Complete Partitioning 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@cnuce.cnr.it, erina.ferro@cnuce.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 resource allocation methods, based on optimized *adaptive complete bandwidth partitioning* strategies, have been studied, simulated and compared between them and with a simple allocation scheme that does not apply any optimization criteria. The allocation methods are studied for a satellite network environment, where a number of earth stations (*traffic stations*) exchange multimedia traffic in weather conditions that may vary from clear sky up to heavy rain, with different levels of fade which affect the transmitted signals. In all the schemes presented, the call admission control (CAC) policy for real-time connections is administered locally in the traffic stations, while a master station is charged to manage the MF-TDMA (multi frequency-time division multiple access) bandwidth allocation policy.

In the first control architecture (*Optimized Centralized, OC*) no explicit requests are issued by the traffic stations for bandwidth allocation; the master computes the allocations by optimizing a cost function which takes into account costs pertaining to the single stations. In the second scheme (*Optimized Proportional, OP*) the master optimizes the allocations according to explicit requests received. The third allocation policy (*Simple Proportional, SP*) allocates the bandwidth proportionally to the requests, without any optimization. This last scheme has been used simply to clearly show that a non-optimal allocation may have a low computational cost, but it may turn out to be less efficient than the others. Figures of merit such as loss, blocking and dropping probabilities are computed for the three methods, in order to evaluate the performance of each system in the same network conditions. Numerical results are provided for a specific network architecture in a real environment, based on the Italsat satellite national coverage payload characteristics.

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

1. Introduction

Resource allocation is one of the main tasks in a network, where different users and services have to 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 the channel quality, the dynamic control of the allocated resources becomes a challenging problem. Typically, control actions need to be exerted over a

widely different range of time scales, to cope with variations that may happen with frequencies ranging from milliseconds to minutes or hours. Other works in the literature (see, e.g., [1] in the satellite environment and [2-4] in different contexts) have already been focused on optimal control choices. Satellite systems not only have to face variable load multimedia traffic but also variable channel conditions with stringent propagation delay as well. The variability in operating conditions is due both to changes in the traffic loads and to the signal attenuation on the satellite links because of bad atmospheric events, which particularly affect the transmissions in the Ka band (20-30 GHz). It is therefore stringent 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 surely the best way to combat the channel variability. This procedure is anyway complex and difficult to obtain, due to the necessary synergies among the levels. In the work here presented we coordinate the actions taken at the physical layer, where the fade countermeasure technique is applied, with the work done at the data link layer, where we try to allocate the bandwidth according to an optimized policy. Though in a different traffic context, our approach is in the same philosophy as [1].

The present work stems from the study initiated in [5], where we assumed that the satellite network consists of a master station, which exerts the control on the access to the common resource, i.e. the satellite bandwidth, and of a number of traffic stations, which have to exchange both real-time (stream) and non-real-time (bulk) traffic. The former is modeled as Continuous Bit Rate (CBR) guaranteed connections (voice or MPG4 video), whereas 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 (ABR traffic). 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 assume to counteract the fade attenuations of the signals due to bad weather conditions by applying adaptive FEC (Forward Error Correction) codes and bit rate variation on data. 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. The work presented in this paper is a case of a wider study, part of which is still in progress. Here we assume that the attenuation seen by each station is independent of the destination of its traffic; this is the case where the fading is of up-link-predominant type, or when all the traffic sent by a station is addressed to destinations affected by the same environmental conditions. Work is in progress to generalize the model used in order to include the case of stations experiencing both up-link and different down-link fading conditions, which requires a more complex numerical analysis of the scenario and a higher computational complexity of the simulation.

In order to avoid too many oscillations in applying the fade countermeasures according to each single fade level variation, we assume that the measured level of the signal attenuation be categorized in a class "*level*", so that the countermeasure strategy adopted remains unchanged for all the levels of signal attenuation belonging to the same class. Thus, for each type of traffic with a given Bit Error Rate (BER) requirement, a fade class aggregates fade levels that require the same data redundancy, expressed, at station i, by

redundancy coefficients $r_{level}^{(i)}$, level=1, 2,..K, where K is the number of levels of fade classes, equal for all the stations. As best effort traffic and real-time traffic have generally different QoS requirements in terms of BER, we will indicate with $r_{level,ng}^{(i)}$ and $r_{level,rt}^{(i)}$ the respective redundancies. 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 in carriers, each one accessed in TDMA. We also assume that a traffic station cannot transmit at different frequencies in the same temporal slot.

The control architecture treated is essentially hierarchical in nature. Basically, the traffic stations exert control actions on a short- and medium-term time scale, by choosing the redundancy to be applied to the data in transmission, and by deciding on the call admission control (CAC) of locally originated real-time traffic, within a bandwidth temporarily assigned to each station by the master. The latter plays the role of a coordinator in the hierarchical setting, by dividing the total available bandwidth among the traffic stations.

In [5] we studied an allocation policy performed by the master 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 that the master had to solve a discrete optimization problem to perform any reallocation at the beginning of the stated time interval; the required computational time imposes a minimum on the length of that interval, which may not be compatible with the frequency of the reallocation requests coming from the traffic stations. In any case, the work done opened the way for further interesting investigations, such as: i) the adoption of different optimized allocation strategies; ii) the introduction of explicit constraints on the minimum and maximum allowable bandwidth assignments in the parametric optimization problem to be solved by the master station; iii) the adoption of on-line gradient descent techniques, of the type considered in [6], 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; iv) the evolution of the network scenario versus more complex up-link and down-link fade combinations; v) the utilization of allocation policies of complete sharing type.

The current work relates to points i) and ii) previously mentioned, in the context of the same satellite network environment used in [5], trying to reduce the master's computational time for the bandwidth allocation algorithm designed in [5]. The master's optimization criterion, modified with the introduction of constraints, is here 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 reallocation procedure, without explicitly indicating a specific amount of bandwidth.

