Available online at www.sciencedirect.com

ScienceDirect

Computer Networks 52 (2008) 275–291

Computer Networks

www.elsevier.com/locate/comnet

Performance evaluation of bandwidth allocation methods in a geostationary satellite channel in the presence of internet traffic

Igor Bisio *, Mario Marchese

DIST – Department of Communication, Computer and System Sciences, University of Genoa, Via Opera Pia 13, 16145 Genova, Italy

Available online 8 September 2007

Abstract

Resource allocation represents an important issue for the next generation TCP/IP Quality of Service-based satellite networks. Many schemes, proposed in the recent literature, consider Internet traffic as the superposition of traffic sources without distinguishing between User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) flows, even if UDP and TCP imply very different traffic characteristics. The basic idea of this work is that a resource allocation algorithm which is conscious of the difference may be more efficient because it can make use of the different behaviour of TCP and UDP in the presence of network congestion. Actually TCP reduces the source flow rate and, as a consequence, also the bandwidth occupancy when there is network congestion. The use of this feature within the bandwidth allocation scheme allows reducing the bandwidth waste due to over provisioning and using the residual bandwidth for other sources. The advantage is particularly clear over satellite channels where fading often affects the communication: having some residual bandwidth available for stations which have experienced fading can improve the satellite network performance.

This paper presents a detailed performance evaluation of a bandwidth allocation scheme, called E-CAP-ABASC and studied for the satellite environment. The bandwidth is assigned to the earth stations that compose the network by a master station on the basis of a cost function whose main part is represented by a closed-form of the packet loss probabilities for the TCP and UDP traffic. The use of two different packet loss probability models for TCP and UDP allows exploiting the different features of the two traffic types, so improving the overall performance either in terms of packet loss or, on the other hand, in terms of the traffic admitted.

The performance evaluation is carried out by varying the link degradation due to fading, the traffic load, and the flow balance between UDP and TCP. The results show a good performance of E-CAP-ABASC, compared with two other schemes. Advantages and drawbacks are discussed.

Keywords: Resource allocation; TCP; Congestion control; Performance evaluation

1. Introduction

TCP/IP based protocols and related networks are the most rapidly spreading technology. Many new applications use these. On the other hand, satellite networks are an essential element in the establishment of long distance communications, and will

* Corresponding author. Tel.: +39 010 3536571/2806; fax: +39 010 3532154.

E-mail addresses: igor.bisio@unige.it (I. Bisio), mario.marchese@unige.it (M. Marchese).

1389-1286/$ - see front matter
doi:10.1016/j.comnet.2007.09.003

have a major role in the implementation of the so called global information infrastructure in the future [1]. Therefore, Internet-based applications represent most traffic over satellite networks. As a consequence adapting a bandwidth allocation scheme which, in satellite environments, has a function as a fading countermeasure, is a key issue.

The reference network for this paper is based on a Geostationary (GEO) satellite accessed by a number (I) of earth stations. Earth stations may be affected by fading. One of them (or the satellite itself, if switching on board is allowed), called the master station, assigns the overall bandwidth among all earth stations. The assignment depends on a cost function whose value is related to the fading level and to the traffic load. The traffic is composed of guaranteed calls (e.g., CBR, Constant Bit Rate, phone calls) and of TCP and UDP flows. The bandwidth assigned to each single station is divided into two portions. One part is dedicated to guaranteed calls (CBR traffic). Once the bandwidth is fixed for them, the maximum number of guaranteed calls that can be accepted in one station is fixed and is used within a CAC scheme. In other words, just to make an example, if the bandwidth portion dedicated to guaranteed calls in one station is 1280 kbit/s and each CBR call has a bit rate of 128 kbit/s, it means that the station can accept 10 CBR calls at most. This rule is applied by the CAC that decides if CBR calls can be admitted in the station or not. The other bandwidth portion is dedicated to TCP/IP traffic, which is the main focus of the paper.

TCP/IP traffic may be divided into TCP and UDP flows. Their behaviour is very different concerning congestion countermeasures. UDP behaviour is independent of what happens in the network. TCP reduces its rate if congestion is detected. Many traffic models and resource allocation schemes do not consider the different reaction to congestion and do not distinguish between TCP and UDP traffic. They model Internet flows as a mere superposition of TCP/UDP sources. The assumption is correct if the number of flows is so large as to be assumed infinite but, in real conditions, the differentiation between TCP and UDP traffic in the bandwidth control scheme is a way to save bandwidth and to have it available in case of need (i.e., for a faded station).

The paper takes its origin from a bandwidth allocation scheme, called CAP-ABASC [2] and identified as CAP-1 in the following, where the TCP and UDP flows are not differentiated for bandwidth allocation and proposes a scheme, called E-CAP-ABASC (Extended-CAP-ABASC), where TCP and UDP are distinguished through two different packet loss probability models appearing in the cost function that determines bandwidth allocation.

Fading is modelled by assigning a probability of channel degradation to each satellite link, along with a weighting coefficient to "measure" the degradation itself. The degradation is seen as a bandwidth reduction. Technically it may be due to the use of fading compensation techniques such as Forward Error Correction (FEC) schemes.

The paper includes a detailed performance evaluation section that allows investigating the main characteristics of E-CAP-ABASC by varying traffic and channel conditions.

The paper is structured as follows: Section 2 contains the description of the network topology, of the channel model and of the general structure of the allocation scheme. Section 3 describes the system architecture and the features of E-CAP-ABASC. Section 4 shows the performance evaluation and Section 5 lists the conclusions.

2. Bandwidth allocation methodology

2.1. Network topology

The GEO satellite network considered is composed of I earth stations connected through a mesh topology. One of them (or, the satellite) is the "master" and controls the bandwidth assignment to the single stations which gather traffic from the users. The study is not linked to a particular bandwidth choice but Ka-band (20–30 GHz) may be the technological reference because here the effect of fading is particularly important.

2.2. The channel model

The fading effect is modelled as bandwidth reduction. Mathematically, this means that the nominal bandwidth $C^{(i)}$ (assigned to a station i) is reduced by using a factor $\beta^{(i)}$ as in (1). $\beta^{(i)}$ is a stochastic parameter uniformly distributed in the real interval $[0,1]$

$$C^{(i)}_{\text{real}} = \beta^{(i)} \cdot C^{(i)}. \tag{1}$$

A specific value $\beta^{(i)}$ corresponds to a fading level measured at the ith station. Each fading level happens with probability $p_f^{(i)}$. Even if describing the fading as bandwidth reduction seems to be rather intuitive, a technical interpretation of the factor $\beta^{(i)}$ may be the presence of a FEC (Forward Error Correction) scheme which extends the bits dedicated to the FEC when the Bit Error Ratio (BER) increases and, in consequence, reduces the portion of the frames dedicated to the transport of information. Considering fading as bandwidth reduction allows making an important assumption: in satellite environments, the link corruption due to noise is regular and the packet loss is due mainly to it; nevertheless, associating link corruption with the bandwidth decrease allows considering each loss as a congestion event. For example, if FEC bits to counteract fading are increased, the errors due to link that experience fading may be neglected but, at the same time, the available bandwidth for information is reduced. This creates possible bottlenecks and consequent losses. In this view, this paper explicitly assumes that all packet losses are due to congestion. The assumption is not too restrictive, within the described framework, and seems a reasonable approximation of the satellite channel behaviour, at least concerning the bandwidth control algorithms.

2.3. Structure of the bandwidth allocation scheme

The general structure of the bandwidth allocation scheme has been introduced in [2]. It is reported in Fig. 1. It is composed of a higher layer called Centralized Bandwidth Allocator (CBA) that distributes the bandwidth capacity among the earth stations identified by the index i and of a lower layer, called Local Controller (LC), which splits the capacity allocated to each station into two contributions: $C_{\mathrm{CBR}}^{(i)}$, assigned to the guaranteed traffic (modelled through Constant Bit Rate (CBR) calls), and $C_{\mathrm{be}}^{(i)}$, assigned to the non-guaranteed TCP-UDP traffic (best-effort Internet flows). Each earth station solves a local optimization problem to share the assigned capacity. It finds a threshold $C_{\min}^{(i)}$, which is the minimum bandwidth that must assigned to the ith station to guarantee a specific QoS level in terms of call blocking probability. The QoS level is modelled by a fixed threshold $\gamma^{(i)}$ (related to the ith station), which upper bounds the call blocking probability.

$$C_{\min}^{(i)} = \arg\min_{X^{(i)}} \left\{ X^{(i)} \in \Re : P_{\mathrm{B}}^{(i)}\left(\left\lfloor \frac{X^{(i)}}{R_{\mathrm{CBR}}^{(i)}} \right\rfloor \right) \leqslant \gamma^{(i)} \right\},$$
$$0 < X^{(i)} \leqslant C_{\mathrm{tot}}. \tag{2}$$

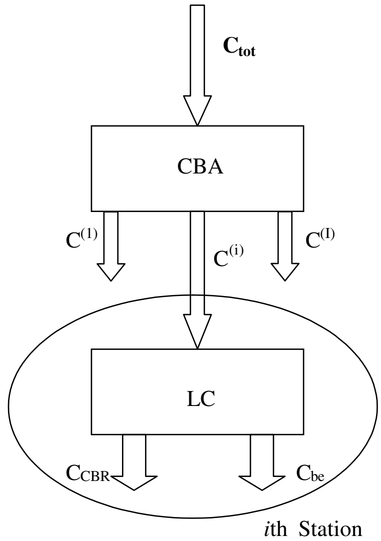

Fig. 1. Bandwidth allocation scheme.

C_{tot} is the overall channel bandwidth. $R_{\mathrm{CBR}}^{(i)}$ is the bit rate at which CBR sources at station i emit data. It is considered the same for all calls. The Call Admission Control for each station, which determines the access of guaranteed traffic, is modelled by an M/M/m/m queuing system whose blocking probability is given by the Erlang B formula. In practice, the call blocking probability of the ith station is given by (3) as a function of the maximum number of available servers $x_{\max}^{(i)}$:

$$P_{\mathrm{B}}(x_{\max}^{(i)}) = \frac{\frac{1}{x_{\max}^{(i)}!}\left(\frac{\lambda^{(i)}}{\mu^{(i)}}\right)^{x_{\max}^{(i)}}}{\sum_{j=0}^{x_{\max}^{(i)}} \frac{1}{j!}\left(\frac{\lambda^{(i)}}{\mu^{(i)}}\right)^{j}}. \tag{3}$$

$\lambda^{(i)}$ and $\mu^{(i)}$ are, respectively, the call arrival rate and the service rate, both of them exponentially distributed.

