ABSTRACT

The statistical multiplexing operation within an ATM network node is considered, with respect to different methods for the allocation of the bandwidth of an outgoing link. Service separation is assumed by dividing the overall traffic flows into classes, homogeneous in terms of performance requirements and statistical characteristics, which share the bandwidth of a link according to some specified policy. This context allows one to clearly define, by means of several existing approaches, a region in the space of connections of the different classes (call space) where quality of service (QoS) requirements at the cell level are satisfied. Within this region, some criteria for allocating the bandwidth of the link to the service classes are proposed, and the resulting allocation and call admission control (CAC) strategies are defined and analyzed. The goal of these operations is to achieve some desired QoS at the call level. Several numerical simulation results are presented, in order to highlight the different performance characteristics of the various methods.

Bandwidth Allocation and Admission Control in ATM Networks with Service Separation

Raffaele Bolla, Franco Davoli, and Mario Marchese, University of Genoa

The multiservice structure of broadband networks, where quality of service (QoS) requirements with possibly very different characteristics must be satisfied for traffic flows with different statistical natures, has led to the application of specific control actions whose main goals consist of maintaining the desired QoS levels and maximizing the utilization of network resources (or some revenue function). This feature, being connected to service integration, is not necessarily related to asynchronous transfer mode (ATM), but it is present in time-division multiplex (TDM) networks as well. As a matter of fact, bandwidth allocation and call admission control (CAC) problems already arise in the integrated services digital network (ISDN) environment (see, e.g., [1–15] for this type of control problem in that context). The control strategies in this case are related to QoS at the call level, in terms of both cost or revenue functions and constraints (e.g., on call blocking probability).

The statistical multiplexing character of ATM adds another control level to this scenario, because it further requires the satisfaction of cell-level constraints. Except by peak bandwidth assignment to each virtual circuit (VC) connection, there is no immediately clear way to identify the amount of resources (buffer space and bandwidth) required by the connection during its lifetime. This has given rise to many investigations on the bandwidth allocation problem (and the related CAC strategies; see [16–29], among others). A number of works deal, in this respect, with the notion of equivalent bandwidth [20–28]; this binds the cell-level QoS guarantees and the statistical characteristics of a source with a bandwidth requirement (within known limits), thus allowing a separation between the cell- and call-level control problems.

Among the various approaches, some impose a certain structuring on the allocation of the resources, where typically services are subdivided into classes, which are homogeneous in terms of performance requirements and/or statistical characteristics of the traffic sources, and bandwidth is allocated accordingly, thereby restricting the statistical multiplexing only within each class; this does not sensibly degrade performance in the case of services with largely different QoS or statistical natures ([1, p. 141 and references therein]; see also [30, 31] for analyses of the multiplexing gains). This approach (service separation) has been adopted in various forms in a number of works in the literature ([1, 32–41], among others). This further eases the decomposition of a very complex overall control task, which is in general characterized by very different time scales and requirements, according to the level where the system dynamics is considered (e.g., cell and call level), into smaller and somehow independent problems. For instance, an essential decoupling between cell and call level is achieved in [34, 35] through the concept of Schedulable Region, whereas a hierarchical decomposition has been used by the authors in previous works [39, 40].

By adopting this philosophy, we distinguish and compare some bandwidth allocation and related CAC strategies at an ATM multiplexer serving a link with a given total transfer capacity, which essentially differ in the way the link bandwidth is allocated among the service classes (and, accordingly, the CAC operation is performed). All the strategies we describe in this article can be interpreted as defining partitions of the region in call space within which QoS constraints at the cell level are satisfied (referred to as the "feasibility region," FR, in the following). Moreover, they belong to the family of complete partitioning strategies [1], where the bandwidth of the link is entirely divided among the classes in a static way. However, in our approach we shall see that quasi-static partitions may be obtained by successive adjustments based on traffic measurements.

Although the focus will be on the above-mentioned partitioning schemes, we will also need an algorithm to ensure the satisfaction of QoS requirements at the cell level within each

0163-6804/97/$10.00 © 1997 IEEE

service class; typically, satisfying these requirements imposes a limitation on the number of acceptable connections for each class, which in turn determines the boundary of the FR. In all our methods but one there is no interaction between this cell-level control and the partitioning and CAC level above it. In general, any methodology (in particular, any based on equivalent bandwidth) could be applied; we will briefly describe the one introduced in [39, 40], which will also be used to obtain most numerical results. It takes into account constraints on cell loss probability and cell delay, and it yields the maximum number of connections that each class can support (i.e., the points on the boundary of the FR).

The article is organized as follows. The next section is devoted to a brief discussion of service separation and related bandwidth allocation and admission control schemes, as well as to the cell-level methodology for evaluating the FR. Four different quasi-static complete partitioning schemes are described in the third section, and their performance is investigated by simulation in the fourth section. The final section contains the conclusions.

Cell- and Call-Level Control Problems under Service Separation

We suppose the traffic in the network to be divided into H classes of bursty (on-off) sources, each class being characterized by statistical parameters like peak rate, average transmission rate, and average burst length, as well as by QoS requirements like cell loss probability and cell delay. We indicate with $B^{(h)}$(cells), $P^{(h)}$(b/s), $M^{(h)}$(b/s) and $b^{(h)} = P^{(h)}/M^{(h)}$, the average burst length, peak bit rate, average bit rate, and burstiness, respectively, of a source of the hth class. Moreover, let $\lambda^{(h)}$ and $1/\mu^{(h)}$ represent the average arrival rate and average duration of connections of class h, respectively, and $\rho^{(h)} = \lambda^{(h)}/\mu^{(h)}$.

At each ATM multiplexer, traffic class h is assigned a separate buffer of length $Q^{(h)}$(cells), whose output is statistically multiplexed on the outgoing link by a scheduler, which substantially divides the global channel capacity C(b/s) into "virtual" partitions $C^{(h)}$ among the classes, whose sum amounts to C. The simplest way to maintain the partitions is by serving the buffer in a weighted round-robin fashion; other possibilities include the assignment of a slot (time to transmit a cell) to a cell of class h randomly, with probability $\Omega^{(h)} = C^{(h)}/C$ [39], or the use of a technique such as generalized processor sharing [42].

Connection requests, which can come from the users directly connected to the node or from other nodes, are also processed on a per-class basis. Given a model for the traffic sources of a class, the cell-level performance requirements (e.g., in terms of average cell loss and delayed cell rate) allow a region to be defined in call space (which, as we mentioned in the introduction, is the FR), where they are certainly satisfied. This region corresponds to the CAC method named "service separation with dynamic partitions" in [1, p. 147]. In a network, one such region can be associated with each link.

Clearly, the points on the boundary of the FR correspond to the maximum numbers of VC connections $[N_{\max}^{(1)}, \ldots, N_{\max}^{(H)}]$ that are compatible with the given cell-level QoS constraints. We can associate each $N_{\max}^{(h)}$ with the minimum amount of bandwidth $C_{\min}^{(h)}$ that is necessary to support that number of connections with the given QoS guarantees.