In addition, we investigate two other policies, referred to as "Optimized Proportional" (OP) and "Simple Proportional" (SP), respectively. In OP the master acts passively, just 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. In SP the master again acts passively, just making assignments proportional to the requests received, but the requests do not take into consideration any specific prediction, being based only on measurements of the traffic intensity. After defining these policies more precisely, we compare their performance in terms of cell loss, call blocking and call dropping probabilities, given a maximum tolerable BER. As the real-time traffic is modeled as CBR calls, the dropping refers to the

case where the bandwidth available is insufficient to maintain all on-going connections with the desired BER. In the three systems that we consider, the master station computes partitions of the available bandwidth that remain fixed for a certain amount of time. Thus the allocation policies fall in the category of *Complete Partitioning* [7] (where, however, the partitions are *adaptively* changed in response to traffic or fading variations).

The reader can found in Section 2 the time scale structure of the system, and in Section 3 the numerical analysis of the OC and OP policies studied. In Section 4 we present the simulation environment where the methods have been compared and the relevant results are commented in Section 5. The conclusions that we derived from this study are summarized in Section 6, which also anticipates the work that is in progress as a completion of this study. In Appendix A the interested reader can found the mathematical description of the constrained dynamic programming method used by the master in OC.

2. The multi-time-scale control structure

In all the three policies, we consider a control architecture that comprises four time scales: a) the master allocation; b) the change of fade class; c) the frame interval, during which each traffic station estimates its fade level; d) the traffic station's data transmission.
The bandwidth assignments made by the master station are upper and lower limited. The constraint relevant to the minimum assignment is specific to each single station; namely, the already outstanding connections of that station must be maintained, in order to avoid that the relevant call dropping probability reaches unacceptable values. The constraint relevant to the maximum assignment is common to all the stations: the bandwidth assigned to a single station cannot exceed the single carrier capacity. The number of supportable call connections depends on the allocated bandwidth and on the blocking and loss probability thresholds.
Let us see now what happens on the four different temporal time scales. Longer time scales are indicated with capital letters; small letters indicate the shorter scales.

2.1 <u>*First level*</u> (on the longest time scale *S*, also called *superframe* time scale).
At each convenient time t, each station i sends the master an information vector $\mathbf{v}^{(i)}(t)$, simply indicated with $\mathbf{v}^{(i)}$, which 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 burst intensity $\lambda_{burst}^{(i)}$ [burst/s] and the mean burst length $\pi^{(i)}$ for best-effort traffic, together with the last applied data redundancy coefficients for the two types of traffic. In both the OP and SP policies, $\mathbf{v}^{(i)}$ represents the total bandwidth request, expressed in minimum bandwidth units (*mbu*, the minimum bandwidth granularity). This value is comprehensive of the needs for the station to accommodate both the real time connections and the best effort traffic within the desired respective QoS requirements (the separate values of the stream request and the datagram request are known individually by the traffic stations only).
In the OC case, on the basis of the last updated $\mathbf{v}^{(i)}$ values, at the (fixed or variable) superframe time, the master computes the bandwidth to be allocated to each station, by minimizing a cost function of the whole

system. The latter, which will be specified in Section 3, takes into account both call blocking and cell loss probabilities of the single stations.

In both the OP and SP cases, the master simply assigns portions of bandwidth to the stations proportionally to the amount of the requests, up to the maximum bandwidth that can be allocated to each station.

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

2.2 <u>*Second level*</u> (on the time scale *E*, also called *fade* time scale).

At each change of fade class, within the received capacity partition $C^{(i)}$, the traffic station i re-computes the threshold $N_{\max}^{(i)}$ on its maximum acceptable connections, given their bandwidth requirements (dependent on the redundancy needed) and the desired upper bound on the call blocking probability. This threshold serves the purpose of CAC, which is performed locally at each traffic station, independently of the bandwidth allocation method used by the master.

At each change of fade class, the traffic station i sends the master two different messages, according to two different cases. Case i) The new fade class is lower than the previous one (lower attenuation): the master station will perform the next synchronous reallocation at the beginning of the new superframe (*fixed superframe time interval*). Reallocations at time frame *S* are performed if at least one notification has been received. Case ii) The new fade class is higher than the previous one (higher attenuation): due to higher redundancy to be applied to data, no residual bandwidth may be available for best-effort traffic or, more generally, a "safety margin" is exceeded in the bandwidth occupied by the ongoing synchronous connections (pre-outage situation). In this case the master will perform an asynchronous reallocation immediately (*variable superframe time interval*).

2.3 <u>*Third level*</u> (on the time scale *f*, also called *frame* time scale):

Every time frame *f*, each traffic station i measures its attenuation value(s) and determines the appropriate class of fade. This procedure is made for each "link" the station has active with any other station.

The *S* time interval is generally longer than *f*. At the beginning of each frame *f* a reference burst is received from the master, which is a copy of the RB sent at time *S*, as far as the allocations are concerned, with updated values relevant to the fades of all stations.

2.4 <u>*Fourth level*</u> (on the shortest time scale *w* within the frame *f*):

By using the last measured attenuation value(s) and the ensuing classification and redundancy(ies), each station transmits its data in its own transmission window *w*, according to the received transmission time plan. Moreover, each station sends the master, piggybacked with the data, the information on the fade level measured in the current frame *f*. In order to allow the evaluation of the link-by-link fade classes, when necessary, the master will redistribute this information within each RB sent at time *f*. It is worth noting that, in general, the stations can experience different down-link fading conditions, according to the data destination.

