Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Control Engineering Practice 11 (2003) 1209–1226
CONTROL ENGINEERING PRACTICE
www.elsevier.com/locate/conengprac

Integration of pricing models between best-effort and guaranteed performance services in telecommunication networks

Marco Baglietto[a], Raffaele Bolla[a], Franco Davoli[a], Mario Marchese[b], Maurizio Mongelli[a,*]

[a] DIST, Department of Communications, Computer and Systems Science, University of Genoa, Via Opera Pia 13, 16145 Genova, Italy
[b] CNIT, Italian National Consortium for Telecommunications, Genoa research unit, Via Opera Pia 13, 16145 Genova, Italy

Accepted 27 February 2003

Abstract

Starting from the mid-1990s, a growing attention has been devoted to more and more sophisticate pricing models for telecommunication services. A number of pricing models have been proposed and analyzed in the context of quality of service (QoS) guaranteed networks and, more recently, also for best effort (BE) environments. Concerning QoS networks, the optimization often influences the call admission control (CAC). In a BE network, where users do not declare QoS parameters and there is no CAC, the pricing policies should be integrated within the flow control and they are different from those adopted in the QoS environments. In this paper we investigate the condition where both BE traffic and traffic explicitly requiring QoS (guaranteed performance, GP) are present. We propose three mechanisms that influence both GP CAC and BE flow control and that are aimed at maximizing the overall revenue for all traffic classes. Moreover, we want to investigate the influence of the BE Pricing scheme on the GP traffic in order to establish a bound for the Internet Service Provider on the prices imposed to the GP users.

Keywords: Call admission control; Pricing; Proportional fairness; Best-effort services; Guaranteed performance services; IP quality of service; ATM; Flow control

1. Introduction

Pricing for telecommunications is an issue widely treated in the literature (DaSilva, 2000; Falkner, Devetsikiotis, & Lambadaris, 2000; Walrand & Varaiya, 2000). In particular, owing to the exponential growth of the Internet and the pervasive diffusion of the TCP/IP paradigm for the transport of both data and real-time traffic, it has become necessary to develop and test pricing methodologies capable of achieving globally optimal utility and fairness.

More specifically, a number of pricing models have been considered and analyzed in the context of quality of service (QoS) guaranteed networks, mainly with respect to the asynchronous transfer mode (ATM) world (see, e.g. DaSilva, 2000; Walrand & Varaiya, 2000; Courcoubetis, Siris, & Stamoulis, 1996; Kelly & Songhurst, 1997; Kelly, 1996; Murphy & Murphy, 1994; Murphy, Murphy, & Posner, 1994). In Murphy and Murphy (1994), a dynamic adaptive priority scheme is proposed, based on the periodic adjustment of prices per unit of bandwidth, associated to each virtual path (VP), performed by the network management; the users decide the bandwidth to be utilized on the corresponding VP and pay accordingly up to the next reallocation. A similar user-network interaction is adopted in Courcoubetis et al. (1996), and used for available bit rate (ABR) services. Kelly (1996) bases the pricing scheme in a QoS context on the concept of effective bandwidth (i.e., the bandwidth that is necessary to satisfy QoS requirements). The users will pay in proportion of the traffic volume and of the call duration, according to a linear law, whose coefficients are the price per unit time, the price per unit volume and a fixed charge per connection; their values are fixed at the time of connection acceptance and depend on the traffic contract of the

*Corresponding author. DIST, University of Genoa, Via Opera Pia 13, 16145 Genova, Italy. Tel.: +39-010-3532806; fax: +39-010-3532154.
E-mail addresses: mbaglietto@dist.unige.it (M. Baglietto), lelus@dist.unige.it (R. Bolla), franco@dist.unige.it (F. Davoli), mario.marchese@cnit.it (M. Marchese), maurizio.mongelli@cnit.it (M. Mongelli).

0967-0661/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0967-0661(03)00055-8

user. In general, the concepts developed in such "guaranteed bandwidth" environments (circuit-switched or ATM) can be applied also in the presence of IP QoS mechanisms (Resource Reservation Protocol (Braden, Zhang, Berson, Herzog, & Jamin, 1997), Integrated Services (Integrated Services Working Group; Wroclawski, 1997) or Differentiated Services (Differentiated Services Working Group) at least as regards the Expedited Forwarding service).

On the other hand, in a BE context (Kelly, Maulloo, & Tan, 1998; Kelly, 2001; Low & Varaiya, 1993; Low & Lapsley, 1999; Malinowski, 2002) where the user does not declare QoS parameters and is not subject to a call admission control (CAC) (i.e., flows are "elastic", as determined by TCP congestion control, or by TCP-friendly mechanisms at the application level), pricing policies should be different from those adopted in the above mentioned QoS environments. First of all, being in the presence of bandwidth fluctuations, a dynamic pricing policy should be implemented; then, it is necessary to introduce prices depending on the service effectively received and on the willingness of the users to pay for it.

Low and Varaiya (1993) have been among the first to propose a network pricing model where prices are periodically adjusted on the basis of a continuous monitoring of the user resource requests, in order to maximize a "global social welfare". The users' requests are formulated according to their traffic parameters and QoS requirements (e.g., packet delay constraints), in order to maximize a "utility function", which depends on their "willingness to pay" for a given service. In this context, pricing becomes strictly related with congestion control in the network, as it is capable of determining the optimal rates of the users' flows that maximize the aggregate source utility (Kelly et al., 1998; Low & Lapsley, 1999).

In this paper, we consider the presence of both BE traffic and of traffic explicitly requiring QoS (guaranteed performance, (GP)). In the literature, this subject is addressed in few papers. In Altman, Artiges, and Traore (1999) the main goal is to evaluate, by means of an analytical model, the decay in the performance of BE traffic as a function of increasing incoming GP traffic. A pricing mechanism based on a Stackelberg game model is also proposed in order to set the prices in such a way that those users who are free to choose between GP or BE service traffic class (the so-called mixed users) would be induced to choose the traffic class that results to be more convenient, so as to reach a global optimality. In our model we do not decide upon the choice of the traffic class; rather, each incoming flow will be declared to belong either to GP or BE beforehand and we take decisions that are explicitly based on the presence of both traffic categories, that dynamically share the available bandwidth. The goal of our model is to establish different strategies in order to maximize the overall revenue for the Internet Service Provider (ISP) and to control the prices imposed to the GP users, by means of the BE prices obtained from the Proportional Fairness Pricing model.

The paper is organized as follows. In the next section, we introduce the main optimization problem for the BE environment (Proportional Fairness Pricing) following Kelly et al. (1998). The third section is devoted to the description of our optimization model. Numerical results are presented in Section 4 and conclusions and future work are drawn in the fifth one.

2. Proportional fairness pricing

The concept of Proportional Fairness Pricing was motivated by the desire to incorporate the notion of fairness into the allocation of network resources (Falkner et al., 2000); it is in fact necessary, for a pricing scheme applied to BE users, to allocate "fairly" a resource, namely, to allocate it in proportion to the users' willingness to pay.

With a notation that slightly differs from that in Kelly et al. (1998) we consider a telecommunication network composed by a set J of unidirectional links; link j has capacity c_j. We call "BE user" a connection established on a specific path, consisting of a non-empty subset of J; R_{BE} is the set of active BE users. We indicate with $A = \{A_{jr},\ j \in J,\ r \in R_{BE}\}$ the matrix assigning resources to BE users ($A_{jr} = 1$ if link j is used by user r, $A_{jr} = 0$ otherwise). Moreover, let x_r be the rate of user r and $U_r(x_r) : [0, +\infty) \to \Re$ the utility function of such user, supposed to be strictly concave, increasing and continuously differentiable over $[0, +\infty)$. Such utility function describes how sensitive user r is to changes in x_r and, in the context of pricing, it is useful to think of it as the amount of money user r is willing to pay for a certain x_r. Finally, let $\mathbf{c} = [c_j,\ j \in J]$, $\mathbf{x} = [x_r,\ r \in R_{BE}]$, $\mathbf{U}(\mathbf{x}) = [U_r(x_r),\ r \in R_{BE}]$ be the aggregate vectorial quantities. The main goal of the ISP can now be stated; it consists of the maximization in the sum of all users' utilities, under the link capacity constraints over the given paths (Kelly et al., 1998; Low & Lapsley, 1999; Malinowski, 2002; Low, 1999):

The SYSTEM Problem $(\boldsymbol{U}(\cdot), \boldsymbol{A}, \boldsymbol{c})$:

$$\boldsymbol{x}^o = \arg\max_{\boldsymbol{x}} \sum_{r \in R_{BE}} U_r(x_r) \tag{1}$$

subject to $\boldsymbol{A} \cdot \boldsymbol{x} \leqslant \boldsymbol{c}$ and $\boldsymbol{x} \geqslant \boldsymbol{0}$.

It is shown in Kelly et al. (1998) that such a problem can be decomposed, by separately considering a network part and a user part. Let w_r, $r \in R_{BE}$, be the price per time unit that user r is willing to pay. By paying w_r money units per time unit, the user receives permission for a flow x_r, determined by the network. We can then

define

$$\lambda_r = \frac{w_r}{x_r} \tag{2}$$

as the price per bandwidth unit, and let $\mathbf{w} = [w_r,\ r \in R_{BE}]$, $\boldsymbol{\lambda} = [\lambda_r,\ r \in R_{BE}]$.

Each user solves the following optimization problem:

The USER$_r$ Problem $(x_r, U_r(\cdot))$:

$$w_r^o = \arg\max_{w_r}\left[U_r\left(\frac{w_r}{\lambda_r}\right) - w_r\right] \tag{3}$$

subject to: $w_r \geqslant 0$.

In practice, for each user r, a software agent periodically contracts with the network the bandwidth allocation x_r, it computes w_r in function of its utility, and sends it to the network. The latter has to solve the following optimization problem:

The NETWORK Problem $(\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{w})$:

$$\boldsymbol{x}^o = \arg\max_{\boldsymbol{x}} \sum_{r \in R_{BE}} w_r \log x_r \tag{4}$$

subject to $\boldsymbol{A} \cdot \boldsymbol{x} \leqslant \boldsymbol{c}$ and $\boldsymbol{x} \geqslant \boldsymbol{0}$.

Given the vector $\boldsymbol{w}$, the network computes $\boldsymbol{x}$, and sends each x_r as a feedback to the flow controller of each user r (Fig. 1).

A similar, though different formulation, based on Lagrangian duality, has been adopted by Low and Lapsley (1999), where the users decide their rates and the network charges them consequently using only local links information, thus eliminating the need for explicit communication (Low, 1999; Athuraliya & Low, 2000). Asynchronous distributed approaches have been developed in several works (Kelly et al., 1998; Low & Lapsley, 1999; Malinowski, 2002; Low, 1999; Athuraliya & Low, 2000). Among others, Malinowski (2002) includes the use of feedback from the real system.

In particular, in Kelly et al. (1998) it is shown that modelling flow control dynamics through suitable differential equations (Kelly, 2001) can yield to an arbitrarily good approximation of the solution of the problem.

