ELSEVIER

Computer Communications 19 (1996) 645–652

computer communications

Simple schemes for traffic integration at call set-up level in ATM networks

Raffaele Bolla, Franco Davoli, Mario Marchese

Department of Communications, Computer and Systems Science (DIST), University of Genoa, Via Opera Pia, 13, 16145 Genova, Italy

Abstract

The paper presents a scheme for admission control at call set-up level at an ATM node. The traffic is considered to be divided into classes, characterized by statistical parameters like peak and average bandwidth, and by Quality of Service (QoS) requirements that allow us to define a 'feasibility region' where such requirements are guaranteed. A model to describe the 'call admission' is proposed. Then, three alternative strategies to get the maximum number of acceptable connections, by taking into account the probability of blocking a call at the call set-up level, are described. The first one is intended to minimize the overall call blocking probability; the second one is aimed at balancing the probability among the different traffic classes; while the last one allows us to enforce a constraint on the call blocking probability. The efficiency of the proposed strategies is tested by simulation and verified by comparing it with other admission control schemes that appear in the literature.

Keywords: Admission control; Call set-up; ATM networks; Traffic integration

1. Introduction

In an ATM environment, broadband integrated services networks carry bursty traffic with different characteristics and Quality of Service (QoS) requirements. Many studies and experiments are currently performed, and many topics investigated, to satisfy the requirements for any application. Fair resource allocation strategies and optimized bandwidth management are a subject widely treated in the literature [1–4]. In [5–10], Call Admission Control (CAC) strategies are investigated and tested, while in [11,12] scheduling algorithms providing satisfactory QoS to each service are proposed. Routing strategies are proposed in [13,14] to obtain a fair resource management. The problems mentioned above are not independent, and, even if each of them is treated separately from the others, they can be viewed in a unified way, as in [15–21]. Other problems considered are the buffer position (input, output or input-output queueing) in an ATM node (discussed in [22]), and the segregation of the traffic in classes, characterized by different parameters, as in [11,18–21,23], among others.

Also the definition of a call-space area ('feasibility region') has been an object of research, to find the optimal feasible space [25,26], or to have a complete separation between the problem of guaranteeing performance quality at the cell level and the problem of accepting a new call in the network [17].

In previous works ([18], for example), the authors proposed a hierarchical two-level control scheme, addressing the last two topics mentioned above, though not achieving a complete separation between them. In this paper, cell level performance requirements for each traffic class are considered to be satisfied through the definition of a 'feasibility region'. The region is computed by using a model already in the literature, but the specific computation mechanism does not constitute a restriction for the strategies introduced in the paper; other computation methods could be used.

An output queueing node model is assumed in the paper; the traffic is divided into H classes, defined by peak and average bandwidth and by QoS requirements (such as cell loss rate and delayed cell rate [18–21], for example). The CAC mechanism is structured by considering a maximum number of acceptable connections that allow us to guarantee performance requirements to each traffic class by retaining the general philosophy proposed in [18–21].

The paper is structured as follows: the next section is dedicated to describing the system model and the CAC scheme, and to the definition of a 'feasibility region'. Section 3 defines the strategies for computing the maximum number of acceptable calls, while simulation results and some comparisons with other CAC strategies are shown in Section 4. Conclusions are provided in Section 5.

0140-3664/96/$15.00 © 1996 Elsevier Science B.V. All rights reserved
PII S0140-3664(96)01097-3

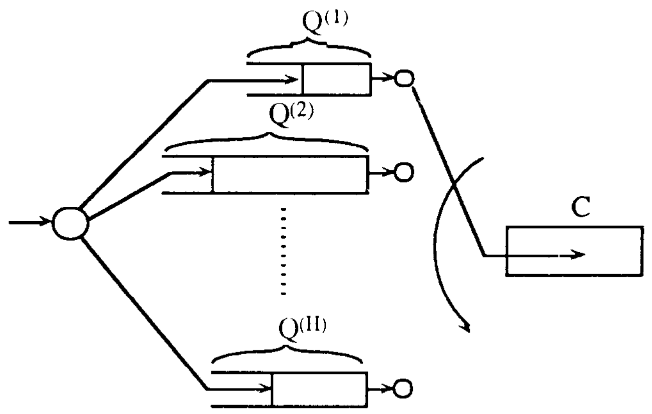

Fig. 1. System model.

2. System model and CAC scheme

The node model is supposed to be an output queueing one, and for simplicity, only one output link is considered (see [20,21] for an extension). The system model is shown in Fig. 1, where the traffic entering the node is queued in buffers, one for each traffic class, with a limited length. The buffer length is fixed to the value $Q^{(h)}$, $h = 1, \ldots, H$, and the capacity of the output link is assumed to be C (Mbits/s).

In correspondence with each output link, a Call Admission Control Block decides if a connection has to be accepted or rejected, while a Scheduler picks up the cells of the output buffer following a strategy (presented in [18,19]), which essentially grants each traffic class a fixed portion of the total bandwidth. The CAC performs the simple strategy explained below. In the following, vectors will be indicated by using a bold type (**A**, for example), and, if **A** and **B** are two vectors of dimension (n), the notation ($\mathbf{A} > \mathbf{B}$) means that the ith element of vector **A** is larger than the ith element of vector **B** ($\forall i,\ 1 \le i \le n$).

