specific criterion, and then computing the maximum number of connections of the other class that the residual capacity can support. For the computations whose results are shown in Fig. 5, we used the same data presented at the beginning of this Section, which will be call "standard data", whereas for those in Fig. 6, we changed the total capacity from 100 Mbits/s to 150 Mbits/s and the delay requirements $D^{(1)}$ and $D^{(2)}$ from 10 and 100 to 50 and 150 slots, respectively; finally, in Fig. 7, also the burstiness has been changed from 2 (class 1) and 5 (class 2) to 5 and 10, respectively.

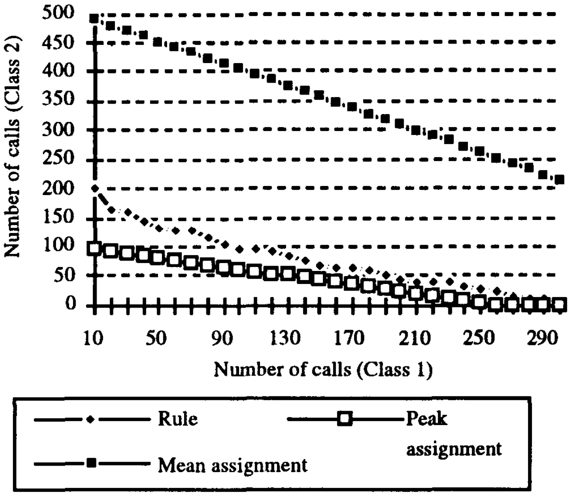

Fig. 5. Maximum number of acceptable calls of class 2 for fixed numbers of active calls of class 1 using the standard data.

The first case (Fig. 5) corresponds to the very restrictive situation that was used so far, and in fact the results obtained by the rule are close to those that would be obtained using the peak assignment; however, we must note that even in the extremely unbalanced situations (i.e., those corresponding to very low values of the number of calls for one of the classes) the allowable number of connections is correctly determined by the admission control rule, whereas the peak assignment would not assure the satisfaction of the requirements.

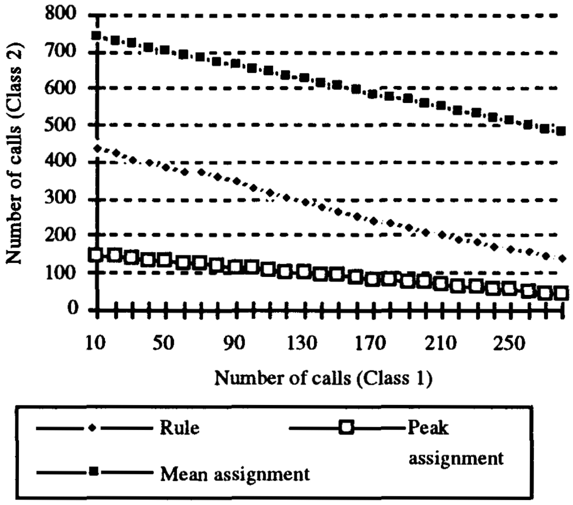

Fig. 6. Maximum number of acceptable calls of class 2 for fixed numbers of active calls of class 1 using C_T = 150 Mbits/s, $D^{(1)} = 50$ and $D^{(2)} = 150$.

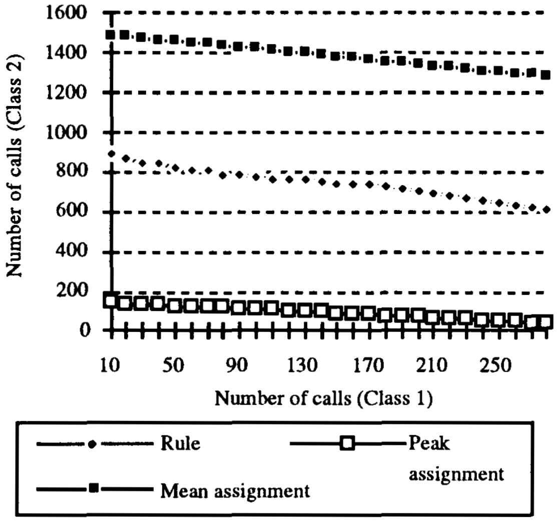

Fig. 7. Maximum number of acceptable calls of class 2 for fixed numbers of active calls of class 1 using C_T = 150 Mbits/s, $D^{(1)} = 50$, $D^{(2)} = 150$, $r^{(1)} = 0.2$, $r^{(2)} = 0.1$, (corresponding to burstiness 5 and 10, respectively).