More specifically, cost functions are defined for each link j, of the type:

$$\mu_j(t) = p_j\left(\sum_{r \in R_{BE}(j)} x_r(t)\right), \tag{5}$$

where the argument of the function $p_j(\cdot)$ represents the total rate on link j and $R_{BE}(j)$ is the subset of BE users whose connections traverse link j. Such functions should set a penalty on an excessive usage of the resource: ideally, it should be $p_j(y) = 0$ if $y \leqslant c_j$, and $p_j(y) = +\infty$ if $y > c_j$; however, it is better to have not too large values in the derivatives, as they compromise the stability of the solution (Kelly et al., 1998). Then, the following dynamic system, including pricing and flow control is considered (Fig. 2). Let $J_{BE}(r)$ be the set of all links used by user r:

$$\frac{\mathrm{d}}{\mathrm{d}t} x_r(t) = \kappa_r\left(w_r - x_r(t) \sum_{j \in J_{BE}(r)} \mu_j(t)\right). \tag{6}$$

The interpretation in terms of flow control is as follows:

1. each link j generates feedback signals according to $p_j(y)$, where y is the flow traversing it,
2. the feedback is interpreted as a congestion indicator by the users, and
3. each user's rate grows with rate w_r and decreases proportionally to the feedback.

On the other hand, in the economic interpretation, $p_j(y)$ is the price imposed per unit of bandwidth, when the link is traversed by flow y; the network suggests the modification of the rates, in order to render the costs $x_r(t) \sum_{j \in J_{BE}(r)} \mu_j(t)$ equal to the target values w_r.

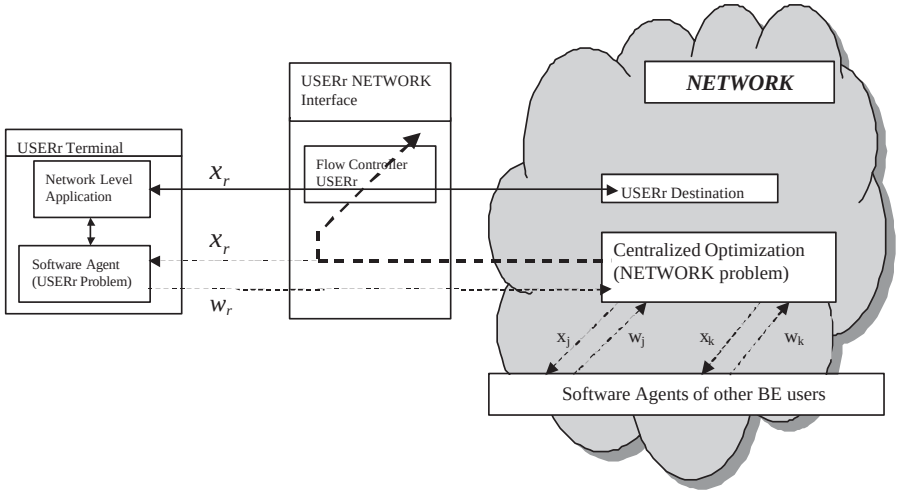

Fig. 1. Proportional Fairness Pricing: decomposition of the SYSTEM optimization problem.

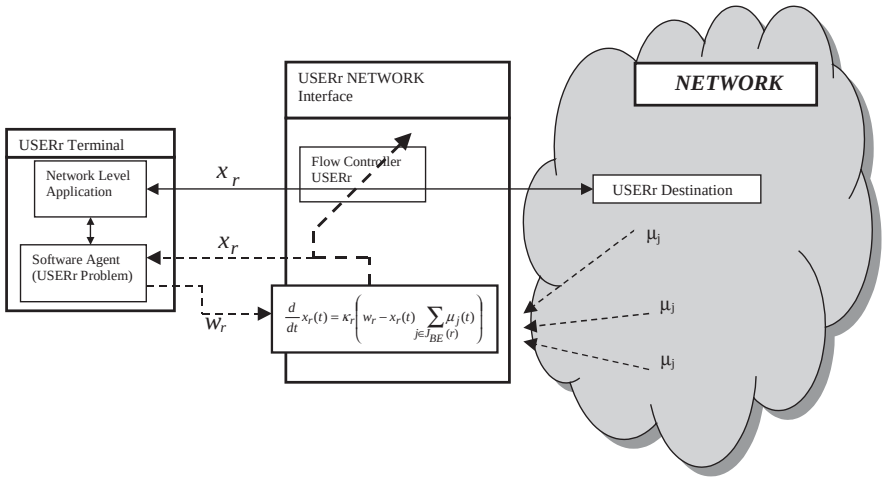

Fig. 2. Proportional Fairness Pricing: a completely decentralized approach.

It can be shown (Kelly et al., 1998) that, under not too restrictive hypotheses on the form of functions $p_j(\cdot)$, the system of differential equations is globally stable and, by adapting the prices w_r according to the solutions of the $USER_r$ problems, the Proportional Fairness Pricing optimum can be reached.

3. GP traffic multiplexed with BE traffic

In this section we shall consider the presence in the network of both BE traffic and of traffic explicitly requiring QoS (GP traffic). In this context, our goal is to influence the BE traffic flow control and to apply a CAC to the GP traffic in order to maximize the ISP's overall revenue. A BE or GP user, in our model, can represent a "big user", i.e. an aggregation of some domestic users or a group of LANs (for example a company, one of its branches, or a university campus (Malowidzki & Malinowski, 2002). Each of these aggregates includes one or more groups of single users that have the same assigned routing paths.

A GP user requests a service with strict QoS requirements, such as with constraints over the end-to-end mean delay or delay jitter or in terms of loss probability of the packets. It is possible, at the time of a GP call, to translate these requirements in terms of the equivalent bandwidth necessary to satisfy the GP user's performance requests (Pitts & Schormans, 2000; Chao & Guo, 2002). If such GP user is accepted in the network, a bandwidth pipe is reserved in order to maintain its traffic performance requirements. This mechanism needs a CAC based on the bandwidth availability. There are several methods to calculate the equivalent bandwidth, based on analytical models or by means of simulation analysis, possibly also on the basis of on-line measurements (see, e.g., Walrand & Varaiya, (2000) or Chao & Guo (2002) for an overview concerning this topic). In the last decade, the telecommunication network traffic has been shown to have a "fractal" statistical behaviour at the packet level (Leland, Taqqu, Willinger, & Wilson, 1994; Garret & Willinger, 1994) and this has a dramatic impact over the resources that must be reserved to guarantee QoS constraints (see, e.g., Pitts & Schormans, 2000; Tsybakov & Georganas, 1998). Also in the presence of such statistical behaviour of the sources, it is possible to use analytic models for the computation of the equivalent bandwidth (Pitts & Schormans, 2000; Tsybakov & Georganas, 1998; Kim & Shroff, 2001). Moreover, as pointed out in Zhang (1995) and Ma and Steenkiste (1997), certain QoS requirements depend also on the specific scheduling algorithm applied in the nodes of the network. With weighted fair queueing (WFQ) scheduling algorithms the end-to-end delay and delay jitter can be translated in terms of equivalent bandwidth. For this reason, in several works, the equivalent bandwidth is considered as the unique parameter used to formulate novel models of CAC or routing in telecommunication networks (see e.g. Walrand & Varaiya, 2000; Ross, 1995; Marbach, Milhatsch, & Tsitsiklis, 2000). We shall follow this approach and we use the bandwidth as the unique QoS metric to manage the GP calls.

With the usual CAC strategy based on the bandwidth availability, the GP traffic tends to prevail: if a huge amount of requests for this kind of traffic arrives, it might completely saturate the network, making the performance of the BE traffic decay, especially in heavy traffic situations (Altman et al., 1999). We propose a strategy to decide how many resources to allocate to the GP traffic, taking into account the performance of the BE traffic. The principle is based on the prices that the GP and BE users pay. Two strategies are firstly illustrated: in the first one a price is decided for all the GP traffic and the network revenue is optimized, deciding whether it is suitable or not to accept the

requests of incoming connections; in the second one, all the requests which pass the CAC Bandwidth Availability check are accepted, but every GP user is assigned a new price. The total ISP's revenue per unit time (for example €/min) G is formed by the sum of two terms, the first one concerning the BE traffic, the second one concerning the GP traffic:

$$G = G_{GP} + G_{BE}. \tag{7}$$

The revenue concerning the BE users is given by the $\mathbf{w}$ vector as the optimal solution of the corresponding proportional fairness pricing problem (3) and (4), while that of the GP traffic is obtained by multiplying the equivalent bandwidth by the assigned charge. The ISP assigns a reserved bandwidth y_r to each user $r \in R_{GP}$, where we denote by R_{GP} the set of active GP users (i.e., all of the GP connections accepted in the network and in progress). Every user r, $r \in R_{GP}$, pays an amount b_r per unit of sent GP traffic data per unit time (e.g., b_r could be €/Mbps per minute). Each time a new GP call/service (i.e., a new user $\tilde{r}$ wants to start up a connection) is requested, the Network is asked for a new amount of bandwidth $y_{\tilde{r}}$. The GP's required bandwidth might be computed, for example, by using a technique like the equivalent bandwidth in ATM, and anyway the feasibility of the request with respect to the available capacity should be tested (CAC with respect to the bandwidth availability).

We suppose the BE traffic to be regulated by a flow control mechanism such as in Section 2, so that the rates x_r, $r \in R_{BE}$, and the prices for time unit w_r, $r \in R_{BE}$, reach the stationary optimal values x_r^o and w_r^o after a finite time period. If the new bandwidth $y_{\tilde{r}}$ will be reserved, the BE traffic rates x_r^o and price w_r^o, $r \in R_{BE}$, will move to the new optimal values $\tilde{x}_r^o$ and $\tilde{w}_r^o$ according to (3) and (4), where the capacity constraint in (1) and (4) becomes:

$$\mathbf{A} \cdot \mathbf{x} \leqslant \tilde{\mathbf{c}}, \tag{8}$$

where $\tilde{\mathbf{c}} = [\tilde{c}_j,\ j \in J]$ is the residual capacity matrix, with $\tilde{c}_j = c_j - \sum_{r \in R_{GP}(j)} y_r$ the residual capacity (capacity not reserved to GP traffic) of link j ($R_{GP}(j)$ is the set of GP connections traversing link j). The revenue's derivative changes with the traffic change; e.g., if the GP traffic increases the GP revenue rate also increases, but the BE traffic rates decrease (less bandwidth available for BE) and so the BE revenue rate decreases, too. In this respect, a possible Acceptance Control Rule for the requests of increasing the GP traffic reserved bandwidth is to accept the new GP bandwidth reservation only if the total instantaneous revenue rate (i.e., the revenue derivative) with the new reservation increases with respect to the current situation. So, in our first proposal, we use the revenue derivative to decide whether to accept a GP bandwidth increase request. In particular, y_r, $r \in R_{GP}$, being the current GP bandwidth reservations, and $y_{\tilde{r}}$ a new bandwidth request for the user $\tilde{r}$ charged by $b_{\tilde{r}}$, we accept the new bandwidth request if

$$\sum_{r \in R_{GP}} b_r y_r + \sum_{r \in R_{BE}} w_r^o \leqslant \sum_{r \in R_{GP}} b_r y_r + b_{\tilde{r}} y_{\tilde{r}} + \sum_{r \in R_{BE}} \tilde{w}_r^o, \tag{9}$$

where $\tilde{w}_r^o$ represents the new optimal values of the BE price in the presence of the new GP allocation $y_{\tilde{r}}$. This means that we accept the request if

$$y_{\tilde{r}} \geqslant \frac{\sum_{r \in R_{BE}} (w_r^o - \tilde{w}_r^o)}{b_{\tilde{r}}}. \tag{10}$$

The resulting scheme is what will be called "CACPricing1".