The traffic entering the node is considered to be segregated into classes, characterized by statistical parameters (for class (h), peak bandwidth $B_p^{(h)}$, average bandwidth $B_m^{(h)}$, and as a consequence, burstiness $b^{(h)} = B_p^{(h)}/B_m^{(h)}$), and by performance requirements that allow us to define a 'feasibility region', i.e. a region in call space where performance requirements are statistically guaranteed. The feasibility region can be obtained by complex analytical models or by simulation, and many studies can be found in the literature (among others, [1,11,24-26], where a survey of classes of control policies and a characterization of the optimal strategy in the specific case of Coordinate Convex Policies is also reported). In our case, an example will be given by Fig. 3 in Section 3. The region depicted in Fig. 3 has been computed using the strategy extensively presented in [18,19], which can be summarized as follows: an ON/OFF model is used for each bursty source, and the probability of generating a cell in the active state (ON) for a call of class h is $B_p^{(h)}/C$. Each traffic class (h) has to be guaranteed with two performance requirements, the cell loss ($P_{\text{loss}}^{(h)}(n^{(h)})$) and the delayed cell rate ($P_{\text{delay}}^{(h)}(n^{(h)})$), supposing $n^{(h)}$ calls in the active state. The probability of having $n^{(h)}$ connections in the active state out of $N^{(h)}$ accepted connections being $v_{n^{(h)},N^{(h)}}$, the performance requirements can be considered to be:

$$\sum_{n^{(h)}=1}^{N^{(h)}} P_{\text{loss}}^{(h)}(n^{(h)})\, v_{n^{(h)},N^{(h)}} \le \epsilon^{(h)} \tag{1}$$

$$\sum_{n^{(h)}=1}^{N^{(h)}} P_{\text{delay}}^{(h)}(n^{(h)})\, v_{n^{(h)},N^{(h)}} \le \delta^{(h)}, \tag{2}$$

where $\epsilon^{(h)}$ is an upper limit on the time-averaged value of the cell loss rate, and $\delta^{(h)}$ has the same meaning for the time-averaged value of cells that suffer a delay longer than a fixed value ($D^{(h)}$ in Section 4).

The two inequalities above ((1) and (2)) define a region in call space where performance requirements are satisfied, that is, they determine a 'feasibility region'. It should be pointed out that the chosen region is just an example, and the particular strategy used to individuate the feasibility region does not affect the overall call acceptance scheme defined in the following and in the next section.

In [26], different classes of control policies are defined, and some possible algorithms are proposed to get a subregion, which allows us to obtain a lower average call blocking probability. In that context, a general Multiple Service, Multiple Resource (MSMR) problem is solved, and a linear model for the definition of the state space is assumed. In our case, a complete modelling and characterization of the call space is quite complex, and it will be the subject of future work. So, here we do not consider the more general class of control policies (named Dynamic Programming), where decisions are based 'not only upon what state admission would place the system in, but also upon what type of service is requested' [26]. Only Coordinate Convex policies (where admission decisions depend only on the state the system would enter if the new request were accepted) are taken into account and, in particular, the case of Complete Partitioning policies is investigated.

This class of policies allows us to define a rectangular sub-region (in the two-classes case, or a N-cube subregion if the general N-classes case is considered) inside the feasibility region. So, the admission strategy can be summarised as follows: a maximum number of acceptable connections $N_{\max}^{(h)}$, whose computation is the object of the next section, shall be defined for each class. $N_c^{(h)}$ being the number of accepted connections, a new call is accepted if the number of calls in progress at the node plus the new one does not exceed the maximum number of acceptable calls for that traffic class. In

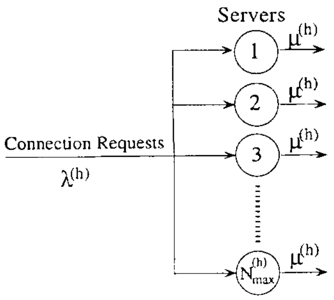

Fig. 2. CAC blocking model.

other terms,

$$\begin{aligned} &\text{if } N_c^{(h)} + 1 \le N_{\max}^{(h)} \rightarrow \text{ connection accepted,} \\ &\text{if } N_c^{(h)} + 1 > N_{\max}^{(h)} \rightarrow \text{ connection rejected.} \end{aligned} \quad (3)$$

Then, the CAC blocking can be modelled, for the generic class (h), as in Fig. 2, where the number of servers is equal to the maximum number of acceptable calls of that class. The quantity $\lambda^{(h)}$ represents the average call arrival rate, according to a Poisson process, while $1/\mu^{(h)}$ is the average duration of a call, which is supposed to be exponentially distributed. The quantity $N_a^{(h)} = \lambda^{(h)}/\mu^{(h)}$ (in Erlangs) is the average traffic intensity for traffic class (h).

Considering the number $N_{\max}^{(h)}$ fixed, the probability of blocking a call is given by the well known Erlang-B formula:

$$P_B^{(h)} = \frac{(N_a^{(h)})^{N_{\max}^{(h)}}/N_{\max}^{(h)}!}{\sum_{l=0}^{N_{\max}^{(h)}} \frac{(N_a^{(h)})^l}{l!}}. \quad (4)$$

This probability is used in section 3, where three alternative strategies to find the maximum number of acceptable connections are presented. It is important to remark that the quantities $N_{\max}^{(h)}$ $h = 1, \ldots, H$, are then considered as variables, and the schemes introduced in the next section have the aim to compute their 'optimal values' (which will be indicated by $N_{\max}^{(h)\text{opt}}$, $h = 1, \ldots, H$), in the sense to be defined.