The computation of the FR has been the object of several studies and can be effected in different ways, either by analysis, given a model of the traffic sources, or by simulation. A more general view includes the characteristics and the optimization of the scheduling process [34], through the related concept of "schedulable region." Using any approach based on equivalent bandwidth (involving, in our service separation context, only homogeneous sources) yields a straightforward boundary of the FR. In any case, it is worth noting that, in the context of the methods to be considered in the next section, the FR itself will just be a tool to describe the CAC schemes. The specific technique to ensure QoS satisfaction at the cell level might be changed (always within the framework of service separation) without affecting the general access control procedure.

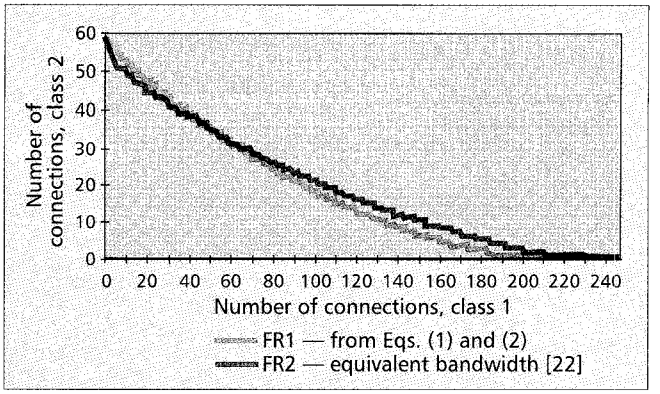


Figure 1. *Feasibility regions for traffic classes 1 and 2.*

However, to fix ideas for the computation of the FR we refer here to the model we have used in [40] and in previous works, where a maximum threshold value is set for the average cell loss rate ($P_{\text{loss}}^{(h)}(n, C^{(h)})$) and for the average delayed cell rate ($P_{\text{delay}}^{(h)}(n, C^{(h)})$), with n calls in the active state out of $N^{(h)}$ accepted calls and a bandwidth $C^{(h)}$ assigned to traffic class h; more specifically,

$$\sum_{n=1}^{N^{(h)}} P_{\text{loss}}^{(h)}(n, C^{(h)}) V_{n,N^{(h)}} \le \varepsilon^{(h)} \quad (1)$$

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

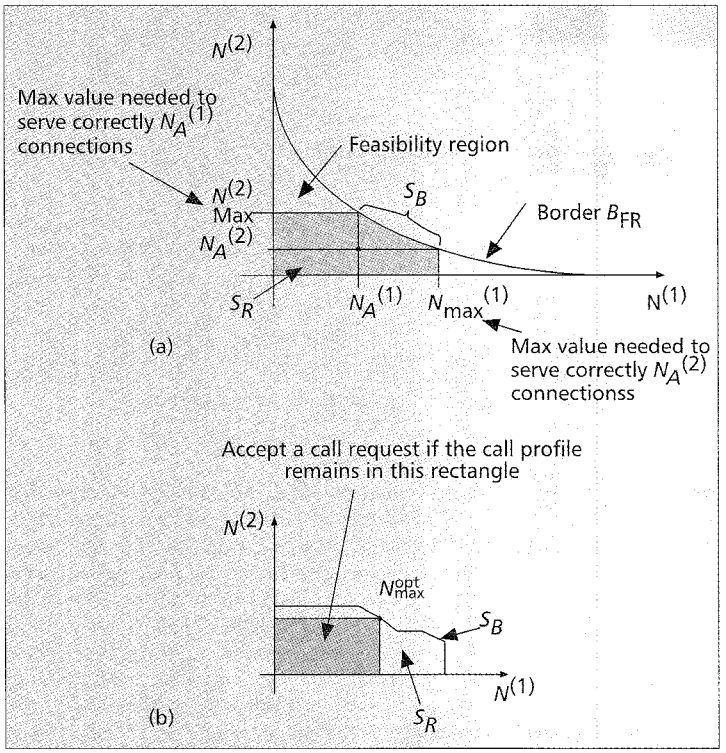
where $\varepsilon^{(h)}$ is an upper limit on the average value of the cell loss rate, and $\delta^{(h)}$ has the same meaning for the average value of cells that suffer a delay longer than a fixed upper bound ($D^{(h)}$ in the fourth section). The quantity $V_{n,N^{(h)}}$ is the probability of having n active connections of class (h), with $N^{(h)}$ accepted connections of the same class, which is given by a binomial distribution; an interrupted Bernoulli process (IBP) [43] is used to model the state of a call, from which $P_{\text{loss}}^{(h)}(n, C^{(h)})$ and $P_{\text{delay}}^{(h)}(n, C^{(h)})$ are derived. $N_{\max}^{(h)}$ and $C_{\min}^{(h)}$, $h = 1, \ldots, H$ can be computed from Eqs. 1 and 2.

As an example, the two FRs derived from the above method and from the equivalent bandwidth approach of [22] (labeled FR1 and FR2, respectively), are represented in Fig. 1 for two traffic classes whose parameter values are given in the fourth section. We stress again that a comparison between the two computation methods is not relevant in our context. Actually, in the particular case of our numerical values, there is a certain advantage in using the equivalent bandwidth approach, whereas other parameter combinations would enhance the previous method; it should also be kept in mind that the two approaches are not completely homogeneous, owing to the presence of constraint Eq. 2.

We let

$$\mathbf{N}_A(k) = \text{col}[N_A^{(h)}(k), h = 1,\ldots,H] \quad (3)$$

where $N_A^{(h)}(k)$ is the number of connections in progress at the generic instant (slot) k for class h. The vector in Eq. 3 repre-



■ **Figure 2.** *a) The effects of the constraints of Eq. 4 on the feasibility region; and b) a corresponding optimization result.*

sents the state of the system at instant k (the VC-profile in [1]).

At this point, we are ready to consider QoS performance measures and constraints at the call level, having essentially decoupled this problem from the lower-level one. Obviously, a possible way of performing the CAC operation is that of accepting an incoming call as long as the VC-profile is within the FR. This, however, does not allow us to take into account specific call-level requirements. Again, there are several methods that can be used; [1, Ch. 4] contains an excellent overview. In the next section, we examine four different criteria, which, however, are limited to the class of policies known as complete partitioning (CP) that define a "rectangular" subregion within the FR. In doing this, we consider the possibility of periodically changing the bandwidth allocated to the various classes (i.e., the vertex of the "rectangle" touching the boundary of the FR) in order to take into account variations in the traffic intensities and their ratios among the classes. We do this in two different ways, which may be viewed as embedding the CP strategy within a two-level hierarchical control scheme and as a certainty equivalent parameter adaptive control strategy, respectively.

