Editorial Manager(tm) for Wireless Networks
Manuscript Draft

Manuscript Number:

Title: Adaptive Cross-Layer Bandwidth Allocation Policies in a Rain-Faded Satellite Environment

Article Type: Manuscript

Section/Category:

Keywords: satellite networks; fade countermeasures; bandwidth allocation policies; call admission control; performance evaluation.

Corresponding Author: Prof. Franco Rino Davoli University of Genoa

First Author: Nedo Celandroni, Dr. Ing.

Order of Authors: Nedo Celandroni, Dr. Ing.; Franco Rino Davoli, Dr. Ing.; Erina Ferro, Dr.; Alberto Gotta, Dr. Ing.

Abstract:

Adaptive Cross-Layer Bandwidth Allocation Policies in a Rain-Faded Satellite Environment(*)

Nedo Celandroni°, Franco Davoli*, Erina Ferro°, Alberto Gotta^

°ISTI-CNR (National Research Council), Area della Ricerca del C.N.R., Via Moruzzi 1, I-56124 Pisa, Italy
nedo.celandroni@isti.cnr.it erina.ferro@isti.cnr.it
*Department of Communications, Computer and Systems Science (DIST)
University of Genova, Via Opera Pia 13, 16145 Genova, Italy
franco@dist.unige.it
^CNIT (Italian National Consortium for Telecommunications) – University of Genoa Research Unit
Via Opera Pia 13, 16145 Genova, Italy
alberto.gotta@cnit.it

Abstract **-** Two adaptive bandwidth allocation methods, called *Optimized Centralized (OC*) and *Optimized Proportional (OP*), respectively, are studied for a satellite network environment, in the presence of both real-time, guaranteed performance, and best-effort traffic flows. In the schemes presented, a number of earth stations (*traffic stations*) operate in different weather conditions, with different levels of fade affecting the transmitted signals. The call admission control (CAC) policy for real-time connections is administered locally at the traffic stations, while a *master station* is charged to manage the MF-TDMA (multi frequency-time division multiple access) bandwidth allocation policy. In both cases, signaling from the traffic stations triggers a new bandwidth partition. In the OC method, a lack of resources is only signaled by the traffic stations to the master, which computes the allocations by minimizing a cost function that takes into account costs pertaining to each individual station. In the OP case, the master performs allocations on the basis of the amounts computed and explicitly requested by the stations. The effect of fade countermeasures, applied at the physical layer, on the bandwidth occupation is always explicitly accounted for. The results of a trivial scheme of bandwidth assignment, which allocates the bandwidth proportionally to the average offered load are shown as well, simply to highlight the improvement in allocation efficiency of the presented methods. For each policy, figures of merit such as loss, blocking and dropping probabilities are computed for a specific real environment based on the Italsat satellite national coverage payload characteristics.

Keywords: satellite networks, fade countermeasures, bandwidth allocation policies, call admission control, performance evaluation.

I. INTRODUCTION

Resource allocation is one of the main tasks of a network, where different users and services must share a pool of common resources. In wireless networks, where bandwidth may be relatively scarce with respect to cabled networks and environmental conditions may affect channel quality, the dynamic control of allocated resources becomes a challenge. Typically, control actions need to be exerted over a wide range of time scales, to cope with events that may occur with frequencies ranging from milliseconds to minutes or hours.

(*) Work supported by MIUR (Ministry of Education, University, and Research) in the framework of the IS-MANET and DIDANET projects.

Other works in the literature ([1] in the satellite environment and [2-4] in different contexts, as an example) have already been focused on optimal control choices. Satellite systems not only have to face variable load multimedia traffic, but also variable channel conditions and large propagation delays. The variability in operating conditions is due both to changes in the traffic loads and to the signal attenuation on the satellite links due to bad atmospheric events, which particularly affect the transmissions in the Ka band (20-30 GHz). It is therefore crucial to make use of adaptive network management and control algorithms to maintain the Quality of Service (QoS) of the transmitted data. The combined action among various layers of the network (from the physical layer up to the application layer) is likely to be the best way to combat channel variability. However, this procedure is complex and difficult to obtain in the widest extent possible, which would imply numerous cross-layer interactions for control purposes and the related exchange of signaling information. In the work presented here, in order to obtain optimized policies for satellite bandwidth allocation, we coordinate the actions taken in a satellite network at the physical layer (where the fade countermeasure technique is applied) with the work which is done at the data link layer (where the satellite bandwidth is allocated), thus obtaining a cross-layer optimization. Though applied to a different traffic context, our approach follows the same philosophy as in [1].

II. PROBLEM CONTEXT AND RELATED WORK

The present work stems from the study initiated in [5], where we assumed that the satellite network consists of a master station, which controls the access to the common resource, i.e. the channel bandwidth, and a number of traffic stations, which have to exchange both real-time (stream) and non-real-time (best-effort) traffic. The former is modeled as Continuous Bit Rate (CBR) guaranteed-bandwidth connections (voice or MPEG4 video), which may be carried within some specific DVB (Digital Video Broadcasting) class [6]. The latter is the aggregation of packet bursts, generated by a high number of sources, which are fragmented into fixed-size cells (typically, to fit in an ATM or DVB payload) and queued in a buffer before their best-effort transmission. This traffic may include TCP/IP “elastic” connections and UDP/IP flows with no particular bandwidth reservation. The fully-meshed satellite network uses the Ka band (20-30 GHz) of a geostationary satellite transponder as a bent-pipe channel, and we counteract the fade attenuations of the signals, due to bad weather conditions, by applying adaptive FEC (Forward Error Correction) codes and bit rates. This means that the fade is countered by applying a redundancy to the data before their transmission to the satellite, according to the detected attenuation level of the signal. In this paper we assume that the attenuation experienced by each station is independent of the destination of its traffic; this is the case where the fading is of the up-link-predominant type, or when all the traffic sent by a station is addressed to destinations affected by the same environmental conditions.

A typical example of such a scenario is presented by the civil protection in case of such a severe disaster that the terrestrial networks are unusable. Several mobile ad-hoc networks (MANETs) may be set up, consisting of teams equipped with devices for specific monitoring and data acquisition, plus a base camp, equipped with a satellite earth station, which is in contact with the operative headquarters via a satellite link. The satellite bandwidth is handled at the base camp (the master station), and distributed among the

MANETs according to optimization criteria. In each MANET, only one piece of equipment (the traffic station) is able to make contacts with the base camp (to request and obtain the bandwidth) and to transmit the data to headquarters, via a satellite link. In this scenario it is clear that all the traffic stations experience the same up-link fade, as they operate in the same geographic area (restricted to a few square kilometers), and the same down-link fade, as all the data are addressed to the same headquarters.

Although realistic, the case studied in this paper is part of a wider ongoing investigation aimed at generalizing the system's model used here, in order to include the case of traffic stations that experience different fading conditions in both the up-link and the down-link.

Since the signal fade may vary in very short time intervals (even less than a second), in order to avoid too many oscillations in applying the fade countermeasures according to each single fade level variation, the measured value of the signal attenuation is categorized in a class "*level*", so that the countermeasure strategy adopted remains unchanged for all those levels of signal attenuation that belong to the same class. Thus, for each type of traffic with a given Bit Error Rate (BER) requirement, a fade class aggregates those fade levels that need the same data redundancy, expressed at station i by redundancy coefficients $r_{level}^{(i)}$, *level*=1, 2,…, *K*, where *K* is the number of fade classes, equal for all the stations. As non-guaranteed traffic and real-time traffic usually have different QoS requirements in terms of BER, we indicate the respective redundancies with $r_{level,ng}^{(i)}$ and $r_{level,rt}^{(i)}$. Moreover, we consider a Multi-Frequency Time Division Multiple Access (MF-TDMA) system, i.e., a network where the total capacity of the satellite transponder is divided into carriers at different frequencies, each one accessed in TDMA. We also assume that a traffic station cannot transmit at different frequencies in the same temporal slot.