Solving (2) by setting $x_{\max}^{(i)} = \left\lfloor \frac{X^{(i)}}{R_{\mathrm{CBR}}^{(i)}} \right\rfloor$, the maximum number of calls; $K_{\max}^{(i)}$, that can be served with $C_{\min}^{(i)}$ is obtained

$$K_{\max}^{(i)} = \left\lfloor \frac{C_{\min}^{(i)}}{R_{\mathrm{CBR}}^{(i)}} \right\rfloor. \tag{4}$$

$C_{\min}^{(i)}$ is not only the minimum bandwidth necessary to satisfy the guaranteed call QoS threshold but, having fixed $K_{\max}^{(i)}$, is also the maximum bandwidth that CBR traffic can use at the station i. In other words, the maximum bandwidth $\widehat{C}_{\mathrm{CBR}}^{(i)}$ that CBR traffic can get at station i is

$$\widehat{C}_{\mathrm{CBR}}^{(i)} = \begin{cases} C_{\min}^{(i)} & \text{if } C^{(i)} > C_{\min}^{(i)}, \\ C^{(i)} & \text{if } C^{(i)} \leqslant C_{\min}^{(i)}. \end{cases} \tag{5}$$

As a consequence, the maximum number of CBR calls that can access station i is

$$\widehat{K}_{\max}^{(i)}(C^{(i)}) = \left\lfloor \frac{\widehat{C}_{\mathrm{CBR}}^{(i)}}{R_{\mathrm{CBR}}^{(i)}} \right\rfloor. \tag{6}$$

CBR traffic does not necessarily take the maximum bandwidth $\widehat{C}_{\mathrm{CBR}}^{(i)}$: let $k^{(i)}(t) \leqslant \widehat{K}_{\max}^{(i)}$ be the number of CBR calls in progress at time t, at station i. $k^{(i)}(t)$ is also the number of busy servers at time t following the model in (3). The bandwidth of CBR traffic at time t is: $C_{\mathrm{CBR}}^{(i)}(t) = R_{\mathrm{CBR}}^{(i)} \cdot k^{(i)}(t)$. The residual capacity $(C_{\mathrm{be}}^{(i)}(t) = C^{(i)} - C_{\mathrm{CBR}}^{(i)}(t))$ is available for Internet traffic at the ith station and time t. This is the bandwidth that E-CAP-ABASC will use. The rationale of the bandwidth allocation algorithm is: assure the *QoS*-guaranteed traffic and, at the same time, help the non-guaranteed traffic to receive the best possible service.

With F being the maximum number of fading levels and $p_f^{(i)}$ the associated probability of the fth fading value at station i, the Centralized Bandwidth Allocator (CBA) assigns the bandwidth portions to the earth stations by minimizing the cost function $J_{CAP}(\cdot)$, defined as

$$J_{\mathrm{CAP}}(X^{(1)}, \ldots, X^{(I)}) = \sum_{i=1}^{I} \sum_{f=1}^{F} p_f^{(i)} J_{\mathrm{CAP}}^{(i)}(\beta_f^{(i)} X^{(i)}), \tag{7}$$

where

$$J_{\mathrm{CAP}}^{(i)}(X^{(i)}) = P_{\mathrm{loss}}^{(i)}(X^{(i)},\ \widehat{K}_{\max}^{(i)}(X^{(i)})) + F_{\mathrm{CAP}}^{(i)}(X^{(i)}). \tag{8}$$

$P_{\mathrm{loss}}^{(i)}$, written as a function of the bandwidth $X^{(i)}$ and of the maximum number of acceptable CBR calls, is the average packet loss probability of the non-guaranteed traffic at the station i. The closed-form expression of $P_{\mathrm{loss}}^{(i)}$ has a great impact on the bandwidth allocation concerning the non-guaranteed traffic and, in consequence, on the service the allocation system can offer to the best-effort flows. A different way to write $P_{\mathrm{loss}}^{(i)}$ differentiates E-CAP-ABASC from CAP-ABASC, as will be clear in the next section.

$F_{\mathrm{CAP}}^{(i)}(\cdot)$ is a penalty function to guarantee the call block probability constraint

$$F_{\mathrm{CAP}}^{(i)}(X^{(i)}) = \begin{cases} 0 & \text{if } \overline{P}_{\mathrm{B}}^{(i)}(\widehat{K}_{\max}^{(i)}(X^{(i)})) \leqslant \gamma^{(i)}, \\ H & \text{if } \overline{P}_{\mathrm{B}}^{(i)}(\widehat{K}_{\max}^{(i)}(X^{(i)})) > \gamma^{(i)}. \end{cases} \tag{9}$$

H is a very large constant, and $\overline{P}_{\mathrm{B}}^{(i)}$ is the call blocking probability averaged over the fading levels

$$\overline{P}_{\mathrm{B}}^{(i)}(\widehat{K}_{\max}^{(i)}(X^{(i)})) = \sum_{f=1}^{F} p_f^{(i)} \cdot P_{\mathrm{B}}^{(i)}(\widehat{K}_{\max}^{(i)}(X^{(i)})). \tag{10}$$

The solution of the minimization problem is

$$\text{if } C_{\mathrm{tot}} \geqslant \sum_{i=1}^{I} C_{\min}^{(i)} \Rightarrow \{C^{(1)}, \ldots, C^{(I)}\} = \arg\min_{X^{(1)},\ldots,X^{(I)}} \{J_{\mathrm{CAP}}(X^{(1)}, \ldots, X^{(I)})\}, \tag{11}$$

$$\text{if } C_{\mathrm{tot}} < \sum_{i=1}^{I} C_{\min}^{(i)} \Rightarrow C^{(i)} = \frac{C_{\mathrm{tot}}}{\sum_{i=1}^{I} C_{\min}^{(i)}} \cdot C_{\min}^{(i)}. \tag{12}$$

3. Bandwidth allocation algorithms and traffic source models

As said, the $P_{\mathrm{loss}}^{(i)}$ expression that appears in (8) actually characterizes the allocation algorithm. Three algorithms are compared in this paper: the first one has been introduced in [2]. The algorithm, called originally CAP-ABASC, is referenced as CAP-1 in the following. $P_{\mathrm{loss}}^{(i)}$ is computed without performing any distinction between TCP and UDP, whose flows are conveyed towards the same buffer.

The second proposal is called CAP-2. It is the first step towards the differentiation of the TCP and UDP flows, which are conveyed to two dedicated buffers. The closed-form expression for $P_{\mathrm{loss}}^{(i)}$ is not different for the two buffers. The distinction is only physical but it helps introduce the third proposal, the algorithm called E-CAP-ABASC. E-CAP-ABASC conveys traffic toward two separate buffers, as in CAP-2, and computes $P_{\mathrm{loss}}^{(i)}$ making use of two different expressions of the packet loss probability: one for TCP flows and one for UDP flows.

3.1. CAP-1

Best-effort traffic is considered as a single component. The overall allocation architecture is shown in Fig. 2.

The quantity $P_{\mathrm{loss}}^{(i)}$ is computed via the Y/D/C_s/Q model proposed in [3]. The rationale of the application of this model is that the packet-based traffic has statistical characteristics similar to *self-similar* processes. The self-similar nature of Ethernet traffic has been analyzed in [4]; the same concepts are

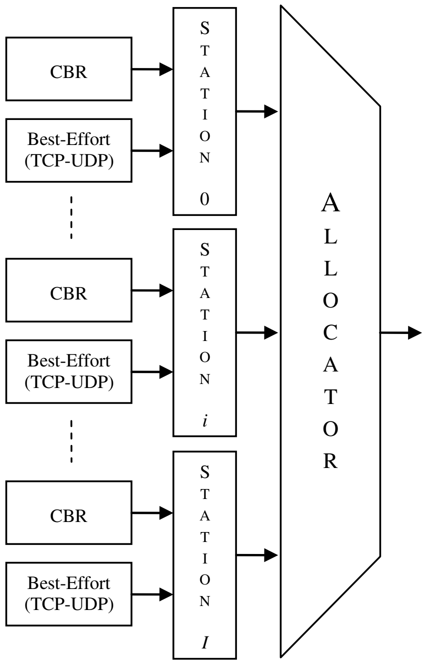

Fig. 2. Bandwidth allocator for CAP-ABASC (CAP-1).

extended to TCP over WANs in [5] not including TCP congestion control algorithms and to *Variable Bit Rate* (VBR) video sources, typically UDP based.

The sources are supposed independent and modelled by an ON–OFF Markov chain. The traffic is generated with constant packet rate only during ON periods, whose lengths are Pareto-distributed with average τ. Any distribution may be assumed for the length of OFF periods because an infinite number of sources (i.e., an infinite number of OFF periods) is assumed to get the closed-form solution of $P_{\text{loss}}^{(i)}$.

The analytical approximation of the packet loss probability for station i is reported in (13).

$$P_{\text{loss}}^{(i)} = \begin{cases} \min\left\{\frac{c^{(i)}\cdot\lambda_s^{(i)}}{\alpha^{(i)}(\alpha^{(i)}-1)(C_s^{(i)}-\lambda_s^{(i)}\tau^{(i)})}\left(Q^{(i)}\right)^{1-\alpha^{(i)}},1\right\} \\ \quad \text{if } C_s^{(i)} > \lambda_s^{(i)}\tau^{(i)}, \\ 1 \\ \quad \text{otherwise.} \end{cases} \tag{13}$$

$C_s^{(i)}$ is the number of servers busy in the Y/D/C_s/Q system and $Q^{(i)}$ is the length (measured in packets of 1500 bytes) of the IP buffer dedicated to best-effort sources. $Q^{(i)}$ is assumed to tend to infinite. $\alpha^{(i)}$is the Pareto parameter ($1<\alpha^{(i)}<2$), $c^{(i)}$ is a normalization constant and $\lambda_s^{(i)} = \lambda_{\text{asy}}^{(i)}\cdot T$ is the arrival rate of sources' supposed i.i.d. Poissonian. T is the source packetization time and $\lambda_{\text{asy}}^{(i)}$ the average number of packets generated by the Internet sources in a second. It is important to specify the relation between $C_s^{(i)}$ and the available transmission bandwidth for this kind of source: if the peak bandwidth for each Internet source is B_p and the average value of packets length is L (both supposed fixed) then T is defined as $\frac{L}{B_p}$ and

$$C_s^{(i)} = \frac{C_{\text{be}}^{(i)}}{L}T = \frac{C_{\text{be}}^{(i)}}{L}\cdot\frac{L}{B_p} = \frac{C_{\text{be}}^{(i)}}{B_p}. \tag{14}$$

It is worth noting that the $P_{\text{loss}}^{(i)}$ expression (and, in consequence, the cost function (8) and the overall allocation strategy) is insensitive to best-effort traffic load variation (e.g., to the number of active Internet sources). It is obvious because the number of sources is supposed infinite. As already said, it also ignores the different characteristics of the UDP and TCP traffic.