3. Computation of the maximum number of acceptable calls

The aim of this section is to present some different cost functions, each one giving rise to a different scheme to compute the maximum number of acceptable calls ($N_{\max}^{(h)}$) for each traffic class. All the cost functions are based on the computation of the call blocking probability (4), which has been shown in the previous section. It should be remembered that the problem is considered in the stationary case; in fact, we suppose $\lambda^{(h)}$ and $\mu^{(h)}$ to remain constant for a long period with respect to the call dynamics. If these values should vary, the optimization procedure would have to be applied again after a certain period.

The first function considered is simply intended to minimize an overall measure of the call blocking rate, namely

$$P_B(\mathbf{N}_{\max}) = \sum_{h=1}^{H} \alpha^{(h)} P_B^{(h)}(N_{\max}^{(h)}), \quad (5)$$

where $\mathbf{N}_{\max} \equiv [N_{\max}^{(h)}, h = 1, \ldots, H]$, $P_B^{(h)}(N_{\max}^{(h)})$ is the call blocking probability of class h, given by (4), and $\alpha^{(h)}$ is a weight that allows attaching a distinct priority level to each traffic class. It is worth noting that, if $\alpha^{(h)}$ is assumed equal to the ratio $\lambda^{(h)}/\sum_{h=1}^{H} \lambda^{(h)}$, the quantity (5) represents the average call blocking probability of the whole system.

The maximum number of acceptable connections $\mathbf{N}_{\max}^{\text{opt}} \equiv [N_{\max}^{(h),\text{opt}}, h = 1, \ldots, H]$ for each traffic class is obtained by the minimization of $P_B(\mathbf{N}_{\max})$ over the feasibility region. Then, by defining the feasibility region as the set F_R of M-tuples $\mathbf{N} \equiv [N^{(h)}, h = 1, \ldots, H]$ that satisfy performance requirements ((1) and (2), in this context), it can be said that:

$$\mathbf{N}_{\max}^{\text{opt}} = \arg \min_{\mathbf{N}_{\max} \subset F_R} P_B(\mathbf{N}_{\max}). \quad (6)$$

The associated CAC scheme will be called the Erlang Scheme (ES). Fig. 3 allows a possible interpretation of the ES: the computation of the maximum number of acceptable calls implies finding a fixed sub-space inside the feasibility region. The sub-space has rectangular shape in Fig. 3, where the situation with two traffic classes is depicted; in general, the sub-region would be hypercube. These types of CAC schemes are indicated as Complete Partitioning Schemes in the literature [26].

The minimization of function (5) allows us to find the minimum value for a possible overall measure of the call blocking rate (or, in a particular case, as said before, the minimum value of the average call blocking probability). But this scheme takes into account neither possible performance requirements for each single class (in terms of blocking probability), nor some balancing criterion among the classes. So, in this case, the solution obtained

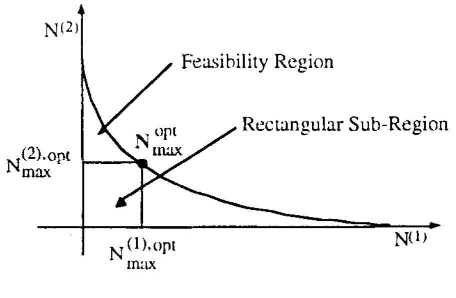

Fig. 3. Example of Complete Partitioning: the ES scheme.

might strongly favour a class with respect to the others, as shown in Section 4. To avoid this problem we have considered two other different cost formulations, which might be more suited for a real application.

Concerning the first one of them, whose corresponding scheme will be called the Balanced Erlang Scheme (BES), the balancing of the call blocking rate among different classes is the main goal; thus, the cost function is defined as

$$P'_B(\mathbf{N}_{\max}) = \max_h \left\{ \alpha^{(h)} \quad P_B^{(h)}(N_{\max}^{(h)}) \right\}. \tag{7}$$

$\mathbf{N}_{\max}^{\text{opt}}$ can be obtained in the same way as in Eq. (6), by substituting the quantity $P_B(\mathbf{N}_{\max})$ with the quantity $P'_B(\mathbf{N}_{\max})$. In formula:

$$\mathbf{N}_{\max}^{\text{opt}} = \arg \min_{\mathbf{N}_{\max} \subset F_R} P'_B(\mathbf{N}_{\max}). \tag{8}$$

The overall call blocking performance given by Eq. (8) is generally worse than that resulting from the application of Eq. (6); however, now all the classes are managed in a fair way (with $\alpha^{(h)} = 1, \ \forall h$), because the allocation of Eq. (8) tends to equalize the blocking probabilities of the different classes. The method of balancing can be simply controlled by changing the weights $\alpha^{(h)}$. Also, this scheme, although very different from the previous one, is a Complete Partitioning Scheme, because it implies, for each class, the computation of a maximum number of calls not varying in time and, as a consequence, the calculation of a fixed sub-region, as in Fig. 3.

The last scheme we present is defined by fixing a constraint on the maximum call blocking probability for each class, that is:

$$P_B^{(h)}(N_{\max}^{(h)}) \le \gamma^{(h)} \qquad h = 1, \ldots, H. \tag{9}$$

Let $\bar{N} = \text{col}\,[\bar{N}^{(h)}, h = 1, \ldots, H]$ be the equality solution of Eq. (9) with respect to $\mathbf{N}_{\max}$. To satisfy the constraints in Eq. (9), we must have

$$N_{\max}^{(h)} \ge \bar{N}^{(h)}, \qquad h = 1, \ldots, H. \tag{10}$$

We can distinguish two cases: $\bar{\mathbf{N}} \in F_R$ and $\bar{\mathbf{N}} \notin F_R$. In the first case, a sub-region which satisfies the constraints (9) (shaped as shown in Fig. 4(a) for a system supporting two classes) can be found. Otherwise, the sub-region does not exist (Fig. 4(b)). On the basis of these considerations, we compute $\mathbf{N}_{\max}^{\text{opt}}$ as

