Control of Resources in Broadband Networks with Quality of Service Guarantees

When quality of service is to be guaranteed, system architects must select robust algorithms that perform well under a wide range of cell and call arrival statistics.

Aurel A. Lazar and Giovanni Pacifici

Integrated telecommunication networks carry traffic of several different classes, including one or more real-time (isochronous) traffic classes, each with its own set of traffic characteristics and performance requirements. Two different approaches have been advanced to deal with this phenomenon. In the circuit-switched approach (synchronous transfer mode), sufficient resources are allocated to each call to handle its maximum utilization; this guarantees that the call will get the quality of service (QOS) it requires, but may be wasteful of system resources. In the packet-switched approach (e.g., asynchronous transfer mode), traffic from all sources is packetized, and statistical multiplexing techniques are used to combine all network traffic through a single switching fabric. This allows higher network utilization, but requires more sophisticated controls to ensure that the appropriate QOS is provided.

In broadband networks, the internodal propagation delays are more significant than node processing delays. This observation has been used in the published literature to argue that dynamic, adaptive feedback, reactive control algorithms, operating within the network, are not suitable for broadband networks [1]. The argument typically states that because of high transmission speeds, by the time a downstream node detects a congestion condition and attempts to signal its neighboring upstream nodes to adjust their behaviors, the large number of cells in-transit could not possibly be affected by the closed-loop feedback controls.

Instead, the advocated control algorithms have been network-edge, congestion-avoidance, preventive algorithms. In this case, it is assumed that there is a prenegotiated contract between the network sources (terminating entities) and the network control architecture, characterizing traffic peak rates, average rates and the burstiness of the traffic stream that each source is allowed to transmit. Mechanisms are put in place in the control architecture to police the actual traffic behavior of a network source to ensure that it does not exceed the limits set forth in the negotiated contract [2]. Other proposed refinements to traffic flow enforcement include mechanisms that mark and discard excess traffic in the presence of network overload and congestion [3].

Notwithstanding this preventive control approach, there have been published results [4] indicating that, whatever the degree of sophistication of the network-edge preventive algorithms, additional reactive controls may be necessary within the network fabric to adequately handle the complex dynamic fluctuations in high-speed traffic interactions. Scheduling, buffer management, and admission control are such candidates.

This paper deals with the design principles of resource control algorithms, together with their interaction and cooperation in a wide area network environment, and presents a framework for evaluating the overall performance of the system. We will focus in particular on two levels of control: scheduling, which mediates the low level competition for service between cells of different classes; and admission control, which regulates the acceptance or blocking of incoming traffic on a call-by-call basis. The performance of the scheduling algorithms will be evaluated based on the schedulable region. The interaction between scheduling and admission control will be quantified using the admission control region. Both the schedulable region and the admission control region are concepts that have been recently introduced in the literature [5, 6].

The resource control algorithms are based on the Asynchronous Time-Sharing (ATS) design principle [7]. ATS is a set of resource allocation principles for the design of broadband packet-switched networks that guarantee QOS. ATS-based networks are similar to those based on Asynchronous Transfer Mode (ATM) in that all traffic offered to the network is in the form of small, fixed-size cells. The primary distinction of ATS is that several classes of traffic with different QOS requirements are considered explicitly at every level of system design, both at the network's edge and core. Therefore, one of the fundamental requirements on ATS systems is that the core of the network makes a distinction between traffic classes.

AUREL A. LAZAR is a professor and director of the Telecommunication Networks Laboratory at Columbia University.

GIOVANNI PACIFICI is a research scientist at the Center for Telecommunications Research at Columbia University.

0163-6804/91/$01.00 1991© IEEE

These design principles have broad applicability, and can help to efficiently provide QOS in many different network settings. They have already been used in the design of two high-speed integrated networks: MAGNET II [8], a testbed for MAN applications, and TeraNet [9], a gigabit/s lightwave network. The introduction of traffic classes into ATM networks, although not in the ATM standard at this time, may be accomplished in a fully compatible manner. For example, traffic class information could be carried in the Virtual Channel Identifier field of the cell header.

This paper is organized as follows. First, the basic concepts of the ATS framework are presented, along with an overview of an architecture for joint scheduling and admission control. The scheduling problem is then introduced. Next, the extension to the networking environment is discussed, followed by the formulation of the admission control problem and its interplay with scheduling. Finally, a reference model for broadband networks is briefly presented. The emphasis herein is on the network control architecture.

Problem Setting

The generic resource allocation problem presented in this paper was originally motivated by requirements on broadband networks with quality of service guarantees. A class of networks based on the concept of Asynchronous Time Sharing was implemented to meet these requirements. The switching architecture of these networks is briefly described. Four traffic classes are introduced via quality of service constraints. Note that, in order to keep the complexity of the network manageable, the QOS for these classes is defined for the network as a whole, rather than for each individual call. The introduction of traffic classes necessitates the introduction of resource allocation algorithms on both the cell and on the call level.