3.2. CAP-2

CAP-2 introduces, as previously said, a physical separation of the buffers respectively dedicated to the UDP and TCP flows. The system architecture is shown in Fig. 3. One best-effort packet is considered lost if at least one of following events is true: (1) one UDP packet is lost; (2) one TCP packet is lost; (3) one UDP and one TCP packet are lost simultaneously. Being $P_{\text{loss}_{\text{TCP}}}^{(i)}$ and $P_{\text{loss}_{\text{UDP}}}^{(i)}$, respectively, the probability to lose one TCP and one UDP packet, at the station i, and supposing UDP and TCP losses to be independent events, the probability $P_{\text{loss}}^{(i)}$ to lose a best-effort packet at station i is:

$$P_{\text{loss}}^{(i)} = P_{\text{loss}_{\text{TCP}}}^{(i)} + P_{\text{loss}_{\text{UDP}}}^{(i)} - P_{\text{loss}_{\text{TCP}}}^{(i)}\cdot P_{\text{loss}_{\text{UDP}}}^{(i)}. \tag{15}$$

CAP-2 still uses (13) to model both $P_{\text{loss}_{\text{TCP}}}^{(i)}$ and $P_{\text{loss}_{\text{UDP}}}^{(i)}$, but it allows assigning a different portion of bandwidth to the two buffers.

It provides a service capacity proportional to the number of active sources. $M^{(i)}$ being the number of UDP sources and $N^{(i)}$ the number of the TCP sources at the station i, the bandwidth allocated to the UDP and TCP traffic flows is, respectively

$$C_{\text{UDP}}^{(i)} = \frac{M^{(i)}}{N^{(i)}+M^{(i)}}\cdot C_{\text{be}}^{(i)} = \rho_M^{(i)}C_{\text{be}}^{(i)}$$

$$\text{where } \rho_M^{(i)} = \frac{M^{(i)}}{N^{(i)}+M^{(i)}}, \tag{16}$$

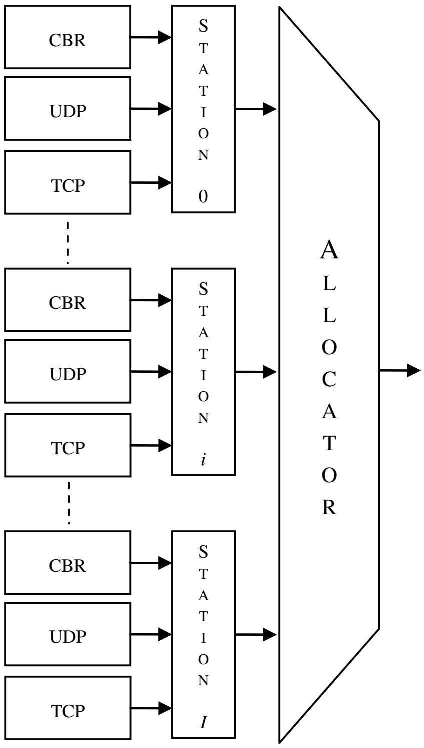

Fig. 3. Bandwidth allocator for CAP-2 and E-CAP.

$$C_{\mathrm{TCP}}^{(i)} = \frac{N^{(i)}}{N^{(i)} + M^{(i)}} \cdot C_{\mathrm{be}}^{(i)} = \rho_N^{(i)} C_{\mathrm{be}}^{(i)} \quad \text{where } \rho_N^{(i)} = \frac{N^{(i)}}{N^{(i)} + M^{(i)}}. \tag{17}$$

Given the $P_{\mathrm{loss}}^{(i)}$ formula (13) and Eq. (14) $\left(C_s^{(i)} = \frac{C_{\mathrm{be}}^{(i)}}{L} T = \frac{C_{\mathrm{be}}^{(i)}}{L} \cdot \frac{L}{B_p} = \frac{C_{\mathrm{be}}^{(i)}}{B_p}\right)$, setting the value of $C_{\mathrm{be}}^{(i)}$ in (14) to the two values defined in (16), valid for UDP, and (17), valid for TCP, two different values for $C_s^{(i)}$ are obtained: one for UDP and one for TCP. Using separately these two $C_s^{(i)}$values in (13), two different $P_{\mathrm{loss}}^{(i)}$ values result, one for UDP and one for TCP. The traffic parameters used in (13), are

– UDP: packet arrival rate $\lambda_{\mathrm{asy}}^{(i)} = \lambda_{\mathrm{asy_{UDP}}}^{(i)}$[burst/s]; UDP buffer size $Q^{(i)} = Q_{\mathrm{UDP}}^{(i)}$ [packets of 1500 bytes]; Pareto parameter $\alpha^{(i)} = \alpha_{\mathrm{UDP}}^{(i)}$. The number of busy servers for UDP is

$$C_s^{(i)} = C_{s_{\mathrm{UDP}}}^{(i)} = \frac{C_{\mathrm{UDP}}^{(i)}}{L} T = \frac{C_{\mathrm{UDP}}^{(i)}}{L} \cdot \frac{L}{B_{p_{\mathrm{UDP}}}} = \frac{\rho_M^{(i)} C_{\mathrm{be}}^{(i)}}{B_{p_{\mathrm{UDP}}}}. \tag{18}$$

– TCP: packet arrival rate $\lambda_{\mathrm{asy}}^{(i)} = \lambda_{\mathrm{asy_{TCP}}}^{(i)}$[burst/s], UDP buffer size $Q^{(i)} = Q_{\mathrm{TCP}}^{(i)}$ [packets of 1500 bytes]; Pareto parameter $\alpha^{(i)} = \alpha_{\mathrm{TCP}}^{(i)}$. The number of busy servers for TCP is

$$C_s^{(i)} = C_{s_{\mathrm{TCP}}}^{(i)} = \frac{C_{\mathrm{TCP}}^{(i)}}{L} T = \frac{C_{\mathrm{TCP}}^{(i)}}{L} \cdot \frac{L}{B_{p_{\mathrm{TCP}}}} = \frac{\rho_N^{(i)} C_{\mathrm{be}}^{(i)}}{B_{p_{\mathrm{TCP}}}} \tag{19}$$

Also in this case the peak bandwidth of the UDP and TCP sources and the average value of the IP packet length are supposed constant.

3.3. E-CAP-ABASC

E-CAP-ABASC (shortened to E-CAP in the performance evaluation section) is aimed at exploiting the features of the expression (15), which models the packet loss probability of a best-effort packet at the station i, as in E-CAP.

The TCP and UDP flows are separated into two different buffers as in CAP-2. The control architecture shown in Fig. 3 is still valid.

The clue is using two different closed-form solutions for $P_{\mathrm{loss_{TCP}}}^{(i)}$ and $P_{\mathrm{loss_{UDP}}}^{(i)}$. The latter is modelled again through (13), by setting the parameters exactly as in CAP-2. $P_{\mathrm{loss_{TCP}}}^{(i)}$ is modelled by considering the scenario shown in Fig. 4 for a single station i.

The following definitions are required. $T_n^{(i)}$ is the Round Trip Time (RTT) at the TCP layer for the nth connection at the station i. It is supposed constant for each packet of the nth connection and it is measured in seconds.

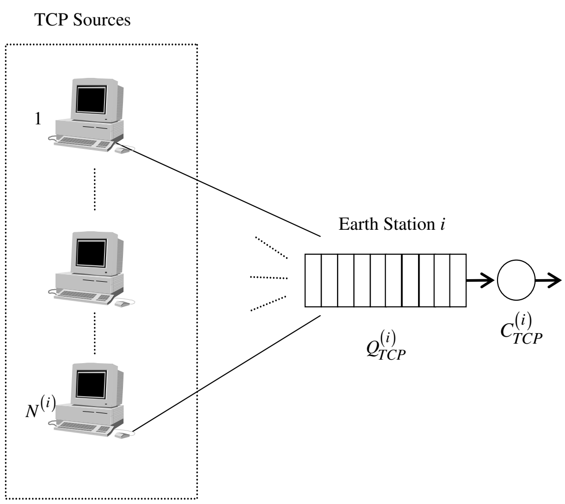

Fig. 4. TCP sources buffer model at station i.

$W_{\mathrm{pipe}}^{(i)}$ is the maximum volume of information that can be transmitted to the system composed of a channel server of capacity $C_{\mathrm{TCP}}^{(i)}$ and of IP buffer size $Q_{\mathrm{TCP}}^{(i)}$. $W_{\mathrm{pipe}}^{(i)}$ and $Q_{\mathrm{TCP}}^{(i)}$ are measured in packets (of length $L = 1500$ bytes); $C_{\mathrm{TCP}}^{(i)}$ in [bps].

Defining $C_{\mathrm{TCP},n}^{(i)}$ and $Q_{\mathrm{TCP},n}^{(i)}$, which are constant over time, respectively, the maximum portion of the capacity $C_{\mathrm{TCP}}^{(i)}$ and of the buffer $Q_{\mathrm{TCP}}^{(i)}$, "seen" by the nth connection, and $W_{\mathrm{pipe},n}^{(i)}$, the maximum volume of information that can be transmitted to the system by the nth connection, it is true that:

$$W_{\mathrm{pipe}}^{(i)} = \sum_{j=1}^{N} W_{\mathrm{pipe},j}^{(i)} = \sum_{j=1}^{N} \left(\frac{C_{\mathrm{TCP},j}^{(i)}}{L} \cdot T_j^{(i)} + Q_{\mathrm{TCP},j}^{(i)} \right), \tag{20}$$

where $N^{(i)}$ is the number of active TCP sources of the station i and, as previously defined, L is the fixed packet length.