If, on the contrary, the GP price is not fixed, but it can be freely assigned every time a connection is accepted, it is possible to assign it in order to leave the total revenue derivative unchanged:

$$\sum_{r \in R_{GP}} b_r y_r + \sum_{r \in R_{BE}} w_r^o = \sum_{r \in R_{GP}} b_r y_r + b_{\tilde{r}} y_{\tilde{r}} + \sum_{r \in R_{BE}} \tilde{w}_r^o, \tag{11}$$

$$b_{\tilde{r}} = \frac{\sum_{r \in R_{BE}} (w_r^o - \tilde{w}_r^o)}{y_{\tilde{r}}}. \tag{12}$$

In this way a price is assigned to every GP connection to exactly equal the revenue that will be lost on the BE traffic. In other words, the request of increment is always accepted if there is enough bandwidth availability, but with a cost $b_{\tilde{r}}$ which depends on the current network traffic conditions. Such a scheme will be called "VariableGPPrice" in the following.

With the "VariableGPPrice" strategy we want to establish a way to calculate the performance of a heterogeneous network, setting the GP prices by only using the current $\mathbf{w}$ vector. In the Proportional Fairness Pricing scheme, the ISP does not control the prices of the BE users; they are in fact only based according to the utility functions and to the current network congestion conditions. The only way for the ISP to control the prices behaviour is to control the state of the network congestion. Applying specific routing strategies, it can introduce fictitious points of congestion in order to increase the values of the $\mathbf{w}$ vector (Malowidzki & Malinowski, 2002), but, anyway, it cannot control directly the equilibrium point of the Proportional Fairness Pricing congestion control. According to this principle, we assign the GP prices on the basis of the current network conditions and Eq. (12) is a reasonable way to do this. The VariableGPPrice can be intended as an instrument to evaluate, during an off-line planning of the network, the static price that the ISP can impose to the GP users. If the ISP assigns the GP users a price in order to guarantee a revenue much higher than the one obtained by the VaribleGPPrice strategy, the only way to justify its choice is the necessity to compensate the higher costs for the establishment and the maintenance

of a GP connection. If such higher costs are not demonstrated, clearly, the ISP has overdimensioned the GP prices. An authority for the control of the market, in general, cannot force directly each ISP to fix the prices according to strict constraints; rather, it can facilitate the competition between them in order to support a general decreasing of the prices (see e.g. Cao, Shen, Milito, & Wirth, 2002 for what concerns the Nash equilibrium between different ISPs under competition). Anyway, recalling the Proportional Fairness Pricing model of the BE traffic, it could be useful to test the performance of a heterogeneous network where the prices are fixed only on the basis of the network traffic conditions. This pricing strategy can be interpreted as an extension to the GP traffic of the concept of fairness applied to the BE traffic. It assigns a price to each new GP connection in function of the decay of performance of the BE traffic when bandwidth $y_{\tilde{r}}$ is no more available for BE traffic.

For both these strategies, every time the CAC block acts, it is necessary to foresee the revenue $\sum_{r\in R_{BE}} \tilde{w}_r^o$ which will be obtained in the future on the BE traffic after the bandwidth reallocation.

As is summarized in Fig. 3 the idea is to use Eq. (1) (SYSTEM Problem) to calculate the new value of the **x** vector after the bandwidth reallocation; then, it is possible to calculate the new BE prices and the new BE revenue and compare the total revenue after the bandwidth reallocation.

The model proposed is clearly based on a centralized approach. Even though it may be advisable to propose decentralized techniques (see for example Gokbayrak & Cassandras, 2002; Baglietto, Parisini, & Zoppoli, 2001), centralized CAC models are not novel in the CAC literature, see for example Barnhart, Wieselthier, and Ephremides (1995), Barnhart, Wieselthier, and Ephremides (1993) and Celandroni, Davoli, and Ferro (to appear). In our model, it is necessary that each node of the network is able to know the "state" of the BE traffic, i.e. the current BE users' routing paths and the utility functions. Only in this way it is possible to apply correctly the algorithms based on the flowchart in Fig. 3. To maintain in each node a perfect knowledge of the network, it is necessary to establish something similar to the Link State (LS) information exchange of the QoS routing in a MPLS environment (Crawley, Nair, Jajagopalan, & Sandick, 1998). In LS routing, network nodes should be aware of the state of the links, possibly located several hops away. This calls for a periodic flooding exchange of LS information, which contributes extra traffic to the network. Among the cost factors of QoS routing, the cost of LS exchange is the dominant contributor (Apostolopoulos, Guerin, Kamat, Orda, & Tripathi, 1999; Apostolopoulos, Guerin, & Tripathi, 1998) and can severely limit the scalability of the QoS routing. Our model needs a periodic exchange of Node State (NS) information concerning the state of the BE users actually present in the network. The reason for this is that, at the time of the CAC of an incoming GP call, it is necessary to evaluate the impact of the new GP bandwidth reservation that can influence, according to the Proportional Fairness Pricing scheme, the bandwidth reservation of all BE users. Due to such NS information exchange, it is necessary to further investigate the possibility of decentralizing the proposed CAC strategies in order to guarantee a higher scalability, for example according to a model based on a team theory framework (Baglietto et al., 2001). Furthermore, in order to avoid periodic exchange of NS information it could be possible to foresee the variability of the BE traffic demands during a certain period and to use on line this knowledge to apply the proposed CAC and possibly to update it by an on-line estimate of the current network traffic conditions (Malowidzki & Malinowski, 2002). Similar considerations can be done concerning the possibility of knowing perfectly the BE users' utility functions. If it is impossible to maintain such knowledge perfectly updated, the CAC mechanism can be updated by an off-line forecast, possibly together with an on-line estimate of the current BE traffic demands (Malowidzki & Malinowski, 2002).

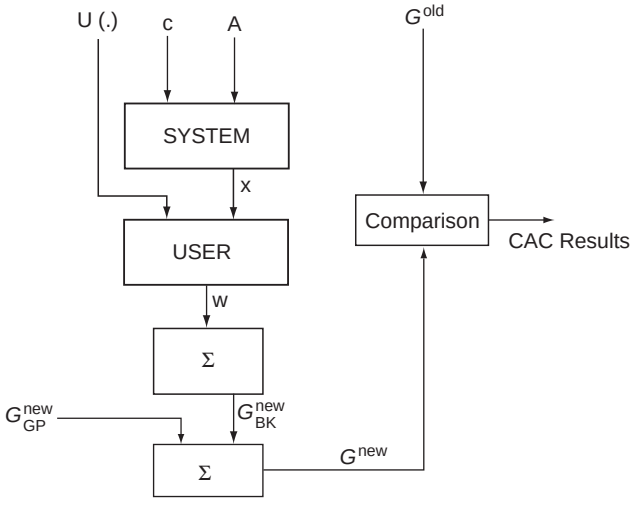

Fig. 3. Flowchart of the model proposed to integrate the GP and BE traffic.

It is important to consider that with the two strategies proposed so far, the revenues per unit time and not the total revenues are compared. These two strategies, in fact, act as Open Loop Feedback Control (OLFC) (Bertsekas, 2001); using a perfect information about the revenues per unit time obtained after the bandwidth reallocation, they ignore what could happen in the future in terms of all of the possible terminations of connections actually present in the network and in terms of the possible arrivals of new connections.

If we had considered the lengths of the connections the situation might have changed and it might have

happened that the best choice was different from the one taken. Namely, it could happen for example that, having accepted a new GP connection, the network is forced to refuse other connections, because it does not have any more bandwidth available for the incoming requests. However, some of these refused connections might contribute to increase the total revenue much more than the accepted one.

To solve this problem it is necessary to explicitly take into account the length of the connection and maximize the total revenue, rather than the revenue per unit time.

Addressing this topic is quite a difficult task; it is necessary to identify all the events possibly happening during each new incoming GP connection lifetime and to solve the SYSTEM problem (Eq. (1)) for each time interval between them. In fact, every time a (BE or GP) connection ends, much more bandwidth becomes available for the BE traffic, so the system converges to a new balance of the total revenue. For the sake of simplicity, to illustrate an optimization strategy that considers this additional information, the arrival of new connections will be initially ignored, only taking into account the terminations. In a second stage a heuristic technique that takes into account also the opening of new connections will be proposed. Let us consider the following example.

Example 1. This example is aimed at pointing out the several reconfigurations of the prices during a connection lifetime. Consider the typical situation shown in Fig. 4, where we have to decide whether the new GP3 connection increases the total revenue.

Let t_a^{GP3} be the instant of time when the new GP3 connection presents its request at the CAC block. The connection will be closed at time instant t_c^{GP3} and it needs a bandwidth allocation of y^{BE3}. Let b be the price for the GP traffic. We consider now all of the possible bandwidth reallocations (and consequently revenue reallocations, too) during the new GP3 connection. We indicate with $t_a^{(\bullet)}$ and $t_c^{(\bullet)}$ the time instants of an arrival and of a termination of a connection, respectively. In this situation, if we want to explicitly compute the total revenue changes during the GP3 connection, it is necessary to calculate all the changes in the w^{BE1} and w^{BE2} values in all the time intervals between the time instant t_a^{GP1} and time instant t_c^{GP3}. Such intervals are highlighted in Fig. 4. The revenue from the BE traffic depends on whether the GP3 connection is refused ($i = 0$) or accepted ($i = 1$):

$$\begin{aligned} G_i^{BE} = &[w_i^{BE1}(t_a^{GP3}; t_c^{GP2}) + w_i^{BE2}(t_a^{GP3}; t_c^{GP2})](t_c^{GP2} - t_a^{GP3}) \\ &+ [w_i^{BE1}(t_c^{GP2}; t_c^{BE2}) + w_i^{BE2}(t_c^{GP2}; t_c^{BE2})](t_c^{BE2} - t_c^{GP2}) \\ &+ w_i^{BE1}(t_c^{BE2}; t_c^{GP1})(t_c^{GP1} - t_c^{BE2}) \\ &+ w_i^{BE1}(t_c^{GP1}; t_c^{GP3})(t_c^{GP3} - t_c^{GP1}), \quad i = 0, 1. \end{aligned}$$

The GP3 connection is finally accepted if and only if

$$G_1^{BE} + y^{GP3} \cdot b \cdot (t_c^{GP3} - t_a^{GP3}) \geqslant G_0^{BE}. \tag{13}$$

The analytical model (based on an average reward dynamic programming problem) proposed in Lin and Shroff (2001) and Paschlidis and Tsitsiklis (2000) deals only with the GP traffic. It is aimed at calculating the optimal GP pricing policy without the presence of the BE traffic. Moreover, as is pointed out in Lin and Shroff (2001), the Proportional Fairness Pricing model does not consider the dynamic evolution of a network, i.e., the interarrival times and the durations of the connections. As is shown by the previous example, the expected revenue of a GP connection in a heterogeneous network is a function of both its equivalent bandwidth and of all the prices reallocations of the BE users multiplexed with such GP connection. For all these reasons, it is quite difficult to foresee with an analytical model the expected revenue of a GP connection in a heterogeneous network. We shall proceed in the following with a heuristic approach.