Architecture and Framework

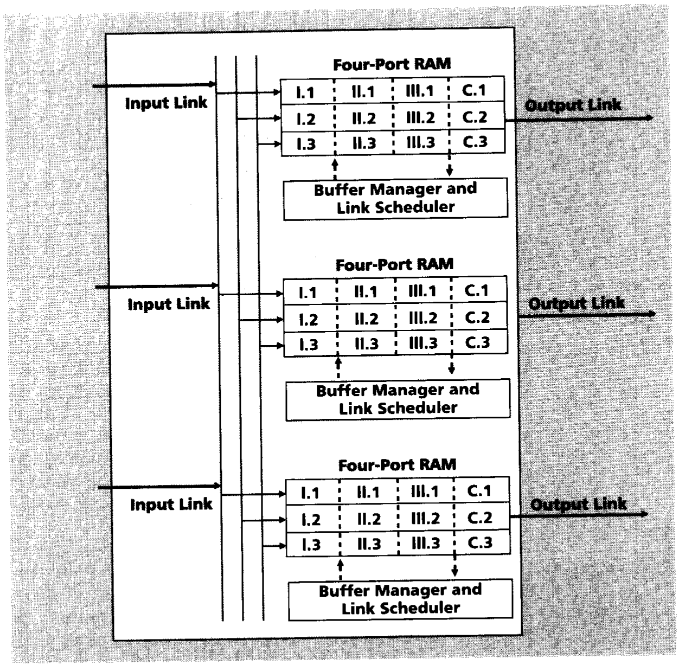
At the heart of the distinction between ATS and ATM is a clear definition of traffic classes based on QOS considerations. Fundamental to any performance analysis is the set of modeling assumptions on which the analysis is based. This section describes these and other key elements of the ATS approach.

We consider a class of networks that guarantee quality of service based on the ATS principle. The basic architecture of the ATS-based switching node has been recently implemented in a new prototype multihop lightwave network called TeraNet [9]. The architecture of the network interface units (switching nodes) is shown in Fig. 1.

Each network interface unit consists of three input links, three output links and a bus-based, nonblocking switch fabric. Congestion may arise only at the output links. The architecture supports four classes of traffic. Traffic arriving at an access point is transferred and stored, according to its class, into one of the multiple buffers. Each group of multiple buffers is connected to an output port. In Fig. 1 the critical resources (buffer space and communication links) are controlled by a Link Scheduler and a Buffer Manager. Scheduling and buffer management policies have hardware support. The switching nodes are interconnected in a mesh-type topology.

What are the four traffic classes that these networks will support? Three of the traffic classes, Classes I, II, and III, transport user traffic and are defined by a set of performance constraints on the cell as well as on the call level. The fourth class, Class C, transports traffic of the network management system, and is not subject to specific QOS constraints. Let's first define the cell level constraints.

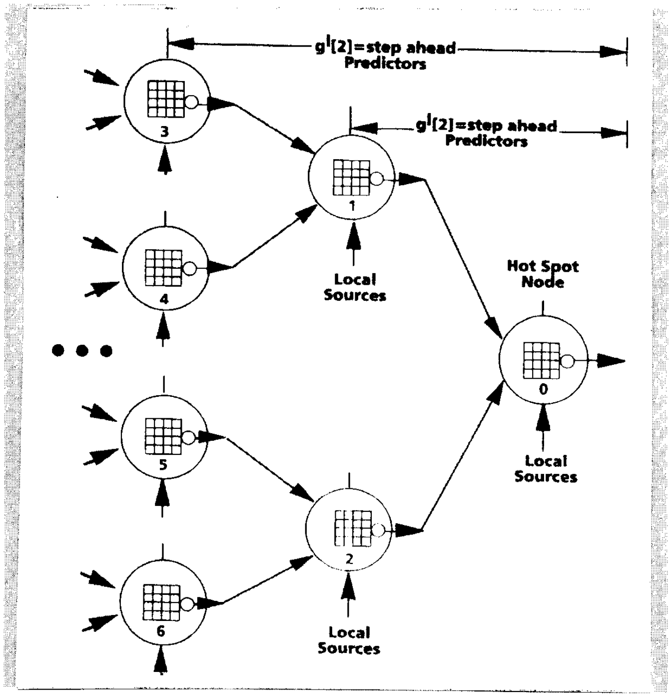
Class I traffic is characterized by 0 percent contention cell loss and an end-to-end time delay distribution with a narrow support. The maximum end-to-end time delay between the source and destination stations is denoted by S^I. Class II traffic is characterized by ε percent contention cell loss and an upper bound, η, on the average number of consecutively lost cells. It also is characterized by an end-to-end time delay distribution with a larger support than Class I. The maximum end-to-end time delay is S^{II}. Here, ε and η are arbitrarily small numbers and $S^I \leq S^{II}$. For Class I and II traffic, there is no retransmission policy for lost cells. Class III traffic is characterized by 0 percent end-to-end cell loss that is achieved with an end-to-end retransmission policy for error correction. If requested, it also is characterized by a minimum average user throughput Γ and a maximum average user time delay T.



■ **Figure 1.** *The architecture of the TeraNet network interface unit*

Finally, the call level QOS is characterized via a set of bounds κ^I, κ^{II}, and κ^{III}, on the probability of call blocking for each traffic class.