The satellite being geostationary, the round trip time may be supposed fixed and equal for all the sources and stations. Mathematically

$$T_j^{(i)} = T_n^{(i)} = RTT \quad \forall j, n \in [1, N^{(i)}] \ \forall i \in [1, I]. \tag{21}$$

Together with the hypothesis of synchronization [6], this results in the fairness condition. Remembering that $\sum_{j=1}^{N^{(i)}} Q_{\mathrm{TCP},j}^{(i)} = Q_{\mathrm{TCP}}^{(i)}$ and $\sum_{j=1}^{N^{(i)}} \frac{C_{\mathrm{TCP},j}^{(i)}}{L} = \frac{C_{\mathrm{TCP}}^{(i)}}{L}$, from Eq. (20), it is true that

$$W_{\mathrm{pipe}}^{(i)} = RTT \cdot \frac{C_{\mathrm{TCP}}^{(i)}}{L} + Q_{\mathrm{TCP}}^{(i)}. \tag{22}$$

Taking TCP Reno as a reference and referring only to the congestion avoidance phase (steady state behaviour), the dimension of the congestion window $W_n^{(i)}$ of a generic source n varies between a minimum and a maximum value as introduced in [7] (TCP-Reno, simplified model). The congestion window $W_n^{(i)}$ size grows to saturate the channel; if a packet is lost, the window decreases its maximum size in dependence of the factor $m^{(i)}$ [8] that varies between 0 and 1 (typically $m^{(i)} = \frac{1}{2}$, as indicated in [8]). The packet loss probability $p_n^{(i)}$ of the nth TCP connection at the station i, the parameters fixed as defined above, is given in (23). b_n is the number of packets covered by one acknowledgement for the nth connection. The detailed computations needed to find it may be found in Ref. [9], where the packet loss probability (23) was originally proposed

$$p_n^{(i)} = \frac{32 (N^{(i)})^2}{3 \cdot b_n \cdot (m^{(i)} + 1)^2 \cdot \left(\frac{C_{\mathrm{TCP}}^{(i)}}{L} \cdot RTT + Q_{\mathrm{TCP}}^{(i)} \right)^2} = \frac{128}{27} \cdot \frac{(N^{(i)})^2}{b_n \cdot \left(\frac{\rho_N^{(i)} C_{\mathrm{be}}^{(i)}}{L} \cdot RTT + Q_{\mathrm{TCP}}^{(i)} \right)^2}. \tag{23}$$

If $b_n = b$, for all n, as assumed in the reminder of this work, the packet loss probability in (23) is independent of the index n (i.e., it is the same for each single source of the ith earth station) and corresponds to the entire TCP aggregate at station i. It means

$$P_{\mathrm{loss\,TCP}}^{(i)} = \frac{128}{27} \cdot \frac{(N^{(i)})^2}{b \cdot \left(\frac{\rho_N^{(i)} C_{\mathrm{be}}^{(i)}}{L} \cdot RTT + Q_{\mathrm{TCP}}^{(i)} \right)^2}. \tag{24}$$

Summarizing, in E-CAP-ABASC: the packet loss probability $P_{\mathrm{loss}}^{(i)}$ used in the cost function is computed from (15), making use of (13) and (24), for the UDP and TCP traffic, respectively. Bandwidth sharing between TCP and UDP is performed by using (16) and (17).

4. Performance comparison

The aim of this evaluation is to compare the performance of the E-CAP-ABASC allocation (called E-CAP in the reminder of the paper) in terms of packet loss probability and bandwidth allocation with CAP-1 and CAP-2. To reach this goal, the allocation strategies have been implemented in C++ language. The minimization algorithm used to solve the allocation problems is based on a dynamic programming procedure [10]. The performance evaluation is carried out by varying traffic and fading conditions.

The packet loss probability values shown in the results are computed analytically through the strategies presented in the previous section for the different schemes (CAP-1, CAP-2 and E-CAP).

4.1. Deterministic fading level with balanced best-effort traffic

The first analysis is based on the variation of the fading levels $\beta^{(i)}$. As introduced in Section 2.2, the fading effect is modelled as a bandwidth reduction introduced by the factors $\beta^{(i)}$, whose technical inter-

Table 1
Fading Class and related β level for the station 3

Fading class	C/N_0 [dB]	β
1	>77.13	1
2	74.63–77.13	0.8333
3	72.63–74.63	0.625
4	69.63–72.63	0.3125
5	66.63–69.63	0.15625
6	<66.63	–

pretation may be the presence of a FEC scheme which extends the number of bits dedicated to FEC when the Bit Error Ratio (BER) increases and, as a consequence, reduces the bandwidth used for the transport of information. The fading state of the channel (the signal to noise ratio) and the possible FEC assignation to counteract fading are associated here. The procedure is taken from [11] and also reported in Table 1. Four earth stations have been considered: stations 0, 1 and 2 are in clear sky conditions ($\beta^{(i)} = 1,\ i = 0,1,2$). The fading level of station 3 varies deterministically by following the values in Table 1. Each of them is supposed fixed in each performed test. The following parameters for each single station i have been used:

CBR: $\lambda^{(i)} = 0.006$ [calls/s]; $\frac{1}{\mu^{(i)}} = 600$ [s]; $R^{(i)}_{\text{CBR}} = 128$ [Kbps]; threshold $\gamma^{(i)} = 0.05$.
UDP: $\lambda^{(i)}_{\text{asy}_{\text{UDP}}} = 10$ [burst/s]; $Q^{(i)}_{\text{UDP}} = 8000$ [packets]; $\alpha^{(i)}_{\text{UDP}} = 1.5$; packet length $L = 1500$ [bytes]; peak rate $B_{p_{\text{UDP}}} = 64$ [Kbps]; and variable number $M^{(i)}$ of UDP sources.
TCP: variable number $N^{(i)}$ of TCP sources; $\text{RTT} = 520$ [ms]; the UDP buffer set to 8000 [packets]; also the TCP buffer has been set to the same length: $Q^{(i)}_{\text{TCP}} = 8000$ [packets]. Concerning CAP-2: it uses two buffers, one for TCP and one for UDP, both using formula (13), as in Section 3.2. The traffic parameters used in (13) are also specified in Section 3.2: the number of busy servers is specified in (18) $\left(C^{(i)}_s = C^{(i)}_{s_{\text{UDP}}} = \frac{C^{(i)}_{\text{UDP}}}{L} T = \frac{C^{(i)}_{\text{UDP}}}{L} \cdot \frac{L}{B_{p_{\text{UDP}}}} = \frac{\rho^{(i)}_M C^{(i)}_{\text{be}}}{B_{p_{\text{UDP}}}}\right)$ for UDP and in (19) $\left(C^{(i)}_s = C^{(i)}_{s_{\text{TCP}}} = \frac{C^{(i)}_{\text{TCP}}}{L} T = \frac{C^{(i)}_{\text{TCP}}}{L} \cdot \frac{L}{B_{p_{\text{TCP}}}} = \frac{\rho^{(i)}_N C^{(i)}_{\text{be}}}{B_{p_{\text{TCP}}}}\right)$ for TCP. So, in the CAP-2 cases, a similar load for UDP and TCP buffers may be obtained by setting the TCP parameters as follows: $\lambda^{(i)}_{\text{asy}_{\text{TCP}}} = 10$ [burst/s]; $\alpha^{(i)}_{\text{TCP}} = 1.5$; packet length $L = 1500$[bytes]; peak rate$B_{p_{\text{TCP}}} = 64$ [Kbps].

To make a fair comparison among the algorithms, the single buffer parameters dedicated to the best-effort traffic in CAP-1 are set to $Q^{(i)} = Q^{(i)}_{\text{UDP}} + Q^{(i)}_{\text{TCP}} = 16000$ [packets] and $\lambda^{(i)}_{\text{asy}} = \lambda^{(i)}_{\text{asy}_{\text{UDP}}} + \lambda^{(i)}_{\text{asy}_{\text{TCP}}} = 20$ [burst/s]. These values are twice the corresponding parameters used in the CAP-2 and E-CAP cases. The parameter choices are directly imposed by the requirements of the model applied for UDP, where the buffer dimension is supposed infinite. The overall number of best-effort active sources is 100 for each station ($M^{(i)} + N^{(i)} = 100$), in all cases. The comparison is performed with a fixed percentage of TCP and UDP traffic loading the system ($\rho^{(i)}_M = 0.5, \rho^{(i)}_N = 0.5$).

The overall bandwidth available C_{tot} is set to 8 [Mbps]. The call blocking probability constraint for the CBR traffic is set to $\gamma^{(i)} = 0.05\ \forall i$. Fig. 5a shows the assigned bandwidth to stations 0, 1 and 2. The bandwidth is the same for all stations in clear sky. Fig. 5b shows the related packet loss probability. Considering station 3 (Figs. 6a and b for allocated bandwidth and packet loss probability, respectively), the allocated bandwidth is larger than the one given to the other stations except for the clear sky case where the allocator provides exactly the same capacity to all the stations. This is due, obviously, to the different fading conditions. Increasing the bandwidth as a reaction to fading is due to the nature of the fading model: the bandwidth reduction factor (Eq. (1)) appears in the denominator of both the packet loss probability models (13) and (23). Decreasing the $\beta^{(i)}$ value implies larger amount of allocated bandwidth to reduce the overall packet loss probability, so compensating the effect of fading.

It is worth noting that all the allocation schemes (CAP-1, CAP-2 and E-CAP) have the same behaviour in terms of allocated bandwidth. The difference in the packet loss probability is due to the different analytical formulae to compute it: Eq. (13) for CAP-1, Eqs. (13) and (15) for CAP-2 and the combination of Eqs. (13), (15) and (24) for E-CAP.

The indication of the results, for now, is the sensitivity to the fading variations obtained by all schemes. If a station is corrupted by fading, it receives a larger quantity of capacity at a cost to the others to counteract and mitigate the fading effect.

It is worth noting that the packet loss probability performance of CAP-1 is better than the one of CAP-2. The reason is related to the traffic model used for both the UDP and the TCP traffic in CAP-2. The packet loss probability defined for the Y/D/C$_s$/Q model is a non-linear function with

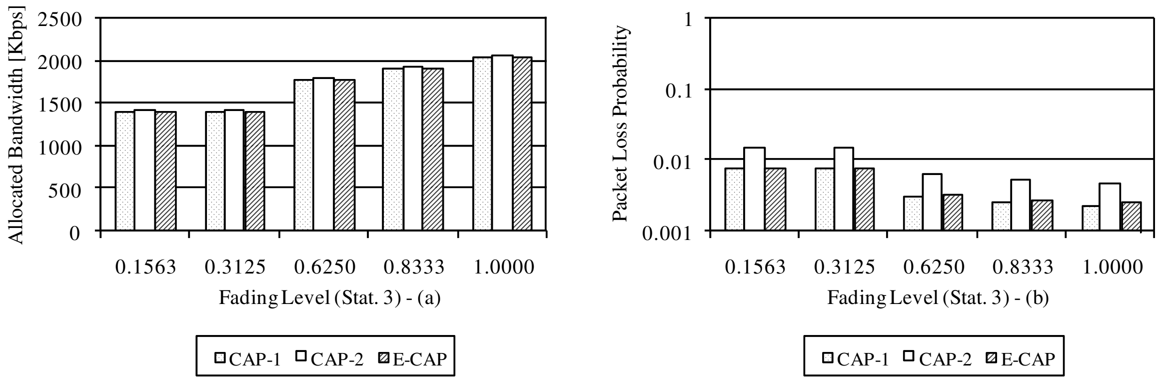

Fig. 5. Allocated bandwidth and Packet Loss Probability of Stations 0, 1, 2 fixing β.