3. The three Complete Partitioning 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 the real-time traffic, the traffic dynamics of interest is at the connection level, and the relevant performance index is the call blocking probability (P_{block}), that is the steady-state probability that an arriving call is refused because the bandwidth devoted to the real-time traffic is all 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). As we assume that all connections of station i belong to a single class, i.e. the same redundancy is always applied (when necessary), we face a very particular single-class case, where the expression of the blocking probability reduces to the classical Erlang B loss formula. From this formula, the maximum number of acceptable calls $N_{\max}^{(i)}$ is easily derived, given the desired upper bound on blocking probability. In general, this is not a very typical situation. In fact, in the reality, a station is more generally in the multiclass case, where the connections that utilize a given bandwidth portion have different statistical parameters and peak rates, and belong to different classes according to the fade classes. This situation occurs in the case where a station experiences different down-link attenuations and therefore must apply, at the same time, different data redundancies. In this multiclass case, the blocking probability results from a stochastic knapsack problem [7], as we are going to analyze in a wider study.

As far as the best effort traffic is concerned, we suppose that it originates from datagram 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, which represents the superposition of on-off sources whose active periods (bursts) have Pareto-distributed 'on' time $\square(\Pr\{\square = \ell\} = c\ell^{\square\square 1}, 1 < \square < 2$ (where $//$ 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 [8, 9]. Actually, there are a number of possible models that can be adopted to approximate the cell loss probability, given the statistical characteristics of burst generation and a fixed rate of extraction of cells from the buffer [9, 10, 11]. 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 non-guaranteed traffic after serving all guaranteed-bandwidth connections in progress at the required peak transmission rate B. Namely,

$$C_{ng}^{(i)}(t) = C^{(i)} \square Br_{level,rt}^{(i)}(t)n^{(i)}(t) \tag{1}$$

where $C^{(i)}$ is the total 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]. 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 already mentioned, the real-time and 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 [14], and at least 10^{-7} for best effort traffic). Consequently, the redundancy to be applied to the latter will be normally higher than the other one. This redundancy factor will further reduce the effective residual capacity. The dependence on the redundancy factors, which are time-varying quantities, deserves a further comment. In fact, all previously discussed calculations regarding the performance indexes are done by considering the current values of these coefficients as lasting forever, and are repeated at each change of fade class. The resulting control scheme is thus a sort of "repetitive control", where the initial time continuously shifts ahead.

3.1 The Optimized Centralized strategy (OC)

In this strategy the traffic stations do not communicate explicit bandwidth requests to the master station, but, at each change of the current fade class, they send the information vector $\mathbf{v}_i$ mentioned in Section 2.1.

At the appropriate time 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, while maintaining the system's constraints satisfied. 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. 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 below the minimum bandwidth necessary to satisfy the 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 taken as a cost. The measured values of the redundancy factors, as known to the master station, are used in the evaluation of the cost. For the sake of clarity, we re-write here the expression of the cost function, which was introduced in [5].

$$J^{(i)}(C^{(i)}) = \begin{cases} \overline{P}_{loss}^{(i)}\left[C_{ng}^{(i)}(S)\right] & \text{if } C^{(i)} \geq C_{rt}^{(i)} \\ \Pi & \text{if } C^{(i)} < C_{rt}^{(i)} \end{cases} \tag{2}$$

where $C^{(i)}$ represents the capacity variable to be allocated to station i, $C_{rt}^{(i)}$ is the minimum amount of capacity that guarantees the satisfaction of the constraint on the blocking probability, and Π is a penalty term (a value of 10 is adequate in most cases). Specifically, if we want that $P_{block}^{(i)} \leq \Delta^{(i)}$ be satisfied at station i, the master has to compute the minimum bandwidth capacity for the real-time traffic of station i:

$$C_{rt}^{(i)} = \min_{Y^{(i)}} \left\{ Y^{(i)} : P_{block}^{(i)}\left(Y^{(i)} / \left(B \cdot r_{level,rt}^{(i)}(S)\right)\right) \leq \Delta^{(i)} \right\} \tag{3}$$

[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, 13].

where

$$P_{block}^{(i)}(M) = \frac{\left(\rho^{(i)}\right)^M / M!}{\sum_{j=0}^{M} \left(\rho^{(i)}\right)^j / j!} \tag{4}$$

and $\lfloor x \rfloor$ is the largest integer less than or equal to x.

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

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

where N is the number of stations in play, and C_c is the maximum allowable information rate for each carrier. The last constraint in (6) stems from the willingness to guarantee continuation of the connections in progress. The minimization of (5) can be efficiently effected by means of a dynamic programming algorithm, as reported in [7]. The algorithm has been modified to take into account the presence of the constraints (6). The modified version is reported in the Appendix A. It is worth noting that the presence of the 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 previous assignments as an upper bound for those stations that did not signal any necessity of reallocation.

The problem admits a solution $C^{(i)} = C_{opt}^{(i)}$, $i = 1,...,N$ if $\sum_{j=1}^{N} C_{rt}^{(j)} < C$, as computed by the master; otherwise, the master computes the allocation as

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

3.2 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 getting the minimum bandwidth necessary to support both the types of traffic of the station, under given QoS constraints on call blocking and cell loss probabilities. The master simply analyzes the requests in FIFO order and assigns the bandwidth proportionally to the requests received.

The problem of how to compute the request is locally solved inside each station. Various methods may be applied. For example, formula (3) 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 on the average loss probability ($\overline{P}_{loss}^{(i)}$ □□$^{(i)}$). Appendix B contains the detailed computations in the case where the expression adopted for $P_{loss}^{(i)}$ is given by the Tsybakov-Georganas formula [9].

3.3 The Simple Proportional strategy (SP)

This strategy is similar to the previous one, just differing in how the requests are computed locally at the single station. No optimization is done nor any constraint (except the one on the maximum allowable capacity) is enforced in this simple strategy. The request is simply computed by multiplying the relevant redundancy factors by the mean non-guaranteed traffic load and the bandwidth corresponding to the number of active connections, respectively, and summing the values. The master computes the allocations proportionally to the requests received.