$$\mathbf{N}_{\max}^{\text{opt}} = \begin{cases} \arg \min\limits_{\substack{\mathbf{N}_{\max} \in F_R \\ \mathbf{N}_{\max} \ge \bar{\mathbf{N}}}} P_B(\mathbf{N}_{\max}) & \bar{\mathbf{N}} \in F_R \\ \arg \min\limits_{\mathbf{N}_{\max} \in F_R} D_{P_B}(\mathbf{N}_{\max}) & \bar{\mathbf{N}} \notin F_R \end{cases} \tag{11}$$

where

$$D_{P_B}(\mathbf{N}_{\max}) = \sum_{h=1}^{H} \left[P_B^{(h)}(N_{\max}^{(h)}) - \gamma^{(h)} \right]^2, \tag{12}$$

and $\mathbf{N}_{\max} \ge \bar{N}$ has the meaning explained at the beginning of section 2.

It should be clear from Eq. (11) that, if $\bar{\mathbf{N}} \in F_R$, we apply the same cost function (Eq. (6)) of the ES method inside the sub-region identified by constraints (10); otherwise, when constraints (10) cannot be satisfied, the cost function (12) is minimized; that is, the configuration at 'minimum distance' is taken as the optimum one. We call this the Rectangular Sub-Region (RSR) scheme. The choice of the cost function (6) if $\bar{\mathbf{N}} \in F_R$ is due to the fact that requirements (9) already guarantee a chosen performance about the call blocking rate of the different traffic classes. So, the choice of a balancing cost function would seem meaningless; the minimization of an overall measure of the call blocking rate could be more significant.

Another different possible approach, whose performance we have reported in section 4 as a comparison with those introduced, but whose investigation is not the subject of this paper, can be found by using a simple Non-Complete Partitioning scheme as a control access method. We will call this CAC method the Complete Sub-Region (CSR) scheme. The rationale can be simply explained as follows: the CSR is the same as the RSR when constrains (10) cannot be satisfied; otherwise, starting from Fig. 4(a), and drawing a perpendicular line from the points X and Y to the $N^{(2)}$ and $N^{(1)}$ axes, respectively, the sub-region S_R, depicted in Fig. 5, can be simply obtained. In this case a new call is accepted if the total number of calls in progress remains inside S_R.

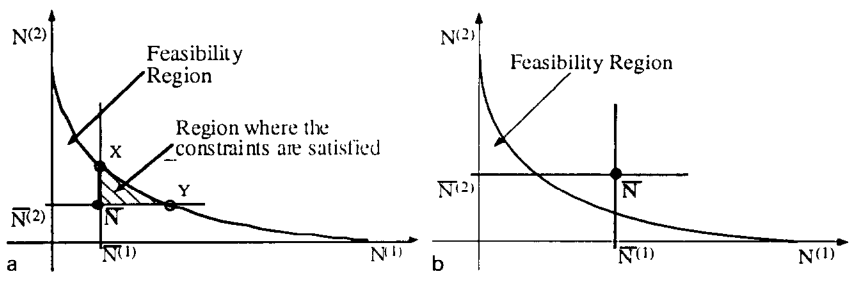

Fig. 4. Effect of constraints (10) on the feasibility region in the case of (a) $\bar{\mathbf{N}} \in F_R$; and (b) $\bar{\mathbf{N}} \notin F_R$.

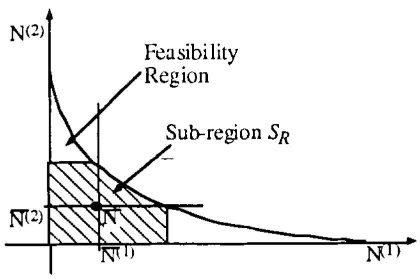

Fig. 5. Sub-region S_R (10) in the case of $\mathbf{N} \in F_R$.

4. Numerical results

The purpose of this section is to investigate the behaviour of the three schemes proposed in the previous section, and to test the efficiency of the overall strategy by comparing the results obtained with other CAC strategies. Four traffic classes with very different characteristics have been chosen to obtain the results in the following. The first traffic class represents medium quality video, and the second bulk data transfer, while the third and the fourth classes denote, respectively, voice and image retrieval, as in [28]. In more detail the parameters which describe each traffic class are listed below:

$B_p^{(1)} = 1$ Mbit/s; $B_p^{(2)} = 10$ Mbits/s; $B_p^{(3)} = 64$ Kbits/s; $B_p^{(4)} = 2$ Mbits/s; (peak bandwidth).

$b^{(1)} = 2$; $b^{(2)} = 10$; $b^{(3)} = 2$; $b^{(4)} = 23$ (burstiness).

$B^{(1)} = 100$; $B^{(2)} = 1000$; $B^{(3)} = 58$; $B^{(4)} = 2604$ cells (average burst length).

$1/\mu^{(1)} = 20$ s; $1/\mu^{(2)} = 25$ s; $1/\mu^{(3)} = 30$ s; $1/\mu^{(4)} = 60$ s (average connection duration).

$\epsilon^{(1)} = \epsilon^{(2)} = 1 \cdot 10^{-4}$; $\epsilon^{(3)} = \epsilon^{(4)} = 1 \cdot 10^{-6}$ (upper limit for the average cell loss rate).