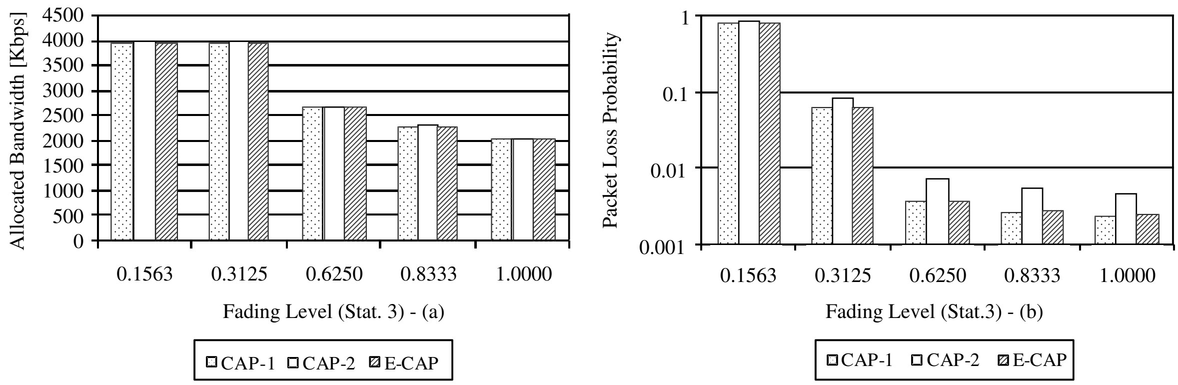

Fig. 6. Allocated bandwidth and Packet Loss Probability of Station 3 fixing β.

respect to buffer size $Q^{(i)}$, $\lambda_{\text{asy}}^{(i)}$ and allocated bandwidth. It implies an increase of the packet loss probability when the bandwidth is split in two portions (obtained through the same model as in CAP-2).

In other words: the combination of the packet loss probability expressed in formula (13) obtained by using Eq. (15) gives a higher overall packet loss probability ($P_{\text{loss}}^{(i)}$) if a portion of the assigned bandwidth (50%, in this case, because $\rho_M^{(i)} = \rho_N^{(i)} = 0.5\ \forall i$) is provided to each component of Eq. (15). The effect disappears for E-CAP where TCP and UDP use different packet loss probability models. Actually E-CAP performance practically overlaps CAP-1 performance in Figs. 5 and 6, where there is no change in the traffic load.

A numerical example directly taken from the results of Fig. 6 may help better understanding: when $\beta^{(3)} = 0.625$ the assigned bandwidth to the station 3 is 2688 [Kbps]: the packet loss probability value for CAP-1 (given by Eq. (13)) is 0.003658. The same value for CAP-2 comes from three components of Eq. (15): $P_{\text{loss}_{\text{TCP}}}^{(i)} = P_{\text{loss}_{\text{UDP}}}^{(i)} = 0.003651$, which are equal because both are modelled by Eq. (13), and $P_{\text{loss}_{\text{UDP}}}^{(i)} \cdot P_{\text{loss}_{\text{TCP}}}^{(i)} = 0.000013329$. The overall value is 0.007302. Also for E-CAP $P_{\text{loss}}^{(i)}$ comes from the three components of Eq. (15): $P_{\text{loss}_{\text{TCP}}}^{(i)} = 0.0001917$, $P_{\text{loss}_{\text{UDP}}}^{(i)} = 0.003651$, which are distinguished, and $P_{\text{loss}_{\text{UDP}}}^{(i)} \cdot P_{\text{loss}_{\text{TCP}}}^{(i)} = 0.7 \cdot 10^{-6}$. The overall value is 0.003842. In E-CAP the contribution due to the TCP flow is smaller than the contribution of UDP, because it considers, through the model in Eq. (24), the effect of the congestion control.

4.2. Unbalanced best-effort traffic

The numerical values previously fixed are unvaried. Only the percentage of UDP and TCP traffic changes. The overall number of active sources is always 100 but $\rho_M^{(i)}$ and $\rho_N^{(i)}$ vary for station 3. The percentage of TCP active sources for station 3 is indicated as $100 \cdot \rho_N^{(3)}$ in Figs. 7–12. The percentage of UDP sources obviously is $\rho_M^{(3)} = 100 \cdot (1 - \rho_N^{(3)})$.

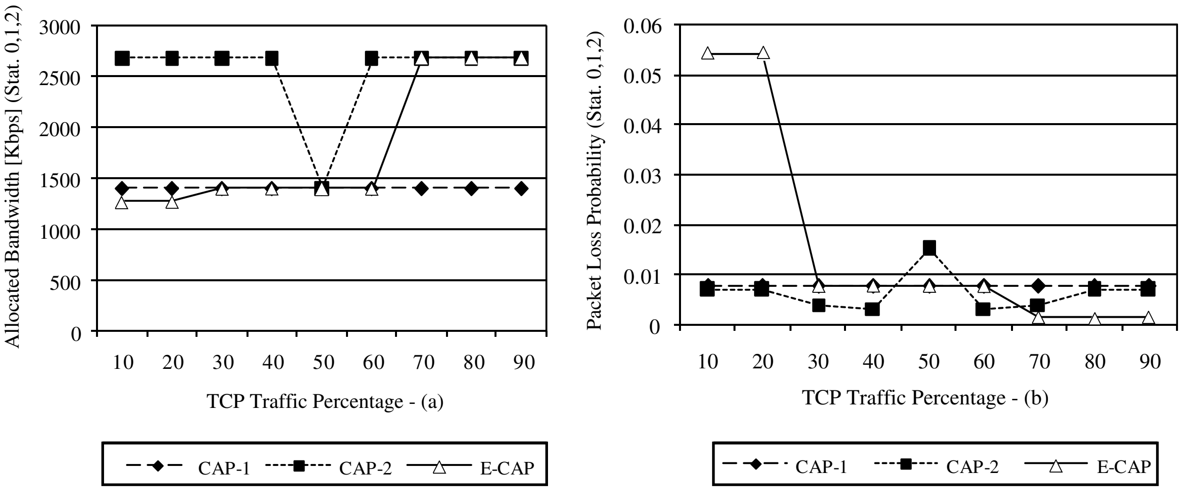

Fig. 7. Allocated bandwidth and Packet Loss Probability of Stations 0, 1, 2 with variable traffic percentage and $\beta = 0.15625$.

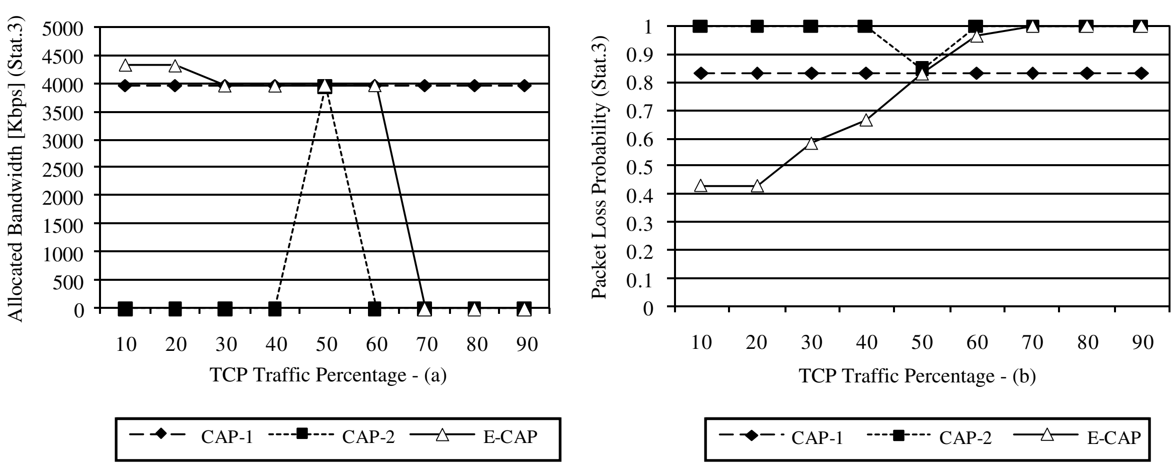

Fig. 8. Allocated bandwidth and Packet Loss Probability of Station 3 with variable traffic percentage and $\beta = 0.15625$.

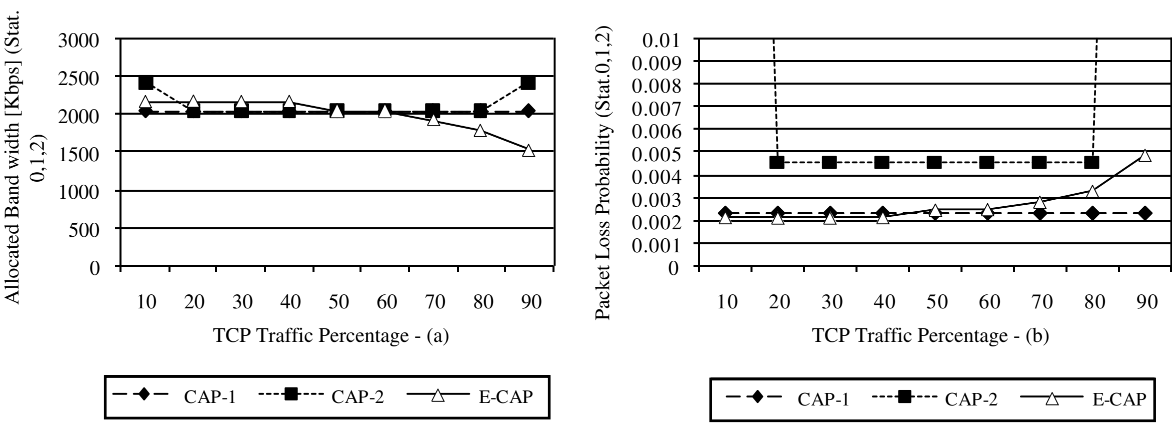

Fig. 9. Allocated bandwidth and Packet Loss Probability of Stations 0, 1, 2 with variable traffic percentage and $\beta = 1$.