In [5] we studied a centralized allocation policy based on the solution of a discrete optimization problem, and we demonstrated that the combination of periodic (synchronous) and event-driven (asynchronous) decisions on the bandwidth allocation gives the best results in terms of call blocking and data loss probabilities of the entire system. The limit we found was on the computational time required by the master to calculate the allocations. As we demonstrate in this paper, this time can be considerably shortened by introducing explicit constraints on the minimum and maximum bandwidth assignments allowable in the optimization problem. The master's optimization criterion, improved with respect to the one presented in [5] by the introduction of constraints, is referred to as "Optimized Centralized" (OC). In OC, the master adopts an allocation policy that is optimized for the whole system; the traffic stations can only trigger a re-allocation procedure, without explicitly indicating a specific amount of bandwidth. In addition to OC, we investigate another allocation policy, referred to as "Optimized Proportional" (OP), where the master acts passively, only making assignments proportional to the requests received. Each of these requests is the result of an optimal policy local to each traffic station, based on predictions derived by traffic models.

Moreover, in this paper, after describing the OP and the OC policies, we compare their performance in terms of cell loss, call blocking and call dropping probabilities, given a maximum tolerable BER. Since the real-time traffic is modeled as CBR calls, "dropping" refers to the case where the bandwidth available is insufficient to maintain all the on-going connections with the desired BER. The allocations' computational times of the OC policy are also compared with those needed by the unconstrained method used in [5].

In many studies of resource allocation, a simple *complete sharing* (CS) policy is used, i.e. connections are admitted simply if sufficient resources are available at the time of the request, without considering the importance of a connection when they are allocated. In the complete sharing policy, the only constraint on the system is the overall capacity *C*. As an almost opposite situation, in the set of policies of *complete partitioning* (CP) type, every class of traffic is allocated a set of resources that can be used only by that class. Other policies have been derived to provide optimized access to resources, and Ross [7] provides an extensive discussion about a number of different solutions. Optimal approaches should be based on Markov decision processes, given a certain cost functional to be minimized (or maximized) as a performance index; however, they must take into detailed account any allowable network state and state transition, which is impractical even for networks of modest complexity. The functional form of the optimal policies is usually unknown. Therefore, a set of generally non-optimal policies with fixed structure (which can be often described by a set of parameters), have been developed, which are simpler to implement and, in some special cases, do correspond to the optimal policy: among others, the above mentioned CP, *trunk reservation* (TR) [8], *guaranteed minimum* (GM) [9], and *upper limit* (UL) policies [9], [10]. Comparisons have been made between these policies and the optimal one. The results indicate that the CP, TR, GM, and UL policies outperform the CS one when significant differences among classes exist in requirements for bandwidth and offered load [11]. Obviously, once one of such fixed-structure policies has been chosen, parametric optimization can be adopted, in order to choose the "best" values of parameters that minimize a given cost function (or maximize a performance index). This is the approach we have taken here. Moreover, whereas the presence of best-effort traffic is most often neglected in this case, we take it explicitly into account when designing the cost functional that determines the bandwidth partitioning.

In both OC and OP strategies, the master partitions the available bandwidth among the stations according to the two types of traffic, and the allocations remain fixed until one or more stations require the reallocations' computation. Our allocation policies fall in the CP category, but it is important to highlight that as the partitions are *adaptively* changed in response to traffic or fading variations, they try to match the traffic load and channel conditions as closely as possible. In this respect, the philosophy of this approach is much in the same line as that of [12]; however, the major novelty introduced here is in the cross-layer optimization, implied by the mechanism of event-driven bandwidth reallocation in response to significant changes in the transmission parameters.

III. THE CONTROL STRUCTURE

We consider a control architecture which comprises three time scales: a) the time interval during which a fade class remains unchanged at station *i* (*fade class* time interval*);* b) the time interval during which each traffic station estimates its fade level (*frame* time interval); c) the duration of a traffic station's data transmission (*transmission window*). The different temporal time scales are depicted in Fig. 1.

The master station recalculates the bandwidth assignments each time at least one traffic station notifies that its fade class has changed, or each time a station enters or leaves the network. The computation is done as if the allocations had to last forever; in principle this is true, because the allocations remain unchanged when system conditions are stable. The allocations are upper- and lower-limited. The constraint relevant to the

minimum assignment is specific to each single station *i*; namely, the already outstanding connections of station *i* must be maintained, in order to prevent the relevant call dropping probability from reaching unacceptable values. The constraint relevant to the maximum assignment is common to all stations: the bandwidth assigned to a single station cannot exceed the single carrier capacity. At station *i*, the number of supportable call connections depends on the bandwidth allocated to the station, and on the call blocking and cell loss probability thresholds the station has imposed.

In order to compute the allocations, the master utilizes the most recently received (at time *t*) vectors $\mathbf{v}^{(i)}(t)$, simply indicated by $\mathbf{v}^{(i)}$, (*i*=1, 2, …, *N*, where *N* is the number of stations), which contain information relevant to each individual traffic station. This vector assumes different meanings, according to the allocation policy adopted. When the OC policy is applied, $\mathbf{v}^{(i)}$ collects the mean real-time traffic intensity $\rho^{(i)}$ [Erlang], the best-effort traffic burst intensity $\lambda_{burst}^{(i)}$ [burst/s] and mean burst length $\bar{\tau}^{(i)}$, together with the last redundancy coefficients applied to data for the two types of traffic, and the current fade class indication. In the OP policy, $\mathbf{v}^{(i)}$ represents the total bandwidth request, expressed in minimum bandwidth units (*mbu*, which represents the minimum bandwidth granularity). This value is comprehensive of the station's need to accommodate both the real-time connections and the best-effort traffic within the desired respective QoS requirements. The separate values of the real-time and non-real-time traffic requests are known individually by the traffic stations only.

In the OC case, on the basis of the last updated $\mathbf{v}^{(i)}$ values, the master computes the bandwidth to be allocated to each station, by minimizing a cost functional relative to the whole system. The latter will be specified in Section IV, and takes into account both call blocking and cell loss probabilities of the single stations.

In the OP case, the master simply assigns portions of bandwidth to the stations proportional to the amount of the requests, up to the maximum bandwidth that can be allocated to each station.

Whatever the allocation policy used, the master sends the capacities $C^{(i)}$, assigned to each station i, in a reference burst (RB), which is transmitted at the beginning of each frame interval *f*. The assigned capacity remains unaltered until a new reallocation is executed.

A. The fade class time interval (E)

The *fade class* time interval is typical of each traffic station. It is the time interval between two consecutive changes in the fade classes. The length of this time interval is unpredictable: it goes from 1 second (chosen as the minimum value) till the duration of the entire satellite session. In this time interval, each traffic station performs its call admission control (CAC) procedure. At each change of fade class, within the received capacity partition $C^{(i)}$, the station i re-computes the threshold $N_{\max}^{(i)}$ on the maximum number of acceptable call connections, given the relevant bandwidth requirements (which depend on the data redundancy needed), and the desired upper bound on the call blocking probability. This threshold serves the purpose of the CAC, which is performed locally at each traffic station, independent of the bandwidth allocation method used by the master. At each change of fade class, station *i* sends the master the previously-mentioned information vector $\mathbf{v}^{(i)}$, thus causing an immediate rescheduling procedure.

B. The frame time interval (f)

The frame *f* is a fixed time interval, in the order of ms. It begins with a reference burst (RB, sent by the master) which contains the current bandwidth allocation plan of all stations in the time-frequency space. In each frame *f* all stations measure their up- and down -link attenuations. In order to filter out the background noise and the major part of the scintillation effect [13], the attenuation values are averaged over an interval of 1 second to make attenuation estimation. The attenuation values are used to compute the fade class of each link (see Section V) and the redundancy coefficients needed by each class of traffic. In the OC case, this information is sent to the master in the vector $\mathbf{v}^{(i)}$. The techniques used to estimate the signal attenuation are beyond the scope of the present paper; a possible method can be found in [14].

C. The transmission window (w)

By using the most recently-measured attenuation value and the ensuing classification and redundancy, each station transmits its data in the assigned transmission window *w*, according to the received transmission time plan. Moreover, each station sends the master, piggybacked with the data, the information on the latest fade level measured. As previously stated, in order to allow the evaluation of the link-by-link fade classes, the master redistributes this information within each RB. In general, the various data flows sent by a station may experience different down-link fading conditions, according to their destination.