4. The simulation environment

We simulated the three strategies by considering a fully meshed satellite network that uses bent-pipe geostationary satellite channels. This means that the satellite performs only the function of a repeater and it does not make any demodulation of data. The system operates in MF-TDMA mode. The master station maintains the system synchronization other than performing capacity allocation to the traffic stations. The master station performance is the same as the others, thus the role of master can be assumed by any station in the system. This assures that the master normally operates in pretty good conditions, because when the current master's attenuation exceeds a given threshold, its role is assumed by another station that is in better conditions. To counteract the signal attenuation the system operates up-link power control, bit and coding rates changing. Traffic stations transmit in temporal slots assigned by the master, each one generally on different TDMA carriers (frequencies). The multi-frequency feature allows us to divide the system capacity into a number of sub-channels, so that the stations can be downsized with respect to a pure TDMA system. Each station cannot transmit simultaneously on different TDMA carriers in the same temporal slot, because it has only one modulator; thus, the transmission capacity per station is limited to that of one carrier. Each station can receive from all others, instead, because we assume it has one demodulator per carrier.

Table I reports the most significant system parameters. In order to compute the link budget we took data from [15], relevant to the transponder #1 of the Italsat national coverage payload (20/30 GHz band), which is no longer operative at the moment, 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 rate is 10^{-7} and 10^{-4}, respectively.

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 datagram (connections)	7 (5) [dB]
BER guaranteed for datagram (connections)	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]

Table 1. Most significant values of the MF-TDMA system considering the Italsat payload.

In order to compute the resulting net values of E_b / N_0 at the earth station's receiver input we used relation (8). 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:

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

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 $A_u - p_r$ (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).

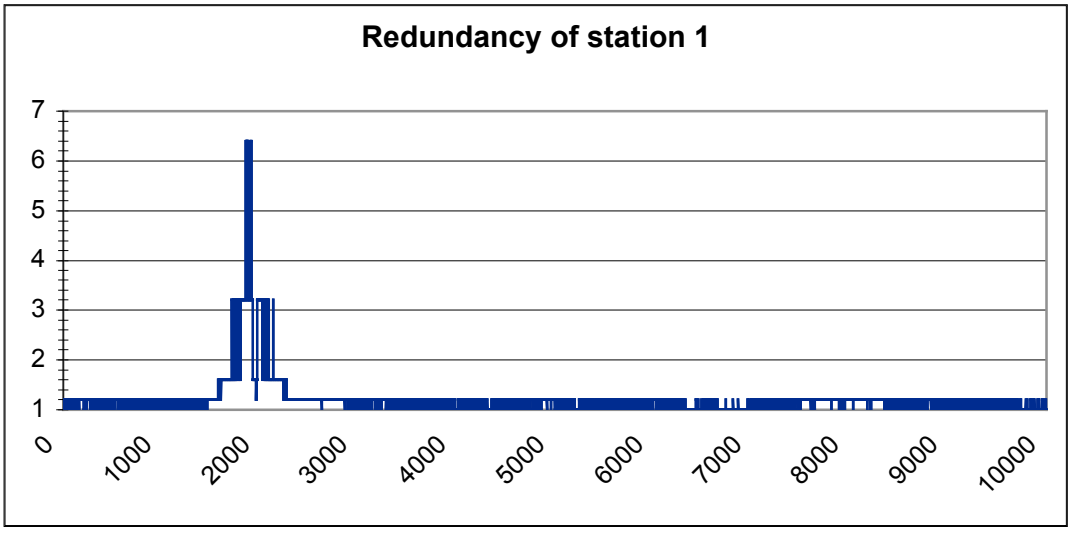
Table II contains the fade levels of the traffic stations, called *fade classes*, as function of the C/N_0 values. Each fade class imposes the adoption of the indicated transmission parameters (and then r_{level} values) to limit the BER below the chosen thresholds.

The system configuration used to make comparisons among the three allocation strategies is made by ten active stations, five of which are in clear sky, and five experience up-link fading. The attenuation patterns of each of the five stations in fade, used for simulation runs, are shown in Fig. 1.

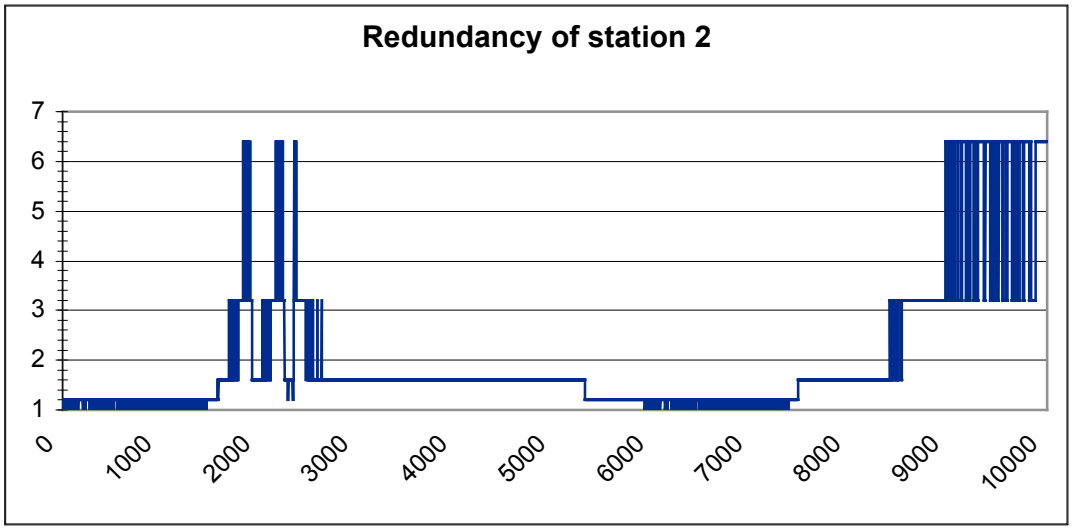
The attenuation data are taken from a data set chosen from the results of the propagation experiment, in Ka band, carried out on the Olympus satellite by the CSTS (Centro Studi sulle Telecomunicazioni Spaziali) Institute, on behalf of the Italian Space Agency (ASI). The up-link (30 GHz) and down-link (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 (North of Italy) station, in September 1992.