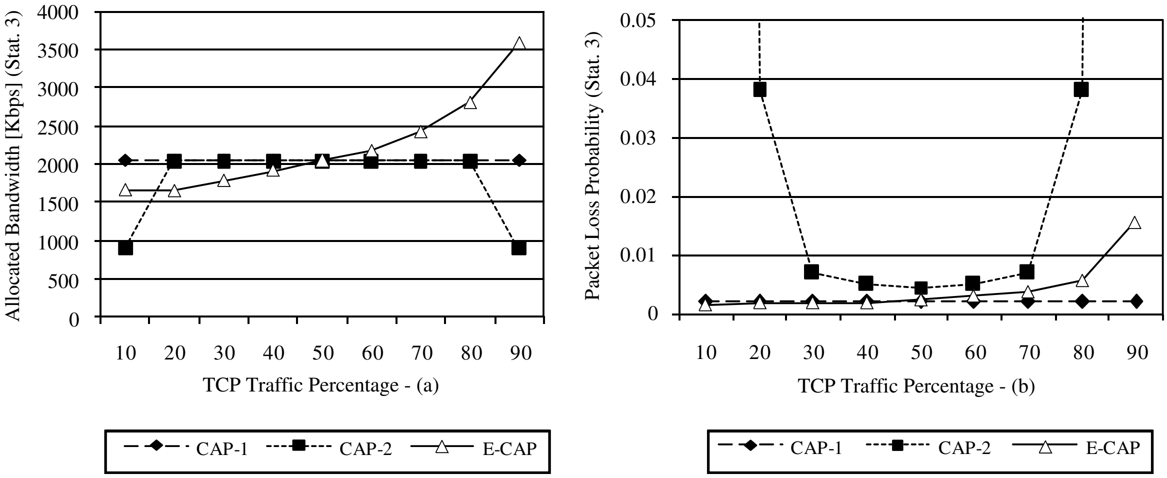

Fig. 10. Allocated bandwidth and Packet Loss Probability of Station 3 with variable traffic percentage and $\beta = 1$.

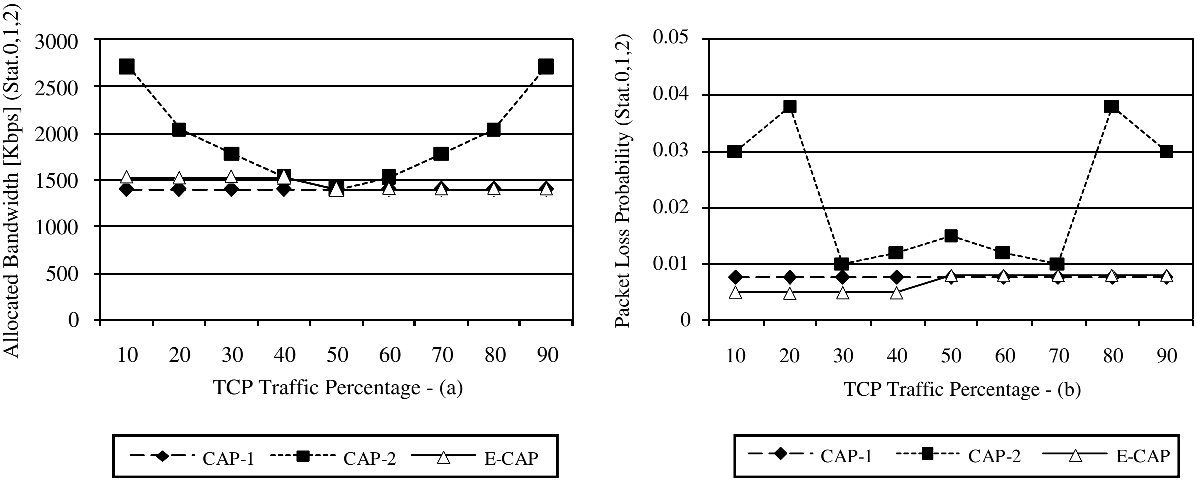

Fig. 11. Allocated bandwidth and Packet Loss Probability of Stations 0, 1, 2 with variable traffic percentage and β stochastic.

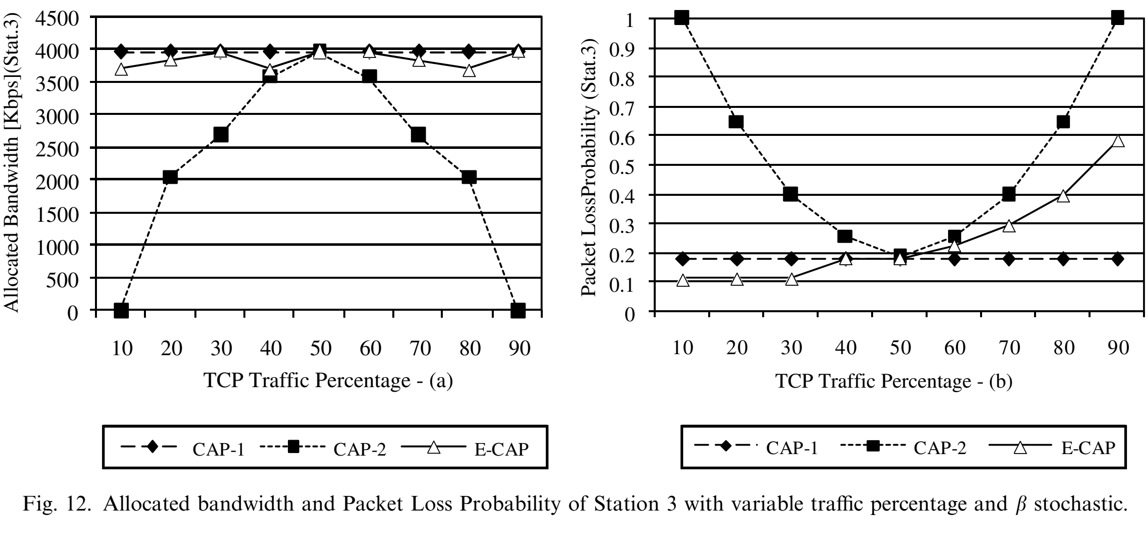

Fig. 12. Allocated bandwidth and Packet Loss Probability of Station 3 with variable traffic percentage and β stochastic.

Figs. 7a and b show, respectively, the comparison among allocation strategies and packet loss probabilities for CAP-1, CAP-2 and E-CAP, for stations 0, 1 and 2, by varying $\rho_N^{(3)}$, $\beta^{(3)} = 0.15625$.

The CAP-1 scheme behaviour is completely independent of the traffic percentage variation. Allocated bandwidth and packet loss probability modelled through Eq. (13) are constant. Actually, in real contexts, a variation of the traffic load in station 3 should modify the bandwidth allocation behaviour because, when $\rho_N^{(3)}$ increases, the number of the traffic sources that use the TCP congestion control increases. This means that the load of station 3 changes as well as the bandwidth used. If there is a reduction in the bandwidth use at station 3, the residual resources may be used by the other stations. This does not happen for the allocation in CAP-1 because the used traffic model does not consider the real offered load to the buffer, but supposes it to be infinite. This assumption is needed to reach the asymptotic self-similar behaviour of the model [3] but if it is applied within the cost function of the shown bandwidth allocation, it implies ignoring the real traffic load offered to a station giving origin to the load-insensitive behaviour shown in Fig. 7.

If CAP-2 is considered, the allocated bandwidth, as well as the packet loss probability, has a symmetric behaviour with respect to the completely balanced case ($\rho_M^{(3)} = 0.5, \rho_N^{(3)} = 0.5$). The symmetric behaviour is due both to the concept of buffer separation (i.e., to the fact of having two separate buffers) and to the use of the same packet loss model for the TCP and UDP traffic. In other words: there are two separate buffers but the same packet loss model is used for them. This makes the two buffers undistinguishable and creates the symmetric behaviour mentioned above. In practice the cost function used for CAP-2 does not distinguish the situation $\rho_N^{(i)} = 0.1$ and $\rho_M^{(i)} = 0.9$ from the case $\rho_N^{(i)} = 0.9$ and $\rho_M^{(i)} = 0.1$ (as well as the other symmetric situations) because the obtained overall cost function values are the same. As clear in Fig. 7a, CAP-2 assigns a large quantity of bandwidth to stations 0, 1, and 2 and heavily penalizes the station experiencing fading (Figs. 8a and b) without getting a significant performance improvement for the packet loss probability of the stations which do not experience any fading (Fig. 7b).

E-CAP, through the use of an explicit analytical model for the TCP sources, exploits the bandwidth reduction due to the TCP congestion control and provides good performance. For large $\rho_N^{(3)}$ values (70%, 80% and 90% of TCP traffic, in Figs. 7 and 8) the bandwidth portion dedicated to the UDP traffic is very low. The component of the E-CAP cost function due to UDP tends to 1 and, consequently, also the overall packet loss probability (Eq. (15)). The hypothesis of infinite active sources made for (13), together with the assignation of small bandwidth portions (i.e., small $C_{\text{be}}^{(i)}$ and, in consequence, small $C_s^{(i)} = \frac{C_{\text{be}}^{(i)}}{L} T = \frac{C_{\text{be}}^{(i)}}{L} \cdot \frac{L}{B_p} = \frac{C_{\text{be}}^{(i)}}{B_p}$), makes the overall packet loss probability value very high. The overall packet loss probability is computed as in Eq. (15) ($P_{\text{loss}}^{(i)} = P_{\text{loss}_{\text{TCP}}}^{(i)} + P_{\text{loss}_{\text{UDP}}}^{(i)} - P_{\text{loss}_{\text{TCP}}}^{(i)} \cdot P_{\text{loss}_{\text{UDP}}}^{(i)}$), which use formula (13) to compute the UDP packet loss. The UDP packet loss being very high for the reason explained above (hypothesis of infinite active sources made for (13), together with the assignation of small bandwidth portions), the overall packet loss in (15) is very high. The behaviour is identical for CAP-2, but, in this case, it also happens for small $\rho_N^{(3)}$ because the two cases cannot be differentiated from the opposite symmetric case (80% and 90% of TCP traffic), already discussed. This is not true for E-CAP that uses formula (24) to compute the TCP packet loss. The overall packet loss probability value is structured into the UDP component, which has the mentioned sensitivity to low bandwidth allocations, and the TCP component, which, considering the TCP rate reduction due to losses, is much less sensitive to small bandwidth allocations. Numerically, when $\rho_N^{(3)}$ is small, the UDP component receives a large bandwidth portion and provides limited values; the TCP component, even if it receives small bandwidth allocations, provides a limited value because of the reasons explained above. Similar comments may be reported for Figs. 9 and 10, which show the same quantities as in Figs. 7 and 8 but with $\beta^{(3)} = 1$. The behaviour, in this case, depends only on the traffic percentage in the station 3 because all stations are in clear sky. The mentioned behaviour due to heavily unbalanced $\rho_N^{(3)}$ and $\rho_M^{(3)}$ is even more evident for CAP-2 than in the previous case because of the very small packet loss probability values of the clear sky case. Actually the packet loss probability when $\rho_N^{(3)} = 0.1$ and $\rho_M^{(3)} = 0.9$ is 0.052. This does not appear in Fig. 9b for a scale factor.