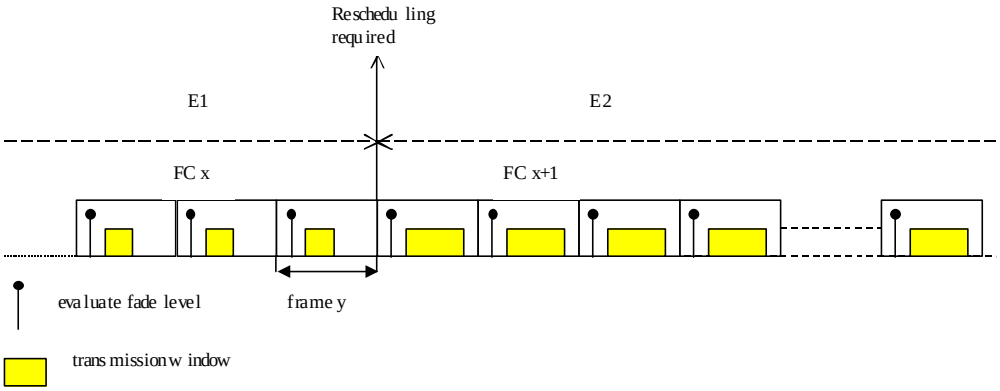

Figure 1. Time intervals at station *i*. FC *x* stands for "Fade Class *x*".

IV. THE OC AND OP POLICIES

Before detailing the control strategies, we need to introduce the analytical expressions of the performance indexes they are based upon, which depend on the traffic models used.

As regards real-time traffic, the traffic dynamics of interest are at the connection level, and the relevant performance index is the call blocking probability (P_{block}); this is the steady-state probability that an arriving call is refused because all the bandwidth devoted to the real-time traffic is busy. Note that we assume that blocked calls are lost (not re-attempted). For this type of traffic we adopt the usual birth-death model with exponential distribution of call inter-arrival and duration times (Poissonian traffic). Since we assume that all connections of station *i* belong to the same fade class, we face a particular single-class case, where the expression of the blocking probability is given by the classical Erlang B loss formula [7]. At station *i*, given

the Erlang traffic intensity $\rho^{(i)} = \lambda^{(i)} / \mu^{(i)}$ (where $\lambda^{(i)}$ [s^{-1}] is the arrival rate of the connection requests, and $1/\mu^{(i)}$ [s] is the average duration of each connection), and the desired upper bound on the blocking probability $\eta^{(i)}$, the maximum number of acceptable calls $N_{\max}^{(i)}$ is then derived as

$$N_{\max}^{(i)} = \max_{M} \left\{ M \in N \quad : P_{block}^{(i)}(M) = \frac{\left(\rho^{(i)}\right)^{M} / M!}{\sum_{j=0}^{M} \left(\rho^{(i)}\right)^{j} / j!} \leq \eta^{(i)} \right\} \tag{1}$$

Generally, a station may be in the *multiclass* case, where the connections that utilize a given bandwidth portion have different statistical parameters and peak rates, and they belong to different fade classes, as they may be addressed to destination stations that experience different down-link attenuations. Therefore the transmission of the data requires the simultaneous application of different data redundancy values. In this multiclass case, which we intend to analyze in a wider forthcoming study, the blocking probability would result from a stochastic knapsack problem [7].

As far as the best-effort traffic is concerned, we assume that it originates from non-real-time data flows, which are fragmented into fixed-size cells (ATM or DVB) before transmission on the satellite channel. At each station *i*, cells are queued in a finite buffer of capacity $Q^{(i)}$. In this context, the quantity of interest is the *cell loss probability* (P_{loss}) in the queue of station i. In order to derive an approximate evaluation of this quantity, we consider a discrete-time self-similar traffic model. This model represents the superposition of on-off sources, whose active periods (bursts) have Pareto-distributed 'on' time ($\Pr\{\tau = \ell\} = c\ell^{-\alpha-1}$, $1 < \alpha < 2$, where α and c are the parameter of the discrete Pareto distribution and its normalization constant, respectively). The detailed description of the model, which yields an upper bound on P_{loss}, can be found in [15, 16]. Actually, various possible models can be adopted to approximate the cell loss probability, given the statistical characteristics of the burst generation and a fixed rate of extraction of cells from the buffer [16, 17, 18]. In our case, we have to take into account that the extraction rate is determined by the residual capacity $C_{ng}^{(i)}(t)$, available for the non-guaranteed traffic after serving all guaranteed-bandwidth connections in progress at the required peak transmission rate *B*. Namely,

$$C_{ng}^{(i)}(t) = C^{(i)} - B r_{level,rt}^{(i)}(t) n^{(i)}(t) \tag{2}$$

where $C^{(i)}$ is the capacity allocated to station i, and $n^{(i)}(t)$ and $r_{level,rt}^{(i)}(t)$ are the number of guaranteed traffic connections in progress at time *t*, and the data redundancy factor applied to them, respectively. Therefore, the residual bandwidth is a random variable; as a consequence, the loss probability at fixed capacity can be considered only as conditional on the number of connections in progress, and its average must be computed with respect to the statistics of the Markov chain that describes the connection dynamics[1].

[1] In computing the average, the fact that the time scales of the guaranteed and non-guaranteed traffic are widely different can be exploited, in order to use independent stationary distributions for both. In other words, the guaranteed traffic process is supposed to be quasi-stationary, so as to use the conditional P_{loss} expression at constant rate, and the non-guaranteed traffic queue is supposed to reach steady-state between successive jumps in the Markov chain [12, 19].

Thus, the P_{loss} that we consider is the average one, from here on indicated as $\overline{P}_{loss}^{(i)}\left[C_{ng}^{(i)}(t)\right]$, to stress the dependence on the residual bandwidth. As previously mentioned, the real-time and the non-real-time traffic generally have different QoS requirements in terms of BER (realistic values of BER indicate about 10^{-4} for real-time voice connections and MPEG4 video [20], and at least 10^{-7} for best-effort traffic). Consequently, the redundancy to be applied to the latter is normally higher than the other one. This redundancy factor further reduces the effective residual capacity. The dependence on the redundancy factors, which are time-varying quantities, deserves further comment. In fact, all previously-discussed calculations regarding the performance indexes were made by considering the current values of these coefficients as lasting forever; they are recomputed at each change of fade class. The resulting control scheme is thus a sort of "open-loop feedback" repetitive control [21], where the initial time continuously shifts ahead.

A. The Optimized Centralized strategy (OC)

In this strategy the traffic stations do not communicate explicit bandwidth requests to the master station, but they send the information vector $\mathbf{v}^{(i)}$, defined in Section III, at each change of the current fade class.

For allocating the portions of bandwidth, the master computes a cost function, which takes into account the costs pertaining to the single stations. The goal is to obtain the "best" bandwidth assignments, which represent the parameters to be optimized, while keeping the system's constraints satisfied. It is worth noting that, under the chosen CP policy, if the overall cost is a separable function of the capacities allocated to the individual stations (e.g., a sum of the individual costs), the assignment can be computed by means of a dynamic programming algorithm. Indeed, the rationale behind the cost function is to set a penalty on bandwidth assignments that would push the partition for real-time traffic at each station *i* below the minimum bandwidth necessary for that station to satisfy its constraint imposed on the call blocking probability. On the other hand, for bandwidth assignments that do not violate this constraint, the loss probability of best-effort traffic is considered as a cost. The measured values of the redundancy factors, as known to the master station, are used in the evaluation of the cost. This allows taking into account the most recent channel characteristics, and effectively binds the "layer 2" decision made here with the physical layer BER control, in a cross-layer optimization, which is one of the main objectives of our approach. For the sake of clarity, we repeat here the expression of the cost function, which was originally introduced in [5].

$$J^{(i)}(C^{(i)}) = \begin{cases} \overline{P}_{loss}^{(i)}\left[C_{ng}^{(i)}(t)\right] & \text{if } C^{(i)} \geq C_{rt}^{(i)} \\ \Theta & \text{if } C^{(i)} < C_{rt}^{(i)} \end{cases} \tag{3}$$

where $C_{rt}^{(i)}$ is the minimum amount of capacity that guarantees the satisfaction of the constraint imposed by station *i* on the blocking probability, $\overline{P}_{loss}^{(i)}\left[C_{ng}^{(i)}(t)\right]$ is given by (B4) in Appendix B, and Θ is a penalty term. Specifically, if we want that station *i*'s blocking probability $P_{block}^{(i)}$ be lower than a threshold $\eta^{(i)}$ ($P_{block}^{(i)} \leq \eta^{(i)}$), the master has to compute the minimum bandwidth capacity for the real-time traffic of station *i*:

$$C_{rt}^{(i)} = \min_{Y^{(i)}} \left\{ Y^{(i)} : \left\lfloor Y^{(i)} \Big/ \left(B \ r_{level,rt}^{(i)}(t) \right) \right\rfloor \leq N_{\max}^{(i)} \right\} \tag{4}$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to *x*, and $N_{\max}^{(i)}$ is obtained from (1). An adequate value for Θ is an upper bound of the sum of costs of all stations. As each cost is a probability (always less than or equal to 1), Θ can be taken equal to the number of stations.