Fade classes	r_{level}	*Coding rate, bit rate [Mbit/s]*	C/N_0 [dB] (best effort)	*Net* E_b/N_0 (best effort)	C/N_0 [dB] (connections)	*Net* E_b/N_0 (connections)
1	1	4/5, 8.192	>77.13	7 dB	>75.13	5 dB
2	1.2	2/3, 8.192	74.63-77.13	4.5 dB	72.63-75.13	2.5 dB
3	1.6	1/2, 8.192	72.63-74.63	4.5 dB	70.63-72.63	2.5 dB
4	3.2	1/2, 4.096	69.63-72.63	2.5 dB	67.63-70.63	0.5 dB
5	6.4	1/2, 2.048	66.63-69.63	2.5 dB	64.63-67.63	0.5 dB
6	-	outage	< 66.63	-	< 64.63	-

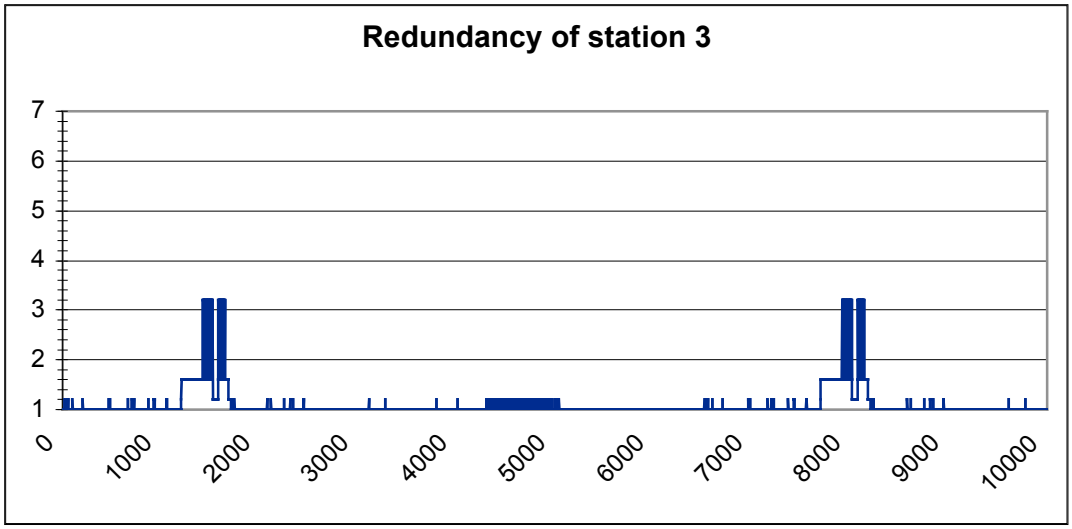
Table II. Redundancy factor r_{level} and signal to noise ratios versus fade classes.



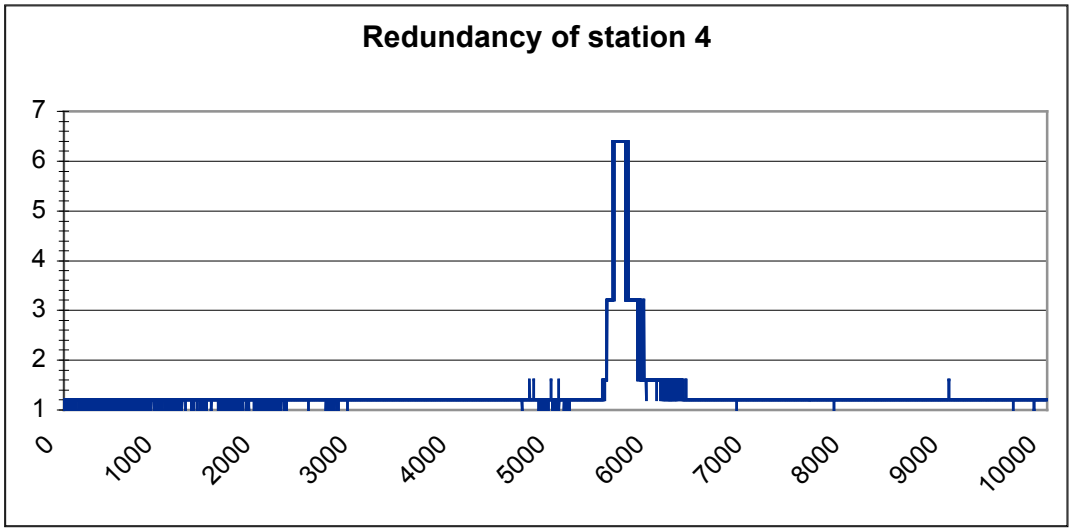
(a)


Redundancy of station 2
7
6
5
4
3
2
1
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