Bandwidth Allocation Strategies

As already mentioned, we propose four different CP schemes, whose common goal is the computation of the optimal (with respect to a different criterion within each scheme) values $\mathbf{N}_{max}^{opt} = col[N_{max}^{(h),opt}, h = 1, ..., H]$ of the points $\mathbf{N}_{max} = col[N_{max}^{(h)}, h = 1, ..., H]$ on the boundary B_{FR} of the feasibility region and of the corresponding capacity vector $\mathbf{C}^{opt} = col[C^{(h),opt}, h = 1, ..., H]$, which characterize the CAC rules. Each method is based on the minimization of a specific cost function, possibly under a set of constraints.

Before describing each method in some detail, we outline the general control architecture within which they are embedded. Let t_k, $k = 0, 1, ...$ represent the initial time instant (in terms of the discrete time unit, or slot) of an "epoch," during which the bandwidth allocated to each class and the parameters of the corresponding CAC rule are kept constant. In practice, if Δ is the number of time slots required by the computation of the bandwidth partitions of a particular method, the same parameter values hold from slot $t_k + \Delta$ to slot $t_{k+1} - 1 + \Delta$. We will refer to the time instants t_k, $k = 0, 1, ...$ as "reallocation instants." We shall distinguish two cases in the following, namely, periodic and asynchronous reallocation instants; in any case, the duration of the interval $t_{k+1} - t_k$ will typically be several orders of magnitude larger than the slot. For the time being, we only note that, whenever a reallocation is performed online, when the system is in operation, there is a set of "dynamic" constraints that must be taken into account. More specifically, in order to ensure the correct service continuation of all the connections already in progress at time t_k we must have

$$N_{max}^{(h)}(t_k + \Delta) \geq N_A^{(h)}(t_k), \quad \begin{matrix} h = 1,...H \\ k = 0,1,... \end{matrix} \tag{4}$$

in the call space or

$$C^{(h)}(t_k + \Delta) \geq C_{lb}^{(h)}(t_k), \quad \begin{matrix} h = 1,...H \\ k = 0,1,... \end{matrix} \tag{5}$$

in the capacity space; here, $C_{lb}^{(h)}(t_k)$ is the lower bound on the capacity needed to ensure the QoS for the $N_A^{(h)}(t_k)$ class h calls in progress at instant t_k, and we have obviously considered all the previously defined quantities as time-dependent. The effect of these constraints is the reduction of the set of values of $\mathbf{N}_{max}$ to be considered in the optimization to a sub-set $S_B(t_k)$ of the boundary B_{FR} of the feasibility region, as shown in Fig. 2 (where the time indices have been dropped for simplicity). Correspondingly, only a subset $S_R(t_k)$ of the feasibility region need be considered.

Having observed this, we can now turn to the definition of the optimization criteria to be used. In the first three methods we describe next, named the Erlang scheme (ES), the balanced Erlang scheme (BES), and the constrained Erlang scheme (CES), we suppose that the interarrival times of the call requests and the duration of connections of class h follow exponential distributions, with average values $1/\lambda^{(h)}$ and $1/\mu^{(h)}$, respectively. Thus, in these cases, as a consequence of the usage of the service separation and the CP policy, we can easily compute, independently for each class h, the stationary blocking probabilities $P_B^{(h)}(N_{max}^{(h)})$, $h = 1, ..., H$ by using the Erlang B formula [1]. By supposing a slow variation of the rates $\lambda^{(h)}$, $\mu^{(h)}$ with respect to the call request dynamics and reallocation interval, we utilize an infinite horizon cost formulation (i.e., averaging according to stationary distributions) and define three different cost functions, based on the quantities $P_B^{(h)}(N_{max}^{(h)})$, $h = 1, ..., H$.

In the ES, the maximum numbers of acceptable calls $\mathbf{N}_{max}^{opt}$ are computed through the minimization of a weighted sum of the per-class blocking probabilities, namely

$$\mathbf{N}_{max}^{opt} = \arg\left\{ \min_{N_{max} \in S_B} P_B(\mathbf{N}_{max}) \right\} \tag{6}$$

where

$$P_B(N_{max}) = \sum_{h=1}^{H} \alpha^{(h)} P_B^{(h)}(N_{max}^{(h)}) \tag{7}$$

and the $\alpha^{(h)}$ are weighting coefficients. If

$$\alpha^{(h)} = \lambda^{(h)} \Big/ \sum_{k=1}^{H} \lambda^{(k)},$$

Eq. 7 represents the average blocking probability of the multi-

plexer. Clearly, in this case, the main goal is to minimize the total blocking probability of the system, giving also the capability to assign a sort of priority to some traffic classes through the weights $\alpha^{(h)}$, $h = 1, \ldots, H$.

On the other hand, in the BES the main goal is the achievement of an equalization of the blocking probabilities over the classes, and this result is obtained by using the following cost function:

$$P_B(\mathbf{N}_{\max}) = \max_h \left\{ \alpha^{(h)} P_B^{(h)}(\mathbf{N}_{\max}^{(h)}) \right\} \quad (8)$$

so that

$$\mathbf{N}_{\max}^{\text{opt}} = \arg \left\{ \min_{\mathbf{N}_{\max} \in S_R} P_B(\mathbf{N}_{\max}) \right\} \quad (9)$$

Again, the weights $\alpha^{(h)}$ offer the opportunity to change the relative "importance" of each class.

The third scheme we present is defined by adding a (stationary) constraint on the maximum call blocking probability for each class; that is,

$$P_B^{(h)}(N_{\max}^{(h)}) \text{ " } \gamma^{(h)}, \; h = 1, \ldots, H \quad (10)$$

Let $\overline{\mathbf{N}} = \text{col}[\overline{N}^{(h)}, h = 1, \ldots, H]$ be the equality solution of Eq. 10 with respect to $\mathbf{N}_{\max}$. To satisfy the constraints in Eq. 10 we must then have

$$N_{\max}^{(h)} \geq \overline{N}^{(h)}, \; h = 1, \ldots, H \quad (11)$$

We can distinguish two cases: $\overline{\mathbf{N}} \in S_R$ and $\overline{\mathbf{N}} \notin S_R$. In the first, a subregion that satisfies the constraints (shaped as shown in Fig. 3a for a system supporting two classes) can be found; in the other case, the subregion does not exist (Fig. 3b). On the basis of these considerations, we compute $\mathbf{N}_{\max}^{\text{opt}}$ as the vector minimizing the blocking probability $P_B(\mathbf{N}_{\max})$ in Eq. 7 within the subregion of S_R, where the constraints of Eq. 10 are satisfied if $\overline{\mathbf{N}} \in S_R$; otherwise, we try to approximate the constraints as closely as possible by choosing $\mathbf{N}_{\max}$ so as to minimize the sum (over the classes) of quadratic deviations between the blocking probability of each class and the corresponding constraint value. In other words, we may interpret this operation as looking for the $\mathbf{N}_{\max}$ at minimum "distance" (in the above sense) from $\overline{\mathbf{N}}$ (Fig. 3b).