The goal is then to minimize

$$J(C^{(1)}, C^{(2)}, ..., C^{(N)}) = \sum_{k=1}^{N} J^{(k)}\left(C^{(k)}\right) \tag{5}$$

subject to the "static" and "dynamic" constraints

$$\begin{cases} \sum_{i=1}^{N} C^{(i)} = C \\ C^{(i)} \le C_c \\ C^{(i)} \ge n^{(i)}(t) r_{level,rt}^{(i)}(t) B \end{cases} \tag{6}$$

where C_c is the maximum allowable information rate for each carrier.

The separability of (5), which represents an overall system cost, stems from the fact that, given the bandwidth partitions and the independence of traffic at the earth stations, the blocking and loss probabilities for each station are independent and can be computed separately. The last constraint in (6) stems from the willingness to guarantee the continuation of connections in progress. The minimization of (5) can be efficiently effected by means of a dynamic programming algorithm, as reported in [7]. The algorithm used in [7] has been modified to take into account the presence of the constraints (6). The modified version is reported in *Appendix A*. It is worth noting that the presence of constraints can greatly reduce the search space, speeding up the computational time of the algorithm. For example, a further reduction may be obtained by imposing the values of the previous assignments as the upper bound for those stations that did not signal any increase of bandwidth need (i.e., no deeper fade or higher traffic load), while the previous assignments can be imposed as lower bounds for stations that signaled an increase of bandwidth need.

The problem admits at least one solution $C^{(i)} = C_{opt}^{(i)}, \quad i = 1, ..., N$, if $\sum_{j=1}^{N} C_{rt}^{(j)} \le C$, where *C* is the total available bandwidth in the system. If $\sum_{j=1}^{N} C_{rt}^{(j)} > C$, the master computes the allocation as

$$C^{(i)} = \min\left\{ C \cdot \frac{C_{rt}^{(i)}}{\sum_{j=1}^{N} C_{rt}^{(j)}} \ , \ C_c \right\} \tag{7}$$

i.e., proportionally to the foreseen bandwidth need of real-time traffic.

B. The Optimized Proportional strategy (OP)

In this strategy, at each time frame the traffic stations send the master requests for explicit bandwidth values. A request is issued for obtaining the minimum bandwidth necessary to support both types of traffic of the station, under given QoS constraints on call blocking and cell loss probabilities. The master simply assigns

the bandwidth proportionally to the requests received. The problem of how to compute the request is solved within each station. Various methods may be applied. For example, formula (4) may be used at station i to compute the bandwidth required for real-time connections, while the bandwidth necessary for the best-effort traffic transmissions can be obtained by considering the constraint imposed on the average loss probability ($\overline{P}_{loss}^{(i)} \le \gamma^{(i)}$). Appendix B contains the detailed computations in the case where the expression adopted for $P_{loss}^{(i)}$ is given by the Tsybakov-Georganas formula [16].

V. THE SIMULATION ENVIRONMENT

We simulated the OC and OP strategies, by considering a real fully-meshed satellite network that uses bent-pipe geostationary satellite channels, as described in the Introduction. Table I reports the most significant parameters of the real satellite system we considered. In order to compute the link budget, we took data from [22], relevant to the transponder #1 of the Italsat national coverage payload (20/30 GHz band), which is currently no longer operating, but still represents a reasonably up-to-date situation. The information rate of 6.554 Mbit/s for each carrier is obtained with a 4/5 punctured convolutional encoder. The net values of 7 and 5 dB of channel E_b / N_0 (bit energy to one-sided noise spectral density ratio) are assumed as the thresholds of the clear sky conditions for best-effort traffic and real-time connections, respectively. At the conditions of the thresholds, after the Viterbi decoder, the bit error rates are 10^{-7} and 10^{-4}, respectively.

Table I. Most significant values of the MF-TDMA system considering the Italsat payload.

Stations' antenna diameter	1.8 m
Stations' power	13 dBW
Satellite *G/T*	5.9 dB/°K
Satellite *E.I.R.P.* (effective isotropic radiation power)	48 dB W
Number of carriers	3
Capacity of each carrier (QPSK modulation)	8.192 [Mbit/s]
Up-link power control range	5 [dB]
Min. net E_b / N_0 in clear sky conditions for non-real-time traffic (real-time traffic)	7 (5) [dB]
BER guaranteed for non-real-time traffic (real-time traffic)	10^{-7} (10^{-4})
Possible data coding rates	4/5 (clear sky), 2/3, 1/2
Total information bit rate in clear sky conditions	19.66 [Mbit/s]
Information bit rate in clear sky conditions after system overhead	18 [Mbit/s]

In order to compute the resulting net values of E_b / N_0 at the earth station's receiver input we used relation (8) below. No automatic gain control feature operates on the transponder. For this reason the attenuation on the up-link affects both the up- and down-link C / N_0 (carrier-to-noise) values.

$$E_b / N_0 = C^{(res)} - 10Log_{10} b_r - m_i \tag{8}$$

where:

$$C^{(res)} = C_r^{(up)} - A_{u_p} + C_r^{(dn)} - A_{u_p} - A_d - 10Log_{10}\left(10^{(C_r^{(up)} - A_{up})/10} + 10^{(C_r^{(dn)} - A_{up} - A_d)/10}\right)$$

is the resulting C/N_0 (carrier power to one-sided noise spectral density ratio) at the earth station receiver,

$C_r^{(up)}$ is the reference (in clear sky) up-link C/N_0 = 80.7 [dBs^{-1}],

$C_r^{(dn)}$ is the reference (in clear sky) down-link C/N_0 = 81.6 [dBs^{-1}],

A_d is the dB down-link attenuation of the receiving station,

A_{u_p} is the dB up-link attenuation of the transmitting station, after up-link power control intervention:

$A_{u_p} = 0$, if the up-link attenuation $A_u \le p_r$ (p_r is the up-link power control range = 5 dB);

$A_{u_p} = A_u - p_r$, if $A_u > p_r$,

b_r is the data bit rate in bit/s,

m_i is the modem implementation margin (assumed equal to 1 dB).

Table II contains the fade classes of the traffic stations as a function of the C/N_0 values. Each fade class imposes the adoption of the indicated transmission parameters (and then r_{level} values) to limit the BER below the chosen thresholds. The system configuration used consists of ten active stations, five of which are in clear sky, whereas the other five experience up-link fading. The attenuation patterns of each of the five stations in fade, used for simulation runs, are shown in Fig. 2. The attenuation data are taken from a real-life data set chosen from the results of the propagation experiment, in Ka band, carried out on the Olympus satellite by the CSTS (Centro Studi sulle Telecomunicazioni Spaziali) Institute, on behalf of the Italian Space Agency (ASI). The up-link (30 GHz) and down-link (20 GHz) samples considered were 1-second averages, expressed in dB, of the signal power attenuation with respect to clear sky conditions. The attenuation samples were recorded at the Spino d'Adda (Northen Italy) station, in September 1992. We have preferred to use real fading traces, rather than relying upon a model for rain fade generation, as no thoroughly satisfactory model has been devised so far.