$\delta^{(1)} = \delta^{(2)} = 1 \cdot 10^{-3}$; $\delta^{(3)} = \delta^{(4)} = 1 \cdot 10^{-5}$ (upper limit for the average delayed cell rate).

$D^{(1)} = 400$; $D^{(2)} = 100$; $D^{(3)} = 200$; $D^{(4)} = 1000$ slots (delay threshold).

$N_a^{(1)} = 80$; $N_a^{(2)} = 40$; $N_a^{(3)} = 250$; $N_a^{(4)} = 92$ Erlangs.

(Global average traffic intensities offered to the network; call arrival processes follow independent Poisson distributions.)

$Q^{(1)} = 20$; $Q^{(2)} = 10$; $Q^{(3)} = 15$; $Q^{(4)} = 15$ cells (buffer length).

To simplify the representation and interpretation of the results, each test has been carried out with a couple of traffic classes, and only the most meaningful results have been depicted.

Figs. 6, 9, 10, 15 and 16 refer to the couple class 1–class 2 (called couple A in the following); a channel capacity $C = 150$ Mbits/s (T_s = slot duration = $2.83 \cdot 10^{-6}$ s (53 bytes/cell)) has been used.

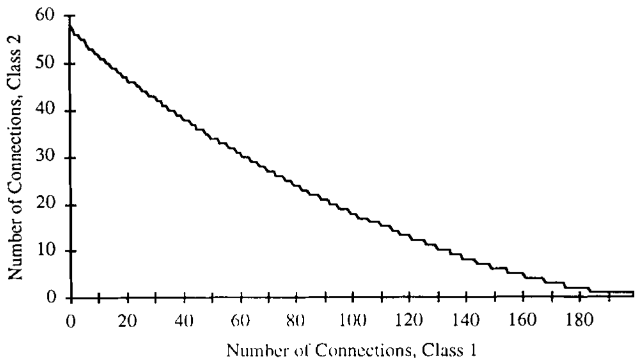

Fig. 6. Feasibility region F_R for classes 1 and 2.

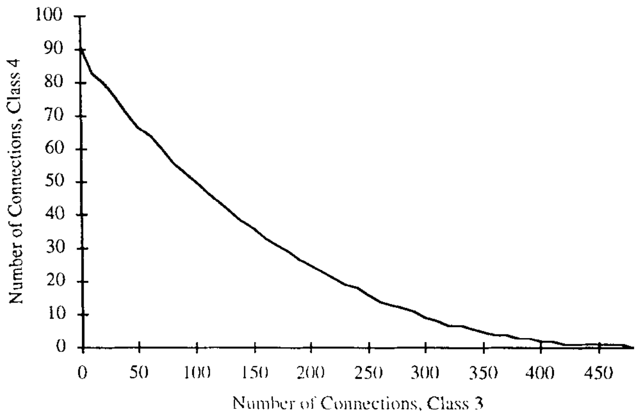

Fig. 7. Feasibility region F_R for classes 3 and 4.

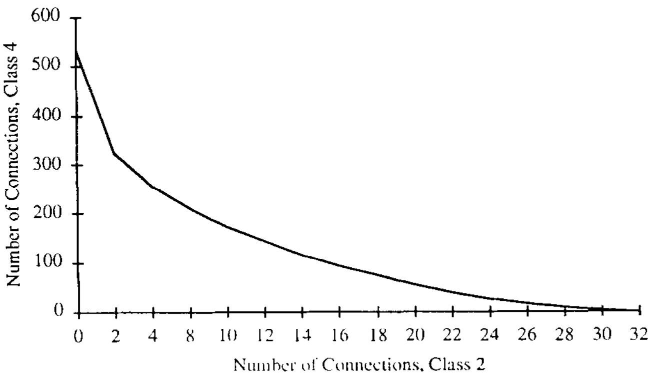

Fig. 8. Feasibility region F_R for classes 2 and 4.

Figs. 7, 11 and 12 are obtained by using the couple class 3–class 4 (called couple B in the following), with a channel capacity $C = 30$ Mbits/s (T_s = slot duration = $14.1 \cdot 10^{-6}$ s (53 bytes/cell)).

Class 2–class 4 (called couple C in the following) and a channel capacity $C = 150$ Mbits/s = $2.83 \cdot 10^{-6}$ s (53 bytes/cell) have been chosen for Figs. 8, 13 and 14.

The duration of the simulations performed, which have been carried out on Sun SPARC 10 and SPARC 20 workstations, correspond to 1 hour and 30 minutes of real network time.

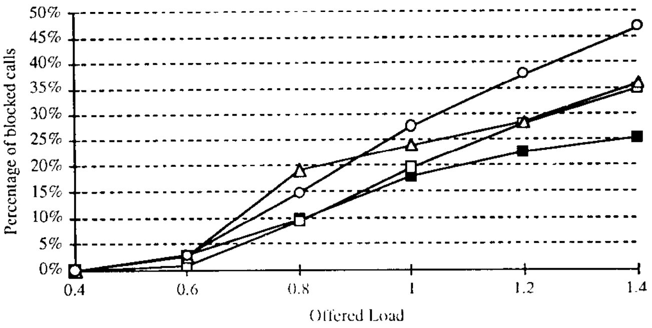

Fig. 9. Overall percentage of blocked cells versus offered load, classes 1 and 2. ■: ES; □: FRS; Δ: RS; O: BES.