At this point, we can specify how the reallocation is performed in conjunction with the three methods above. In this case, the parameters $\lambda^{(h)}$ and $\mu^{(h)}$ of the connections can be continuously estimated and, whenever a significant change is observed, the reallocation can be called upon to compute new values of $\mathbf{N}_{\max}^{\text{opt}}$ and corresponding bandwidth partitions. Thus, the overall scheme can be viewed as a parameter adaptive certainty equivalent control [44], and the reallocation is normally effected at asynchronous instants.

Our fourth method, which will be referred to as DRS (dynamic reallocation scheme) in the following, differs from the above-described ones in some respects. First of all, no assumption is made on the statistical nature of the call arrivals and durations; this fact does not allow defining any cost function averaged over a stationary probability distribution of the call process. We therefore change our view of the cost function, and also the general organization of the control architecture. The reallocation is made periodic, and the bandwidth partitions can be interpreted as coordination variables in a two-level hierarchical scheme with repetitive control [39, 40] (the "lower" level consists of the per-class admission control).

In this respect, the cost function to be used in the bandwidth allocation process is now based on the same quantities that have been introduced in the previous section in conjunction with the cell-level control problem. More specifically, it involves an estimate of the overall cell loss rate that would be generated by the total offered load, measured over the previous interval. The basis for this computation is given by the expression on the left side of Eq. 1. However, a linear combination of two different sums is constructed:

- One extending up to the average number of connections in the system over the previous interval
- Another extending over the total number of connection requests (either accepted or rejected), still averaged over the previous interval

The aim is to obtain a quantity that reflects the variation in cell loss rate which would have been incurred by accepting all incoming calls and, in turn, to allow detection of changes in the load proportions among the classes. The detailed expression of the cost function is not reported here (see [39, 40]). The length of the reallocation period can be chosen as a compromise between two requirements: that the interval be long enough to contain a reasonable number of sample points (of the call arrival process) with a given probability, but not "too long," in order to allow a fast reaction to sudden load variations. In any case, its length will typically be comparable to the call time scale, which is some orders of magnitude larger than the cell scale.

Two facts are worth noting that are common to all the methods we have described: i) the reallocation operation gives rise to different complete partitioning "rectangular" regions, with a vertex moving on the boundary of the FR to adapt to load variations (see Fig. 4); and ii) the admission control rule for each class h is very simple, involving only a comparison of the current number of connections plus the new one with the current value of $\mathbf{N}_{\max}^{(h),\text{opt}}$; new values of these quantities have to be computed only at the beginning of a reallocation period or epoch. From the point of view of implementation, this computation is the heaviest one, and would deserve careful investigation.

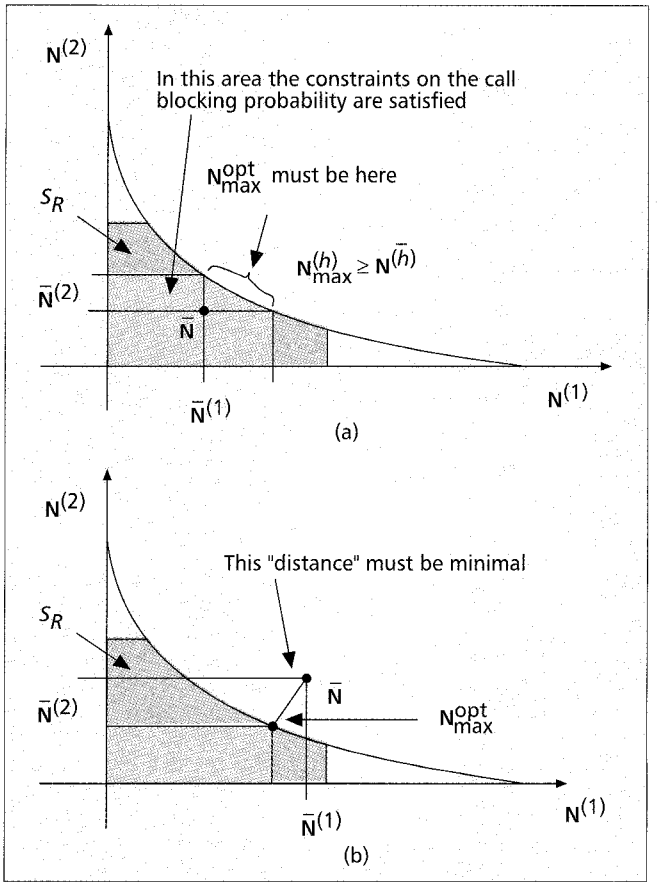
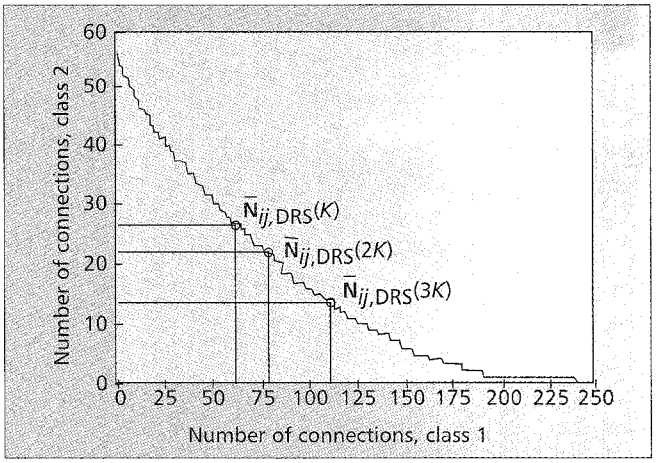


Figure 3. *The effect of the constraints in Eq. 10 on the acceptance region a) in the case $\overline{N} \in S_R$; and b) in the case $\overline{N} \notin S_R$.*



■ **Figure 4.** *Acceptance regions in three different reallocation instants.*

Traffic class: h	$h = 1$	$h = 2$
Peak bandwidth: $\rho^{(h)}$	2 Mb/s	10 Mb/s
Burstiness: $b^{(h)}$	5	10
Average burst length: $B^{(h)}$	500 cells	1000 cells
Average connection duration	15 s	25 s
P_{loss} upper bound: $\varepsilon^{(h)}$	0.0001	0.0001
P_{delay} upper bound: $\delta^{(h)}$	0.001	0.001
Delay constraint: $D^{(h)}$	200 slots	1000 slots
Buffer length: $Q^{(h)}$	15 cells	10 cells

■ **Table 1.** *Parameter values.*

In all schemes considered, it requires performing a number of descending steps of a mathematical programming method for the parametric minimization of the specific cost function. In this respect, the DRS cost function evaluation is relatively more complex than those of the other methods. Nevertheless, it has already been noted that the length of the reallocation periods should at least be comparable to the average time between successive events in the call processes, or even span several such time intervals in most cases; moreover, call requests coming in during the computation may be temporarily held until its completion, and several other possibilities exist, along similar lines to [15, App. C], to improve the feasibility of this procedure.