Table II. Redundancy factors and signal-to-noise ratios versus fade classes

Fade classes	$r_{level,ng}$	$r_{level,rt}$	*Coding rate, bit rate [Mbit/s] (non-real-time traffic)*	*Coding rate, bit rate [Mbit/s] (real-time traffic)*	C/N_0 [dB]
1	1	1	4/5, 8.192	4/5, 8.192	>77.13
2	1.2	1	2/3, 8.192	4/5, 8.192	77.13 – 75.13
3	1.2	1.2	2/3, 8.192	2/3, 8.192	75.13 –74.63
4	1.6	1.2	1/2, 8.192	2/3, 8.192	74.63 – 72.63
5	3.2	1.6	1/2, 4.096	1/2, 8.192	72.63 – 70.63
7	3.2	3.2	1/2, 4.096	1/2, 4.096	70.63 – 69.63
8	6.4	3.2	1/2, 2.048	1/2, 4.096	69.63 – 67.63
9	6.4	6.4	1/2, 2.048	1/2, 2.048	67.63 – 66.63
10	outage	6.4	----------	1/2, 2.048	66.63 – 64.63
11	outage	outage	----------	----------	<64.63

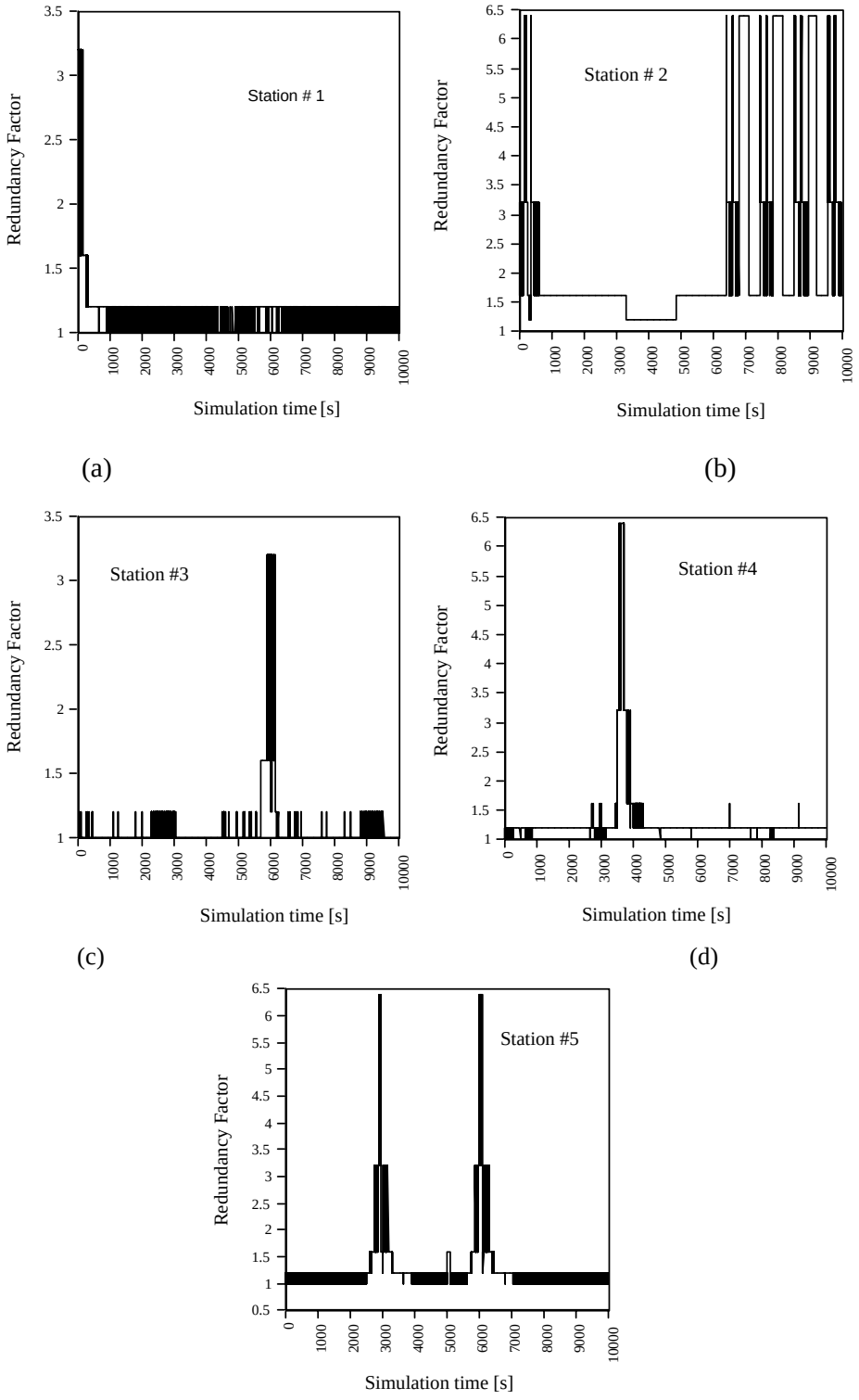

Fig. 2. Redundancy values vs time [s] represented for the five faded stations (Figs. a-e).

VI. SIMULATION RESULTS

We report in the following the results of a simulative analysis of the proposed adaptive strategies. According to (8) and Table II, the attenuation level determines the attribution of each station's traffic type to a certain fade class. In order to avoid too many oscillations in the “instantaneous” bandwidth assignment of a station,

we introduced a sort of hysteresis mechanism, whereby a station's traffic type remains in the same fading class, unless the corresponding attenuation value exhibits a change above a given threshold (1 dB, in our case) for more than 3 seconds. On the other hand, as far as the station's outage is concerned, we adopted the definition of "unavailable time" given in ITU-T Recommendation G.821 [20] for connections.

Table III. Data used in the simulation.

Station connection generation rate for real-time traffic	λ_s = 0.2 [request/s]
Average real-time connection duration	$1/\mu = 60$ [s]
Shape parameter of the Pareto distribution	$\alpha = 1.5$
Number of cells generated in one slot time (*see Appendix B)*	R = 10
Cell payload (*see Appendix B)*	L = 384 [bit]
Non-real-time traffic sources' peak rate	B_{ng} = 256 [kbit/s]
Slot time duration (*see Appendix B*)	T=15 [ms]
Average non-real-time traffic burst duration	$\overline{\tau} = 1.945$ [slot] or 29.175 [ms]
Non-real-time traffic burst generation intensity (equal for all stations)	λ_{burst} = 16.67 [burts/s]
Buffer dimension	8000 [cell]
P_{block} threshold (equal for all stations)	η = 5%
P_{loss} threshold (equal for all stations)	γ = 1%

The data reported in Table III have been used for the generation of the traffic. The minimum bandwidth unit that can be allocated has been taken as equal to 8 kbit/s. Each simulation run covers a 10,000s time span. The final values are obtained by averaging the results on a number of simulation runs sufficient to produce a 5% confidence interval at 99% level.

In the graphs produced, a "SP (Simple Proportional) policy" also appears, where the master acts passively, only making assignments proportional to the bandwidth requests received; moreover, the requests do not take into consideration any specific prediction and cross-layer interaction, but they are simply based on measurements of the traffic intensity. The SP policy has been presented in the graphs as a comparison with a very simple and straightforward method for satellite resource allocation.

The results are divided into two sets. The first one (Figs. 3-4) shows the call blocking, call dropping and cell loss probabilities, averaged over all stations in the system, and over a time window of 1000 s. The second set (Figs. 5-7) shows the same quantities for each station, averaged over the entire simulation time (10,000 s).

As regards the blocking probability, it can be seen that all methods keep the average overall system blocking probability below the 5% threshold. However, the *individual* (per station) blocking probability is highly unbalanced for the SP, whereas both OC and OP tend to essentially equalize this value: as their goal is to respect the constraint, they do not waste bandwidth in favor of the more privileged stations (the ones that are not in fade); this saving can be dedicated to keeping the number of packets dropped at the faded stations at a lower value.

