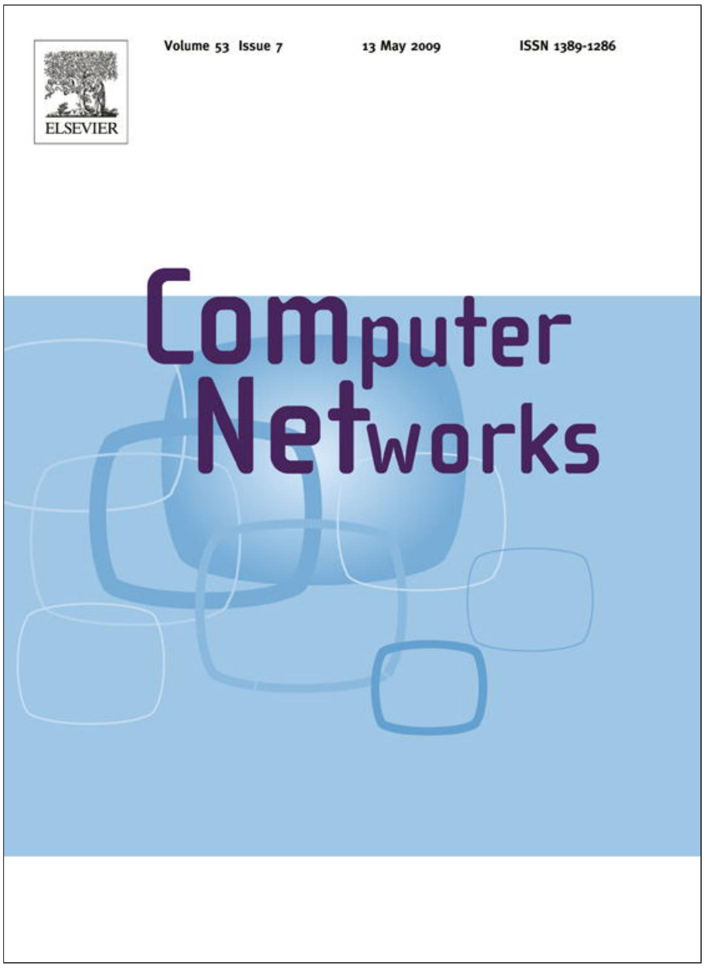
Volume 53 Issue 7
13 May 2009
ISSN 1389-1286
ELSEVIER
Computer Networks

Computer Networks 53 (2009) 973–988

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Maximizing the capacity of mobile cellular networks with heterogeneous traffic ☆

Jorge Martinez-Bauset [a], David Garcia-Roger [b,*], Mª Jose Domenech-Benlloch [a], Vicent Pla [a]

[a] GIRBA-ITACA, Universidad Politecnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain
[b] Dpto. de Ingeniería Telemática, Universidad Carlos III de Madrid, Av. Universidad, 30, 28911 Leganés, Madrid, Spain

ARTICLE INFO

Article history:
Received 27 July 2007
Received in revised form 25 November 2008
Accepted 2 December 2008
Available online 13 December 2008

Responsible Editor: C.B. Westphall

Keywords:
Cellular multiservice networks
Streaming and elastic traffic
Adaptive admission control
QoS objectives
Stationary and non-stationary regimes

ABSTRACT

We propose a novel adaptive reservation scheme that handles, in an integrated way, heterogeneous traffic of two types: streaming and elastic. The scheme adjusts the rates of streaming sessions to meet the QoS objective, adapting to any mix of traffic and enforcing a differentiated treatment among services, in both fixed and variable capacity systems. The resource utilization achieved by streaming traffic is close to the one obtained by an optimal policy, while the efficiency in the use of resources achieved by elastic traffic is greatly improved by limiting the abandonment probability. The performance evaluation carried out verifies that the QoS objective is met with an excellent precision and that the scheme converges rapidly to new operating conditions. We also compare the new adaptive scheme with two previous ones verifying that ours performs better in terms of carried traffic and convergence rate. The proposed scheme has low complexity which makes it practical for real cellular networks.

1. Introduction

One of the fundamental problems in mobile communications is that the radio resources are scarce and need to be managed efficiently in order to provide an acceptable level of quality of service (QoS) to the subscribers. The mobility of terminals makes it very difficult to insure that the resources available at session setup will also be available along the session lifetime, as the terminal moves from one cell to another, unless proper mechanisms are in place. Efficient session admission control (AC) strategies optimize the amount of resources that need to be reserved in each cell for handovers coming from neighboring cells, which allows to increase the carried traffic while meeting the QoS objective. The design of AC policies must take into consideration not only packet related parameters like maximum delay, jitter or losses, but also session related parameters like setup request blocking and forced termination probabilities. Blocking probability is a major QoS parameter in fixed networks [1] but in mobile networks is crucial, as the forced termination probability of admitted sessions is related to the blocking probability of handover requests [2,3]. Additionally, it is now accepted that, compared to scheduling, AC might be a more appropriate traffic management mechanism to provide service differentiation, particularly in wireless links [4].