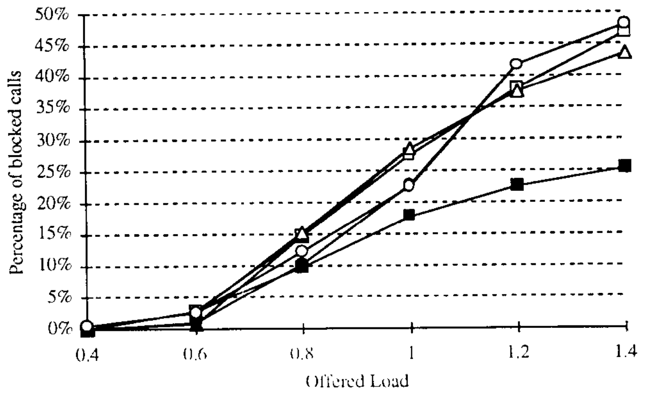

Fig. 10. Overall percentage of blocked cells versus offered load, classes 1 and 2. ■: ES; □: BES; ▲: CSR, a = 0.2; Δ: RSR, a = 0.2; ●: CSR, a = 0.4; O: RSR a = 0.4.

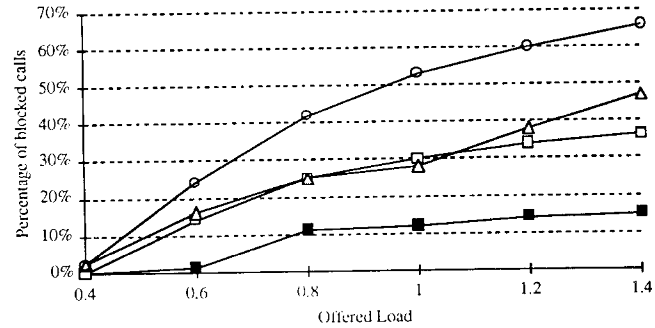

Fig. 11. Overall percentage of blocked calls versus offered load, classes 3 and 4 (for key see Fig. 9).

Figs. 6, 7 and 8 depict the feasibility region for couples A, B and C, respectively. The regions have been obtained by using inequalities (1) and (2) and the data reported above; however, any other feasibility region, obtained using different techniques, could be utilized without affecting the global strategy.

The traffic flow generated by the above data is considered to be a normalized offered load of value 1; an offered load x corresponds to the same data, except for the global average traffic intensities offered to the network ($N_a^{(h)}$, $h = 1, 2, 3, 4$), which are multiplied by x.

The overall percentage of blocked calls versus the

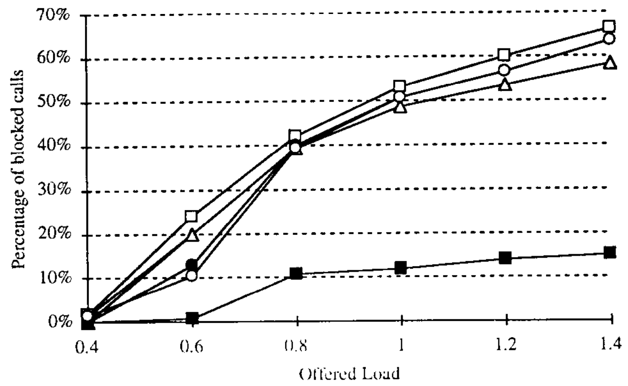

Fig. 12. Overall percentage of blocked calls versus offered load, classes 3 and 4 (for key see Fig. 10).

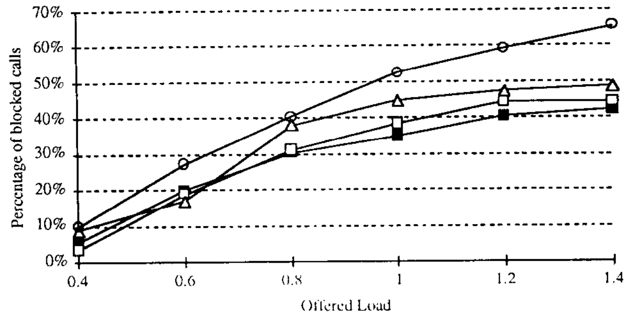

Fig. 13. Overall percentage of blocked calls versus offered load, classes 2 and 4 (for key see Fig. 9).

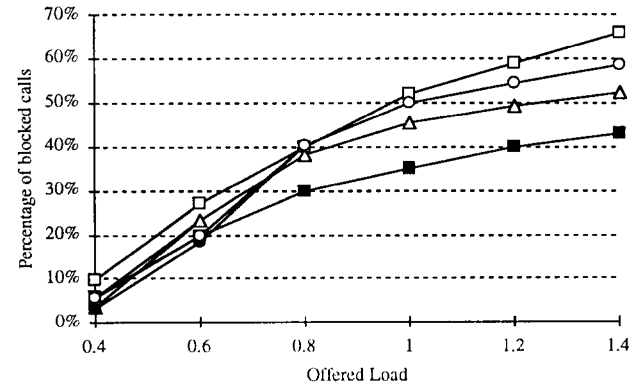

Fig. 14. Overall percentage of blocked calls versus offered load, classes 2 and 4 (for key see Fig. 10).

offered load is depicted in Figs. 9–14. In detail, couple A has been used for Figs. 9 and 10; couple B for Figs. 11 and 12; and couple C for Figs. 13 and 14. Moreover, Figs. 9, 11 and 13 show a comparison of the results obtained using ES and BES with the results obtained using two other strategies. The first of these, called the Feasibility Region Scheme (FRS) here, does not compute a maximum number of acceptable connections, but always accepts a call, with the only constraint of the feasibility region; that is, if a new call in the system keeps the total number of connections within F_R, the call is accepted. The second is a strategy reported in [19], called the Reallocation Scheme (RS) here. The improvement in the percentage of blocked calls by using ES is noticeable; in fact, cost function (5) has the only purpose of minimizing an overall measure of the call blocking probability, without consideration of balancing among classes. In Fig. 10 (concerning couple A), Fig. 12 (couple B) and Fig. 14 (couple C), the overall percentage of blocked calls versus the offered load is also reported; but, in this case, the schemes ES, BES, RSR and CSR (Complete Sub-Region) have been utilized. The simulations have been performed considering $\alpha^{(h)} = a$, $h = 1, 2, 3, 4$; two values of a ($a = 0.2$ and $a = 0.4$) have been presented.