A more evident difference is shown by the probability of call dropping, both in the overall system average and in the individual cases. We recall that a call is dropped at a station whenever the applied redundancy (needed in response to a change in fading class) is such that the sum of the bandwidths of the calls in

progress overcomes the maximum amount of bandwidth temporarily allocated to the station. This is a quantity over which we have no direct control, as we have assumed that the stream traffic does not tolerate a reduction in the transmission speed. A different scenario could be envisaged, in the presence of a certain degree of elasticity in the stream service, or in the presence of Variable Bit Rate (VBR) coding (see, e.g., [1, 24]). In this scenario, the dropping rates experienced by OP and OC could be further reduced by the adoption of suitable rate adaptation techniques. Anyway, the call blocking and dropping probabilities are loosely related, as a more cautious acceptance behavior implies a smaller likelihood of dropping in severe fading conditions. In this sense, the OC shows slightly more robustness than the OP (see Fig. 3).

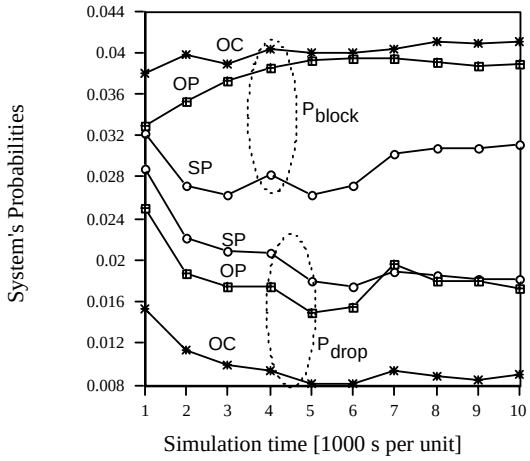

Fig. 3. Average system's call blocking and dropping probabilities vs simulation time

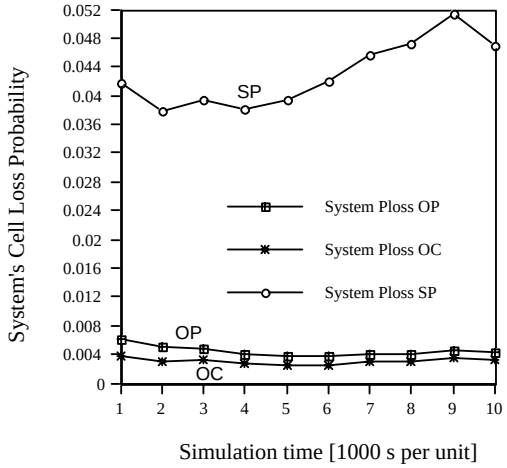

Fig. 4. Average system's cell loss probability vs simulation time

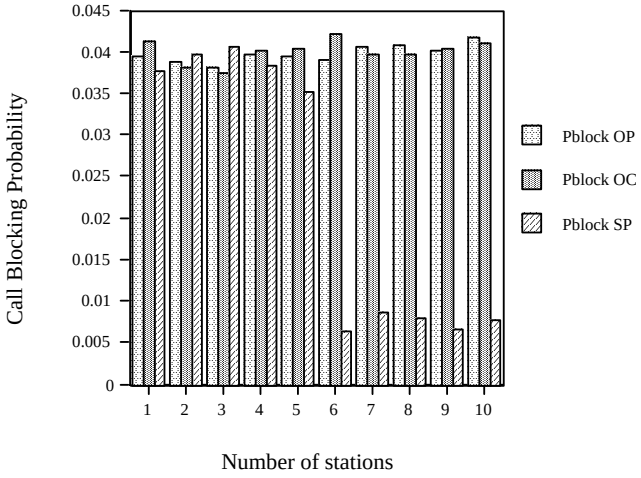

Fig. 5. Average call blocking probability per station.

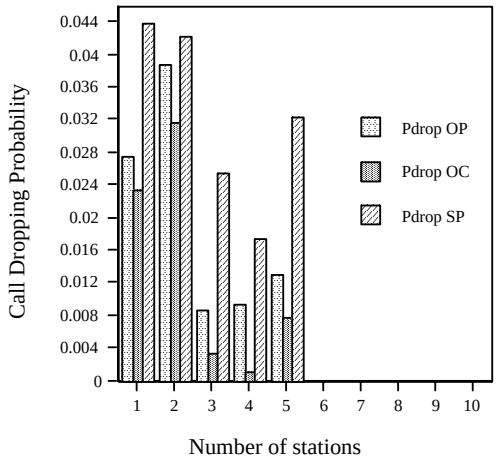

Fig. 6. Average call dropping probability per station. Stations 6-10 do not experience fades.

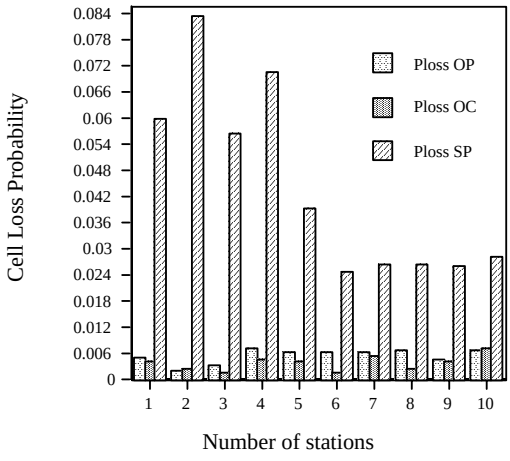

Fig. 7. Average cell loss probability per station.

A remarkable gain is obtained by the cross-layer adaptive policies as regards the cell loss probability. Both the system's averages (Fig. 4) and the per-station ones (Fig. 7), are kept under the 1% threshold. However, the difference with respect to the OP policy is less remarkable than what one might expect, owing to the more decentralized nature of the latter. In this respect, it must be noted that both criteria are based on the same traffic models and essentially operate the bandwidth assignment in a centralized way. The main difference lies in the amount of signaling and in the computational times required to implement them.

Some considerations on the computational times

As is observed in Appendix B, the calculations needed to compute the bandwidth allocations in the OP strategy (which are essentially the same that yield the cost-per-station components in the OC) can be performed in advance for all possible values of the fading coefficients and of the discrete capacity allocations, and stored in look-up tables. Therefore, the main computational burden that is left for on-line execution to the master station is the actual calculation of the bandwidth partitions. In the OP case, the computational time is practically negligible; on the other hand, it may rise to significant values in the OC, where the dynamic programming algorithm of Appendix A must be implemented. Figure 8 shows that, in the case of 10 stations considered here, even the highest values are still quite manageable; at any rate, the presence of constraints on the minimum and maximum numbers of *mbu* that can be allocated per station may greatly help in reducing them. Figure 8 also presents the computational times in the case of a total capacity of 10000 mbu, and 3333 mbu per carrier. Both the computational times refer to the execution on a computer running Windows XP, with AMD CPU and a clock of 1800 MHz.

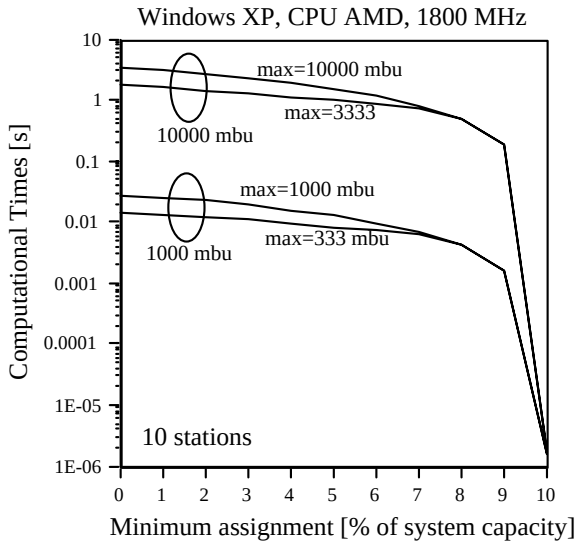

Fig. 8. Computational times for the dynamic programming algorithm, depending on constraints.