Applications expected to produce the bulk of traffic in the future multiservice Internet can be broadly categorized as streaming or elastic [5]. Streaming traffic requires a minimum transfer rate in order to work properly as well as some time related requirements such as bounded delay and jitter. Elastic traffic has loose time requirements and can adapt to the available resources. In the light of the above arguments it seems natural to give priority to streaming traffic and leave elastic traffic use the remaining capacity (a small amount of resources might be reserved for the elastic traffic to prevent starvation in case of

☆ A shorter version of this work was presented at ITC20, 2007, Ottawa, Canada.

* Corresponding author. Tel.: +34 916248803.
E-mail address: dgarciaroger@gmail.com (D. Garcia-Roger).

1389-1286/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2008.12.002

overload of the streaming traffic). Elastic flows are generally transported over TCP which takes care of rate adaptation and bandwidth sharing among the different flows. If the total traffic demand of elastic flows exceeds the available capacity some flows might be aborted due to impatience. Flow impatience due to a very low throughput can arise from human impatience or because TCP or higher layer protocols interpret that the connection is broken. Abandonments are useful to cope with overload and serve to stabilize the system but, on the other hand, this phenomenon will have a negative impact on the efficiency because capacity is wasted by non-completed flows [5]. This drop of efficiency led the authors of [5] to claim that AC should also be enforced for elastic traffic.

In this paper we propose a novel adaptive AC scheme for mobile wireless cellular networks, that handles in an integrated way both streaming and elastic traffic and tries to maximize the carried traffic while meeting certain QoS objectives. The QoS objective for streaming traffic is expressed as upper bounds for the blocking probabilities of new and handover requests, while for elastic traffic is defined as a bound for the abandonment probability. The new scheme is adaptive in the sense that if the offered load or the number of resource units changes, or both simultaneously, the AC system will react trying to meet the QoS objective for as many services as possible. Therefore the proposed scheme might be deployed in both fixed capacity systems (e.g. FDMA/TDMA) and systems limited by interference where capacity is variable (e.g. CDMA).

Our work is motivated in part by the fact that previous adaptive proposals like [6–10] deploy long measurement windows to estimate system parameters, which make the convergence time too long to cope with real operating conditions, or do not provide explicit indication of how the time window must be configured [11–13]. Another motivation is the fact that most of the studies devoted to adaptive schemes only consider the stationary regime and no evidence is provided about their behavior in the transient regime. Therefore, we consider that a fundamental characteristic of an adaptive scheme like its convergence rate to new operating conditions has not been sufficiently explored.

Our scheme does not rely on measurement intervals to estimate the value of system parameters. It generalizes the novel adaptive AC strategy introduced in [14], which operates in coordination with the well-known trunk reservation policy named multiple guard channel (MGC). It has been shown that deploying trunk reservation policies instead of the complete-sharing (CS) policy in mobile networks allows the operator to achieve higher system capacity, i.e. to carry more traffic while meeting certain QoS objectives [15].

Many multimedia applications are adaptive in the sense that a satisfactory playback experience may be obtained over a wide range of compression levels. When the number of streaming sessions that share a common link increases, if the system is left uncontrolled, the perceived quality is reduced significantly. However, proper admission control coupled with an adequate rate-adaptation policy has the potential of guaranteeing acceptable quality of reception while limiting the blocking probability and achieving a higher system utilization. Our AC scheme also integrates a rate-adaptation (RA) policy that adjusts the rate of sessions in order to meet the blocking probability objectives, while taking into account the QoS perceived by users. We assume that the perceptual QoS (PQoS) is mostly affected by the frequency of rate changes rather than by the absolute value of the rate [16–18]. In this sense, our RA policy has been designed to request adaptations to a given session, at most, each time the session is handed over to a new cell.

There exists a considerable proportion of related literature where the bandwidth sharing problem is studied by deploying a full sharing approach. In these studies, the rate of ongoing sessions is adapted downwards when a new or handover session cannot be accepted with the minimum rate or upwards after the departure of a session. For multiservice scenarios, the authors of [19] deploy the CS policy as the AC policy, while in [20] a singular version of the multiple fractional guard channel (MFGC) policy is deployed. In both papers the full sharing approach is used for the RA policy, but a slightly more sophisticated version is proposed in [20] where the rate assigned to sessions during overload is proportional to the difference between the maximum and minimum allowed rates of each service. As full sharing approaches suffer from high rate-adaptation frequency and require a considerable signaling load their practical applicability seems to be limited.