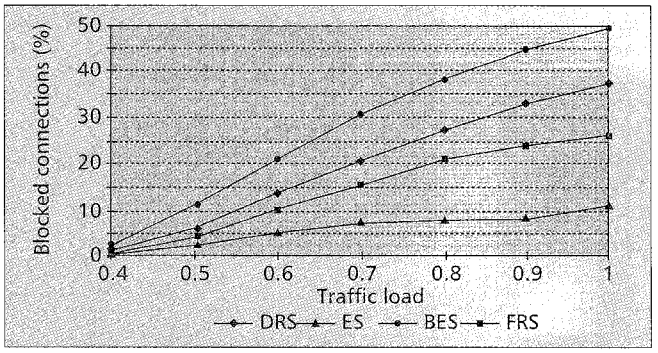
Simulation Results

The first part of this section is aimed at showing the differences among the various strategies introduced previously. The schemes are tested with two service classes, over a range of the traffic loads $\rho^{(h)}$, $h = 1, 2$. The overall channel capacity is $C = 150$ Mb/s.

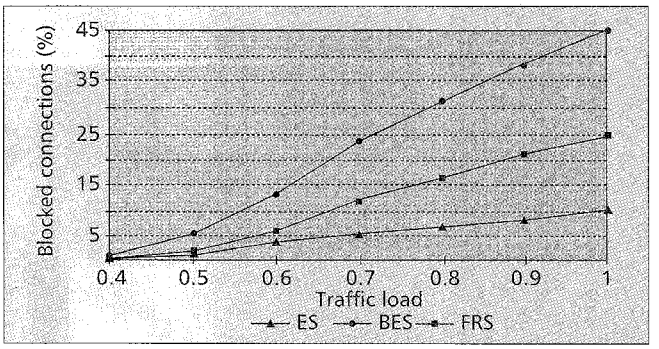
We normalize the traffic flow to that generated with the data specified in Table 1, when $\rho^{(1)} = 200$, $\rho^{(2)} = 30$; we refer to this load as unity. An offered load "x" corresponds to the same data, except for the traffic intensities, which are multiplied by x. The bandwidth reallocation for the DRS is performed at fixed instants (t_k), every 905.6 s (about 15 minutes). A simulation run for each constant traffic load value lasts 48 of these reallocation intervals (obviously, in this case, the reallocations are performed only for the DRS). So, the duration of the overall simulation is 43,468.8 s (i.e., about 12 hr network time). This duration guarantees a high level of reliability of the measured values; more specifically, the results have a 95 percent confidence interval around 3 percent of the estimated value, or less.

Figures 5–9 are dedicated to this first part. The second part of the results has a different aim. Specifically, the object is to simulate time-varying traffic conditions, as may happen in a real situation; therefore, the traffic load is changed during the simulation every 8 reallocation intervals (2 hr network time), as it might do in a real network depending on the time of day. The goal is to get some indications on the response time of each mechanism. In this context, the bandwidth assigned at each reallocation is also shown. As in the previous case, the duration of each simulation corresponds to about 12 hr network time. This division of time is well suited, in this case, to approximating the behavior during a day. Figures 10–14 are related to this second part. The data used for both parts are the same (except for the traffic loads) and are summarized in Table 1.

Figure 5 shows the global percentage of blocked connections for the strategies DRS, ES, BES, and FRS. The last acronym stands for "feasibility region scheme," which is a simple strategy where a connection is accepted whenever the system's state remains within the FR after acceptance. It is presented here just to allow a further comparison among the methods. FR1, stemming from Eqs. 1 and 2, has been used. As can be seen from the figure, the ES provides the best results, and the overall percentage of blocked calls is relatively small even at high load. On the other hand, the worst result is obtained by using the BES scheme, whose aim is not the minimization of the global percentage of blocked connections, but rather balancing the blocking probability between the classes. The results concerning the CES mechanism are not reported



■ **Figure 5.** *Overall percentage of blocked connections (DRS, ES, BES, FRS), in conjunction with FR1.*



■ **Figure 6.** *Overall percentage of blocked connections (ES, BES, FRS) in conjunction with FR2.*

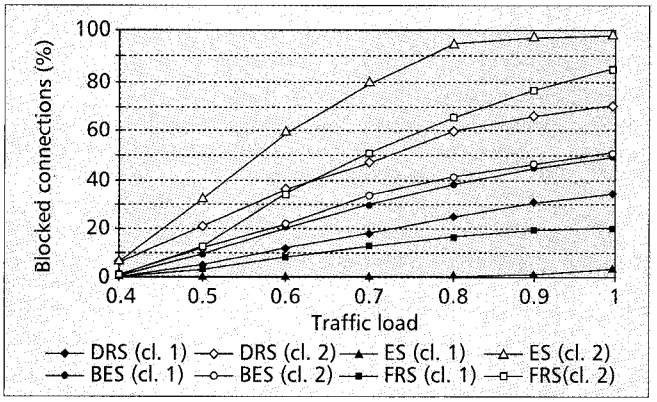


Figure 7. *Percentage of blocked connections for each class (DRS, ES, BES, FRS) in conjunction with FR1.*

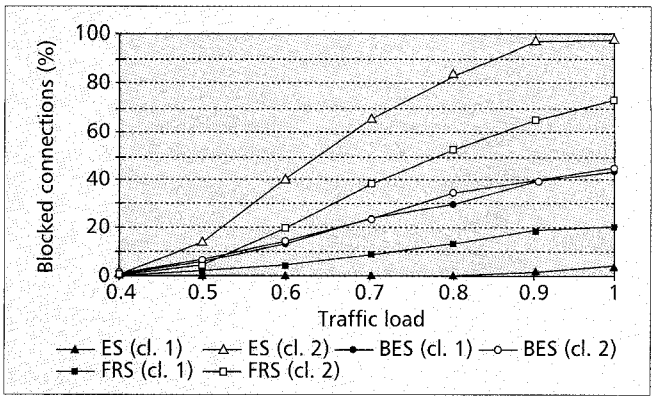


Figure 8. *Percentage of blocked connections for each class (ES, BES, FRS) in conjunction with FR2.*

in this figure, as they would not be meaningful in this form, due the fact that the constraints are referred to each traffic class and not to the overall percentage.

To see the effect of a change in the shape of the FR on the CAC strategies, the graphs referring to ES, BES, and FRS are represented again in Fig. 6, this time by using FR2 produced by equivalent bandwidth. The DRS is not reported in this case, since its call-level cost is anyway intimately connected with Eqs. 1 and 2, as noted in the previous section. It can be seen from this figure and the preceding one that the shape of the FR (depending on the accuracy of the specific computation method and of the source models) obviously affects the performance of the CAC strategies. However, two facts are worth noting:

- The qualitative behavior with increasing load is the same.
- As far as CP strategies are concerned, this effect is limited to the point of the boundary where the vertex of the "rectangle" is positioned.