VII. CONCLUSIONS

This study presented addresses the resource allocation problem in rain-faded satellite networks at different levels in the protocol architecture, under different time scales. As a matter of fact, actions taken on transmission parameters (e.g., data coding rate) in response to fading variations affect the results and the

effectiveness of call admission control and bandwidth allocation policies, which take place at the data link and network control layers. Intuitively, it is reasonable to expect that the upper control layers would benefit by being aware of the presence of the underlying ones. With this multi-layer approach in mind, the main goal of this paper was to evaluate and compare two centralized bandwidth allocation policies. The first one (OC) is based on the parametric optimization of a cost function, which the master station constructs by taking into account traffic models, whose parameters are provided by the traffic stations. The second one (OP) assigns the bandwidth on the basis of explicit requests issued by the stations, after each of them has locally computed its bandwidth requirements, according to the satisfaction of performance constraints derived by the same traffic models used also in the OC method. The performance evaluation, conducted by simulation, using real fading traces, shows that both methods are capable of reacting to variations in the traffic load and balance, as well as in fading levels, and are effective in keeping call blocking probability below a given threshold, while at the same time ensuring a small amount of dropped calls at the stations in fading and a low DVB packet loss probability, evenly distributed among the stations. Their comparison indicates that OC slightly outperforms OP, at the expense of a higher, but still manageable, computational complexity. The comparison with the simple SP method has shown that the cross-layer adaptive policies can yield a significant performance improvement in all quantities of interest (by reducing cell loss and call dropping, and by keeping call blocking probabilities evenly distributed among the traffic stations).

Further work is in progress, along the following lines: i) to extend the policies to stations that experience both up- and down-link fades towards multiple destinations, which create a multirate environment, due to the presence of different data redundancies at the same station; ii) to investigate the effect of the adoption of on-line gradient descent techniques of the type considered in [25], to relax the discrete integer optimization problem to a continuous one, whose solution can be spread over time, instead of being concentrated at the beginning of a fixed time interval; iii) to compare complete partitioning, complete sharing, and hybrid allocation policies, always taking into account the presence of two basic traffic types; iv) to include models of TCP elastic traffic and their related performance optimization.

APPENDIX A. The dynamic programming algorithm with constraints used in the OC strategy

A complete partitioning policy is optimal for loss systems when the traffic is very heavy. In particular, we address the problem of determining the optimal complete partitioning policy under general traffic conditions. Let N be the number of stations among which the total capacity C (expressed in multiples of *mbu*) must be divided, and let $J_k(i)$ be the cost function relevant to station k when i *mbu*s are assigned to it. Moreover, let $C^{(i)}$,$\{1 \le i \le N\}$, be the assignment vector in order to obtain the $\min\left\{\sum_{k=1}^{N} J_k(C^{(k)})\right\}$ with the following two constraints: 1) $C = \sum_{i=1}^{N} C^{(i)}$; and 2) $C_m^{(i)} \le C^{(i)} \le C_M^{(i)}$; $1 \le i \le N$, where $C_m^{(i)}$ is the minimum assignment imposed to station i, and $C_M^{(i)} = \min\left\{C_{MAX}^{(i)},\ C_{CC}^{(i)},\ C - \sum_{j=1,\, j\neq i}^{N} C_m^{(j)}\right\}$, $C_{MAX}^{(i)}$ being the maximum assignment which we want to impose to station i, and $C_{CC}^{(i)}$ being the maximum assignable capacity. The latter constraint

is imposed by the maximum capacity of a single carrier. The assignments must be done in such a way that $C \le \sum_{n=1}^{N} C_M^{(n)}$ must be satisfied. Dynamic programming can solve the problem of finding the required solution.

The unconstrained algorithm reported in [7] is modified below to take into account the presence of the constraints.

Let $h_k(i)$ be the minimum of the sum of the first *k* stations' relative costs, when the total capacity *i* is allocated to these *k* stations.

The corresponding dynamic programming equations are:

$$h_1(i) = J_1(i), \quad C_m^{(1)} \le i \le C_M^{(1)},$$

$$h_2(i) = \min\left\{J_2(j) + h_1(i-j);\ \max\left\{C_m^{(2)}, i - C_M^{(1)}\right\} \le j \le \min\left\{C_M^{(2)}, i - C_m^{(1)}\right\}\right\}$$

$$\text{where } C_m^{(1)} + C_m^{(2)} \le i \le \min\left\{C,\ C_M^{(1)} + C_M^{(2)}\right\}$$

$$h_k(i) = \min\left\{J_k(j) + h_{k-1}(i-j);\ \max\left\{C_m^{(k)}, i - \sum_{n=1}^{k-1} C_M^{(n)}\right\} \le j \le \min\left\{C_M^{(k)}, i - \sum_{n=1}^{k-1} C_m^{(n)}\right\}\right\}; \qquad 2 \le k \le N-1$$

$$\text{where } \sum_{n=1}^{k} C_m^{(n)} \le i \le \min\left\{C,\ \sum_{n=1}^{k} C_M^{(n)}\right\}.$$

Once we have solved the dynamic programming equations, we obtain an optimal partitioning policy as follows.

$$C^{(N)} = \arg\min\left\{J_N(j) + h_{N-1}(\min\left\{C,\ \sum_{n=1}^{N-1} C_M^{(n)}\right\} - j);\ C_m^{(N)} \le j \le \min\left\{C_M^{(N)}, C - \sum_{n=1}^{N-1} C_m^{(n)}\right\}\right\},$$

$$C^{(N-1)} = \arg\min\left\{\begin{array}{l} J_{N-1}(j) + h_{N-2}(\min\left\{C - C^{(N)},\ \sum_{n=1}^{N-2} C_M^{(n)}\right\} - j); \\ \max\left\{C_m^{(N-1)}, C - C^{(N)} - \sum_{n=1}^{N-2} C_M^{(n)}\right\} \le j \le \min\left\{C_M^{(N-1)}, C - C^{(N)} - \sum_{n=1}^{N-2} C_m^{(n)}\right\} \end{array}\right\},$$

$$C^{(N-m)} = \arg\min\left\{\begin{array}{l} J_{N-m}(j) + h_{N-m-1}(\min\left\{C - \sum_{n=0}^{m-1} C^{(N-n)},\ \sum_{n=1}^{N-m-1} C_M^{(n)}\right\} - j); \\ \max\left\{C_m^{(N-m)}, C - \sum_{n=0}^{m-1} C^{(N-n)} - \sum_{n=1}^{N-m-1} C_M^{(n)}\right\} \le j \le \min\left\{C_M^{(N-m)}, C - \sum_{n=0}^{m-1} C^{(N-n)} - \sum_{n=1}^{N-m-1} C_m^{(n)}\right\} \end{array}\right\};$$

$$1 \le m \le N-2$$

The last assignment is then easily computed as $C^{(1)} = C - \sum_{n=2}^{N} C^{(n)}$.

APPENDIX B. Bandwidth request calculation in the OP policy

Let $C_{rt}^{(i)}$, $C_{nrt}^{(i)}$ and $C_{req}^{(i)}$ be the minimum bandwidth for the satisfaction of the constraint on call blocking and cell loss probability and the overall bandwidth request of station i, respectively. First, $C_{rt}^{(i)}$ is computed from (1) and (4). Then, after substituting $C_{rt}^{(i)} + C_{nrt}^{(i)}$ in lieu of $C^{(i)}$ in the expression (2) of $C_{ng}^{(i)}$ (for the sake of simplicity, we omit the time dependence in $C_{ng}^{(i)}(t)$), let us consider the following formula [16], which gives

an asymptotic (in the buffer length $Q^{(i)}$) upper bound to the cell loss probability in a buffer loaded with the self-similar traffic introduced in Section IV:

$$P_{loss}^{(i)}(X^{(i)}) = \begin{cases} \min\left\{ \dfrac{c \cdot \lambda_{ng}^{(i)} R^{\alpha}}{\alpha \cdot (\alpha - 1) \cdot \left(X^{(i)} - \lambda_{ng}^{(i)} R \bar{\tau} \right)} \cdot (Q^{(i)})^{-\alpha+1},\ 1 \right\} & \text{if } X^{(i)} > \lambda_{ng}^{(i)} R \bar{\tau} \\ 1 & \textit{otherwise} \end{cases} \tag{B1}$$

Some of the parameters appearing in (B1) have been previously defined; the others are explained as follows. Let *T* be a reference time interval (*slot*), to which we will refer all the relevant parameters of the cell queue. The slot also represents the minimum duration of a burst, and the burst length τ is expressed as an integer number of slots. Let B_{ng} be the peak generation rate of each asynchronous source [bits/s], and *L* the number of bits in a cell. Then, $R = \lceil T / (L / B_{ng}) \rceil = \lceil TB_{ng} / L \rceil$ is the number of cells generated by an active burst in a slot ($\lceil x \rceil$ being the smallest integer greater than or equal to x). The numbers of new sources becoming active in each slot are i.i.d. Poissonian with parameter $\lambda_{ng}^{(i)} = \lambda_{burst}^{(i)} \cdot T$. If *H* is the cell's header length in bits, then $X^{(i)} = \left\lfloor \dfrac{C_{ng}^{(i)}}{L+H} \cdot T \right\rfloor$ represents the bandwidth $C_{ng}^{(i)}$ expressed in cells per slot ($\lfloor x \rfloor$ being the largest integer less than or equal to *x*).