A measurement-based AC for single service scenarios is proposed in [17] to limit the degradation ratio (fraction of users with degraded service) and the degradation degree. An analytical methodology to determine the degradation ratio (the fraction of time a user receives degraded QoS) and the upgrade/degrade frequency for single service scenarios is proposed in [2]. In [21] a predictive RA scheme is proposed to limit the degradation ratio in a single service scenario, where the AC takes into account the state of neighboring cells. Unfortunately, no extensions of these schemes to multiservice scenarios have been proposed.

An alternative approach to determine the RA policy is by optimizing a revenue function, that typically is a function (linear or convex) of the assigned rate. Revenue functions can also be interpreted in terms of utility functions [18]. In [22] two optimization problems are formulated to maximize independently either revenue and PQoS or revenue and fairness. They deploy an AC policy of the upper limit type [15], which is coordinate-convex and produces a product-form solution, but is not integrated with the RA policy. Finally, it is suggested in [23] that a more realistic analysis of the PQoS should take into account what happened in all cells visited by a session during its lifetime, as opposed to considering only a single cell analysis. Under some assumptions, its authors obtain a product-form solution for the stationary distribution of a set of cells and discuss how PQoS parameters can be obtained. Unfortunately, as pointed out in the paper, the assumptions do not hold in common mobile networks.

Adaptive AC mechanisms have also been studied, for example in [8–10,3,24,25], both in single service and multiservice scenarios, but in a context that is somewhat different to the one of this paper. There, the adjustment of the AC policy configuration is based on estimates of the mobility pattern and of handover arrival rates derived from the current number of ongoing sessions in neighboring cells. It is expected that the performance of our scheme will improve when provided with such predictive information but this is left for future study. The work in [25] is particularly relevant in the context of this paper. Its authors propose an AC scheme that supports multiple streaming services. The scheme adapts the transmission rate of ongoing sessions according to the state of neighboring cells, being able to limit the forced termination probabilities and to minimize the adaptation rate. The RA policy is determined following an optimization approach. Nevertheless, there are many differences among the scheme proposed there and here, for example: its performance in the transient regime is not evaluated, the adaptation frequency is minimized but it is not bounded, only exponential distributions for the session duration and residence time are considered, and no support is provided for elastic traffic.

The main characteristics of our scheme can be summarized as follows:

1. It is self-adaptive and does not require any configuration parameters beyond the blocking probability objectives, being able to operate in both fixed capacity systems (e.g. FDMA/TDMA) and in systems limited by interference where capacity is variable (e.g. CDMA).
2. It can operate with any arrival process and any distribution of the session duration and residence time.
3. It handles in an integrated way both streaming and elastic traffic. For the streaming traffic, it handles multiple services.
4. For the streaming traffic, it also integrates a rate-adaptation policy that adjusts the bandwidth of sessions in order to meet the blocking probability objectives, while taking into account the QoS perceived by users (PQoS).
5. For streaming services, it allows the operator to freely configure different QoS objectives in terms of bounds for the blocking probabilities of new and handover requests. Besides, the operator can also freely configure a prioritization order that guarantees that during overload episodes higher priority services will be able to meet their QoS objective, possibly at the expense of lower priority ones.
6. For streaming traffic, the resource utilization is close to the one obtained by an optimum MFGC policy, which in turn has a performance very close to the one of the optimal policy [15]. By an optimum MFGC policy we refer to a static policy that is designed for each arrival rate and knowing all system parameters (e.g. session duration rates, cell residence rates, etc.). In practice, real-time estimation of system parameters is a challenging task. Besides, the precision with which system parameters are determined have a mayor impact on the performance of the MFGC policy [15]. Our scheme does not require any configuration parameter nor does it require knowing any system parameter.
7. For streaming traffic, it has a remarkable fast and oscillation-free transient response when compared to previous proposals. The paper discusses a novel performance perspective of AC schemes as it is the study of the transient regime, which was not addressed by previous studies.
8. For elastic traffic, it integrates a flow admission control scheme not considered in previous proposals that only dealt with streaming traffic. This scheme is able to limit the abandonment probability and in turn improve system efficiency by avoiding the waste of resources associated to non-completed flows.

Given that the operation of the AC scheme when handling streaming traffic is independent of the elastic traffic because the former has higher priority than the latter, we describe first the operation of the AC scheme and evaluate its performance only with streaming traffic. The adaptive scheme with streaming traffic is first introduced in a scenario with the RA policy disabled. This allows to understand more clearly its operation. Section 2 describes the fundamentals of the adaptive scheme, introducing the adjustment strategy and how multiple streaming services are handled. In Section 3 we present the performance evaluation of the proposed adaptive scheme when handling streaming traffic in different scenarios, both under stationary and non-stationary traffic conditions. In Section 4 we introduce the RA policy and evaluate its performance. Section 5 describes the operation of the scheme when handling elastic traffic and evaluates its performance. Finally, Section 6 concludes the paper.