Actually, the BES, whose vertex sticks to the right of the FR (where FR2 has the advantage, as can be seen from Fig. 1) at high load, exhibits a gain in Fig. 6 over Fig. 5; on the other hand, the ES, which is placed almost at the crosspoint of the two boundaries in Fig. 1, maintains an almost equal call-level performance.

The small number of blocked calls for the ES scheme is "paid" for by a strong unbalance between the traffic classes. This can be seen in Fig. 7, where the percentage of blocked connections is shown for the same strategies as in Fig. 5, but where the values referred to each traffic class are presented. The BES strategy provides really balanced results, as was the aim of this scheme. The same comments as above can be made regarding Fig. 8, where the same situation is reported with respect to FR2 (except the DRS).

Having established the comparison with respect to the choice of FR, to fix ideas from now on we will refer to the specific choice of FR1 and concentrate on the different behavior of the various CAC and bandwidth allocation methods we have introduced. The same quantities as in Fig. 7 are reported in Fig. 9 for various configurations of the CES scheme. Some explanation of the notation is necessary: CES ($x\% - y\%$) means that the upper bounds for the percentage of blocked connections are x (concerning class 1) and y (concerning class 2), respectively. Referring to the theory in the previous section, this means that $\gamma^{(1)} = 0.01 \cdot x$ and $\gamma^{(2)} = 0.01 \cdot y$ (recall that $\gamma^{(h)}$ is the limit on the call blocking probability of class h).

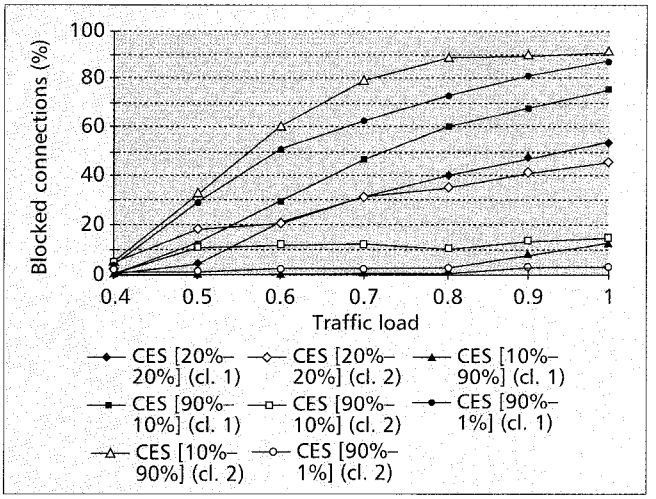


Figure 9. *Percentage of blocked connections for each class (CES).*

It can be seen that, concerning CES (20% – 20%), the constraints are respected as far as possible (load 0.4 and 0.5); then the strategy tends to maintain the losses of the two classes near the constraint value: the final effect is an actual balance between the two classes. Concerning the other configurations, the constraints are basically respected. In order to appreciate this property even at very high load, one of the traffic classes has been given a very "loose" constraint (90 percent); in this case, some room is always left to satisfy the more stringent constraint of the other.

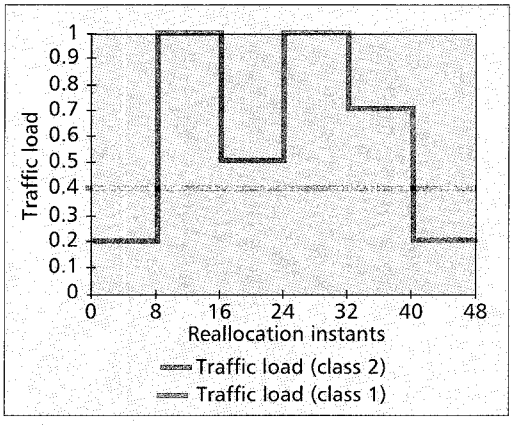


Figure 10. *Traffic configuration.*

We now turn to the second part of results, under the time-varying load configuration, which is shown in Fig. 10. The load of class 2 remains constant for the entire duration of the simulation, whereas the load of class 1 varies every 8 reallocation instants (2 hr). For homogeneity and better comparison, the adaptive schemes (ES, BES, CES) are also updated regularly at each reallocation interval of the DRS.

Figures 11 and 12 are aimed at showing the reaction of the system to traffic variations. Figure 11 shows

the values of the maximum number of acceptable connections for each traffic class at each reallocation instant (for DRS, ES, and BES). Clearly, the values are different for every strategy, but the response of the system to a load step variation is noticeable. The same behavior can be found in Fig. 12, where the bandwidth allocated to each class at each reallocation instant is shown for the same strategies. A similar behavior can be obtained for the CES scheme.

The last figure (Fig. 13) shows the average percentage of blocked calls under the time-varying load configuration for both the single classes and the overall number. As in Fig. 5, even in this case the aim is to show the different behavior of each strategy. Because the traffic conditions are so varying, the global behavior is different than in Fig. 5, but the same general considerations can be made. The balancing effect of BES, DRS, and CES (20% – 20%) is noticeable, as well as the low total number of blocked connections and the large unbalance of ES. The CES strategy is really effective (CES (90% – 10%) and (90% – 1%), in Fig. 13), if the attainment of a precise bound on the call blocking probability of each class is allowed, even at very high load; both stringent requirements (10% and 1%) are, in fact, respected.