In short, the performance determining factors are the fading level at a given station and the bandwidth offered to the TCP and to UDP buffer (whose model is more sensitive to small capacity values).

4.3. Stochastic fading level with unbalanced best-effort traffic

All possible fading values (Table 1) are considered and each of them has a statistical characterization. The allocation is performed by minimizing the packet loss probability averaged over all possible fading levels. The framework presented in this section is ideal to test the algorithms for possible use in network planning, where the fading value may be only forecast with a given probability. Again the TCP traffic percentage is varied while the other parameters are unchanged. The results are reported in Figs. 11 and 12. All considerations are still applicable but E-CAP deserves one more comment. It always assures the lowest packet loss value for the station in clear sky and for the faded station up to a TCP traffic percentage below 50%. When $\rho_N^{(3)} \geqslant 0.6$, small allocations to the UDP traffic affect the overall performance as extensively explained above. Actually the UDP traffic is not controlled at all. A possible idea for future work may be bandwidth adaptation for UDP traffic sources.

The role of the number of TCP and UDP sources may be better seen from the results of the next subsection.

4.4. Effects of a variable number of UDP sources

Keeping the other parameters unchanged, the number of TCP sources for station 3 is fixed to $N^{(3)} = 50$ and the number of UDP sources $M^{(3)}$ is

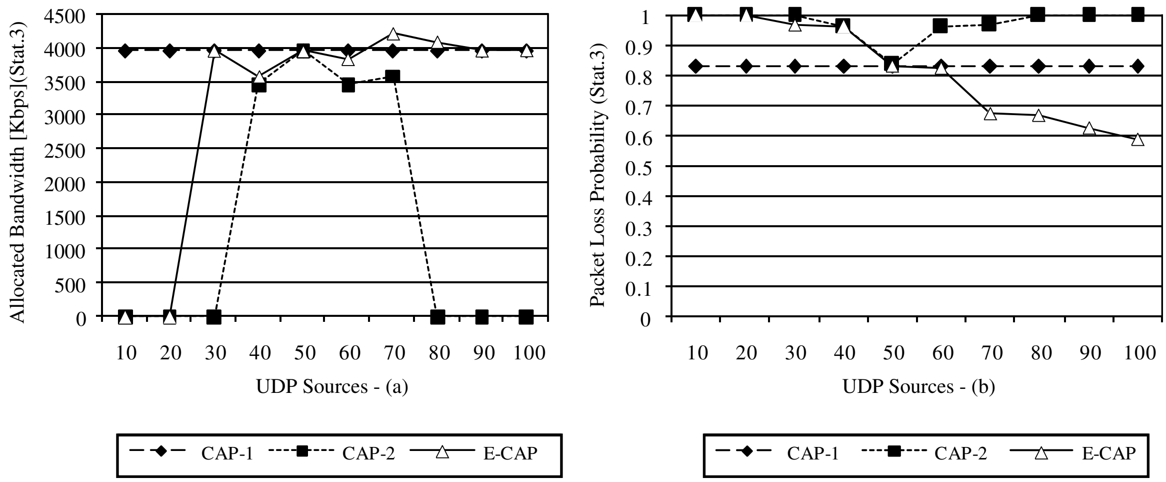

Fig. 13. Allocated bandwidth and Packet Loss Probability of Station 3 with $\beta = 0.15625$ and variable number of UDP sources.

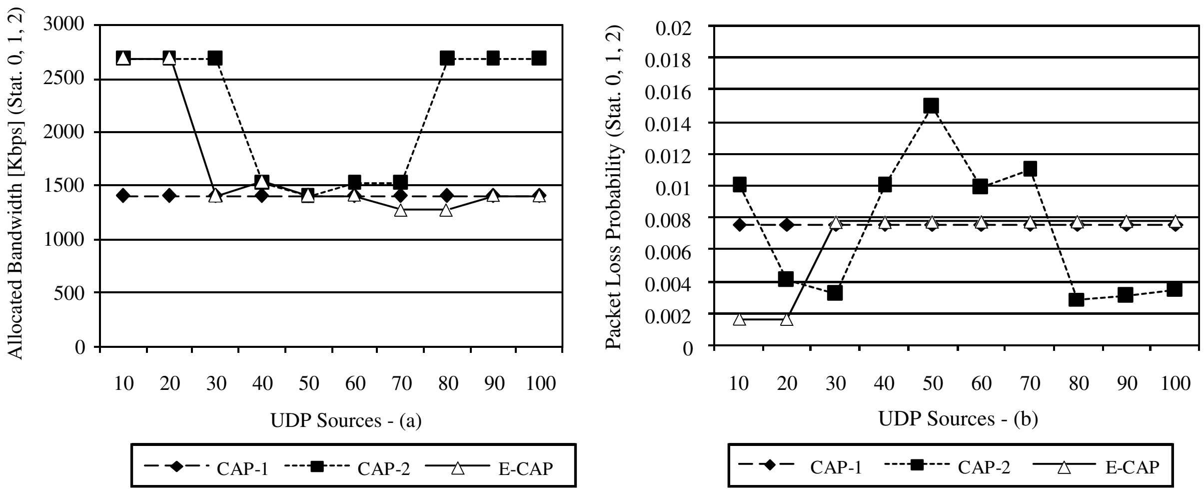

Fig. 14. Allocated bandwidth and Packet Loss Probability of Stations 0, 1, 2 with $\beta = 0.15625$ and variable number of UDP sources.

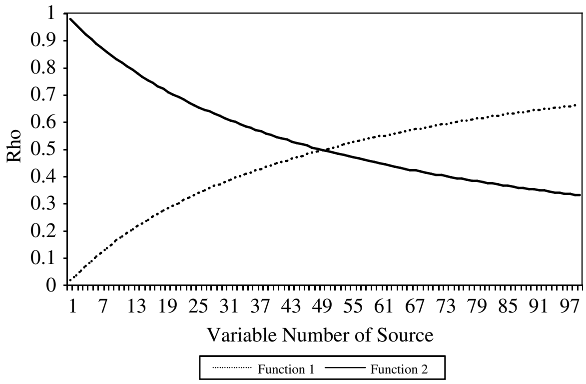

Fig. 15. Shape of "Function 1" and of "Function 2".

variable from 10 to 100 in Figs. 13 and 14. For stations 0, 1 and 2 $M^{(i)} = N^{(i)} = 50$. The results are aimed at evaluating the impact on the performance of a strong variation of the UDP traffic with no variation in the TCP sources. Again $\beta^{(3)} = 0.15625$ and $\beta^{(i)} = 1$, if $i = 0$, 1 and 2.

The key to the analysis is the shape of $\rho_N^{(3)}$ and $\rho_M^{(3)}$ over the number of sources. When the number of UDP sources is variable, the behaviour of $\rho_M^{(3)}$ has the form of "function1" while $\rho_N^{(3)}$ has the form of "function2", both shown in Fig. 15. Figs. 13a and b show the performance of the station with fading. When the number of UDP sources ($M^{(3)}$) is small, E-CAP does not allocate bandwidth to station 3. This behaviour is due to the small quantity of capacity provided to the UDP buffer: the values assumed by $\rho_M^{(3)}$ are very small as shown in Fig. 15 ("function1") and the overall packet loss probability of the station 3 tends to 1 because the values of the UDP component in the cost function tend to 1, as extensively explained before. When $\rho_M^{(3)}$ grows, $\rho_N^{(3)}$ decreases (Fig. 15, "function2") and the performance, in terms of packet loss probability, tends to improve: more bandwidth to UDP buffer reduces the component of the packet loss probability due to the UDP traffic, while less bandwidth given to the TCP buffer increases its packet loss probability but the increase is not particularly meaningful because of the self-regulatory nature of TCP which is fully considered in the applied analytical model. In practice, the overall packet loss probability is globally decreasing with $\rho_M^{(3)}$. If CAP-2 is considered, the behaviour is similar for small and large values of $\rho_M^{(3)}$, having the same model for UDP and TCP. When "function1" intersects "function2", the traffic load is balanced between TCP and UDP. The performance of E-CAP and CAP-2 is the same at the station 3. CAP-1 is totally insensitive to the traffic variations. If the stations 0, 1 and 2 are considered (Figs. 14a and b), the performance is obviously complementary with respect to the case described above for all methods considered. CAP-2 has a packet loss probability peak where the allocator assigns much bandwidth to station 3.

The same comments hold when fixing the UDP sources and varying the number of TCP sources.

4.5. Saving bandwidth

A very important issue in the satellite resource allocation context is surely saving bandwidth. This is the capacity saved by using one allocation strategy with respect to another, if a Quality of Service (e.g., Packet Loss Probability, Delay, Delay Jitter) requirement has been fixed. In other words, if a "degree" of Quality of Service is also requested for Internet traffic, it is interesting to investigate how much capacity C_{tot} is necessary to match it. For each single station:

$$X^{(i)} \geqslant X_{thr}^{(i)} \quad \forall i \in I \tag{25}$$

with

$$X_{thr}^{(i)} : P_{\text{loss}}^{(i)}(X_{thr}^{(i)}) \leqslant thr \quad \forall i \in I. \tag{26}$$

$X_{thr}^{(i)}$ is the bandwidth needed at the station i with respect to the fixed QoS requirement in terms of packet loss probability for the UDP and TCP traffic. $X^{(i)}$ is the bandwidth allocated to the ith station. This analysis "opens the door" not only to the addition of QoS requirements for Internet traffic but also to the introduction of classes of service with different QoS requirements. Internet traffic may be further differentiated by using, for example, a specific buffer for each traffic class. The minimum overall bandwidth to satisfy the Internet traffic requirement is

$$C_{thr} = \sum_{i=0}^{I} X_{thr}^{(i)}. \tag{27}$$

Its value depends on the allocation scheme. It is called "threshold bandwidth". Figs. 16a and b contain a comparison of the threshold bandwidth to get a packet loss probability for each station lower than thr, a generic fixed threshold considered equal, in this specific case, for all stations. The fading value of the station 3 is set to 0.15625 in Fig. 16a and to 1 in Fig. 16b. Three cases have been considered by changing the percentage of TCP and UDP traffic: "10% TCP 90% UDP", "50% TCP 50% UDP",