2. Operation of the adaptive AC scheme

Throughout the paper we use the term *service* to refer to the same concept referred to as *QoS class* in UMTS or as *service class* in IEEE 802.16 (WiMAX), i.e. a radio bearer service defined between a base station and a mobile terminal [26,27]. For simplicity, we refer to all delay-sensitive services as streaming services and to all non delay-sensitive services as elastic. Additionally, in mobile networks it is common practice to define arrival classes. This allows the system to treat differently new and handover requests of the same service class. This is important because blocking objectives are different for different arrival classes [28,3].

We consider that in each cell a set of R different streaming services contend for C resource units, where the meaning of a unit of resource depends on the specific implementation of the radio interface. For each streaming service, new and handover arrival requests are distinguished, which defines $2R$ arrival classes. For convenience, we denote by s_i the ith arrival class, $1 \leqslant i \leqslant 2R$. Additionally, we denote by s_r^n (s_r^h) the arrival class associated to new (handover) requests of the streaming service r, being $s_r^n = s_r$ and $s_r^h = s_{r+R}$, $1 \leqslant r \leqslant R$. Please refer to Table A.1 for a description of the symbols deployed in the paper. For brevity, when we refer to a service or to a class we mean a streaming service or a streaming arrival class respectively. Elastic traffic is discussed in Section 5.

Service r requests require d_r resource units per session. As each service has two associated arrival classes, if we denote by c_i the amount of resource units that an arrival class requires for each session, then $d_r = c_r = c_{r+R}$, $1 \leqslant r \leqslant R$. For variable bit rate sources, d_r resource units denotes the effective bandwidth of the session. Note that performance objectives at the packet level, like delay and loss rate, can be accounted for by dimensioning appropriately the effective bandwidth [29,30].

We denote by P_i, $1 \leqslant i \leqslant 2R$, the blocking probability perceived by s_i requests and by $P_r^n = P_r$ ($P_r^h = P_{R+r}$) the blocking probability perceived by new (handover) requests of service r. The QoS objective is expressed as upper bounds for the blocking probabilities of each arrival class. Thus, we denote by B_r^n (B_r^h) the bound for new (handover) blocking probabilities. Note that the scheme can handle service classes with different rate requirements but the same blocking objective or vice versa, simply by defining additional classes. Furthermore, the basis of the adaptive scheme holds even when the rate of sessions belonging to a given streaming class is not fixed, i.e. it varies across sessions of the same class. This is demonstrated in Section 4, that is dedicated to the rate-adaptation policy.

Let the ongoing sessions vector be $\boldsymbol{n} := (n_1, \ldots, n_R)$, where n_r is the number of sessions of service r in progress initiated as new or handover requests in the cell. We denote by $c(\boldsymbol{n}) = \sum_{r=1}^{R} n_r d_r$ the number of busy resource units in state $\boldsymbol{n}$. The definition of the MGC policy is as follows. One threshold parameter is associated with each class s_i, $l_i \in \mathbb{N}$. An arrival of s_i in state $\boldsymbol{n}$ is accepted if $c(\boldsymbol{n}) + c_i \leqslant l_i$ and blocked otherwise. Therefore, l_i is the amount of resources that s_i has access to and increasing (decreasing) it reduces (augments) P_i. Number based SAC, which is a common technique in systems which capacity is limited by blocking, has also been considered a good approach for those systems whose capacity is limited by interference, see for example [31] and references therein.

Most of the adaptive schemes proposed for single service scenarios deploy a reservation strategy based on *guard channels*, increasing its number when the QoS objective of the handover arrival class is not met. The extension of this heuristic to a scenario with multiple services is much more difficult to manage because the adjustment of the threshold parameter l_i has an impact not only on the QoS perceived by class s_i but also on the QoS perceived by the rest of classes. Our scheme has been designed to handle this difficulty. As a first step to handle this difficulty, we classify the different arrival classes into two generic categories: (i) several *protected* classes, for which specific QoS objectives must be met and (ii) one *Best-Effort Class* (BEC), with no specific QoS objective. Additionally, the operator can define priorities for the protected classes at its convenience in order to give greater protection to the most important classes. Note that BEC arrival requests perceive an unpredictable blocking probability but those sessions accepted are allocated a constant amount of resources during their lifetime.