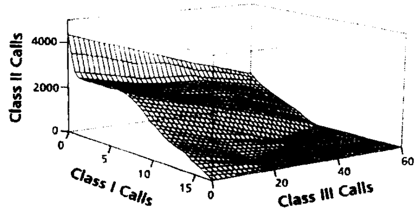
In order to satisfy the QOS constraints at the cell level, each traffic class will require a specific portion of network resources (communication bandwidth and buffer space). To efficiently cope with statistical variations of the traffic flow, scheduling and buffer management algorithms that dynami-



■ **Figure 2.** *A directed rooted tree associated with a hot spot node*

cally allocate the network resources are present at each contention point. Scheduling affects the order in which packets are served at an output link as shown in Fig. 1. Therefore, the effect of such algorithms is local. Note that in a wide area network environment, with switching nodes interconnected in a mesh-type topology, the delay on the communication links is one of the main limitations in congestion control.

Let us assume that a link scheduler at a particular node in the network cannot meet its prespecified QOS requirements. The link scheduler is then said to be in congestion or overload condition. The node which contains a congested link is said to be a hot spot in the network. From the graph model of the network, one can construct a tree (called the feed-in tree) rooted at the hot spot node. This tree consists of the sources and intermediate nodes and links whose traffic flows contribute to the hot spot condition. An example of such a directed, rooted feed-in tree is illustrated in Fig. 2. The goal is to remove the hot spot condition by distributed cooperation among the nodes in the feed-in tree. The difficulty here is due to the large delays between the different switching nodes. For example, the hot spot node cannot instantly signal its upstream neighbors that it is experiencing a congestion situation.



■ **Figure 3.** *Schedulable region for MARS, with QOS=[2 ms, 4 ms, 0.001, 5.0, 8 ms]*

The cell level QOS may be trivially guaranteed by any scheduling mechanism if a conservative admission control policy is used to limit utilization to sufficiently low levels. Guarantees of cell-level QOS to admitted calls is not sufficient if it comes at the cost of unreasonably high rates of call blocking. Thus, there is a need to simultaneously guarantee a certain quality of service at the call level as well. Later an admission control policy that guarantees call and cell-level QOS and maximizes the expected system utility is introduced.

In the design of controls for the ATS architecture, robust algorithms were sought, which would perform well under a wide range of cell arrival statistics corresponding to diverse real-world traffic sources. To this end, a conscious decision was made to eschew the traditional assumption of Poisson cell arrivals in favor of more complex models, such as variable bit rate and constant bit rate video sources, together with an On-Off model for voice sources [10]. At the call level, however, Poisson call arrivals and exponentially distributed holding times are assumed [6].

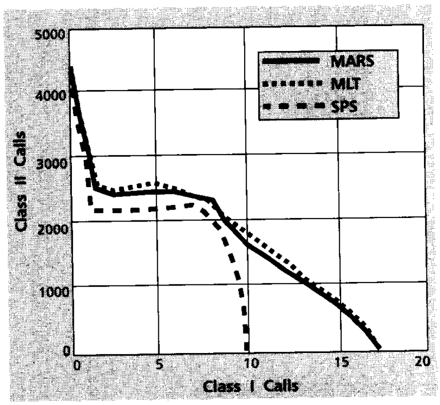
Real-Time Scheduling

As discussed above, dynamic bandwidth allocation generally leads to higher complexity. Is this complexity warranted? To answer this question, from a strict performance point of view, the concept of the schedulable region was introduced [5].

Intuitively, the schedulable region *S* of a queueing system is the set of points in the space of possible loads for which the quality of service at the cell level is guaranteed. As such, this concept is a generalization of the concept of the stability region. Recall that the general concept of stability calls for finding the region in the space of loads for which the average time delay is finite. In our case the set of constraints that determine the schedulable region is defined by the QOS constraints at the cell level. Examples of constraints were given previously and include: hard time delay constraints, probability of blocking and average gap constraints, average throughput, and average time delay constraints. Note that the schedulable region might be finite even for the case of a queueing system with finite buffer size. This is because the QOS constraints at the cell level might restrict the loading on the system before the finite buffer size does.

In ATS, transmission resources are time-shared between traffic classes according to a cycle scheme 7]. The MAGNET II Real-time Scheduling MARS) algorithm [5] is a mechanism for adapively setting the parameters that govern this cycle cheme, based on observations of cell arrivals and epartures. The scheduling algorithm is based on e intuition that in order to achieve high throughut, each cycle should serve only the cells whose ansmission cannot be further delayed to satisfy e QOS requirements.

The schedulable region for the MARS algoithm is depicted in Fig. 3. The axes in this figure how the load for each traffic class measured in number of calls. Each of the traffic classes carries information of a very specific type. Class I is assumed to



■ **Figure 4:** *Schedulable region for highly correlated traffic sources, QOS=[1 ms, 1 ms, 0.001, 5.0, 1 ms]*

consist of video calls, Class II of voice calls, and Class III of data sources [5]. The region in the three-dimensional space below the shaded surface represents the schedulable region. The size of this region depends on the scheduling algorithm used, the values of the QOS parameters, and the traffic load's statistics. (These dependencies and how to increase the schedulable region through the use of scheduling algorithms is discussed in detail in reference [5].)

The schedulable regions for MARS, Static Priority Scheduling (SPS), a frequently suggested mechanism for scheduling real-time packet traffic, and a variant of the Minimum Laxity Threshold (MLT) policy are shown in Fig. 4. To facilitate the comparison, the two-dimensional projections of schedulable regions onto the plane Class III calls = 0 is illustrated.