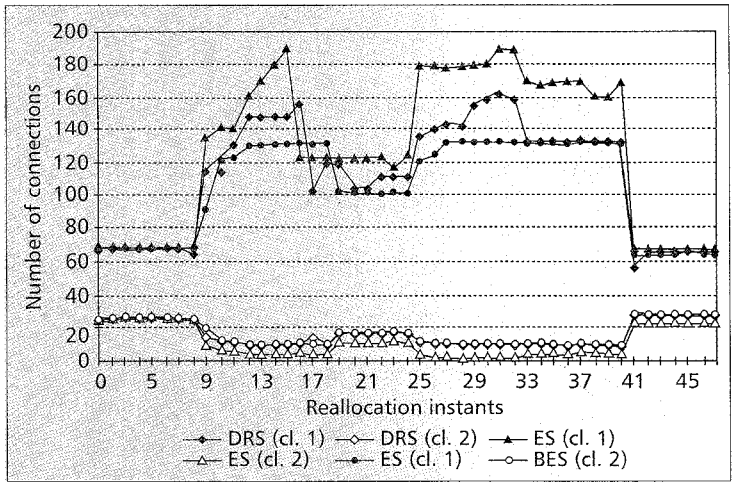


Figure 11. *Maximum number of connections vs. reallocation instants.*

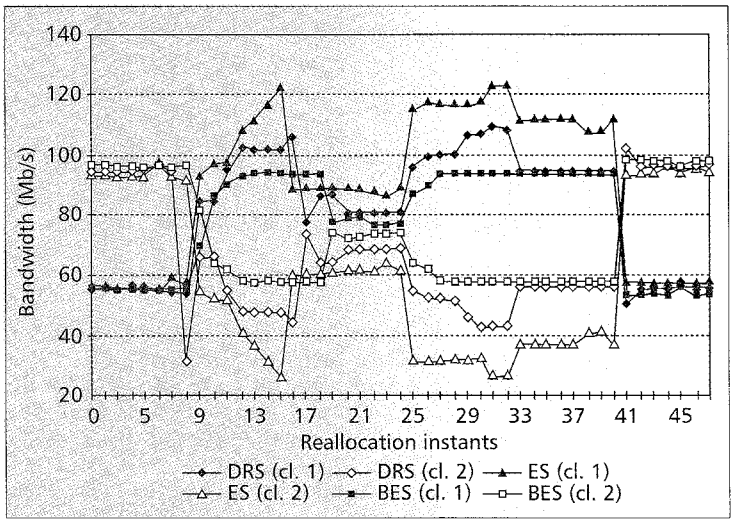


Figure 12. *Allocated bandwidth vs. reallocation instants.*

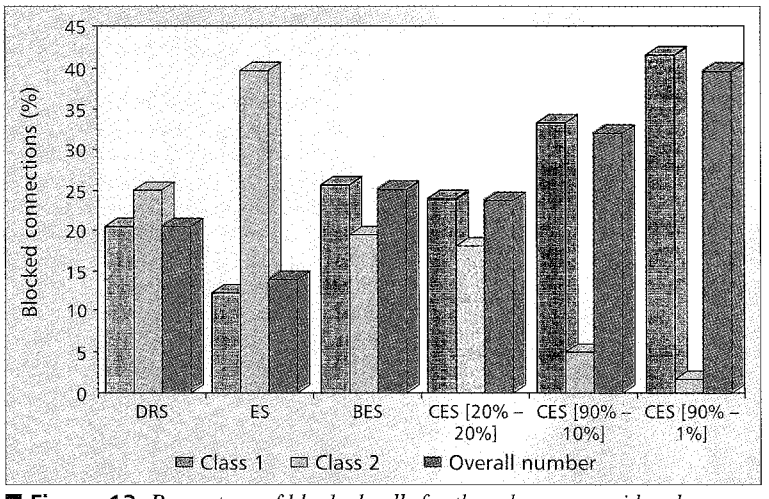


Figure 13. *Percentage of blocked calls for the schemes considered.*

Conclusions

We have examined and compared different bandwidth allocation and call admission control strategies for ATM networks. Our approach has been in the context of service separation, where traffic sources have been grouped into classes, homogeneous in terms of performance requirements and statistical characteristics. Given the cell-level constraints for each class (namely, upper bounds on cell loss and delay probabilities or rates), a region in the space of connections ("call space") of each class can be easily defined, where a point represents a combination of the numbers of connections of each class that satisfy the constraints. Within this "feasibility region" (FR), we have considered different bandwidth allocation (and the corresponding CAC) schemes, all belonging to the family of complete partitioning strategies, which define a "rectangular" subregion with a vertex on the boundary of the FR.

All schemes have been embedded within an overall adaptive control architecture, whereby their parameters are recomputed, either periodically or at asynchronous instants, at a rate comparable with the time scale of the connection requests. The computational complexity of the overall procedure has been briefly discussed. The different performance characteristics have been compared by simulation, in both cases of static and time-varying traffic parameters. The results show very satisfactory behaviors in accordance with the intended goals.

Acknowledgment

This work was supported by the Italian Ministry of University and Scientific and Technological Research (MURST).

References

[1] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, London: Springer Verlag, 1995.

[2] B. S. Maglaris, M. Schwartz, "Optimal Fixed Frame Multiplexing in Integrated Line- and Packet-Switched Communication Networks," *IEEE Trans. Info. Theory*, vol. IT-28, Mar. 1982, pp. 283–73.

[3] B. Kraimeche and M. Schwartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks," *IEEE Trans. Commun.*, vol. COM-33, Oct. 1985, pp. 1085–93.