Let now GP^{new} be the new incoming GP connection, which is subject to the CAC based on the total revenue comparison in a more general situation. As done in Example 1, for calculating the two terms G_{with}^{BE} and $G_{without}^{BE}$ it is necessary to break in the same way as it was previously mentioned the interval between the beginning and the end of the GP^{new} connection in all of the intervals where there are no changes in the $\boldsymbol{w}$ and $\boldsymbol{x}$ vectors. To take into account also the arrival of new requests from GP and BE traffic during the GP^{new}

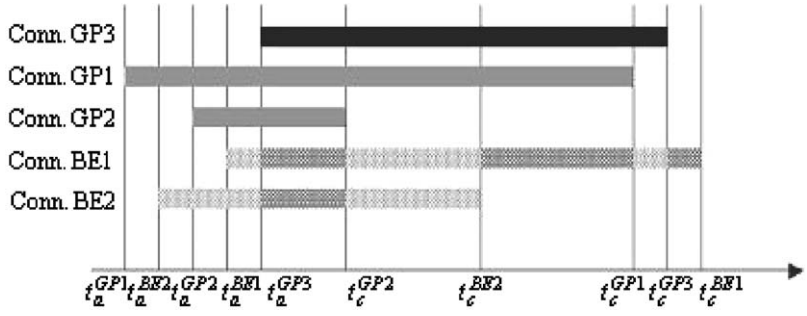

Fig. 4. Typical sequence of events during a GP connection; each of these determines a bandwidth reallocation for the BE traffic.

connection, we have used a heuristic approach based on simulation. Namely, for every new request of GP connection, GP^{new}, we generate n different simulation runs for the length of the GP^{new} connection. At the end of each simulation we calculate the overall revenue considering the terminations of the BE and GP connections within the GP^{new} call and the arrivals of new BE and GP connections within the same time interval. In this way, all of the bandwidth reallocations that happen during the duration of the GP^{new} connection are considered. As to the new GP connections starting during the length of GP^{new}, a CAC strategy based only on the bandwidth availability is applied. Then we determine an estimate of the expectation of the revenue by means of n simulations considering two situations: in the case that the new GP connection is accepted (G^{BE}_{with}), and in the case that the new GP connection is not accepted ($G^{BE}_{without}$). The final choice is to accept the incoming GP^{new} only if it increases the estimate of the mean value of the total revenue obtained at the end of the recursive procedure. In the appendix we present the pseudocode of the procedure used to decide whether to accept or not the incoming GP^{new} connection by applying this strategy.

This strategy (called "CACPricing2" in the following) can be referred to the family of the so-called Receding Horizon techniques. A performance index (the revenue) that is referred to a finite temporal window (the duration of GP^{new}) is maximized. The perfect information on the termination instants of all the connections that will end during the GP^{new} call is exploited and a Montecarlo simulation is performed to take into account arrivals of the new GP and BE connections. All of this additional information leads to an estimation of the expectation of the overall revenue that could be obtained accepting or refusing the GP^{new} connection.

Let $\hat{G}_{without}$ and $\hat{G}_{with}$ be the estimation of the expectation of the revenue when the GP^{new} terminates in the case GP^{new} is refused or accepted, respectively. In spite of the fact that the "CACPricing2" technique is based on a heuristic simulation approach, it is necessary to estimate the standard deviation of $\hat{G}_{without}$ and $\hat{G}_{with}$ values during the simulation procedure used by "CACPricing2". As is known from the theory of the Confidence Intervals (see e.g. Lind & Mason, 1994), the reliability of a simulation measure of a mean value is based on the following equations:

$$\hat{x} - \Delta_{\hat{x}} \leqslant \mu \leqslant \hat{x} + \Delta_{\hat{x}},$$
$$\Delta_{\hat{x}} = z\frac{s}{\sqrt{n}},$$
$$s = \sqrt{\frac{\sum_{i=1}^{n-1}(x_i - \hat{x})^2}{n-1}}, \tag{14}$$

where $\hat{x}$ is the estimation of the mean value, n the number of samples, s the standard deviation of the population samples, μ the true value of the mean, and z is a constant that relates to the probability that μ lies in the confidence interval constructed by means of Eq. (14). In our model, there are two confidence intervals, one for $\hat{G}_{without}$ and one for $\hat{G}_{with}$. We decide to accept the incoming GP^{new} connection if

$$\hat{G}_{with} - \Delta_{\hat{G}_{with}} \geqslant \hat{G}_{without} + \Delta_{\hat{G}_{without}} \tag{15}$$

and refuse it if

$$\hat{G}_{without} - \Delta_{\hat{G}_{without}} > \hat{G}_{with} + \Delta_{\hat{G}_{with}}. \tag{16}$$

Thus, we apply the CAC decision based on a Montecarlo simulation only if the values returned by the simulation procedure are validated by a confidence interval, namely one of the Eqs. (15) or (16) are satisfied. In the other cases the CAC decisions are taken following the "CACPricing1" technique (Fig. 5).

Clearly, to effectively test the model proposed by the "CACPricing2" strategy, it will be necessary to accurately tuning, by means of simulations, the parameter n (i.e. the number of simulations of the "future" network evolution) in order to guarantee as closely as possible the application of the CAC rules based on Eqs. (15) or (16).

It is clear that this approach is very time consuming and cannot be applied in a real scenario where the CAC block acts on line, but it could be very useful to test the performance of the previous techniques, where an optimization is applied based only on the revenue per unit time.

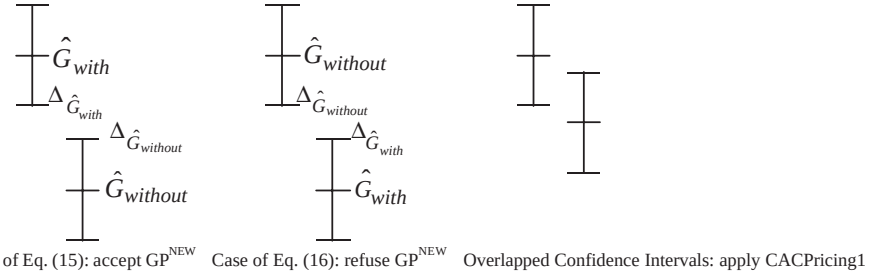

Fig. 5. CACPricing2 decisions: validation by confidence intervals.

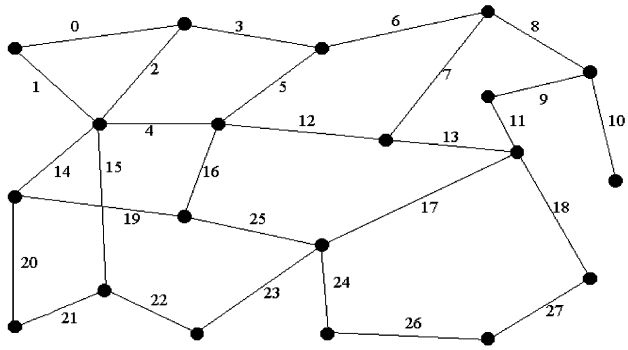

Fig. 6. Topology of the test network.

4. Numerical results

We have developed a simulation tool that describes the behavior of the network at call (GP traffic) and flow (BE traffic) level, to verify the performance of the proposed price-based CAC mechanism. The simulator does not model the packet level. The test network, composed of 28 links and 20 nodes, is shown in Fig. 6. We consider a subset of ten active routes, where each active route can generate both BE and GP traffic connections:

Route 1 : {0, 3, 6, 8, 10}
Route 2 : {4, 5, 6, 8, 10}
Route 3 : {19, 25, 17, 11, 9, 10}
Route 4 : {19, 25, 24, 26}
Route 5 : {2, 15, 21}
Route 6 : {21, 22, 23, 17, 13, 7}
Route 7 : {25, 17}
Route 8 : {12, 13, 18}
Route 9 : {1, 15}
Route 10 : {16, 5}

We have chosen this network topology in order to have a sufficient number of nodes to establish a complex network scenario (i.e., with different traffic routes; see, e.g., the 16-Node Network used in Marbach et al., 2000) and, according to Magoni and Pansiot (2001), we have fixed the average node degree δ to 2.7. It is in fact pointed out in Magoni and Pansiot (2001) that a typical value of δ found in the Internet is 2.0 for the sparse network topologies and 2.9 for the dense network topologies.

In the following simulation results, we shall use the term "connection" to address the establishment of a new (GP or BE) service in the network and to avoid confusion with the term "user". In fact, as we have pointed out at the beginning of Section 3, a "user" that can generate a call in a network node is not generally intended as a single one, but rather as the aggregation of the traffic of different single users. They have in common the same routing paths and the same utility functions, if they require a BE service, or the same equivalent bandwidth if they require a GP service.

We have imposed a probability distribution over all of the significant variables of the problem (namely, interarrival times of the BE and GP users, required bandwidths and utility functions), to produce variable traffic conditions. We have defined as "static" a scenario where each parameter follows a probability distribution with a fixed mean value. This corresponds to a real situation where the users' behavior, for each traffic class, remains the same during the simulation (for example in terms of mean arrival rate or mean duration time of the calls). On the contrary, we have defined as "dynamic" a scenario where each parameter follows a probability distribution with different mean values; namely, the users' behavior can change during the simulation. Both scenarios have been used to test the effectiveness of the proposed techniques.

4.1. The static scenario

Calls are generated following Poisson distributions with mean rates $\lambda_r^{(BE)}$ and $\lambda_r^{(GP)}$ for each route r, for BE and GP traffic, respectively. The call durations follow an exponential distribution with mean value $1/\mu_r^{(BE)}$ and $1/\mu_r^{(GP)}$. In general, the CAC models and the GP Pricing techniques proposed in the literature are tested using Poisson distributed interarrival and service times (see e.g. Lin & Shroff, 2001; Paschlidis & Tsitsiklis, 2000; Ross, 1995; Gokbayrak & Cassandras, 2002). The bit rate of the BE traffic is controlled according to the Proportional Fairness Pricing scheme, using Eq. (1) and Eq. (3), as is shown in Fig. 3. Each GP call requires a bandwidth generated with exponential distribution with mean value γ. We use the following utility function for the BE traffic:

$$U(x) = \alpha\sqrt{x}, \tag{17}$$

where the parameter α is generated with an exponential distribution with mean value 1.

The simulation data are summarized in the following:

- $\lambda_r^{(BE)} = \lambda_r^{(GP)} = \lambda = 10$ calls per minute $\forall r \in \{1, \ldots, 10\}$,
- $1/\mu_r^{(BE)} = 1/\mu_r^{(GP)} = 1/\mu = 1$ minute $\forall r \in \{1, \ldots, 10\}$,
- $c_j = c = 5$ Mbps (link capacity), $\forall j \in \{0, \ldots, 19\}$,
- $\gamma = 1$ Mbps (average bandwidth required by a GP call),
- Time of simulation: 100 min and
- n: number of the internal simulations used by the CACPricing2 strategy: 110.

The number of the internal simulations used in the "CACPricing2" procedure has been fixed by means of preliminary simulation analysis. Setting the confidence interval with $z = 1.96$ and $n = 110$, if the "CACPricing2" strategy is applied, the 90–95% of the calls at the

Table 1
Static simulation scenario #1

	BE revenue (€)	GP revenue (€)	Total revenue (€)
AlwaysAccept	493.451	72.7033	566.1543
HalfAccept	549.5345	41.0011	590.5356
NeverAccept	647.9359	0	647.9359
CACPricing1	610.5655	31.6206	642.1861
VariableGPPrice	493.4514	228.5182	721.9696
CACPricing2	601.0774	98.968	700.0454

Total revenue and its components for the case $b = 0.1$.

Table 2
Static simulation scenario #1

	BE revenue (€)	GP revenue (€)	Total revenue (€)
AlwaysAccept	493.451	727.0332	1220.4842
HalfAccept	549.5345	410.0111	959.5456
NeverAccept	647.9359	0	647.9359
CACPricing1	505.108	714.5679	1219.6759
VariableGPPrice	493.4514	228.5182	721.9696
CACPricing2	493.5101	796.6199	1290.13

Total revenue and its components for the case $b = 1.0$.