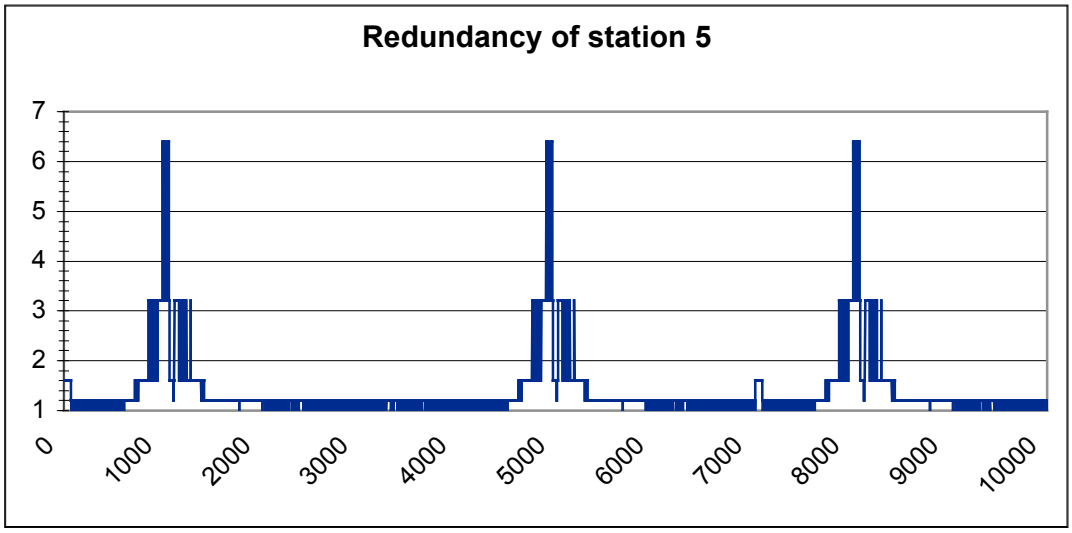
(b)


Redundancy of station 3
7
6
5
4
3
2
1
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

(c)


Redundancy of station 4
7
6
5
4
3
2
1
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

(d)



(e)

Fig. 1. Redundancy values (y axis) vs time [sec] (x axis) represented for the five faded stations (Figs. a-e).

5. The simulation results

We report in the following the results of a simulative analysis of the proposed adaptive strategies. The attenuation level determines, according to (8) and Table II, the attribution of each station's traffic type to a certain fade class. Actually, the value of attenuation (and, consequently, of $r_{level}^{(i)}$) to be used over a reallocation period is determined by averaging the current estimation with a few previous values, weighted with a forgetting factor. In order to avoid too many oscillations in the "instantaneous" bandwidth assignment of a station, we have 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 outage is concerned, we adopted the definition of "unavailable time" given in ITU-T Recommendation G.821 [16] for connections.

Station connection generation rate for stream traffic	$\lambda_s = 0.2$ [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-guaranteed traffic source peak rate	$B_{ng} = 256$ [kbit/s]
Slot time duration (*see Appendix B*)	T=15 [ms]
Average non-guaranteed traffic burst duration	$\beta = 1.945$ [slot] or 29.175 [ms]
Non-guaranteed traffic burst generation intensity (equal for all stations)	$\lambda_{burst} = 16.67$ [burts/s]
Buffer dimension	8000 [cell]
P_{block} threshold (equal for all stations)	$\Delta_1 = 5\%$
P_{loss} threshold (equal for all stations)	$\Delta_2 = 1\%$

Table III. Data used in the simulation.

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 equal to 8 kbit/s. The stream source rate is assumed to be 64 kbit/s for all sources. The reallocation interval is 1000 s. Each simulation run per station gives values averaged over a 12,000 s 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.
The results are divided into two sets. The first one (Figs. 2-4) 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 (corresponding to a synchronous reallocation interval). The second set (Figs. 5-7) shows the same quantities for each station, averaged over the whole simulation time (12,000 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, and overcomes the threshold for one of the faded stations (# 2), whereas both OC and OP tend to essentially equalize the blocking.

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, 17]). In this scenario, the dropping rates experienced by OP and OC could be further reduced by the adoption of suitable rate adaptation techniques. The blocking and dropping probabilities are anyway loosely related, as a more cautious acceptance behaviour implies a lesser likelihood of dropping in severe fading conditions. In this sense, the OC shows slightly more robustness than the OP (see Figs. 2-3 and 5-6).

A remarkable gain is obtained by the optimized policies as regards the cell loss probability. Both the system averages (Fig. 4) and the individual ones (Fig. 7), are kept under the 1% threshold (with the sole exception of the first 1000-s interval in Fig. 4 for the OP case).
In general, the performance of the OC case tends to be better in all quantities of interest. However, the difference with respect to the OP policy is less remarkable than what one might have expected at first sight, 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 that is required to implement them.

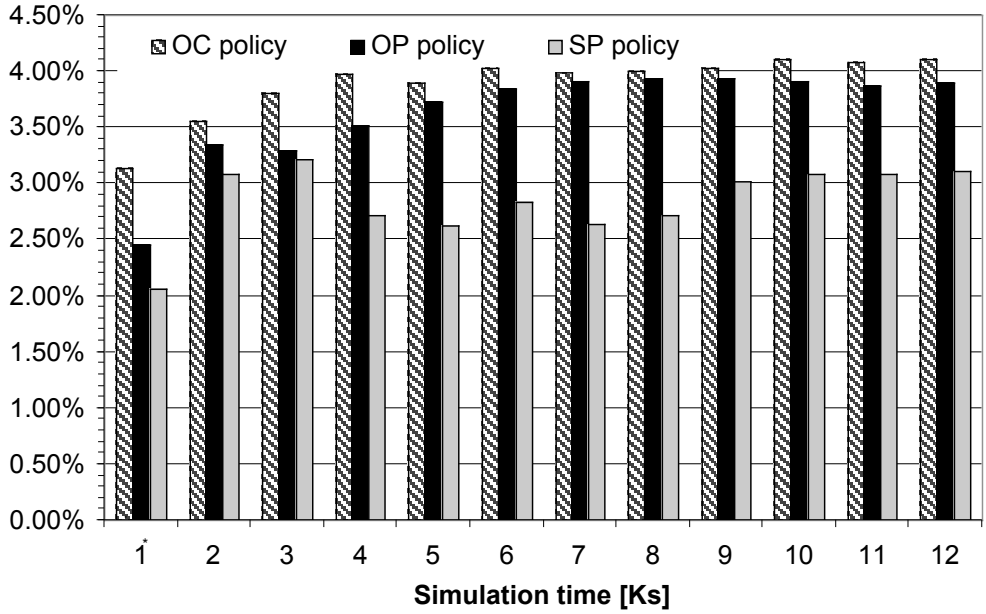


Fig. 2. Average system blocking probability vs simulation time.