The second and third cases (Figs. 6 and 7, respectively) show that the efficiency increases if we relax the constraints and if we put the statistical multiplexer in more favorable working conditions by increasing the total capacity and the burstiness of the two traffic classes.

4. PERFORMANCE EVALUATION OF THE OVERALL CONTROL ARCHITECTURE BY SIMULATION

In this section we want to evaluate the behavior of the complete system, to estimate the effects of the new access rule inside the overall control architecture. We have used the same standard data shown in the previous Section.

In the simulations, no lost and delayed cells were obtained. These results confirm the correct behavior of our access rule.

Additionally, the simulation results we present in the following show the behavior of the system in reaction to sudden variations in the offered load, especially as regards the bandwidth allocation and the influence on the call-level performance parameters.

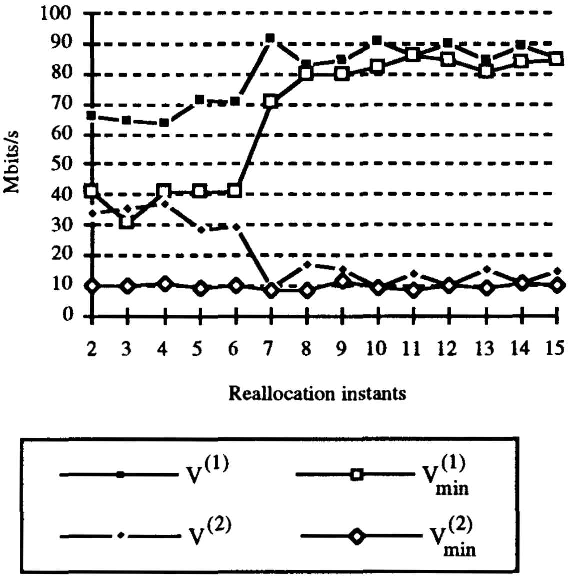

Fig. 8. Dynamic response of the system (in terms of bandwidth allocation) to a step variation in the intensity of class 1 connection requests.

In Fig. 8 we show the behavior of the bandwidth allocation as a function of time, measured in terms of reallocation intervals. We plot the bandwidth allocated to both classes, as well as the minimum bandwidth required to ensure quality of service to the already accepted connections. Initially, before the jump in the offered load of class 1, there is a certain difference in the allocated and minimum bandwidth of both classes. After the jump, which takes place between reallocation instants 6 and 7, the minimum capacity for class 1 increases up to the previously allocated bandwidth (due to the accepted connections), and new bandwidth is assigned at the expense of that of class 2, which is pushed down to the minimum. In just another interval, offered connections that had been previously refused are accepted (up to the maximum allowed by the total capacity), and the situation remains stationary.

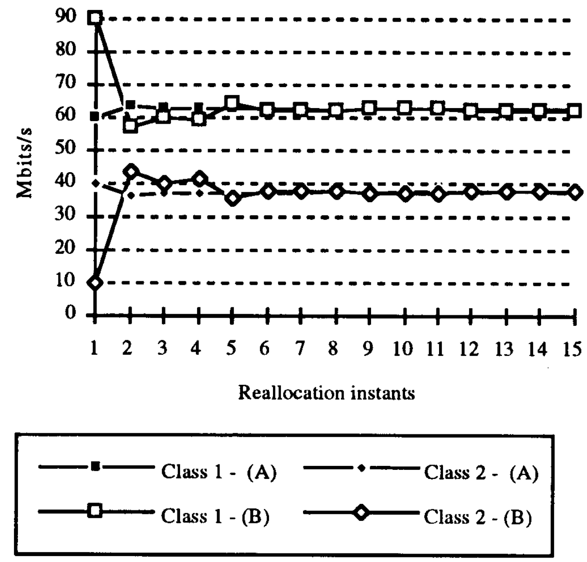

Fig. 9. Dynamic response of the system (in terms of bandwidth allocation) starting from the two different initial conditions (A) and (B).

Finally, Fig. 9 shows, in a situation with constant parameters, that the bandwidth allocations determined by the control system tend to be independent of the initial conditions.