The simulation results reported [5] showed that when SPS is used, Class I traffic always experiences a small maximum delay, while with MARS the Class I maximum delay approaches S^I. This allows Class II and III to have more resources allocated to them, and thus the multiplexer has a greater link utilization factor. MLT scheduling achieves the largest schedulable region among the studied algorithms. Note that improved performance of a scheduling algorithm corresponds to an increase in complexity. Finally, note that the size of the schedulable region is a prime factor in determining the admission control policy for ATS-based networks, as discussed later.

Distributed Scheduling

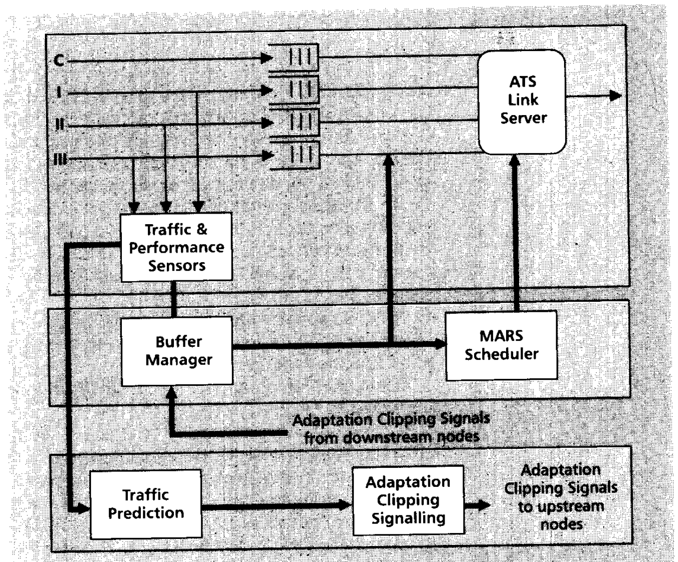
The role of resource allocation algorithms for an ATS node taken in isolation has been discussed. To efficiently cope with congestion in a wide area network environment, however, an interaction among the different resource allocation algorithms is required. Thus, we will explore the design principles of cooperative distributed algorithms for wide area integrated networks, with substantial delays on the communication links.

With cooperative distributed algorithms the quality of service, for all the network nodes, is met through coordination. This coordination involves the following actions: Each node predicts the traffic streams of any upstream neighbors on a horizon equal to twice the propagation delay between the node and each of its upstream neighbors. Feedback signals to upstream nodes are triggered by comparing the quality of service parameters of estimated queueing dynamics, derived from the traffic prediction entities, with threshold values. By using traffic prediction a node can anticipate (local) network overload and congestion and still have ample time to send feedback signals to affect upstream cell transmissions at the times when congestion is expected to occur. Whenever an upstream node receives an adaptive feedback signal, it discards cells of some of its traffic classes up to the limits that enable it to guarantee the node's prespecified QOS requirements.

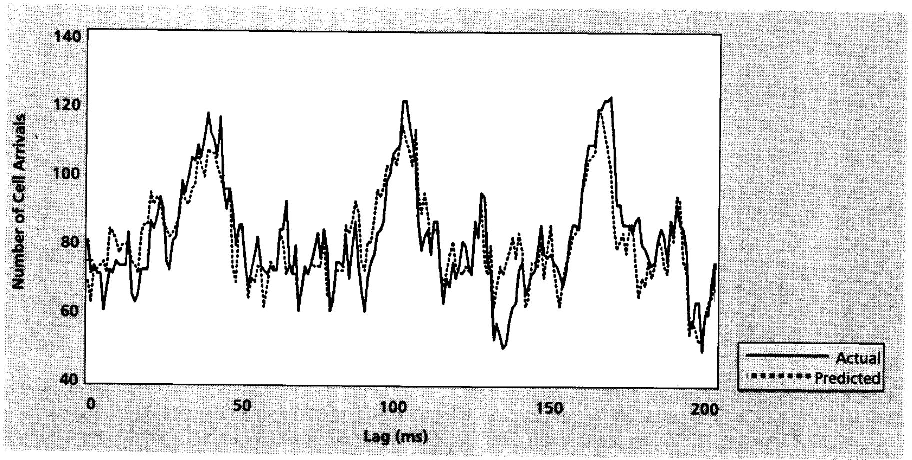
A cooperative distributed scheduling algorithm called D_MARS was introduced in [11]. An overview of the D_MARS functional modules at a switching node is presented in Fig. 5. As shown in the diagram, user traffic cell arrivals into a buffer group at a node are used by Traffic Prediction modules for load estimation on a horizon greater than twice the propagation delay between the link scheduler and its neighbor link schedulers in the upstream nodes of the feed-in tree (Fig. 2).

Traffic prediction exploits the correlation between arriving cells at a given traffic port. The cell arrival processes, for multiplexed video and voice sources, are modeled as seasonal auto-regressive processes [11]. The prediction of multiplexed video sources with the actual traffic is compared in Fig. 6. It can be seen that the predictor tracks reasonably well the jumps in the arrival process.