CAC module were solved making the CAC decision on the basis of a comparison validated by the confidence interval (i.e., one of the Eqs. (15) and (16) is satisfied), while for the remaining 10–5% of the calls it was necessary to apply the "CACPricing1" strategy.

The results obtained with the proposed CAC with price optimization are compared with fixed CAC rules, which accept a constant percentage p of calls that do not violate the bandwidth constraints. This means that a CAC checking bandwidth availability is applied when a call enters the network. Among the calls that respect this rule, only the percentage p is really accepted. We have considered three different percentages: $p = 100\%$ ("AlwaysAccept" strategy), $p = 50\%$ ("HalfAccept" strategy) and $p = 0\%$ ("NeverAccept" strategy). These fixed CAC rules are aimed at representing two extreme conditions, and an average one as well, concerning the acceptance of the incoming GP requests; in this way the contribution of the GP traffic to the total revenue is highlighted.

Tables 1 and 2 show the total revenue (Fig. 7) and the incoming revenue from each type of traffic class, for two different choices of the price b, which is the amount of money per unit of sent GP traffic data the user pays for

- *Case* 1: $b = 0.1$ €/Mbps/min.
- *Case* 2: $b = 1$ €/Mbps/min.

The corresponding GP traffic blocking probability is shown in Fig. 8.

The performed simulations fall in the category of the so-called "finite time horizon" or "terminating" simulations (Pawlikowski, 1990). The pricing strategies are compared in terms of the total revenue and blocking probability of GP traffic. For the simulation times of the following results the variability of the performance parameters using independent replications (i.e., using different random seeds) is quite low. For computational time reasons (in particular for the "CACPricing2" strategy) the Independent Replications technique for the analysis of stochastic simulation systems (Pawlikowski, 1990) (i.e., the repetition of the same simulation with different pseudorandom number generators, until a confidence interval is reached for the performance parameter) could not be applied. For this reason, a fixed sequence of realizations for the stochastic processes involved in the problem has been used. In this way it is guaranteed that the characteristics of the requests of all the traffic classes are identical in each simulation where a different Pricing technique is applied.

For $b = 0.1$, the BE traffic represents the greatest source of revenue for the ISP, while the GP traffic

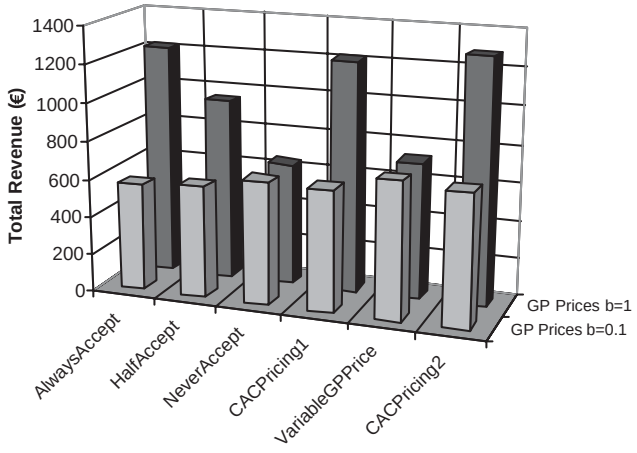

Fig. 7. Static simulation scenario #1. Total revenue [€] with two values of GP prices.

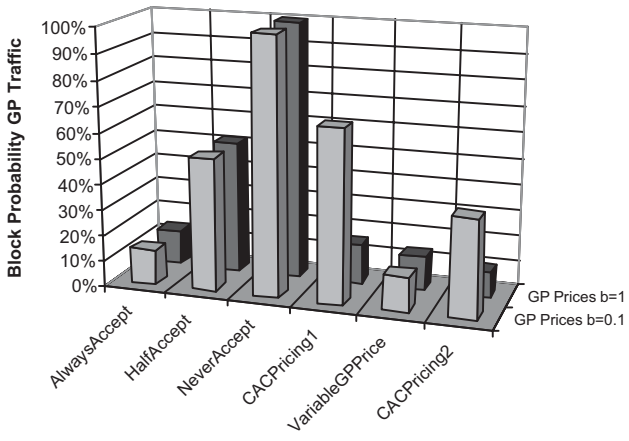

Fig. 8. Static simulation scenario #1. GP traffic Blocking Probability, with two values of GP prices.

contributes for the largest part of the revenue for the $b = 1.0$ case. Consistently to this situation, the best fixed strategies are the "AlwaysAccept" if the GP price is 1.0 and the "NeverAccept" strategy if it is 0.1. This shows that if the price paid by the GP users is low, it may be convenient to refuse all of the GP calls, leaving all the bandwidth to the BE traffic and, in the opposite situation, accepting all GP calls may be much more convenient. In both cases, the "CACPricing1" maximizes the total revenue and offers good performance if $b = 1$, while in the $b = 0.1$ case, more than 67% of the GP connections are refused. However, it is important to remind that from the point of view of the revenue, in this situation, the best solution was "NeverAccept", so "CACPricing1" obtains similar revenue, but with a much lower blocking probability.

Concerning the "VariableGPPrice" performance, in spite of the fact that this technique is indifferent to the change in b and that its revenue performance turns out to be the best one when $b = 0.1$, while it appears quite poor if $b = 1.0$, we could note that $b = 0.1$ is a too low value (GP users pay less than they would have to), while $b = 1.0$ is slightly too high. Moreover, "VariableGPPrice" guarantees a very low blocking probability (the same as "AlwaysAccept"), because it accepts all the connections that pass the first CAC level based on the bandwidth availability.

As regards the difference between the "CACPricing1" and "CACPricing2" performance, clearly, the latter, due to its estimation of the future, has been able to guarantee an improvement in terms of the obtained total revenue and a much lower blocking probability, especially for the $b = 0.1$ case. Anyway, it must pointed out that the "CACPricing1" technique (which is based only on the maximization of the actual revenue per unit time) already works well and the inaccuracy that is made ignoring the opening and the termination of new connections in the future is not so significant, in particular under the revenue performance point of view.

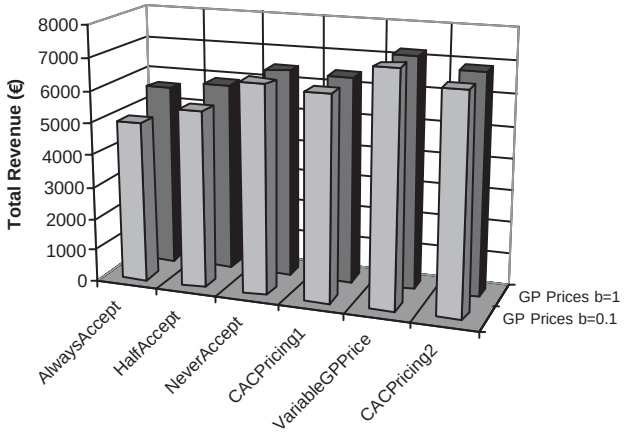

Fig. 9. Static simulation scenario #2. Total revenue (€) with two values of GP prices.

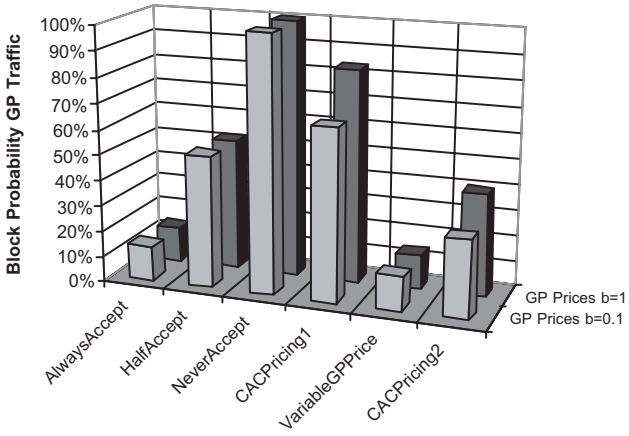

Fig. 10. Static simulation scenario #2. Blocking Probability for GP traffic, with two values of GP prices.

We consider now a scenario where the willingness to pay of the BE users is increased by an order of magnitude; namely, the parameter α of Eq. (17) is generated with an exponential distribution with average value 10. The results are summarized in Figs. 9 and 10. In this situation, for both $b = 0.1$ and $b = 1$, not accepting the GP calls results to be the best choice, because the tariffs and the associate revenue regarding the GP traffic are very low. The BE users' availability to pay dominates. Again the proposed pricing strategies determine the maximum revenue and a lower blocking probability, too.

The proposed CAC mechanisms offer a lower blocking probability than the fixed strategy that maximizes the overall revenue (in particular, when $b = 0.1$). Therefore, they allow, with the same revenue, to satisfy a greater number of users. This is a useful property if, for instance, a fixed tariff is added to the price imposed to every user that enters the network for the first time; in this case the revenue would draw benefit from the lower blocking probability. We shall return to this topic again, at the end of the analysis of the dynamic scenario.

4.2. The static scenario. CACPricing1 and CACPricing2 performance in a non-Markovian environment

From all the results presented in the previous section, we can see that "CACPricing2" provides higher revenue and a lower blocking probability than the "CACPricing1". Clearly, this is due to the receding horizon estimation of the future applied by the "CACPricing2". In this section we want to investigate the performance differences between these two techniques with a different stochastic behaviour of the sources. Until now, we have supposed that the calls of all (GP and BE) users are generated following a Poisson distribution. This fact implies the so-called PASTA property (Poisson Arrivals See Time Averages), namely, Poisson arrivals see the time average behaviour of the system. In our test

scenarios the PASTA property could help the "CACPricing1" to guarantee performance very close to the "CACPricing2" one, but, in a non-Markovian environment, the difference between these two techniques might increase. So, as is done by Lin and Shroff in Lin and Shroff (2001), we investigate the performance of our strategies in a non-Markovian environment. To this aim we repeat the simulation performed in the latter static scenarios, by applying a Pareto distribution over the GP and BE users' interarrival times and call durations. As is known, the Pareto distribution allows an infinite variance over the mean of a stochastic process (Pitts & Schormans, 2000). From Figs. 11–14 it is clear that the Pareto distribution provides an increase in the revenue and in the blocking probability, but, as regards the performance difference between the "CACPricing1" and "CACPricing2" strategies in a non-Markovian environment, we can see that such difference is quite similar to the one obtained in the Markovian environment.

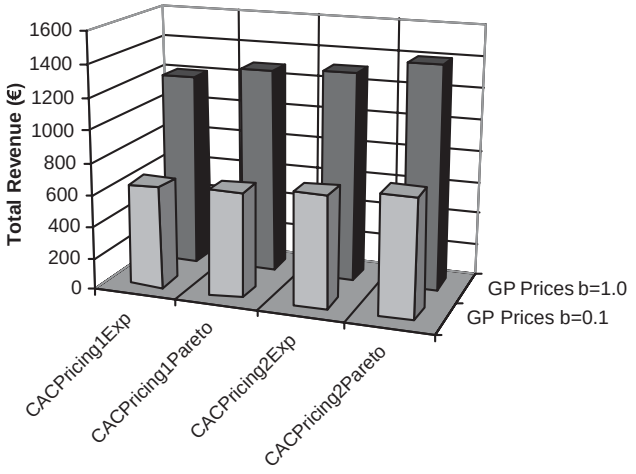

Fig. 11. Static simulation scenario #1, total revenue (€). CACPricing 1 & 2 performance in a Markovian and in a non-Markovian environment.