Since $n^{(i)}(t)$ in (2) can assume only discrete values from 0 to $N_{\max}^{(i)}$, as a consequence, $C_{ng}^{(i)}(t)$ only takes on discrete values with certain probabilities, depending on the probability of having $n^{(i)}(t)$ connections in progress at time t at station *i*. If we indicate by $X_j^{(i)}$ the realization of the variable $X^{(i)}$, corresponding to $n^{(i)}(t) = j$, and define *D = (L+H)/T*, we have:

$$X_j^{(i)} = \left\lfloor \frac{C_{rt}^{(i)} + C_{nrt}^{(i)} - jBr_{level,rt}^{(i)}(t)}{D \cdot r_{level,ng}^{(i)}(t)} \right\rfloor, \quad j = 0, 1, \ldots, N_{\max}^{(i)} \tag{B2}$$

and

$$\Pr\left\{ X^{(i)} = X_j^{(i)} \right\} = \Pr\left\{ n^{(i)}(t) = j \right\} \tag{B3}$$

where $\Pr\left\{ n^{(i)}(t) = j \right\}$ is given by the stationary distribution of a $M / M / N_{\max}^{(i)} / N_{\max}^{(i)}$ queueing system.

We assume as an indication of the cell loss rate at station *i* the quantity defined in (B1), averaged over the number of guaranteed bandwidth connections; thus we have:

$$\overline{P}_{loss}^{(i)}\left(C_{ng}^{(i)} \right) = \sum_{j=0}^{N_{\max}^{(i)}} P_{loss}^{(i)}(X_j^{(i)}) \cdot \Pr\left\{ n^{(i)}(t) = j \right\} \tag{B4}$$

Then, we obtain $C_{nrt}^{(i)}$ as

$$C_{nrt}^{(i)} = \min_{Z^{(i)}} \left\{ Z^{(i)} : \overline{P}_{loss}^{(i)}\left(C_{ng}^{(i)} \right) = \sum_{j=0}^{N_{\max}^{(i)}} P_{loss}^{(i)}\left(\left\lfloor \frac{C_{rt}^{(i)} + Z^{(i)} - jBr_{level,rt}^{(i)}(t)}{D \cdot r_{level,ng}^{(i)}(t)} \right\rfloor \right) \cdot \Pr\left\{ n^{(i)}(t) = j \right\} \le \gamma^{(i)} \right\} \tag{B5}$$

Finally, since $C_{nrt}^{(i)}$ may turn out to be negative or null, due to the fact that the residual bandwidth might be sufficient to satisfy the constraint on the loss probability, we must have $C_{req}^{(i)} = \max\left\{C_{rt}^{(i)} + C_{nrt}^{(i)}, C_{rt}^{(i)}\right\}$.

It is worth noting that the computations in (1), (4) and (B1)-(B5) can be performed in advance for all possible values of $r_{level,rt}^{(i)}$ and $r_{level,ng}^{(i)}$, and the results may be stored in a lookup table. The values should be recomputed on-line only upon changes in the traffic statistical parameters (a situation that happens on a relatively longer time scale).

REFERENCES

[1] F. Alagoz, D. Walters, A. AlRustamani, B. Vojcic, R. Pickholtz, "Adaptive rate control and QoS provisioning in direct broadcast satellite networks", *Wireless Networks*, vol. 7, no. 3, pp. 269-261**,** 2001.

[2] R.Gibbens, F. Kelly, P. Key, "A decision theoretic approach to call admission control in ATM networks", *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1101-1114, Aug. 1995.

[3] E.W. Knightly, N.B. Shroff, "Admission control for statistical QoS: theory and practice", *IEEE Network Mag.*, vol. 13, no. 2, pp. 20–29, March/April 1999.

[4] M. Naghshineh, M. Schwartz, "Distributed call admission control in mobile/wireless networks", *IEEE J. Select. Areas Commun.*, vol.14 , no. 4 , pp. 711-717, May 1996.

[5] N. Celandroni, F. Davoli, E. Ferro, "Static and dynamic resource allocation in a multiservice satellite network with fading", *Internat. J. of Satellite Commun.*, Special Issue on Satellite IP Quality of Service, vol. 21, no. 4-5, pp. 469-487, July-Oct. 2003.

[6] *ETSI EN 300 421 V1.1.2*, Digital Video Broadcasting, frame structure, channel coding and modulation for 11-12 GHz satellite services, 1997.

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

[8] P.B.Key, "Optimal control and trunk reservation in loss networks", *Prob. Eng. Inform. Sci.*, vol. 4, pp. 203-242, 1990.

[9] G. Choudhury, K. Leung, W. Whitt, "An algorithm to compute blocking probabilities in multi-rate, multi-class multi-resource loss models", *Adv. Appl. Prob.*, vol. 27, pp. 1104-1143, 1995.

[10] C. C. Beard, V. S. Frost, "Prioritized resource allocation for stressed networks", *IEEE/ACM Trans. Networking*, vol. 9, no. 5, pp. 618-633, 2001.

[11] S. B. Biswas, B. Sengupta, "Call admissibility for multirate traffic in wireless ATM networks", in *Proc. IEEE INFOCOM*, vol. 2, Kobe, Japan, pp. 649-657, 1997.

[12] R. Bolla, F. Davoli, "Control of multirate synchronous streams in hybrid TDM access networks", *IEEE/ACM Trans. Networking*, vol. 5, no. 2, pp. 291-304, 1997.

[13] M. Filip, E. Vilar, "Optimum utilization of the channel capacity of a satellite link in the presence of amplitude scintillations and rain fade", *IEEE Trans. Commun*. vol. 38, no. 11, Nov. 1990.

[14] N. Celandroni, F. Potortì, "Fade countermeasure using signal degradation estimation for demand-assignment satellite systems", *J. of Commun. and Networks*, vol. 2, no. 3, pp. 230-238, Sept. 2000.

[15] B. Tsybakov, N.D. Georganas, "On self-similar traffic in ATM queues: definition, overflow probability bound, and cell delay distribution", *IEEE/ACM Trans. Networking*, vol. 5, no. 3, pp. 397-409, 1997.

[16] B. Tsybakov, N.D. Georganas, "Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue", *Performance Evaluation*, vol. 32, pp. 57-80, 1998.

[17] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks", *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, 1995.

[18] H.S. Kim, N.B. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers", *IEEE Trans. Networking*, vol. 9, no. 6, pp. 755-768, 2001.

[19] S. Ghani, M. Schwartz, "A decomposition approximation for the analysis of voice/data integration", *IEEE Trans. Commun.*,vol. 42, no. 7, pp. 2441-2452, 1994.

[20] ISO/IEC JTC1/SC29/WG11 N4668, " Overview of the MPEG-4 Standard", March 2002.

[21] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, 1995.
[22] F. Carducci, M. Francesi, "The Italsat satellite system", *Internat. J. Satellite Commun.*, vol. 13, pp. 49-81, 1995.
[23] ITU-T Recommendation G.821, vol. III - fasc. III.3, Malaga-Torremolinos Plenary Assembly, Oct. 1984.
[24] N. Celandroni, E. Ferro, F. Potortì, A. Chimienti, M. Lucenteforte, "Dynamic rate shaping on MPEG-2 video streams for bandwidth saving on a faded satellite channel", *European Trans. Telecommun.*, vol. 2, no. 4, pp. 363-372, 2000.
[25] K. Gokbayrak, C.G. Cassandras, "Online surrogate problem methodology for stochastic discrete resource allocation problems", *J. Optim. Theory Appl*., vol. 108, no. 2, pp. 349-376, Feb. 2001.