For simplicity, we assume that indices associated to the arrival classes define their priority relationship and therefore $\{s_1, \ldots, s_{2R}\}$ defines the prioritization order chosen by the operator. Note that our scheme allows to define as prioritization order any permutation of the arrival classes. Then s_1 is called the *Highest-Priority Class* (HPC) and s_{2R} the *Lowest-Priority Class* (LPC). If there is a BEC, this class will be the LPC. We study two implementations, one in which the LPC is treated as a protected class and one in which the LPC is the BEC.

For the sake of clarity, the operation of our scheme is described assuming that arrival processes are stationary and the system is in steady state. In practice, we can assume without loss of generality that the QoS objective for s_i can be expressed as $B_i = b_i/o_i$, where $b_i, o_i \in \mathbb{N}$. When $P_i = B_i$, it is expected that s_i will experience, in average, b_i rejected requests and $o_i - b_i$ accepted requests, out of o_i offered requests. For example, if the QoS objective for s_i is $B_i = 1/100$, then $b_i = 1$ and $o_i = 100$. It seems intuitive to think that the adaptive scheme should not change the threshold parameters of those arrival classes meeting their QoS objective and, on the contrary, adjust them on the required direction if the perceived QoS is different from the objective.

Therefore, given that the MGC policy deploys integer values for its threshold parameters, we propose to perform a probabilistic adjustment each time a request is processed in the following way: (i) if accepted, do $\{l_i \leftarrow (l_i - \Delta l)\}$ with probability $1/(o_i - b_i)$; (ii) if rejected, do $\{l_i \leftarrow (l_i + \Delta l)\}$ with probability $1/b_i$, where $\Delta l \in \mathbb{N}$ is the adjustment step for the threshold parameters. Under stationary traffic, if $P^i = B^i$ then, on average, l^i is increased by Δl and decreased by Δl every o^i offered requests, i.e. its mean value is kept constant. When the traffic is non-stationary the adaptive scheme will continuously adjust the thresholds in order to meet the QoS objective if possible, adapting to any mix of traffic. Note also that in the operation of this simple scheme no assumptions have been made concerning the arrival processes or the distribution of the session duration and cell residence times. As in our scheme the thresholds are not configuration parameters but internal variables handled by the AC system, from now on we will refer to them simply as thresholds.

Fig. 1 shows the operation of the AC policy and the adaptive scheme in more detail. As shown, to admit a s_i request it is first checked that at least c_i free resource units are available. Note that once this is verified, HPC requests are always admitted, while the rest of classes must also fulfill the admission condition imposed by the AC policy. Once the admission decision has been taken, the adaptive scheme performs the probabilistic adjustment of the corresponding threshold. The probabilistic adjustment is described in subroutines SR1 and SR2 that are shown in Fig. 2. In a simplified manner, the adaptive scheme can be perceived as composed of one individual adaptive scheme per arrival class. Typically, these individual adaptive schemes operate independently, except when one of the arrival classes is suffering from congestion. When this happens, the adaptive schemes of lower-priority classes become under control of the adaptive scheme of the class suffering from congestion. We say that adaptive schemes of these lower-priority classes get disabled. Note that the threshold of the BEC is not updated when admission decisions are made for arrivals of this class.

When the threshold associated to s_i has to be increased, which is an indication that this class requires more resources to meet its QoS objective, two different strategies are deployed in subroutine SR2 of Fig. 2. The *direct* way

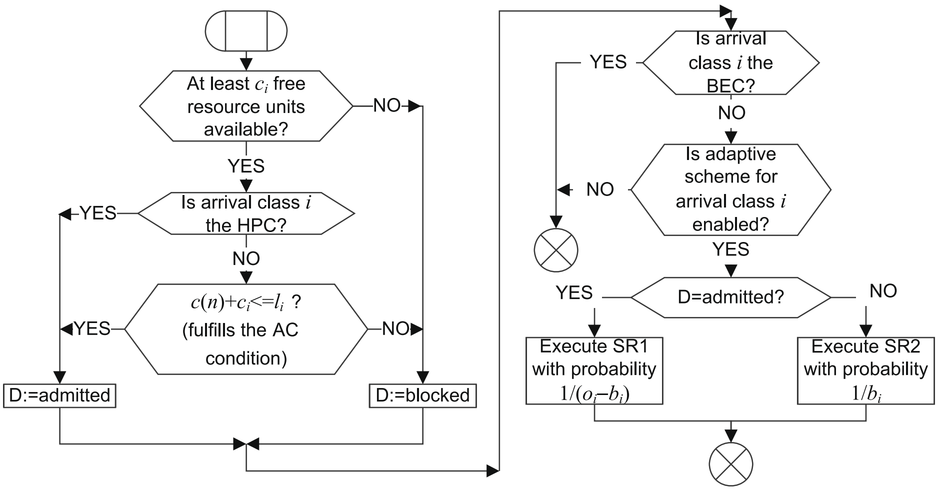
At least c_i free resource units available?
NO
YES
Is arrival class i the HPC?
YES
NO
$c(n)+c_i<=l_i$?
(fulfills the AC condition)
YES
NO
D:=admitted
D:=blocked
Is arrival class i the BEC?
YES
NO
Is adaptive scheme for arrival class i enabled?
NO
YES
D=admitted?
YES
NO
Execute SR1 with probability $1/(o_i-b_i)$
Execute SR2 with probability $1/b_i$

Fig. 1. Operation of the adaptive AC scheme.