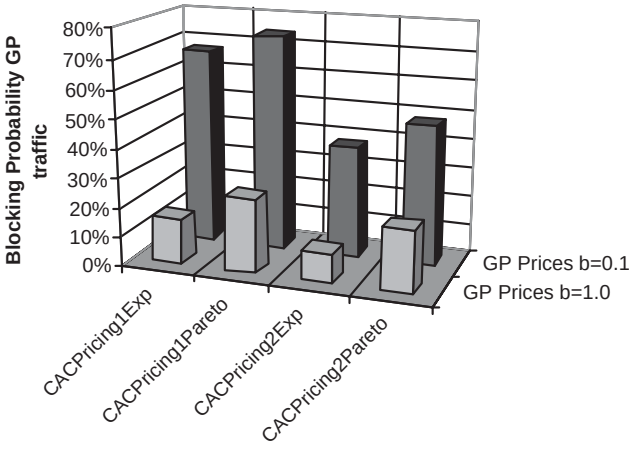

Fig. 12. Static simulation scenario #1, Blocking Probability GP traffic. CACPricing 1 & 2 performance in a Markovian and in a non-Markovian environment.

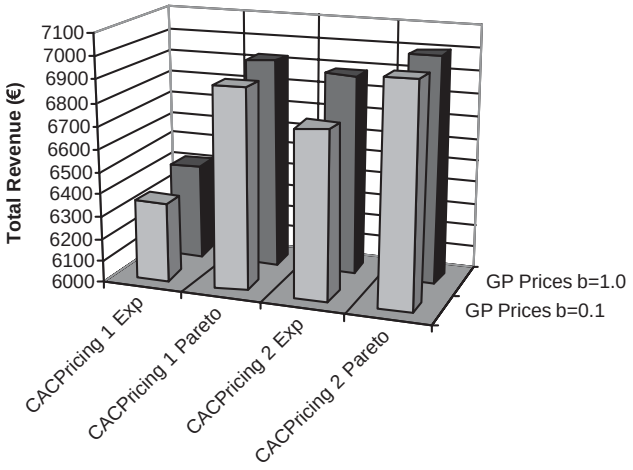

Fig. 13. Static simulation scenario #2, total revenue (€). CACPricing 1 & 2 performance in a Markovian and in a non-Markovian environment.

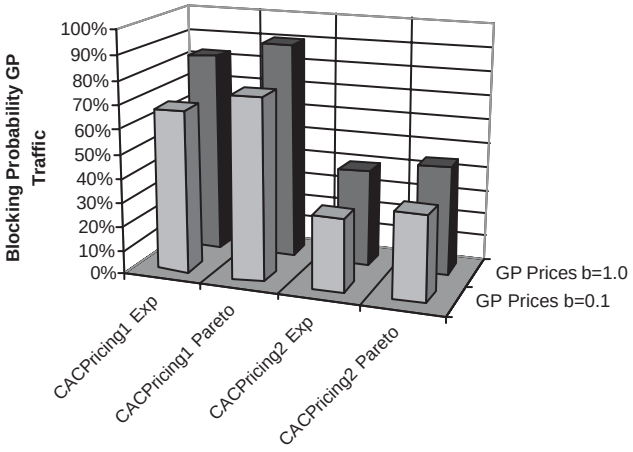

Fig. 14. Static simulation scenario #2, Blocking Probability GP traffic. CACPricing 1 & 2 performance in a Markovian and in a non-Markovian environment.

Moreover, we have found out that setting the parameter n (i.e., the number of simulations of the "future" network evolution in the "CACPricing2" technique) to 110, only the 60–65% of the calls at the CAC module were resolved taking the CAC decision on the basis of Eqs. (15) or Eq. (16), while, for the Markovian environment, these percent values, as we have mentioned at the beginning of Section 4.1, were around the 90–95%. This is clearly due to the stronger variability of a stochastic process when it follows a Pareto distribution. For the same reason in fact, typically, the transient period of a simulation test is much longer if some Pareto distributed variables are involved (see e.g. Pawlikowski, 1990).

4.3. The static scenario: the case of a feedback of the GP prices on the GP calls interarrival times

So far, we have considered the interarrival times of the GP calls a variable independent of the price imposed

by the ISP to the GP users. As is known, the demand $\lambda^{(GP)}(\cdot)$, which determines the arrival rate of the GP calls, is a function of the price b. There exists a price b_{MAX} beyond which the demand $\lambda^{(GP)}(b)$ becomes zero and the function $\lambda^{(GP)}(b)$ is continuous and strictly decreasing in the range $b \in [0, b_{MAX}]$ (Lin & Shroff, 2001; Paschlidis & Tsitsiklis, 2000; Gallego & Van Ryzin, 1997). We want to explicitly take into account the feedback of the price on the GP calls and we repeat the analysis performed for the first static scenario, by verifying if different sensitivities to the GP calls over the GP price can influence the previous results. We use three different parameterisations of the $\lambda^{(GP)}(b)$ function, following the formula used in Paschlidis and Tsitsiklis (2000):

$$\lambda^{(GP)}(b) = \lambda_0^{(GP)}\left(1 - \frac{b}{b_{MAX}}\right),$$

(a) $\lambda_0^{(GP)} = \frac{100}{9}$ and $b_{MAX} = 10$,

(b) $\lambda_0^{(GP)} = \frac{1000}{99}$ and $b_{MAX} = 100$, and

(c) $\lambda_0^{(GP)} = \frac{100}{5}$ and $b_{MAX} = 2$.

For all of these cases, if the ISP sets $b = 1$, $\lambda^{(GP)}$ is always 10 calls/min, so, for the $b = 1$ case, the simulation results are the same as the ones obtained in Tables 1 and 2 and Figs. 7 and 8. $\lambda^{(GP)}$ is quite similar to 10 calls/min for case (b), it is in fact 10.1 calls/min. For case (b), the simulation results are the same as those in Tables 1 and 2 and Fig. 7, in the $b = 0.1$ case.

On the contrary, for the (a) and (b) cases, $\lambda^{(GP)}$ is quite different if the price for the GP calls is set to 0.1. If $b = 0.1$, we have in fact $\lambda^{(GP)} = 11$ in case (a) and a $\lambda^{(GP)} = 19$ in case (c). In Figs. 15 and 16 we compare the simulation results obtained in the first static scenario for $b = 0.1$ among $\lambda^{(GP)} = 10$, $\lambda^{(GP)} = 11$ and $\lambda^{(GP)} = 19$. We can see that, in spite of the change in $\lambda^{(GP)}$, the simulation results are quite similar. The best fixed strategy is the "NeverAccept" and the "CACPricing1"

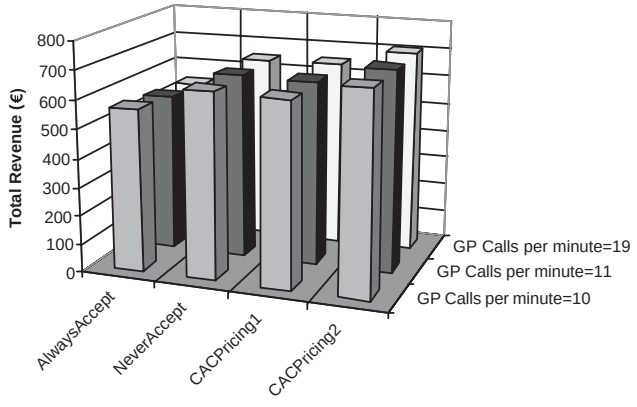

Fig. 15. Static simulation scenario #1, total revenue (€) for the $b = 0.1$ case with different values of the GP calls interarrival times.

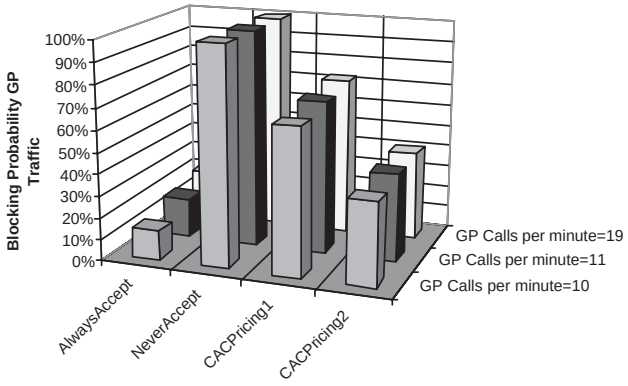

Fig. 16. Static simulation scenario #1, Blocking Probability GP traffic for the $b = 0.1$ case with different values of the GP calls interarrival times.

and "CACPricing2" strategies perform well in both the obtained revenues and the blocking probabilities. There is only a change in the "AlwaysAccept" performance from $\lambda^{(GP)} = 10$ to $\lambda^{(GP)} = 19$, i.e. a little decrease in the obtained revenue and a little increase in the blocking probability. So, the comments proposed for the analysis of the simulation results of the static scenarios are confirmed also taking into account different sensitivities in the GP calls as regards the GP prices.

4.4. The dynamic scenario

Now we consider a situation in which the volume of the traffic and the users' behavior can change within the same simulation. It is well known in fact, in the context of communication networks, that the traffic profile depends on the period of the day (Ben-Ameur, 2002; Ben-Ameur, Gourdin, Liau, & Michel, 2000): for example, most of the traffic carried during the day is professional traffic (e.g. among companies), so it is prevalently GP traffic, while BE traffic (for example residential traffic) dominates in the evening. Furthermore, in Lin and Shroff (2001) it is pointed out that, in large telecommunication networks, the revenue guaranteed by an optimal dynamic pricing strategy, where the prices are optimized during the network evolution in function of the current utilization of the resources, can be always reached by an optimal static pricing strategy if the statistics of the sources are quite "regular" (i.e., stationarity and ergodicity of the interarrival times of the calls in function of the prices). We have found out something similar in our static scenarios too, where there always exists a fixed strategy that maximizes the overall revenue. It could be interesting to analyze the performance of an heterogeneous network where the average behavior of the sources changes, namely, the stationarity and the ergodicity of the processes is verified only in some time periods but not during all the network lifetime.

From the previous results within the static scenario we could note that the best revenue strategy changes when it passes from the first static scenario with $b = 1$ to the second static scenario again with $b = 1$.

In both circumstances the proposed strategies provide the maximum revenue, so it is reasonable to expect that, in dynamic conditions, they are able to provide better CAC choice than the fixed techniques.

We consider now a situation in which the willingness to pay of BE users changes a number of times within the same period of simulation, e.g., the utility functions for the BE traffic are generated according to (17), where the parameter α is generated with an exponential distribution with increasing mean value from 1 (first static scenario) to 10 (second static scenario) (see Fig. 17).

The other simulation data are almost the same as in the previous static scenario (only the mean call duration is increased), but the simulation time is increased to 2000 min:

- $\lambda_r^{(BE)} = \lambda_r^{(GP)} = \lambda = 10$ calls/min $\forall r \in \{1, \ldots, 10\}$,
- $1/\mu_r^{(BE)} = 1/\mu_r^{(GP)} = 1/\mu = 5$ min $\forall r \in \{1, \ldots, 10\}$,
- $c_j = c = 5$ Mbps (link capacity), $\forall j \in \{0, \ldots, 27\}$,
- $\gamma = 1$ Mbps (average bandwidth required by a GP call),
- GP prices $b = 1$ €/Mbps/min,
- Time of simulation: 2000 min, and
- n: number of simulations of the procedure used by the CACPricing2 strategy: 110.