Should an overload condition be predicted, a Class II adaptation submodule triggers an indication of cell dropping that must be performed by each upstream node in order to relieve the projected congestion. These messages are sent as adaptation clipping (feedback control) signals to upstream nodes. Note that when a link scheduler at an upstream node receives an adaptation clipping signal, it will discard cells only up to its own prespecified QOS limit



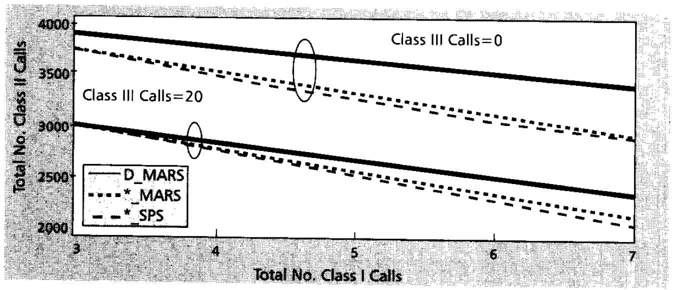
■ **Figure 5.** *Block diagram representation of the D_MARS functional modules*



■ **Figure 6.** *Predictor performance for video sources*

ε. Instructions to turn on (or off) Adaptation Clipping are signalled to upstream nodes as Class C cells. Such feedback signals are received by the Buffer Manager module of the upstream links. Each Buffer Manager discards (or stops discarding) cells of the affected user traffic class.

The cooperative distributed scheduling algorithm D_MARS has been extensively compared against the noncooperative algorithms *_MARS and *_SPS [11]. The latter algorithms are based on the MARS and SPS algorithms, respectively. Each link scheduler operates a local copy of these algorithms, and there is no coordination among the schedulers in the network. It has been shown that the D_MARS algorithm always results in a nondecreasing area of the schedulable region compared to the other two schemes (see Fig. 7). While, in this example, the *_MARS and *_SPS algorithms have almost overlapping schedulable regions, the D_MARS algorithm has a larger area for its schedulable region. The same effect can be observed for other loading conditions [11].

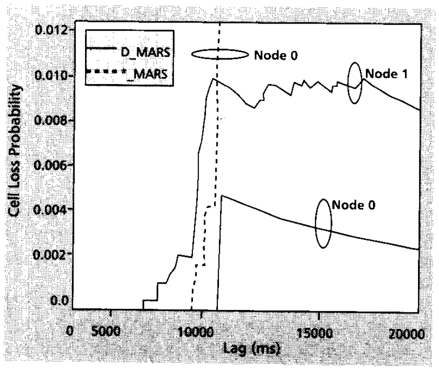


■ **Figure 7.** *Schedulable regions at the hot spot node for D_MARS, *_MARS and *_SPS*

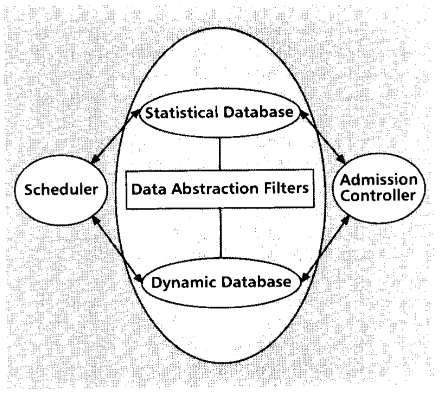
How can this increase in the schedulable region be explained? The trajectory of the cell loss probability at the root node (node 0) and one of its upstream nodes (node 1) when the D_MARS and *_MARS are, respectively, employed is shown in Fig. 8. Several features about these graphs are noted. In the noncooperative case, the hot spot node (node 0) exceeds the QOS constraint of $\varepsilon = 0.01$ after 10.5 s. This is because node 0 is unable to cope with the burst it receives at 10.4 s. The cell loss probability at node 1 is zero since this node is not congested. In the cooperative case, the D_MARS algorithm displays a stress balancing effect. This effect will now be discussed. Between 9.7 s and 10.5 s, the cell dropping algorithm at node 1, controlled by the proactive Adaptation Clipping module at node 0, significantly adjusts the arrival traffic statistics into node 0. At the same time, node 1 is able to stay within its QOS constraints. Thus, the cooperation between the two nodes handles the traffic burst and allows the hot spot node to overcome the overload condition without transferring the congestion condition to node 1.

Admission Control

Now let's focus exclusively on the interplay between scheduling and admission control at a switching node. The tasks of these two modules, and their interrelationship, are described herein and depicted in Fig. 9. The scheduler, as described above, controls the high-speed flow of cells through the switch. The dynamic data base temporarily stores the relevant cell-level information for use by the scheduler in making its decisions, as well as the resultant cell-level performance. Meanwhile, data abstraction filters digest this vast information flow, and continually update a statistical database, which stores various time-averaged and estimated quan-



■ **Figure 8.** *Cell loss probabilities in *_MARS and D_MARS*



■ **Figure 9.** ***The architecture for joint scheduling and admission control***

tities for use by other network management entities. Specifically, the schedulable region is stored here for use by the admission controller.