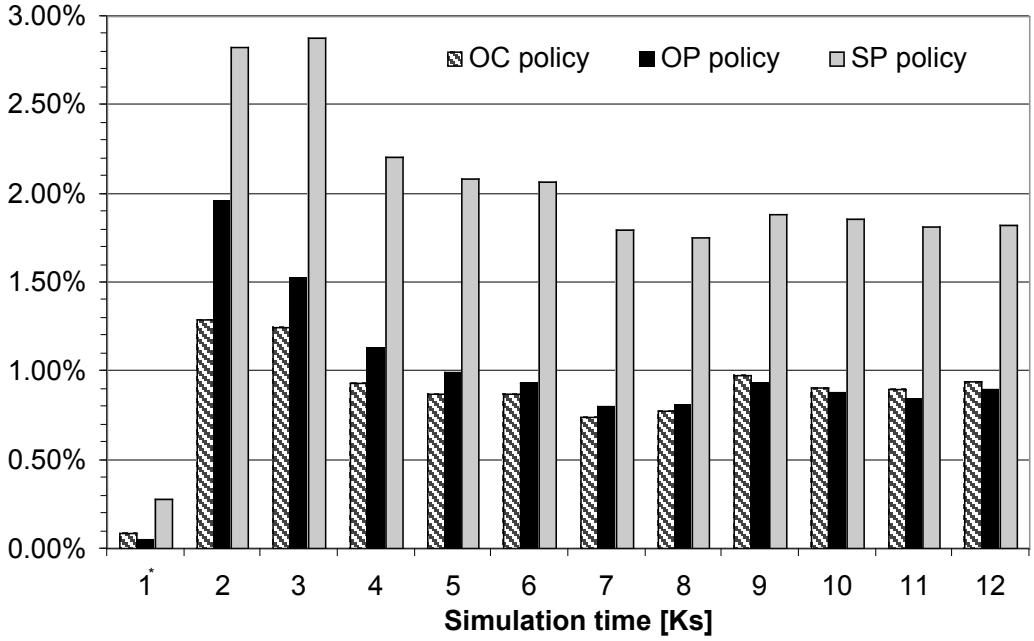


Fig. 3. Average system dropping probability vs simulation time.

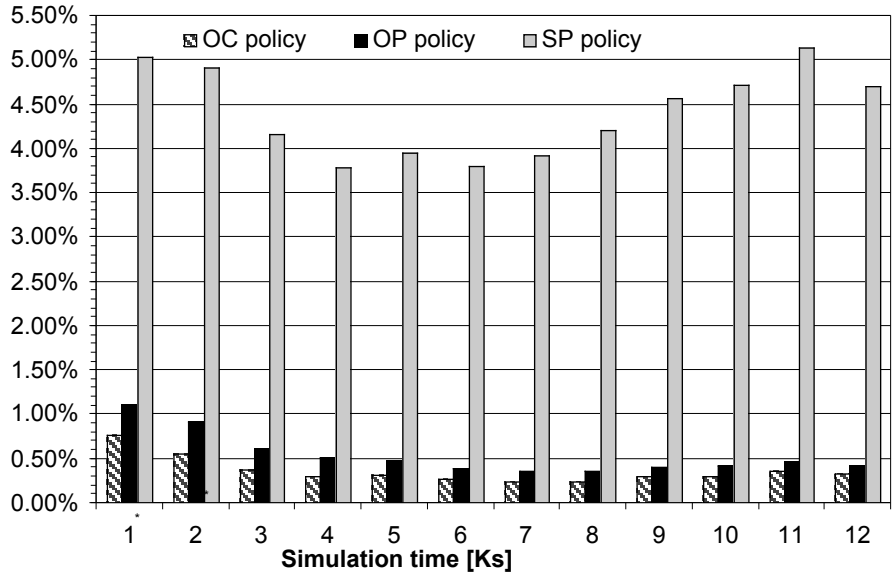


Fig. 4. Average system loss probability vs simulation time.

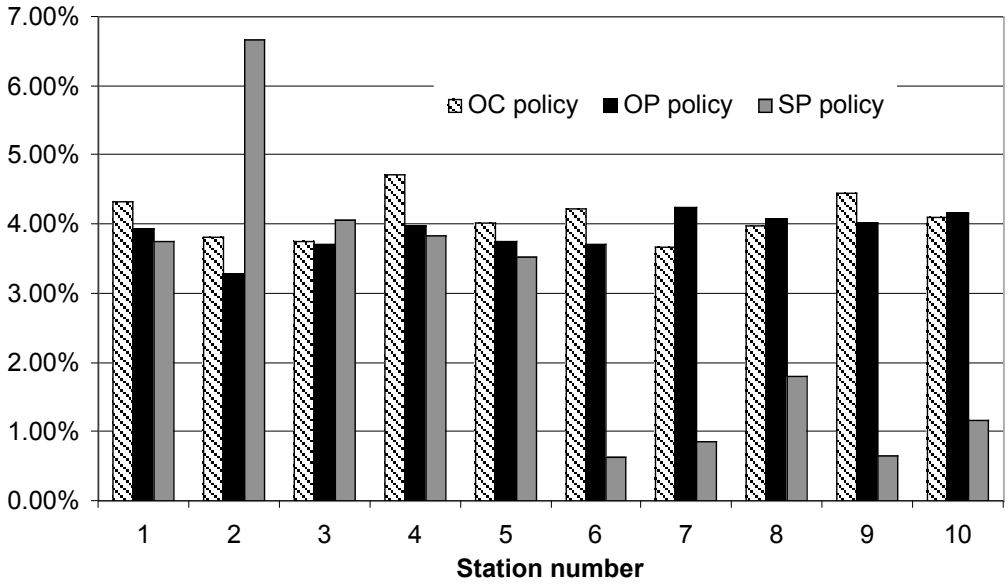


Fig. 5. Average blocking probability per station.