The results are summarized in Figs. 18 and 19. Observing the values of the revenue obtained at the

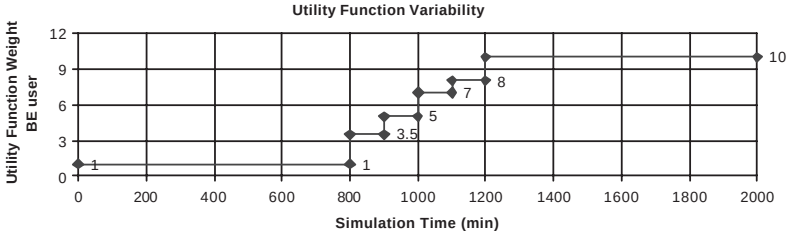

Fig. 17. Dynamic simulation scenario #1. Utility Function Variability.

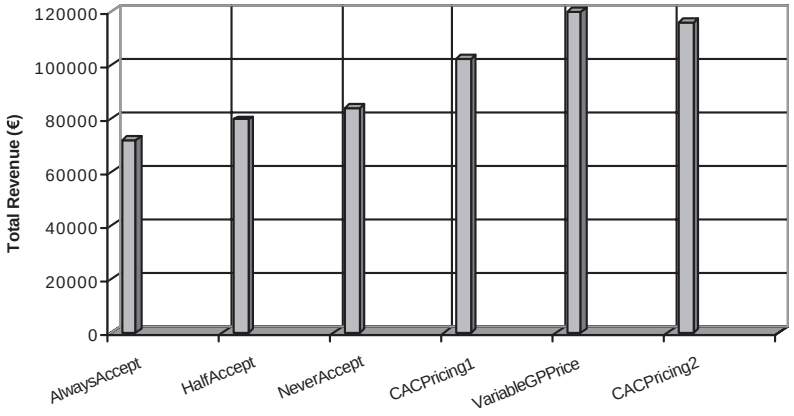

Fig. 18. Dynamic simulation scenario #1. Total Revenue (€).

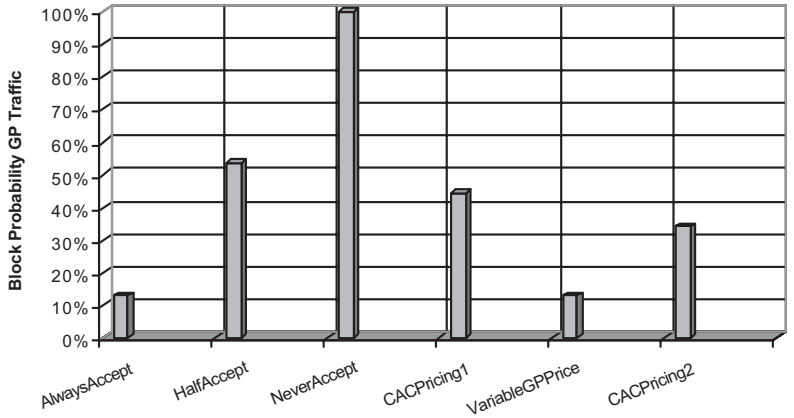

Fig. 19. Dynamic simulation scenario #1. GP Traffic Blocking Probability.

end of the simulation period it is clear that, in dynamic conditions, the fixed strategies do not succeed in optimizing the overall revenue. The proposed strategies can better adapt to dynamic traffic conditions.

We now consider a different dynamic situation, where the willingness to pay remains the same, but there is a strong increase in the interrarival frequency of the BE users from 1 call/min in the first 800 min to 20 calls/min in the last 800 min (Fig. 20).

The utility functions for the BE traffic are again

$$U(x) = \alpha\sqrt{x},$$

where the parameter α is generated with an exponential distribution with mean value 1 and the other simulation data are almost the same as in the previous static scenario (only the mean call duration is increased), but again with an increase in the simulation period of 2000 min:

- $\lambda_r^{(GP)} = 10$ calls/min $\forall r \in \{1, \ldots, 10\}$
- $1/\mu_r^{(BE)} = 1/\mu_r^{(GP)} = 1/\mu = 5$ min $\forall r \in \{1, \ldots, 10\}$
- $c_j = c = 5$ Mbps (link capacity), $\forall j \in \{1, \ldots, 28\}$
- $\gamma = 1$ Mbps (average bandwidth required by a GP call)
- GP prices $b = 0.1$ €/Mbps per minute
- Time of simulation: 2000 min
- n: number of simulations of the procedure used by the CACPricing2 strategy: 110.

We can see (Figs. 21 and 22) that the proposed pricing strategies optimize the overall revenue. It is clear from

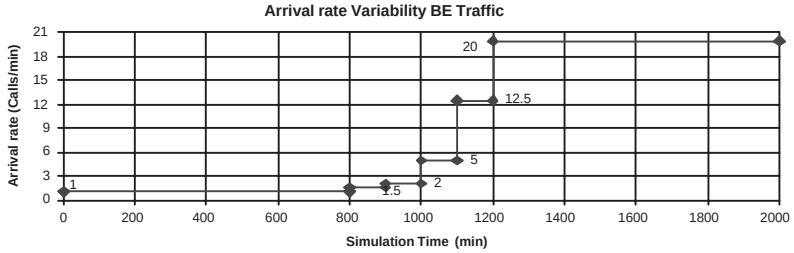

Fig. 20. Dynamic simulation scenario #2. Arrival Rate Variability (BE Traffic).

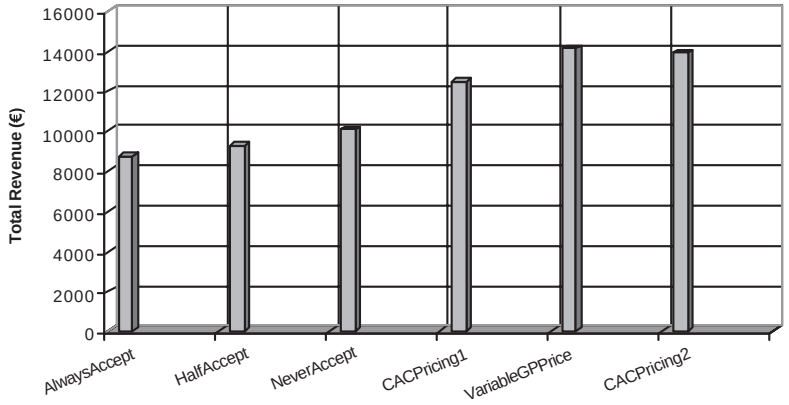

Fig. 21. Dynamic simulation scenario #2. Total Revenue (€).

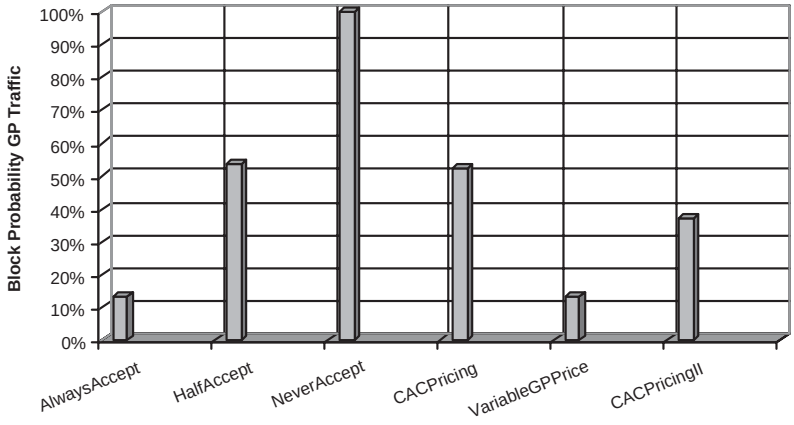

Fig. 22. Dynamic simulation scenario #2. Blocking Probability (GP Traffic).

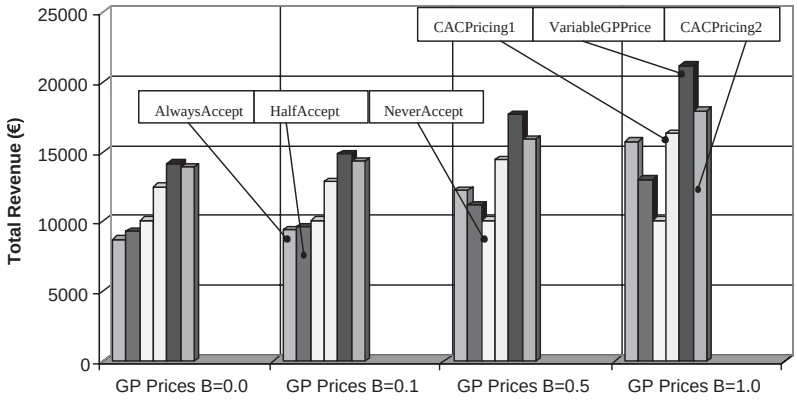

Fig. 23. Dynamic simulation scenario #2. Total Revenue with an additional price for the establishment of a GP call.

these results that at the increase of the traffic variations during the same period of simulation, the difference between the performance obtained with the fixed strategies and the CAC with Pricing optimization strategies increases too.

In a dynamic scenario, an ISP adopting only fixed strategies could reach the best revenue only with a perfect estimate of the traffic variability and calculating off line the best fixed strategy that has to be used in every time interval where the traffic conditions reach an overall stability.

On the contrary, adopting the proposed strategies and keeping them updated with an estimate of the current traffic conditions, it is sufficient to have the CAC block always supply the best choice in a dynamic scenario.

Now we return more in detail to the topic regarding a more complex tariff imposed to GP users, in order to evaluate our strategies with a revenue influenced by the Blocking Probability performance. So far, in fact, we have treated the revenue and the blocking probability as independent variables; now we want to establish a first simple relationship between them. When the proposed pricing mechanisms offer a lower blocking probability, they also allow, at the same revenue, to satisfy a greater number of users. If, for instance, a fixed tariff were added to the price imposed to each user that enters the network (e.g., a tariff in function of the cost of the ISP to signal the establishment of the new connection by using the RSVP protocol (Wroclawski, 1997), also the revenue would draw benefit from a lower blocking probability.

Considering the results presented in the latter dynamic scenario, we introduce a further fixed price that a GP user must pay in order to enter the network for the first time. The user is charged a price $b = 0.1$ [€/Mbps/min] again, but also a fixed price B [€]. In spite of the fact that, in a telecommunication network, the revenue from the maintenance of a call should be larger than the establishment one, a reasonable value of B should be $B \leqslant 0.5$€ (0.5€ is the mean revenue for a GP call on the basis of the price b). In Fig. 23 the total revenue with different values of B around the 0.5€ value is presented.

It can be noticed that because of the increasing in the B price, the best fixed strategy changes from "NeverAccept" for the $B = 0.0$ case to "AlwaysAccept" for the $B = 1.0$ case. The other proposed strategies, especially the "VariableGPPrice", maintain an overall optimality in function of the variability of the B price.