[4] I. Viniotis and A. Ephremides, "Optimal Switching of Voice and Data at a Network Node," *Proc. 26th IEEE Conf. Decision and Control*, Los Angeles, CA, Dec. 1987, pp. 1504–7.
[5] K. W. Ross and B. Chen, "Optimal Scheduling of Interactive and Noninteractive Traffic in Telecommunication Systems," *IEEE Trans. Auto. Control*, vol. 33, Mar. 1988, pp. 261–67.
[6] K. W. Ross and D. Tsang, "The Stochastic Knapsack Problem," *IEEE Trans. Commun.*, vol. 37, 1989, pp. 740–47.
[7] M. Zukerman, "Circuit Allocation and Overload Control in a Hybrid Switching System," *Comp. Networks and ISDN Sys.*, vol. 16, 1989, pp. 281–98.
[8] M. Zukerman, "Bandwidth Allocation for Bursty Isochronous Traffic in a Hybrid Switching System," *IEEE Trans. Commun.*, vol. COM-37, Dec. 1989, pp. 1367–71.
[9] K. W. Ross and D. H. K. Tsang, "Optimal Circuit Access Policies in an ISDN Environment: A Markov Decision Approach," *IEEE Trans. Commun.*, vol. COM-37, Sept. 1989, pp. 934–39.
[10] G. Meempat and M. Sundareshan, "Optimal Channel Allocation Policies for Access Control of Circuit-Switched Traffic in ISDN Environments," *IEEE Trans. Commun.*, vol. 41, Feb. 1993, pp. 338–50.
[11] M. Aicardi *et al.*, "A Parametric Optimization Approach to Admission Control and Bandwidth Assignment in Hybrid TDM Networks," *Int'l. J. Digital and Analog Commun. Sys.*, vol. 6, Jan. 1993, pp. 15–27.
[12] R. Bolla and F. Davoli, "Dynamic Hierarchical Control of Resource Allocation in an Integrated Services Broadband Network," *Comp. Networks and ISDN Sys.*, vol. 25, no. 10, May 1993, pp. 1079–87.
[13] A. Gavious and Z. Rosberg, "A Restricted Complete Sharing Policy for a Stochastic Knapsack Problem in a B-ISDN," *IEEE Trans. Commun.*, vol. 42, 1994, pp. 2375–79.
[14] R. Bolla and F. Davoli, "Call Admission Control and Bandwidth Allocation in a Multiservice DQDB Network," *Comp. Commun.*, vol. 18, no. 8, Aug. 1995, pp. 537–44.
[15] R. Bolla and F. Davoli, "Control of Multirate Synchronous Streams in Hybrid TDM Access Networks," to appear, *ACM/IEEE Trans. Networking*, 1997.
[16] G. Gallassi, G. Rigolio, and L. Fratta, "ATM: Bandwidth Assignment and Bandwidth Enforcement Policies," *Proc. GLOBECOM '89*, Dallas, TX, Nov. 1989, pp. 1788–93.
[17] G. M. Woodruff and R. Kositpaiboon, "Multimedia traffic management principles for guaranteed ATM network performance," *IEEE JSAC*, vol. 8, no. 3, April 1990.
[18] M. Decina and T. Toniatti, "On Bandwidth Allocation to Bursty Virtual Connections in ATM Networks," *Proc. ICC '90*, Atlanta, GA, Apr. 1990, pp. 844–51.
[19] J. A. S. Monteiro, M. Gerla, and L. Fratta, "Statistical Multiplexing in ATM Networks," *Proc. 4th Int'l. Conf. Data Commun. Sys. and Their Performance*, Barcelona, Spain, June 1990.
[20] F. P. Kelly, "Effective Bandwidths at Multi-Class Queues," *Queuing Sys.*, vol. 9, 1991, pp. 5–15.
[21] R. J. Gibbens and P. J. Hunt, "Effective Bandwidths for the Multi-Type UAS Channel," *Queuing Sys.*, vol. 9, 1991, pp. 17–27.
[22] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High Speed Networks," *IEEE JSAC*, vol. 9, no. 7, Sept. 1991, pp. 968–81.
[23] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Trans. Networking*, vol. 1, June 1993, pp. 329–43.
[24] G. Kesidis, J. Walrand, and C. S. Chang, "Effective Bandwidths for Multi-Class Markov Fluid and Other ATM Sources," *IEEE/ACM Trans. Networking*, vol. 1, Aug. 1993, pp. 424–28.
[25] L. Gün, V.G. Kulkarni, and A. Narayanan, "Bandwidth Allocation and Access Control in High-Speed Networks," *Ann. Op. Res.*, vol. 49, 1994, pp. 161–83.
[26] G. de Veciana, G. Kesidis, and J. Walrand, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidths," *IEEE JSAC*, vol. 13, Aug. 1995, pp. 1081–90.
[27] I. Hsu and J. Walrand, "Admission Control for ATM Networks," F. P. Kelly, R. J. Williams, eds., *Stochastic Networks, The IMA Volumes in Mathematics and its Application*, vol. 71, New York: Springer-Verlag, 1995, pp. 411–27.
[28] A. Elwalid, D. Mitra, and R. H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node," *IEEE JSAC*, vol. 13, Aug. 1995, pp. 1115–27.
[29] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Commun. Mag.*, vol. 34, no. 11, Nov. 1996, pp. 82–91.
[30] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the Most Out of ATM," *IEEE Trans. Commun.*, vol. 447, Feb. 1996, pp. 203–21.
[31] I. Sidhu and S. Jordan, "Multiplexing Gains in Bit Stream Multiplexors," *IEEE/ACM Trans. Networking*, vol. 3, Dec. 1995, pp. 785–97.
[32] Y. Takagi, S. Hino, and T. Takahashi, "Priority Assignment Control of ATM Line Buffers with Multiple QOS Classes," *IEEE JSAC*, vol. 9, Sept. 1991, pp. 1078–92.
[33] A. A. Lazar, A. Temple, and R. Gidron, "An architecture for Integrated Networks that Guarantees Quality of Service," *Int'l. J. Digital and Analog Commun. Sys.*, vol. 3, July 1990, pp. 229–38.
[34] J. M. Hyman, A. A. Lazar, and G. Pacifici, "Real Time Scheduling with Quality of Service Constraints," *IEEE JSAC*, vol. 9, Sept. 1991, pp. 1052–63.
[35] J. M. Hyman, A. A. Lazar, and G. Pacifici, "A Separation Principle between Scheduling and Admission Control for Broadband Switching," *IEEE JSAC*, vol. 11, May 1993, pp. 605–16.
[36] G. Gallassi, G. Rigolio, and L. Verri, "Resource Management and Dimensioning in ATM Networks," *IEEE Network*, May 1990, pp. 8–17.
[37] K. Sriram, "Methodologies for Bandwidth Allocation, Transmission Scheduling, and Congestion Avoidance in Broadband ATM Networks," *Comp. Networks and ISDN Sys.*, vol. 26, 1993, pp. 43–59.
[38] A. Gupta and D. Ferrari, "Resource Partitioning for Real-Time Communication," *IEEE/ACM Trans. Networking*, vol. 3, Oct. 1995, pp. 501–8.
[39] R. Bolla *et al.*, "Hierarchical Dynamic Control of Multiple Traffic Classes in ATM Networks," *Euro. Trans. Telecommun.*, vol. 5, no. 6, Nov. 1994, pp. 747–55.
[40] R. Bolla, F. Davoli, and M. Marchese, "A Global Control System for Integrated Admission Control and Routing in ATM Networks," *Proc. IEEE GLOBECOM '95*, Singapore, Nov. 1995, pp. 437–43.
[41] R. Bolla, F. Davoli, and M. Marchese, "Simple Schemes for Traffic Integration at Call Set-up Level in ATM Networks," *Comp. Commun.*, vol. 19, 1996, pp. 645–52.
[42] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case," *IEEE/ACM Trans. Networking*, vol. 1, 1993, pp. 344–57.
[43] J. P. Cosmas *et al.*, "A Review of Voice, Data and Video Traffic Models for ATM," *Euro. Trans. Telecommun.*, vol. 5, no. 2, Mar.-Apr. 1994, pp. 11–26.
[44] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. I, Belmont, MA: Athena Scientific, 1996.

BIOGRAPHIES

RAFFAELE BOLLA [M] received a "laurea" degree in electronic engineering from the University of Genoa, Italy, in 1989 and a Ph.D. degree in telecommunications at the Department of Communications, Computer and Systems Science (DIST) of the University of Genoa in 1994. From 1994 to 1996 he was a post-doctoral research fellow at DIST. Since November 1996 he has been an assistant professor in the same department. His current research interests are in management and control of STM and ATM networks, multimedia communications, and multiple access in integrated mobile radio networks.

FRANCO DAVOLI [M] received a "laurea" degree in electronic engineering from the University of Genoa, Italy, in 1975. From 1985 to 1990 he was an associate professor and, since 1990, he has been a full professor of telecommunication networks at the University of Genoa, where he is with the Department of Communications, Computer and Systems Science (DIST). From 1989 to 1991 and from 1994 to 1996, he was also with the University of Parma, Italy, where he taught a class in telecommunication networks by means of interactive distance learning over ISDN. His current research interests are in bandwidth allocation, admission control and routing in STM and ATM networks, multimedia communications and services, and integrated services mobile radio networks.

MARIO MARCHESE [M] received a "laurea" degree in electronic engineering from the University of Genoa, Italy, in 1992. He has concluded his Ph.D. in telecommunications at the Department of Communications, Computer and Systems Science (DIST), University of Genoa. His research interests include traffic modeling, admission control, and routing in ATM-based networks.