The task of the admission controller is to accept or reject arriving calls in order to maximize a utility function based on the weighted average throughput. It is constrained by the need for the network to guarantee the required QOS at the cell level to all calls admitted into service, and by limits on the call blocking probabilities as well. The information available to the admission controller includes the boundaries of the schedulable region S specified by the scheduler, the call arrival and departure rates associated with each type of service, and the weights used in the utility function. Thus, the admission controller is shielded from the flood of detailed cell-level information, and is able to guarantee QOS at the cell level by keeping the number of calls within the schedulable region. The detailed shape of the schedulable region may have a strong impact on the choice of the admission control policy. (The admission control problem is formally presented and solved in reference [6].)

For network service providers concerned with proper dimensioning to meet projected demands, the important question to be asked of a given system is: Under what loading conditions can all QOS constraints be satisfied? The answer to this question may be represented for our multiclass system as a region in the space of call arrival rates of each class. For a given schedulable region S, given call holding times, and given QOS bounds on the call blocking rates, the *admissible load region A* is defined by the region in the space of call arrival rates where the admission controller guarantees the cell and call-level QOS for all classes. This region, called the admissible load region, is highly dependent on the schedulable region for the switch, as well as the call holding times for the various classes; it provides a useful characterization of the joint performance of the scheduling and admission control algorithms. It can play a role not only in evaluating the merits of one admission control algorithm relative to another under a given scheduler, but also in comparing scheduling algorithms in terms of their impact on final system performance under load.

For a given schedulable region S, call holding times and QOS bounds on the call blocking rates, the admissible load region A is defined by the region in the space of call arrival rates where the admission controller guarantees the cell and call-level QOS for all classes.

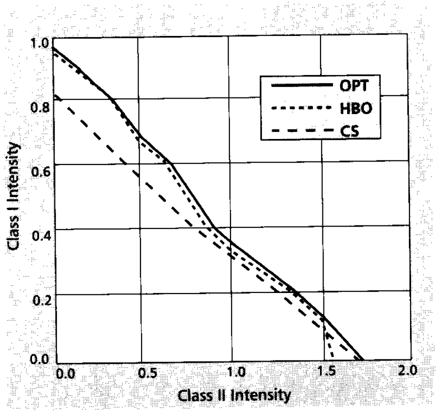
The optimal admission control policy for the case of a two-class system has been presented, along with two different heuristic control policies for comparison: Complete Sharing (CS) and Hold-Back-One (HBO) [6]. The optimal policy was achieved by reformulating the optimization problem as a linear program. The CS policy always admits incoming calls, except in cases where doing so would lead to a state outside of the schedulable region. The HBO policy is similar to the CS policy, but, in this case, a slot is always held open for the anticipated future arrival of a Class I call. In order to quantify the performance gain achieved by the optimal policy, each of these policies, in turn, was evaluated using the MARS scheduling algorithm.

It has been shown that, when the optimal policy is used, two distinct types of gain can be identified: an increase in the admissible load region, allowing the system to operate at higher offered loads; and an increase in utility at a given offered load. The admissible load regions achieved by the CS, HBO and optimal admission control policies are shown in Fig. 10. The optimal admission control is seen to achieve a gain of more than 15 percent over the CS policy when Class I traffic dominates. The HBO policy does almost as well as the optimal policy in this case, but performs very poorly when Class II traffic dominates.

In addition to exploring the boundaries of the admissible load region, it also is instructive to evaluate the gain in utility achieved by imposing various controls. In the cases examined in [6], utility increases are mainly due to extending the admissible load region to yield higher utility. The utility gain offered for the experiment of Fig.10 is pictured in Fig. 11.

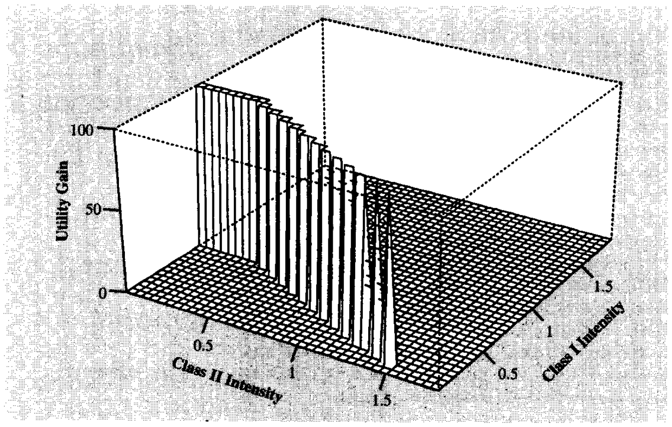
Figure 12 shows the utility gain of the optimal policy over the CS policy with no call-blocking constraints, using MARS scheduling and a utility weighting vector reflecting the assumption that the Class II calls were three times as valuable (per unit bandwidth) as the Class I calls. When the utility was considered proportional to the average bandwidth, similar effects were observed; however, the magnitude of the gain was smaller.

Broadband networks will require congestion-tolerant control algorithms embedded in the network switch fabric itself.