The performance of RSR appears to be similar to that of CSR (the schemes are the same above a certain load threshold, because the sub-region where the constraints are satisfied does not exist and, as a consequence, the S_R area does not exist). Moreover, it can be noted that, if the load is below (or equal to) a certain threshold (0.8 concerning the $a = 0.2$ case; 1 concerning the $a = 0.4$ case), the overall call blocking rate is really under the fixed threshold. For higher values of the load, the behaviour of RSR is similar to BES; in fact, the threshold being the same for each class, the global effect will be a balancing among the classes.

This behaviour can be better explained by observing Figs. 15 and 16 (couple A), where the percentage of blocked calls is shown versus the offered load for each traffic class. The ES, BES and FRS schemes have been used in Fig. 15; the BES and RSR schemes, with $a = 0.2$ and $a = 0.4$, have been utilized to obtain Fig. 16. It can be seen that the blocking percentages of the two classes are completely unbalanced for the FRS scheme and, in particular, for the ES case. So the lower overall call blocking percentage of the ES scheme is paid for with a conspicuous unbalancing effect.

On the contrary, the utilization of BES does not guarantee the lowest overall percentage of blocked calls, but it assures a fair division among the classes. Concerning a

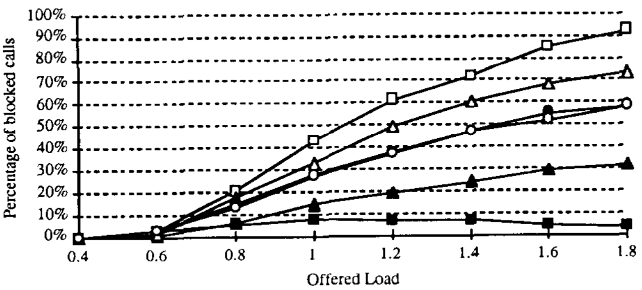

Fig. 15. Percentage of blocked calls for each traffic class vs. offered load, using the ES, FRS and BES for couple A. ■: ES Class 1; □: ES Class 2; ▲: FRS Class 1; △: FRS Class 2; ●: BES Class 1; ○: BES Class 2.

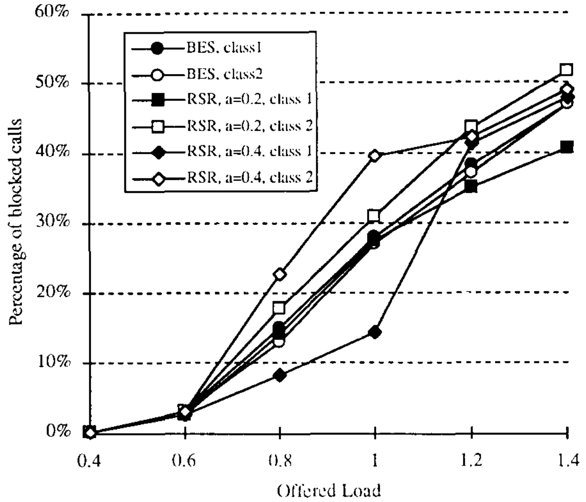

Fig. 16. Percentage of blocked calls for each traffic class vs. offered load, using the BES, RSR with $a = 0.2$ and RSR with $a = 0.4$, for couple A.

lower load situation, where the sub-region S_R exists, the RSR scheme shows a certain unfairness, which is due to the use of a cost (the cost used in the ES scheme, even if over a smaller area) aimed at minimizing an overall measure of the total call blocking probability, without taking into account any possible balancing. For higher loads, where S_R does not exist, the minimization of Eq. (12) gives results similar to those obtained with the BES scheme (this clearly happens only when all the $\alpha^{(h)}$ have the same value). It is worth noting that the real aim of the RSR scheme is to fix a threshold on the call blocking probability for each traffic class. Then, if it is not possible to find an area in the call space where the constraints are satisfied, the aim is to be as close as possible to the fixed thresholds. This behaviour can be noticed in Fig. 16, where, concerning the case ($a = 0.2$), the constraints can be satisfied (namely an area in the call space can be found) up to a load of 0.8. As far as the case ($a = 0.4$) is concerned, the constraints can be satisfied up to a load of 1; for a load higher than 1, it is not possible to maintain the call blocking probability below the value 0.4 and, so, the aim is to be as near as possible to this value. The threshold value being the same for each class, the global effect (for load 1.2 and 1.4) is a fair balancing.

5. Conclusions

In this paper, some simple CAC schemes for traffic integration in ATM networks at the call set-up level are proposed. The first scheme, called the Erlang Scheme (ES), aims to minimize the overall call blocking probability; the second, the Balanced Erlang Scheme (BES), has the main purpose of balancing the number of blocked calls among the traffic classes. The last scheme takes into account a set of constraints on the minimum call blocking probability, and applies the ES scheme to a sub-region (if it exists), where these constraints are satisfied; otherwise it tries to minimize the distance from the maximum call blocking probability. In all the schemes, the minimization is performed by taking into account the constraint of the feasibility region.