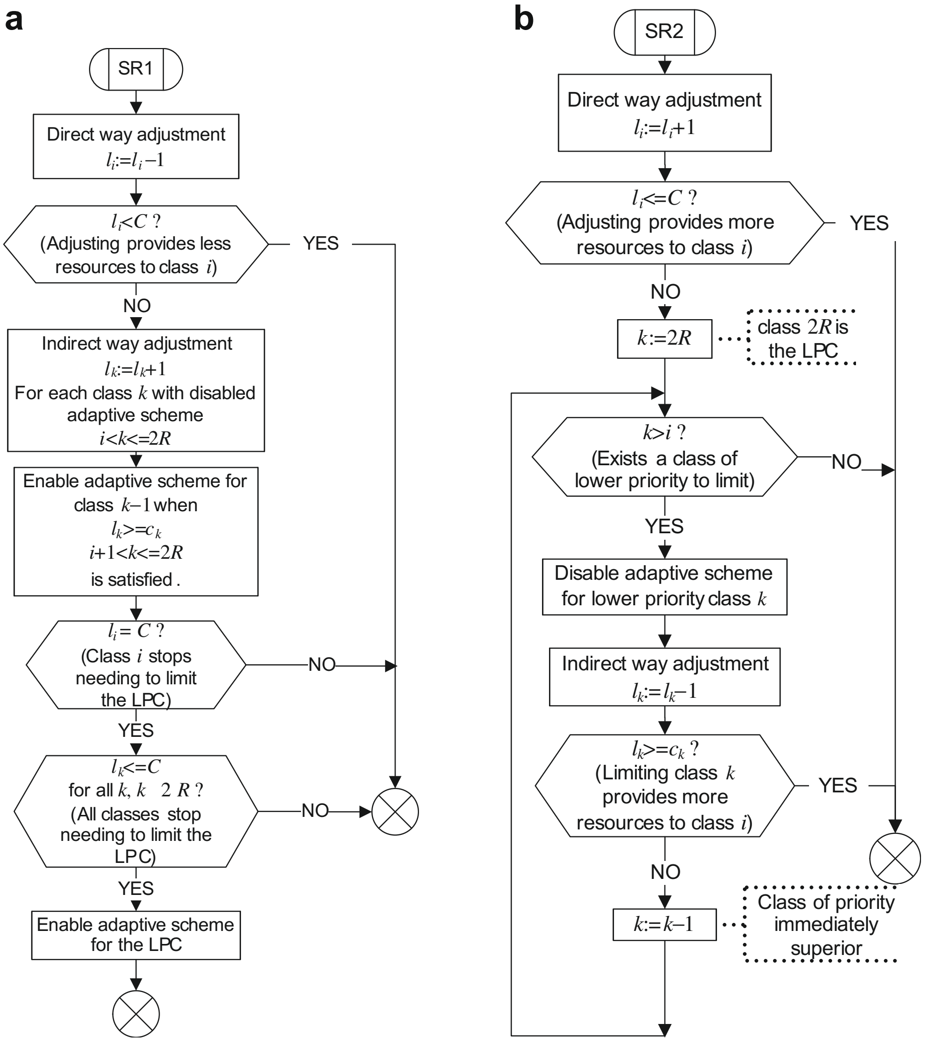
a
SR1
Direct way adjustment
$l_i:=l_i-1$
$l_i<C$?
(Adjusting provides less resources to class i)
YES
NO
Indirect way adjustment
$l_k:=l_k+1$
For each class k with disabled adaptive scheme
$i<k<=2R$
Enable adaptive scheme for class $k-1$ when
$l_k>=c_k$
$i+1<k<=2R$
is satisfied .
$l_i= C$?
(Class i stops needing to limit the LPC)
NO
YES
$l_k<=C$
for all k, k 2 R ?
(All classes stop needing to limit the LPC)
NO
YES
Enable adaptive scheme for the LPC
b
SR2
Direct way adjustment
$l_i:=l_i+1$
$l_i<=C$?
(Adjusting provides more resources to class i)
YES
NO
$k:=2R$
class $2R$ is the LPC
$k>i$?
(Exists a class of lower priority to limit)
NO
YES
Disable adaptive scheme for lower priority class k
Indirect way adjustment
$l_k:=l_k-1$
$l_k>=c_k$?
(Limiting class k provides more resources to class i)
YES
NO
$k:=k-1$
Class of priority immediately superior