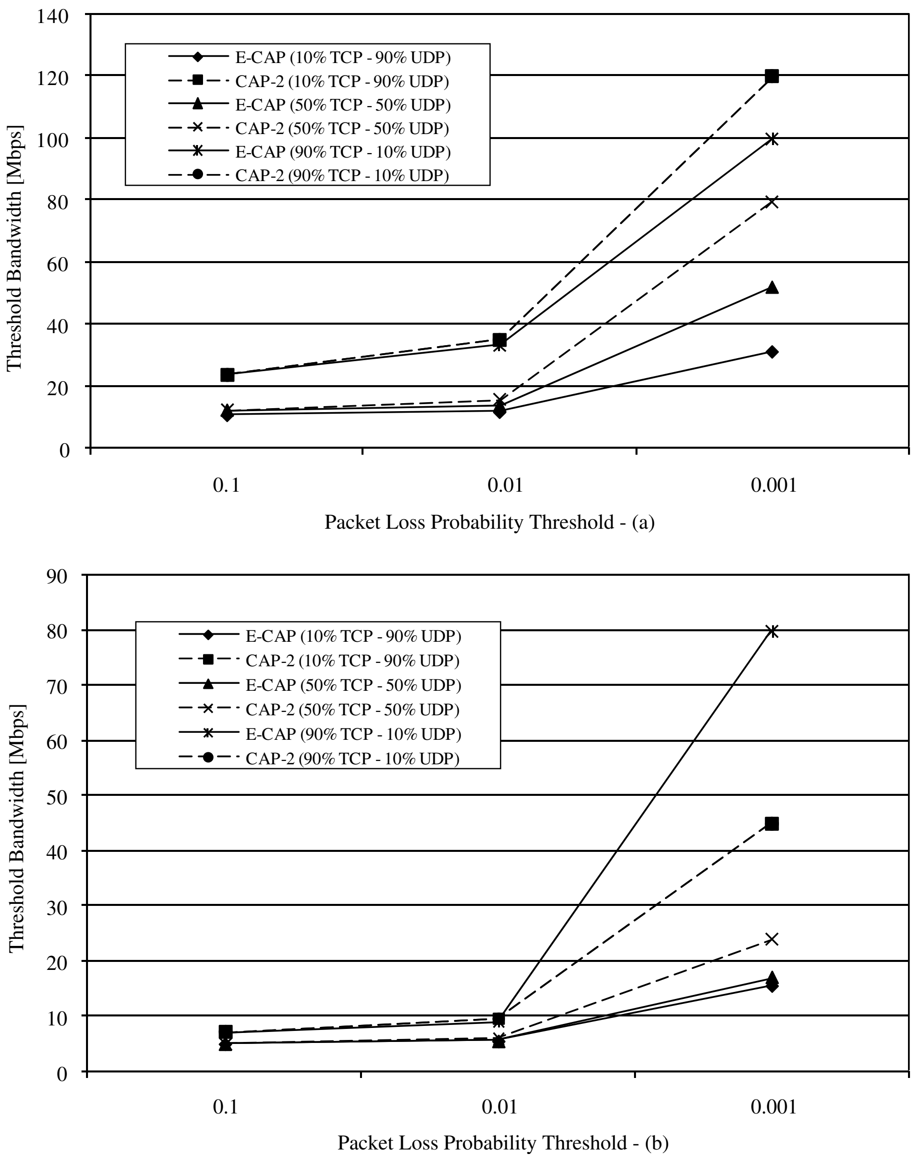

Fig. 16. Threshold bandwidth to guarantee fixed Packet Loss Probability level for each Station, $\beta = 0.15625$ (a) and $\beta = 1$ (b).

and "90% TCP 10% UDP". The other parameters of the systems are the same as before. Only the E-CAP and CAP-2 mechanisms have been considered. CAP-1 is not responsive to traffic variations and does not differentiate between TCP and UDP.

Case "10% TCP 90% UDP": the bandwidth needed by E-CAP to guarantee the packet loss probability constraint is much lower than the bandwidth needed by using CAP-2. The difference is really relevant if the threshold considered is particularly strict ($thr = 10^{-3}$), both when the station 3 experiences fading (Fig. 16a) and in clear sky.

Case "50% TCP 50% UDP": the E-CAP advantage is also clear even if it is reduced for the clear sky case.

Case "90% TCP 10% UDP": the "10% TCP 90% UDP" and "90% TCP 10% UDP" cases are the same if CAP-2 is considered. The amount of saved bandwidth which is guaranteed by E-CAP is clear but more limited than in the previous case because of the small capacity assigned to the UDP buffer (10% of the global bandwidth), which affects the computation of the overall packet loss probability.

To conclude: E-CAP is very efficient; it allows obtaining a fixed performance by using less bandwidth than the CAP-2 technique. The advantage is due to the consideration of real TCP traffic source rate adaptation.

5. Conclusions

The paper presents three bandwidth allocation schemes for the satellite environment. One of them (CAP-1) considers UDP and TCP traffic in undifferentiated fashion. CAP-2 separates the two Internet traffic types but uses the same traffic model. E-CAP exploits traffic separation and applies a different model to compute the TCP packet loss probability. The performance evaluation contains an extensive comparison of the three approaches. The full distinction of TCP and UDP traffic gives evident advantages in the bandwidth allocation because it allows using the automatic source rate reduction resulting from TCP congestion control. The residual bandwidth is used by other sources in different earth stations. E-CAP is particularly efficient if the percentage of TCP traffic is below 50%. It adapts its behaviour to the fading conditions and is suited also for network planning (e.g., when the fading level is not known "a priori"). It allows saving a significant amount of resources when a minimum bandwidth is required to satisfy *QoS* requests for the Internet traffic. This last feature allows forecasting a future application for VoIP and IP differentiated services environments.

References

[1] A. Jamalipour, T. Tung, The role of satellites in global IT: Trends and implications, IEEE Personal Communications 8 (3) (2001) 5–11.

[2] R. Bolla, F. Davoli, M. Marchese, Adaptive bandwidth allocation methods in the satellite environment, in: Proceedings of IEEE international conference on communications (ICC 2001), Helsinki, Finland, 2001, pp. 3183–3190.

[3] B. Tsybakov, N.D. Georganas, On self-similar traffic in ATM queues: Definition, overflow probability bound, and cell delay distribution, IEEE/ACM Transaction on Networking 5 (3) (1997) 397–409.

[4] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of ethernet traffic, IEEE/ACM Transaction on Networking 2 (1) (1994) 1–15.

[5] V. Paxson, S. Floyd, Wide-area traffic: The failure of poisson modeling, IEEE/ACM Transaction on Networking 3 (3) (1995) 226–244.

[6] E. Altman, C. Barakat, E. Laborde, P. Brown, D. Collange, Fairness analysis of TCP/IP, in: Proceedings of Conference on Decision and Control (ICDC), Sydney, Australia, 2000, pp. 231–242.

[7] T. Lakshman, U. Madhow, The performance of TCP/IP for networks with high bandwidth-delay products and random loss, IEEE/ACM Transactions on Networking 5 (3) (1997) 336–350.

[8] J.P. Hespanha, S. Bohacek, K. Obraczka, J. Lee, Hybrid Modeling of TCP Congestion Control, Proc. HSCC International Workshop, Rome, Italy, 2001, pp. 291–304.

[9] I. Bisio, M. Marchese, Analytical expression and performance evaluation of TCP packet loss probability over geostationary satellite channels, IEEE Communications Letter 8 (4) (2004) 232–234.

[10] D. Bertsekas, Dynamic Programming and Optimal Control, vol. I, Athena Scientific, Belmont, MA, USA, 1995.

[11] N. Celandroni, F. Davoli, E. Ferro, Static and dynamic resource allocation in a multiservice satellite network with fading, International Journal of Satellite Communications 21 (4–5) (2003) 469–488.

Igor Bisio was born in Novi Ligure (Alessandria), Italy in 1978. He got his "Laurea" degree at the University of Genoa, Italy in 2002. and his Ph.D. degree in "Information and Communication Sciences" at the University of Genoa in 2006. His Ph.D. course was funded by the Italian Consortium of Telecommunications (CNIT). He worked for both CNIT University of Genoa Research Unit and Selex Communications s.p.a., from 2002 to 2006 as research consultant. He is currently in a Research Assistant Position and he is member of the Telecommunication Networking Research Group and, in particular, of the Satellite Communications and Networking Laboratory research staff at the University of Genoa. He is also an IEEE Student Member since 2004 and IEEE Satellite and Space Communications Technical Committee Member since 2005. He is author of more than 30 scientific papers and he is recipient of the "Globecom 2006 Best Student Paper Award" in the Physical Communications Systems Category. His main research activity concerns: Resource Allocation and Management for Satellite Communication systems, Optimization Algorithms and Architectures for Satellite Sensor Networks, Traffic Modelling.

Mario Marchese was born in Genoa, Italy in 1967. He got his "Laurea" degree cum laude at the University of Genoa, Italy in 1992 and the Qualification as Professional Engineer in April 1992. He obtained his Ph.D. (Italian "Dottorato di Ricerca") degree in "Telecommunications" at the University of Genoa in 1996. From 1999 to 2004, he worked with the Italian Consortium of Telecommunications (CNIT), by the University of Genoa Research Unit, where he was Head of Research. From February 2005 he has been Associate Professor at the University of Genoa, Department of Communication, Computer

and Systems Science (DIST). He is the founder and still the technical responsible of CNIT/DIST Satellite Communications and Networking Laboratory (SCNL) by the University of Genoa, which contains high value devices and tools and implies the management of different units of specialized scientific and technical personnel. He was the Official Representative of CNIT within the European Telecommunications Standard Institute (ETSI) from 1999 to 2005, he is active member in the Satellite Earth Station (SES) Broadband Satellite Multimedia (BSM). He is Chair of the IEEE Satellite and Space Communications Technical Committee and Senior Member of the IEEE. He is author and co-author of more than 150 scientific works, including international magazines, international conferences and book chapters. He is the author of the book "Quality of Service over Heterogeneous Networks", John Wiley & Sons, Chichester, 2007. He is Associate Editor of the International Journal of Communication Systems (Wiley) and Technical Committee Co-Chair of various international conferences, including Globecom and ICC. His main research activity concerns: Satellite and Radio Networks, Transport Layer over Satellite and Wireless Networks, Quality of Service over ATM, IP and MPLS, Data Transport over Heterogeneous Networks, Emulation and Simulation of Telecommunication Networks.