Four traffic classes, with different characteristics, are considered to test the proposed strategies. Simulation results have shown the difference among the proposed schemes, and verified the efficiency of ES, BES and RSR with respect to other CAC strategies. More specifically, ES allows the lowest overall call blocking probability, BES offers quite a fair balancing among the different traffic classes, while RSR can be considered a good compromise for a real implementation. The extreme simplicity of the cost functions makes the computation very fast and the algorithms well suited for control mechanisms.

References

[1] R. Guérin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high speed networks, IEEE J. Select. Areas Comm., 9(7) (September 1991) 968–981.

[2] F.P. Kelly, Effective bandwidths at multi-class queues, Queueing Syst., 9 (1991) 5–15.

[3] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS channel, Queueing Syst., 9 (1991) 17–27.

[4] A. Suruagy Monteiro and M. Gerla, Bandwidth allocation in ATM networks, Ann. Op. Res., 49 (1994) 25–50.

[5] H. Saito, Teletraffic Technologies in ATM Networks, Artech House, 1994.

[6] R.O. Onvural, Asynchronous Transfer Mode Networks: Performance Issues, Artech House, 1994.

[7] T. Kamitake and T. Suda, Evaluation of an admission control scheme for an ATM network considering fluctuations in cell loss rate, Proc. IEEE Glovecom '89, Dallas, TX, November 1989, pp. 1774–1780.

[8] L. Tassilulas, Y.C. Hung and S.S. Panwar, Optimal buffer control during congestion in an ATM network node, IEEE/ACM Trans. Networking, 2 (August 1994) 374–386.

[9] M. Abdelaziz and I. Stravrakakis, Some optimal traffic regulation schemes for ATM networks: a Markov decision approach, IEEE/ACM Trans. Networking, 2 (October 1994) 508–519.

[10] D. Hong and T. Suda, Congestion control and prevention in ATM networks, IEEE Network Mag., 5(4) (July 1991) 10–16.

[11] J. Takagi, S. Hino and T. Takahashi, Priority assignment control of ATM line buffers with multiple QOS classes, IEEE J. Select. Areas Comm., 9 (September 1991) 1078–1092.

[12] J.M. Hyman, A.A. Lazar and G. Pacifici, Real time scheduling with quality of service constraints, IEEE J. Select. Areas Comm., 9 (September 1991) 1052–1063.

[13] S. Gupta, K.W. Ross and M. El Zarki, Routing in virtual path based ATM networks, Proc. IEEE Globecom '92, Orlando, FL, December 1992, pp. 571–575.

[14] S. Gupta and P.P. Gandhi, Dynamic routing in multi-class non-hierarchical networks, Proc. IEEE Internat. Conf. Commun., ICC '94, New Orleans, LA, May 1994, pp. 1390–1394.

[15] L. Gün, V.G. Kulkarni and A. Narayanan, Bandwidth allocation and access control in high speed networks, Ann. Op. Res., 49 (1994) 161–183.

[16] Z. Dziong, J. Choquette, K.Q. Liao and L. Mason, Admission control and routing in ATM networks, Comput. Networks and ISDN Syst., 20 (December 1990) 189–196.

[17] J.M. Hyman, A.A. Lazar and G. Pacifici, A separation principle between scheduling and admission control for broadband switching, IEEE J. Select. Areas Comm., (May 1993) 605–616.

[18] R. Bolla, F. Danovaro, F. Davoli and M. Marchese, An integrated dynamic resource allocation scheme for ATM networks, Proc. IEEE Infocom '93, San Francisco, CA, March 1993, pp. 1289–1297.

[19] R. Bolla, F. Davoli, A. Lombardo, S. Palazzo and D. Panno, Hierarchical dynamic control of multiple traffic classes in ATM networks, Euro. Trans. Telecomm., 5(6) (November 1994) 747–755.

[20] R. Bolla, F. Davoli and M. Marchese, A distributed routing and access control scheme for ATM networks, Proc. IEEE Int. Conf. Commun., ICC '94, New Orleans, LA, May 1994, pp. 44–50.

[21] R. Bolla, F. Davoli and M. Marchese, Performance of hop-by-hop distributed routing and resource allocation in an ATM network, Proc. Inter. Conf. on Comput. Commun. and Networks, ICCCN '94, San Francisco, CA, September 1994, pp. 235–241.

[22] M. De Prycker, Asynchronous Transfer Mode: Solution for Broadband ISDN (2nd ed), Ellis Horwood, 1993.

[23] C.C. Huang and A. Leon-Garcia, Separation principle of dynamic transmission and enqueueing priorities for real- and non-real-time traffic in ATM multiplexers, IEEE/ACM Trans. Networking, 2 (December 1994) 588–601.

[24] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE/ACM Trans. Networking, 1 (June 1993) 329–343.

[25] S. Jordan and P. Varaiya, Throughput in multiple service, multiple resource communication networks, IEEE Trans. Comm., 39 (August 1991) 1216–1222.

[26] S. Jordan and P. Varaiya, Control of multiple service, multiple resource communication networks, IEEE Trans. Comm., 42, (November 1994) 2979–2988.

[27] M. Ritter and P. Tran-Gia (eds), Multi-rate models for dimensioning and performance evaluation of ATM networks, COST 242 Interim Report, June 1994.

[28] T. Yang and D.H.K. Tsang, A novel approach to estimating the cell loss probability in an ATM multiplexer loaded with homogeneous on-off sources, IEEE Trans. Comm., 43 (January 1995) 117–126.