Fig. 2. Adjustment algorithm of the adaptive scheme. (a) Adjustment algorithm after an admission decision. (b) Adjustment algorithm after a rejection decision.

is to increase the threshold l_i, but its maximum value is C, i.e. when $l_i = C$ full access to resources is provided to s_i and setting $l_i > C$ does not provide additional benefits. In these cases, an *indirect* way to help s_i is to limit the access to resources of lower-priority classes by reducing their associated thresholds. It is clear that when a higher priority class s_i needs to adjust the threshold of a lower-priority class s_j, then the adaptive scheme must adjust l_j only when arrivals from s_i occur, while no adjustments must be carried out when arrivals from s_j occur. When this happens, we say that the adaptive scheme associated to s_j is disabled. Therefore, when a protected class resorts to the indirect adjustment, this event is an indication that it is experiencing congestion. Note that although HPC arrivals are always accepted, l_1 is updated to detect when the HPC becomes congested.

3. Performance evaluation with streaming flows

In this section the performance of the adaptive scheme when handling streaming traffic is evaluated. We first describe the teletraffic model that is used for the evaluation. Next, we compare the performance of the proposed scheme with the performance of the schemes reported in [6,7], which use a similar probability-based adjustment scheme. The evaluation is performed in a single service scenario because this is the scenario for which these schemes were conceived. Finally, we evaluate the performance of our scheme in a multiservice scenario. The evaluation in the single service and multiservice scenarios is performed for both stationary and non-stationary traffic.

3.1. System model

We consider the homogeneous case where all cells are statistically identical and independent. Consequently the global performance of the system can be analyzed focusing on a single cell. Nevertheless, given that the proposed scheme is adaptive it could also be deployed in non-homogeneous scenarios.

For mathematical tractability we are making the common assumption of modeling the inter-arrival time of handover requests as an exponential distribution, which is considered a good approximation [32]. Therefore, new (handover) requests of service r arrive according to a Poisson process with rate λ_r^n (λ_r^h). Besides, although our scheme does not require any relationship between λ_r^h and λ_r^n, for simplicity we suppose that λ_r^h is a constant fraction of λ_r^n [33,29].

For a service-r session, both its duration and its cell residence (dwell) time are also assumed to be exponentially distributed with rates μ_r^s and μ_r^d. Hence, the resource holding time for a service r session in a cell is also exponentially distributed with rate $\mu_r = \mu_r^s + \mu_r^d$. We also study the impact on performance of having a resource holding time distribution different from the exponential one. Note that the proposed scheme can easily take into account terminals moving at different speeds by defining additional arrival classes for any service. Note also that the exponential assumption also represents a good approximation for the cell dwell time (essentially, only its average matters), when the performance of the system is defined in terms of the blocking probabilities [34]. It should be highlighted that the operation of our scheme is based on simple balance equations described in Section 2, which hold for any arrival process and holding time distribution. Hence the basis of the adaptive scheme holds beyond the assumptions made only for modeling purposes.

Finally, we denote by λ_{max} the system capacity, i.e. the maximum λ that can be offered to the system while meeting the QoS objectives, where λ is the aggregated arrival rate of new requests $\lambda = \sum_{r=1}^{R} \lambda_r^n$, $\lambda_r^n = f_r \lambda$ and $\sum_{r=1}^{R} f_r = 1$. Defining service penetrations (f_r) is a common approach when studying these systems [29]. As mentioned before, our scheme can adapt to any mix of traffic, i.e. to an aggregated traffic with any penetration factors f_i, even when these change with time.

We evaluate the performance of the proposed adaptive AC scheme by solving the continuous-time Markov chain (CTMC) that describes its operation, both in the stationary and transient regimes. When this is not possible we resort to simulation. In both regimes P_i is determined as the fraction of time an arrival request from s_i would be rejected.