5. Conclusions and future work

In this paper we have proposed three CAC techniques taking into account the price mechanisms that operate in networks where there are both Best Effort and Guaranteed Performance traffics. The mechanism includes a flow control method (Proportional Fairness Pricing) for the Best Effort traffic and an original CAC scheme for Guaranteed Performance calls. The first CAC strategy proposed ("CACPricing1") is based on the comparison of the instantaneous revenue of the whole network in two consecutive time instants (before and after the CAC decision instant) and it is compared with the third one ("CACPricing2") where the obtained total revenue is calculated taking also into account the future events that can happen after the CAC decision. "CACPricing2", due to its receding horizon estimate of the future events, effectively guarantees higher revenue and lower blocking probability than those of the "CACPricing1". But the difference between these two techniques is not so significant in terms of the obtained revenue. So "CACPricing1" can be effectively applied on line in order to optimize the Internet Service Provider's revenue. The second CAC strategy proposed ("VariableGPPrice") is aimed at evaluating the Internet Service Provider's revenue, by setting the GP pricing only on the basis of the utilization of the network resource. The simulation results presented show that the proposed mechanism adapts well to traffic and price

changes, in particular in a dynamic scenario, where the average users' behavior can change during the network lifetime. It is reasonable to expect that an Authority for the control of the market could impose the ISP to fix the price for the Guaranteed Performance traffic as a function of the establishment and maintenance costs of a connection, plus a further taxable price based on a strategy quite similar to the "VariableGPPrice" (in order to increase the price of a Guaranteed Performance call in function of the variability of the service received by the Best Effort users).

Future work could include the development of the proposed CAC techniques under a complete distributed mechanism, possibly taking into account also the dynamics of the Proportional Fairness Pricing optimum in the presence of fluctuations in the bandwidth allocation of the BE traffic. A model for the maximization of a unified social welfare for both the GP and the BE users is under investigation, too.

Appendix

Pseudocode of the procedure used by the CACPricing2 strategy

CACPricing2(GP^{new}) {
  $t_a^{GP^{new}}$: *arrival time of* GP^{new} *connection*;
  $t_c^{GP^{new}}$: *termination time of* GP^{new} *connection*;
  n: *number of simulations of the network evolution*;
  $T = [t_a^{GP^{new}}, t_c^{GP^{new}}]$: *length of the n simulations*;
  $\hat{G}_{without}$: *estimation of the expectation of the revenue at the end of* T *in the case* GP^{new} *is refused*;
  $\hat{G}_{with}$: *estimation of the expectation of the revenue at the end of* T *in the case* GP^{new} *is accepted*;
  T_EV: *set of the time instants of all of the GP and BE connection terminations inside* T;

  for $i = 1$ *to* n {
    B_EV^i: *ith event list of BE and GP births inside* T *and corresponding death time*
    $EV^i = T\_EV \cup B\_EV^i$;
    On the basis of EV^i:
      - *Simulate the evolution of the network over* T *in the case* GP^{new} *is refused*;
      - $G^i_{without}$ = *resulting overall revenue at the end of* T;
      - *Simulate the evolution of the network over* T *in the case* GP^{new} *is accepted*;
      - G^i_{with} = *resulting overall revenue at the end of* T;
  }
  $\hat{G}_{without} = \dfrac{\sum_{i=1}^{n} G^i_{without}}{n}$,
  $\hat{G}_{with} = \dfrac{\sum_{i=1}^{n} G^i_{with}}{n}$.
  Decide to accept or refuse GP^{new} *on the basis of the* $\hat{G}_{without}$ *and* $\hat{G}_{with}$ *values*
}

References

Altman, E., Artiges, D., & Traore, K. (1999). On the integration of best-effort and guaranteed performance services. *European Transactions on Telecommunications* (Special Issue on Architectures, Protocols and Quality of Service for the Internet of the Future) *10*(2), 125–134.

Apostolopoulos, G., Guerin, R., Kamat, S., Orda, A., & Tripathi, S. K. (1999). Intradomain QoS routing in IP networks: A fisibility and cost/benefit analysis. *IEEE Network*, *13*(5), 42–54.

Apostolopoulos, G., Guerin, R., & Tripathi, S. K. (1998). QoS routing: A performance perspective. *Proceedings of ACM SIGCOMM'98*, Vancouver, BC, September 1998.

Athuraliya, S., & Low, S. H. (2000). Optimization flow control with Newton-like algorithm. *Journal of Telecommunication Systems*, *15*(3/4), 345–358.

Baglietto, M., Parisini, T., & Zoppoli, T. (2001). Distributed-information neural control: The case of dynamic routing in traffic networks. *IEEE Transactions on Neural Networks*, *12*(3), 485–502.

Barnhart, C. M., Wieselthier, J. E., & Ephremides, A. (1993). An approach to voice admission control in multihop wireless networks. *Proceedings of INFOCOM*, San Francisco, 1993, pp. 246–255.

Barnhart, C. M., Wieselthier, J. E., & Ephremides, A. (1995). Admission control policies for multihop wireless networks. *Wireless Networks*, *1*(4), 373–387.

Ben-Ameur, W. (2002). Multi-hour design of survivable classical IP networks. *International Journal of Communication Systems*, *15*(6), 553–572.

Ben-Ameur, W., Gourdin, E., Liau, B., & Michel, N. (2000). Dimensioning of internet networks. *Proceedings of DRCN2000*, Munich April 2000 (pp. 56–61).

Bertsekas, D. (2001). *Dynamic programming and optimal control* (2nd ed.). Belmont, MA: Athena Scientific.

Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997). *Resource reservation protocol (RSVP)—Version 1 functional specification*. IETF, RFC 2205, September 1997.

Cao, X., Shen, H., Milito, R., & Wirth, P. (2002). Internet pricing with a game theoretical approach: Concept and examples. *IEEE/ACM Transactions of Networking*, *10*(2), 208–216.

Celandroni, N., Davoli, F., & Ferro, E. Static and dynamic resource allocation in a multiservice satellite network with fading. *International Journal of Satellite Communications* (Special Issue on Quality of Service (QoS) for Satellite IP Networks), to appear.

Chao, H. J., & Guo, X. (2002). *Quality of service control in high-speed networks*. New York: Wiley.

Courcoubetis, C., Siris, V. A., & Stamoulis, G. (1996). Integration of pricing and flow control for available bit rate services in ATM networks. *Proceedings of IEEE globecom*, London, UK, 1996 (pp. 644–648).

Crawley, E., Nair, R., Jajagopalan, B., & Sandick, H. (1998). *A framework for QoS-based routing in the internet*. RFC 2386, IETF, August 1998.

DaSilva, L. A. (2000). Pricing for QoS-enabled networks: A survey. *IEEE Communication Surveys (2nd Quarterly)*, *3*(2), 2–8.

Differentiated Services Working Group, http://www.ietf.org/html.charters/diffserv-charter.html.

Falkner, M., Devetsikiotis, M., & Lambadaris, I. (2000). An overview of pricing concepts for broadband IP networks. *IEEE Communication Surveys (2nd Quarterly)*, *3*(2), 9–20.

Gallego, G., & Van Ryzin, G. (1997). A multiproduct dynamic pricing problem and its applications to network yield management. *Operation Research*, *45*(1), 24–41.

Garret, M. W., & Willinger, W. (1994). Analysis, modelling and generation of self-similar VBR video traffic. *Proceedings of SIGCOMM'94*, London (pp. 269–280).

Gokbayrak, K., & Cassandras, C. (2002). Adaptive call admission control in circuit-switched networks. *IEEE Transactions on Automatic Control*, *47*(6), 1234–1248.

Integrated Services Working Group, http://www.ietf.org/html.charters/intserv-charter.html.

Kelly, F. (1996). Charging and accounting for bursty connections. In L. W. McKnight, & J. P. Bailey (Eds.), *Internet economics*. Cambridge, MA: MIT Press.

Kelly, F. (2001). Mathematical modelling of the internet. In B. Engquist, & W. Schmid (Eds.), *Mathematics unlimited-2001 and beyond* (pp. 685–702). Berlin: Springer.

Kelly, F. P., Maulloo, A. K., & Tan, D. K. H. (1998). Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of Operational Research Society*, *49*(3), 237–252.

Kelly, F., & Songhurst, D. (1997). Charging schemes for multiservice networks. In V. Ramaswami, & P. Wirth (Eds.), *Teletraffic contributions for the information age*. Amsterdam: Elsevier.

Kim, H., & Shroff, N. (2001). Loss probability calculation and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Transactions on Networking*, *9*(6), 755–768.

Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1994). On the self-similar nature of ethernet traffic. *IEE/ACM Transactions on Networking*, *2*(1), 1–15.

Lin, X., & Shroff, B. (2001). Pricing-based control of large networks. *Proceedings of the evolutionary trends of the Internet 2001 Tyrrhenian international workshop on digital communications*, *IWDC 2001*, Taormina, Italy.

Lind, D., & Mason, R. (1994). Basic statistics for business and economics, IRWIN books series in statistics. Homewood: IL Irwin.

Low, S. (1999). Optimization flow control with on-line measurement or multiple paths. *16th international teletraffic congress*, Edinburgh, UK, June 1999.

Low, S. H., & Lapsley, D. E. (1999). Optimization flow control, I: Basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, *7*(6), 861–874.

Low, S. H., & Varaiya, P. P. (1993). A new approach to service provisioning in ATM networks. *IEEE/ACM Transactions on Networking*, *1*(5), 547–553.

Ma, Q., & Steenkiste, P. (1997). Quality of service routing for traffic with performance guarantees. *Proceedings of the IFIP fifth international workshop on quality of service*, New York, May 1997 (pp. 115–126).

Magoni, D., & Pansiot, J. J. (2001). *Comparative study of internet-like topology generators*. Technical Report, LSIIT Laboratory, Universite Louis Pasteur.

Malinowski, K. (2002). Optimization network flow control and price coordination with feedback; proposal of a new distributed algorithm. *Computer Communications*, *25*, 1028–1036.

Malowidzki, M., & Malinowski, K. (2002). Optimization flow control in IP networks: Distributed pricing algorithms and reality oriented simulation. *Proceedings of the international symposium on performance evaluation of computer and telecommunication systems* (*SPECTS 2002*), San Diego, CA, July 2002 (pp. 403–413).

Marbach, P., Milhatsch, O., & Tsitsiklis, J. (2000). Call admission control and routing in integrated services networks using neuro-dynamic programming. *IEEE Journal of Selected Areas of Communication*, *18*(2), 197–208.

Murphy, J., & Murphy, L. (1994). *Bandwidth allocation by pricing in ATM networks*. Technical Report, Dublin City University.

Murphy, J., Murphy, L., & Posner, E. (1994). Distributed pricing for embedded ATM networks. *Proceedings of international IFIP conference on broadband communication* (*BB-94*), Paris France, March 1994.

Paschlidis, I. C., & Tsitsiklis, J. N. (2000). Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking*, *8*, 171–184.

Pawlikowski, K. (1990). Steady state simulation of queueing processes: A survey of basic problems and solutions. *ACM Computing Surveys*, *22*(2), 123–170.

Pitts, J. M., & Schormans, J. A. (2000). *Introduction to IP and ATM design and performance* (2nd ed.). New York: Wiley.

Ross, K. (1995). *Multiservice loss models for broadband telecommunication networks*. Berlin: Springer.

Tsybakov, B., & Georganas, N. D. (1998). Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue. *Performance Evaluation*, *32*, 57–80.

Walrand, J., & Varaiya, P. (2000). *High-performance communication networks* (2nd ed.). San Francisco, CA: Morgan-Kaufmann.

Wroclawski, J. (1997). *The use of RSVP with IETF integrated services*. IETF, RFC 2210, September.

Zhang, H. (1995). Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE*, *83*, 1373–1396.