■ **Figure 10.** ***Admissible load regions for the optimal, HBO, CS admission control policies, with call blocking constraints*** $\kappa^I = \kappa^{II} = 0.1$

Generally, it was found that under low to mod-



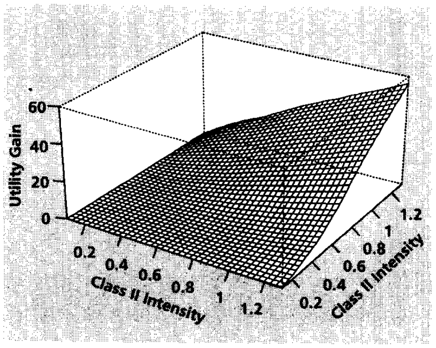
■ **Figure 11.** *Utility gain function for the optimal policy over CS with call blocking constraints*

erate loading conditions, the CS policy was nearly optimal. At higher offered loads, however, the CS policy suffers from an instability effect. As the offered loads are increased, there is a sudden increase in the blocking for Class II calls, leading to a drop in utility. The HBO policy ameliorates this problem somewhat, but can suffer from the same problem as loads continue to increase. By avoiding this instability, the optimal control can achieve significant gains in utility.

Network Control Architecture

We have already discussed the design principle of scheduling and admission control, emphasizing the problem of how to distribute resource allocation algorithms and how to ensure effective communication among them. In order to more formally identify the issues arising in congestion control of high-speed networks and to demonstrate how the ideas presented herein for two classes of control strategies can be generalized, an Integrated Reference Model (IRM) for broadband networks is briefly presented.

The network architecture contains the primitives for controls, communications, and management. These are organized in the Traffic Control, Information Transport, and Management Architectures. The subdivision of the IRM into the TCA and ITA is based on the principle of separation between communications and controls [12].



■ **Figure 12.** *Utility gain function for the optimal policy over CS with no call blocking constraints*

The separation between the MA and TCA is due primarily to the different time scales on which these architectures operate. Note, however, that in addition to control functions, management also includes tasks such as fault management.

The traffic control and the information transport architectures are logically divided into a set of vertical planes, a number of horizontal layers, and modules (Fig. 13). The purpose of this division is to facilitate the identification of the main issues when implementing the network architecture. The vertical subdivision corresponds to the main control and communications tasks. The control and communications tasks are originated, respectively, in the resource management and control (M), resource monitoring and management (D), connection management and control (C), and user transport (U) planes. The first three planes are part of the Traffic Control Architecture. The fourth plane is part of the Information Transport Architecture.

A plane is characterized by a set of entities and their relationships. The (M)-plane has the entities and algorithms responsible for resource management and control. The (D)-plane contains the entities and algorithms for monitoring and management. (The data about the network is stored in a Knowledge Database.) The (C)-plane contains the entities and algorithms responsible for connection management and control. The (U)-plane models the user transport of information. All entities and algorithms that support or are part of information transport are organized in this plane. The (U)- and (C)-planes are horizontally layered. The horizontal subdivision corresponds to the layering concept originally introduced by the OSI RM.

Recursive application of the OSI Service Model consisting of a service provider and multiple service users is the basis for layering the (U)- and (C)-planes. The (D)- and (M)-planes consist of a number of objects or modules. There are five classes of algorithms whose performance affects the efficiency of the information transfer task. These are configuration control, scheduling and buffer management, routing, flow control, and admission control. The Principle of Asynchronous Resource Management and Control applies to the last four classes [12]. The operating point of the network is achieved via an asynchronous algorithm among these four classes of algorithms. The algorithms are organized as modules in the (M)-plane.

The IRM incorporates two different time scales. The Network Management Architecture runs on the slow time scale. Both the Traffic Control and the Information Transport Architecture run on the fast time scale.

Conclusions

Broadband networks will require congestion-tolerant, control algorithms embedded in the network fabric itself, in addition to any network-edge preventive algorithms that are applied to police the traffic streams entering the network. The fundamental issue of the dominance of propagation delays in the face of very high-speed cell transfers can be obviated by relying on traffic prediction to anticipate future congestion conditions. Once such traffic prediction is feasible, the design of network controls reduces to the design of distributed, asynchronous control algorithms.

In the design of resource allocation algorithms the concepts of schedulable region and admissible load region have been demonstrated to be a powerful tool for investigating the capacity of a link or switch. The schedulable region depends on the scheduling algorithm employed, the QOS parameters, and the traffic statistics. This observation has many practical consequences. First, because of the many parameters that influence it, a universal approximation of the schedulable region does not appear to be easy to extract. Under certain traffic loads and profile, as well as under certain QOS constraints, such an approximation might lead to a substantial under/over utilization of resources. Second, as the traffic mix can have a major impact on the schedulable region, the latter can only be estimated as well as one can estimate the traffic mix. Thus, in practice it is advisable to specify an operating region that allows some margin around the boundary of the schedulable region, and an admission control policy that restricts the network load to within the operating region. Third, *adaptive scheduling algorithms* of the type described in this paper are called for in networks with unknown traffic statistics. This will guarantee a large schedulable region under many different operating conditions.

Finally, the schedulable region can be used for admission control. By estimating its size, the admission controller can decide whether a call should be accepted or rejected. Due to the computational constraints, we cannot hope to solve for the optimal policy for networks of realistic size, and heuristic policies must be chosen. By restricting our study to a smaller problem that could be fully solved, we have been able to gain an understanding into the behavior of the optimal admission control policy, and when various heuristics can approximate it. Even in the context of the limited control problem described herein, however, it does not seem feasible to calculate the optimal controls in real time. Rather, it is anticipated that the admission control specialist would have a stored table describing the admission policy in effect, allowing admit/reject decisions for individual call requests to be handled quickly by table lookup.