5. CONCLUSIONS

A control strategy for dynamic bandwidth management in ATM networks, based upon the same philosophy introduced in [12, 13] has been considered in the paper. The global scheme relies upon an architecture organized in two control levels, consisting in a central bandwidth allocation controller, and as many call admission controllers as the service classes envisaged for characterization of the user traffic.

The admission control rules (one for each class) operate according to virtual capacity shares that are assigned to the various service classes by the allocation controller. The rules are synthesized with the aim of maintainig quality of service constraints, expressed in terms of cell loss probability and cell delay. Actually, a major difference between the approach used in this paper, with respect to the previous works [12, 13], consists in the fact that both the above mentioned performance measures have been expressed here in terms of the same basic quantities.

Analytical and simulative performace analyses of the present admission control rules have been reported, together with a comparison with those used in [13], showing a rather efficient behaviour and the satisfaction of the quality of service constraints in all operating conditions. Simulations showing the dynamic behaviour of the overall control scheme have been also reported and discussed.

ACKNOWLEDGMENT

This work was partially supported by the Italian National Council of Research (C.N.R.), under the Telecommunications Program, and by the Ministry of University and Scientific Research (M.U.R.S.T.).

REFERENCES

[1] D.Hong, T.Suda, "Congestion control and prevention in ATM networks", *IEEE Network Mag.*, vol. 5, no. 4, pp. 10-16, July 1991.

[2] *IEEE J. Select. Areas Commun.*, Special Issue on Congestion Control in High-Speed Packet Switched Networks, Sept. 1991.

[3] F. C. Schoute, "Simple decision rules for acceptance of mixed traffic streams", *Proc. ITC 12*, Torino, Italy, June 1988.

[4] G.Gallassi, G.Rigoglio, L.Fratta, "ATM: bandwidth assignment and bandwidth enforcement policies", *Proc. GLOBECOM '89*, Dallas, TX, Nov. 1989, pp. 1788-1793.

[5] T. Kamitake, T. Suda, "Evaluation of an admission control scheme for an ATM network considering fluctuations in cell loss rate", *Proc. GLOBECOM '89*, Dallas, TX, Nov. 1989, pp. 1774-1780.

[6] M.Dècina, T.Toniatti, "On bandwidth allocation to bursty virtual connections in ATM networks",

Proc. ICC '90, Atlanta, GA, April 1990, pp. 844-851.

[7] C. Rasmussen, J. H. Sørensen, K. S. Kvols, S. B. Jacobsen, "Source independent call acceptance procedures in ATM networks", *IEEE J. Select. Areas Commun.*, vol. 9, n. 4, pp. 351-358, April 1991.

[8] G. M. Woodruff, R. Kositpaiboon, "Multimedia traffic management principles for guaranteed ATM network performance", *IEEE J. Select. Areas Commun.*, vol. 8, n. 3, April 1990.

[9] J.A.Suruagy Monteiro, M.Gerla, L.Fratta, "Statistical multiplexing in ATM networks", *Proc. 4th Internat. Conf. on Data Commun. Syst. and their Performance*, Barcelona, Spain, June 1990.

[10] A. Lombardo, S. Palazzo, D. Panno, "Admission control over mixed traffic in ATM networks", *Internat. J. of Digital and Analog Commun. Syst.*, vol. 3, Aug. 1990.

[11] Z.Dziong, J.Choquette, K.Q.Liao, L.Mason, "Admission control and routing in ATM networks", *Comput. Networks and ISDN Syst.*, vol. 20, pp. 189-196, Dec. 1990.

[12] R.Bolla, F.Davoli, A.Lombardo, S.Palazzo, D.Panno, "Adaptive access control of multiple traffic classes in ATM networks", *Proc. GLOBECOM '91*, Phoenix, AZ, Dec. 1991, vol. 2, pp. 331-338.

[13] R.Bolla, F.Davoli, A.Lombardo, S.Palazzo, D.Panno, "Adaptive bandwidth allocation by hierarchical control of multiple ATM traffic classes", *Proc. INFOCOM '92*, Florence, Italy, May 1992, pp. 30-38.

[14] Y.Takagi, S.Hino, T.Takahashi, "Priority assignment control of ATM line buffers with multiple QOS classes", *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1078-1092, Sept. 1991.