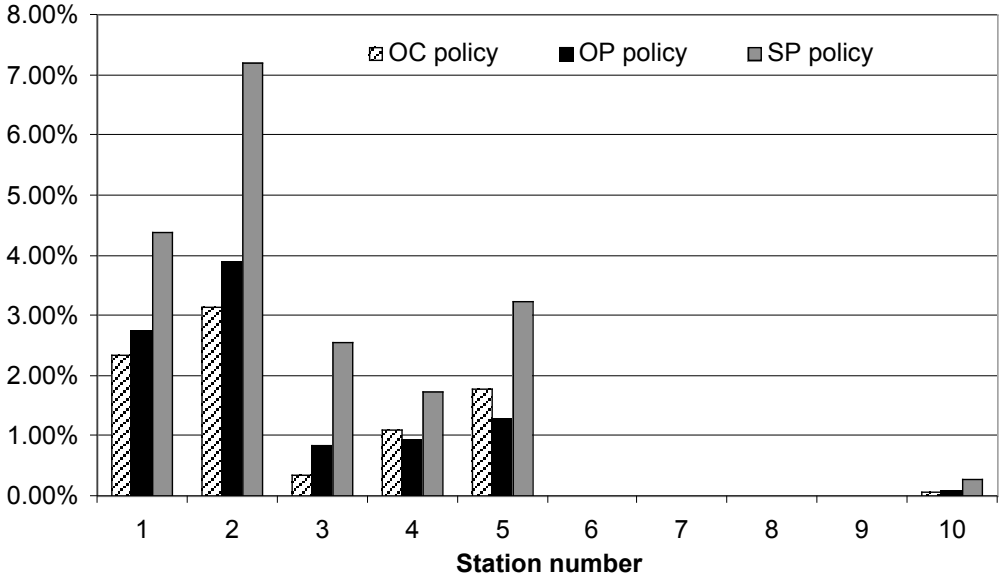


Fig. 6. Average dropping probability per station.

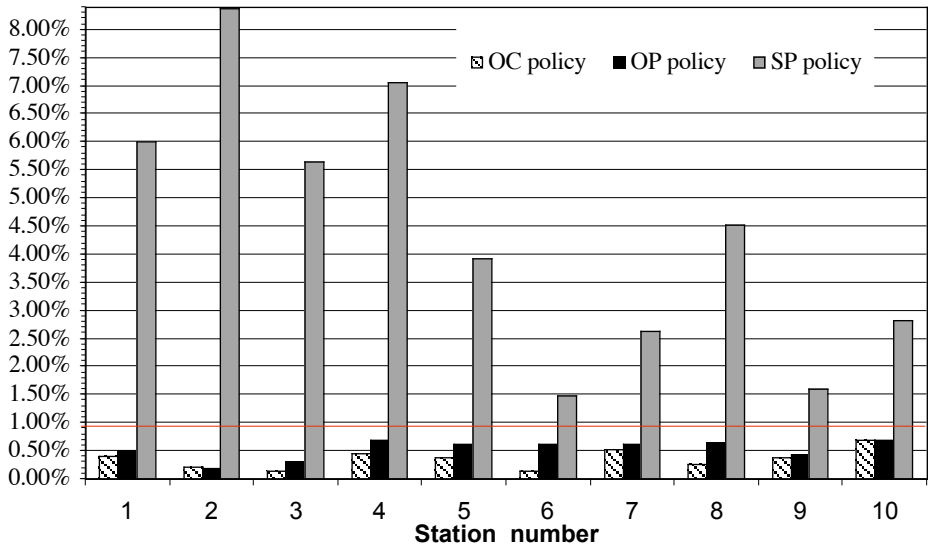


Fig. 7. Average loss probability per station. The horizontal line indicates the 1% threshold.

6. Conclusions

The work presented has addressed 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 the transmission parameters (e.g., data code 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 the present paper was to 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 that 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 comparison indicates that OC is preferable with respect to OP, at the expense of a higher computational complexity. The comparison with the simple SP method has shown that the optimized policies can yield a significant performance improvement in all quantities of interest (cell loss, call dropping and call blocking probabilities).

Further work is in progress 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. Moreover, investigation is in progress about the comparison among complete partitioning, complete sharing, and hybrid allocation policies.

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 tp 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:

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

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

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

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

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

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

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

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

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

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

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 (3) and (4). Then, after substituting $C_{rt}^{(i)} + C_{nrt}^{(i)}$ in lieu of $C^{(i)}$ in the expression (1) of $C_{ng}^{(i)}$ (for the sake of simplicity, we omit the time dependence in $C_{ng}^{(i)}(t)$), let's consider the following formula [9], 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 3:

$$P_{loss}^{(i)}(X^{(i)}) = \begin{cases} \min\left[\dfrac{c \cdot \rho_{ng}^{(i)} R}{\theta \cdot (\theta - 1) \cdot \left(X^{(i)} - \rho_{ng}^{(i)} R \right)} \cdot (Q^{(i)})^{-\theta + 1},\ 1 \right] & \text{if } X^{(i)} > \rho_{ng}^{(i)} R \\ 1 & \text{otherwise} \end{cases} \tag{B1}$$

Some of the parameters appearing in (B1) have been already defined; the others are explained in the following. 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 $1/\beta$ 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 $\rho_{ng}^{(i)} = \rho_{burst}^{(i)} \cdot T$. 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 (1) 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 packet 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

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

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 (3), (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”, to appear in *Internat. J. on Satellite Commun.*, Special Issue on QoS on Satellite IP Networks.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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] 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.

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

[15] F. Carducci, M. Francesi, “The Italsat satellite system”, *Internat. J. Satellite Commun.,* vol. 13, pp. 49-81, 1995.

[16] ITU-T Recommendation G.821, vol. III - fasc. III.3, Malaga-Torremolinos Plenary Assembly, October 1984.

[17] 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.