References

[1] G. Woodruff, R. Rogers, P. Richards, "A Congestion Control Framework for High Speed Integrated Packetized Transport," in Proceed. of the IEEE Global Telecommun. Conf. (Hollywood, FL), pp. 203-07, November 1988.

[2] E. Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks," IEEE J. on Sel. Areas in Commun., vol. 9, pp. 325-34, April 1991.

[3] A. Eckberg, D. Luan, D. Lucatoni, "An Approach to Controlling Congestion in ATM Networks," Int'l. J. of Digital and Analog Commun. Sys., vol. 3, pp.199-209, April-June 1990.

[4] A. DeSimone, "Generating Burstiness in Networks: A Simulation Study of Correlation Effects in Networks of Queues," Computer. Commun. Rev., vol. 21, pp. 24-31, January 1991.

[5] J. M. Hyman, A. A. Lazar, G. Pacifici, "Real-Time Scheduling with Quality of Service Constraints," IEEE J. on Sel. Areas in Commun., 1991. To appear.

[6] J. M. Hyman, A. A. Lazar, G. Pacifici, "Joint Scheduling and Admission Control for ATS-based Switching Nodes," IEEE J. on Sel. Areas in Commun., 1992. To appear.

[7] A. A. Lazar, A. T. Temple, R. Gidron, "An Architecture for Integrated Networks That Guarantees Quality of Service," Int'l J. of Digital and Analog Commun. Sys., vol. 3, pp. 229-38, April-June 1990.

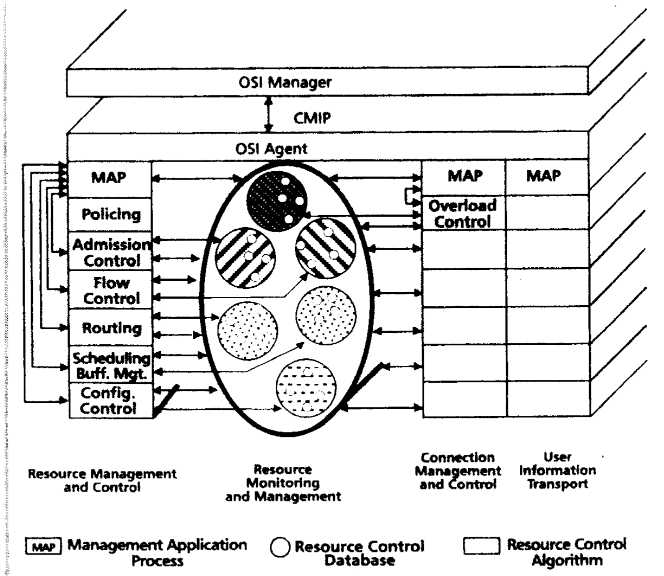
[8] A. A. Lazar, A. T. Temple, R. Gidron, "MAGNET II: A Metropolitan Area Network Based on Asynchronous Time Sharing," IEEE J. on Sel. Areas in Commun., vol. SAC-8, pp. 1582-94, October 1990.

[9] R. Gidron, A. T. Temple, "TeraNet: A Multihop Multichannel Lightwave Network," in Proceed. of the IEEE Int'l. Conf. on Commun. (Denver, CO), pp. 602-08, June 1991.

[10] A. A. Lazar, G. Pacifici, J. S. White, "Real-Time Traffic Measurements on MAGNET II," IEEE J. on Sel. Areas in Commun., vol. SAC-8, pp. 467-83, April 1990.

[11] J.-T. Amenyo, A. A. Lazar, G. Pacifici, "Cooperative Distributed Scheduling for ATS-Based Broadband Integrated Networks," CTR Technical Report 240-91-21, Center for Telecommunications Research, Columbia University, August 1991.

[12] A. A. Lazar, "An Architecture for Real-Time Control of Broadband Networks," in Proceed. of the IEEE Global Telecommun. Conf.(Phoenix, AZ), December 2-5, 1991.



■ **Figure 13.** *The integrated reference model*

Biography

AUREL A. LAZAR received a Dipl.-Ing. degree in communications engineering (Nachrichtentechnik) from the Technische Hochschule Darmstadt, Darmstadt, Germany, in 1976, and a Ph.D. degree in information sciences and systems from Princeton University, New Jersey, in 1980. In 1980 he joined the faculty in the Department of Electrical Engineering of Columbia University as an assistant professor. Since 1988 he has been a professor and director of the Telecommunication Networks Laboratory. He is an editor of the Springer Verlag monograph series on Telecommunication Networks and Computer Systems and area editor for Network Management of the IEEE Transactions on Communications. He is a founding member of the Center for Telecommunications Research at Columbia University, and a member of IEEE and ACM.

GIOVANNI PACIFICI received Dr.Eng. and Ph.D degrees from the Department of Information and Communication Technology at the University of Rome La Sapienza, Italy, in 1984 and 1989, respectively. In the course of his studies, he was a Visiting Scholar at the Center for Telecommunications Research, Columbia University, from 1987-88. In 1989, he joined the staff of the Center for Telecommunications Research as a research scientist.