In general, the system can be modeled by a multidimensional CTMC, where the state vector is given by $(n_1, \ldots, n_R, l_1, \ldots, l_{2R})$. Recall that n_r is the number of sessions in progress of service r in the cell initiated as new or handover requests and $l_i \in \mathbb{N}$ is the threshold associated with arrival class s_i. We allow l_i to take positive and negative values as a means to remember past adjustments and to identify the adjustment type the scheme uses (direct or indirect). Given that the general multidimensional diagram is difficult to represent, in Fig. 3 we show a bidimensional CTMC as an example. This system has only one service and therefore two arrival classes, s^h and s^n, with $d = 1$, C resource units and $\Delta l = 1$. It is assumed that s^h is the HPC and therefore its requests are always accepted (if free resources are available), while s^n is the BEC. The system state vector is defined as (n, l^h), where n is the number of resource units occupied. In this system, l^h is adjusted following the probabilistic adjustment rule described previously and $l^n = C - \max\{0, (l^h - C)\}$. Note that during underload episodes $l^n = C$, but during overload episodes s^h might have to resort to the indirect adjustment in which case l^n is decreased accordingly.

Tables 1 and 2 show the transition rates for a system in which the LPC is a protected class and for a system in which the LPC is the BEC, respectively, with $d = 1$ and $\Delta l = 1$. The state diagram of the second system is shown

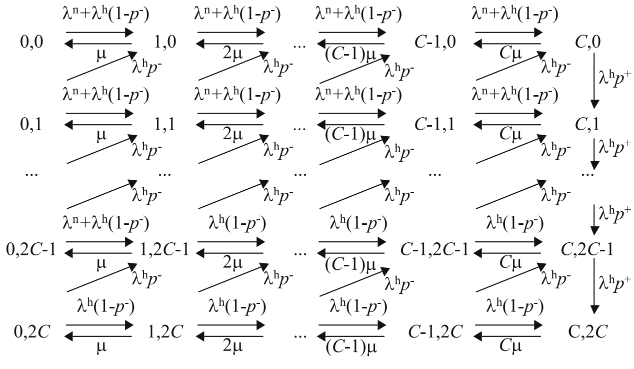

Fig. 3. State diagram of the CTMC in a scenario with two classes.

Table 1
LPC is a protected class. Current state is $\mathbf{x} = (n^n, n^h, l^n, l^h)$.

Next state	Transition rate
$(n^n + 1, n^h, l^n, l^h)$	$\lambda^n \cdot \alpha_n(\mathbf{x}) \cdot ((1 - p_n^-) \cdot \beta(\mathbf{x}) + (1 - \beta(\mathbf{x})))$
$(n^n + 1, n^h, l^n - \Delta l, l^h)$	$\lambda^n \cdot p_n^- \cdot \alpha_n(\mathbf{x}) \cdot \beta(\mathbf{x})$
$(n^n, n^h, l^n + \Delta l, l^h)$	$\lambda^n \cdot p_n^+ \cdot (1 - \alpha_n(\mathbf{x})) \cdot \beta(\mathbf{x})$
$(n^n, n^h + 1, l^n, l^h)$	$\lambda^h \cdot (1 - p_h^-) \cdot \alpha_h(\mathbf{x})$
$(n^n, n^h + 1, l^n, l^h - \Delta l)$	$\lambda^h \cdot p_h^- \cdot \alpha_h(\mathbf{x}) \cdot \beta(\mathbf{x})$
$(n^n, n^h, l^n, l^h + \Delta l)$	$\lambda^h \cdot p_h^+ \cdot (1 - \alpha_h(\mathbf{x})) \cdot \beta(\mathbf{x})$
$(n^n, n^h + 1, l^n + \Delta l, l^h - \Delta l)$	$\lambda^h \cdot p_h^- \cdot \alpha_h(\mathbf{x}) \cdot (1 - \beta(\mathbf{x}))$
$(n^n, n^h, l^n - \Delta l, l^h + \Delta l)$	$\lambda^h \cdot p_h^+ \cdot (1 - \alpha_h(\mathbf{x})) \cdot (1 - \beta(\mathbf{x}))$
$(n^n - 1, n^h, l^n, l^h)$	$n^n \mu$
$(n^n, n^h - 1, l^n, l^h)$	$n^h \mu$

Table 2
LPC is the BEC. Current state is $\mathbf{x} = (n^n, n^h, l^n, l^h)$.

Next state	Transition rate
$(n^n + 1, n^h, l^n, l^h)$	$\lambda^n \cdot \alpha_n(\mathbf{x})$
$(n^n, n^h + 1, l^n, l^h)$	$\lambda^h \cdot (1 - p_h^-) \cdot \alpha_h(\mathbf{x})$
$(n^n, n^h + 1, l^n, l^h - \Delta l)$	$\lambda^h \cdot p_h^- \cdot \alpha_h(\mathbf{x}) \cdot \beta(\mathbf{x})$
$(n^n, n^h + 1, l^n + \Delta l, l^h - \Delta l)$	$\lambda^h \cdot p_h^- \cdot \alpha_h(\mathbf{x}) \cdot (1 - \beta(\mathbf{x}))$
$(n^n, n^h, l^n, l^h + \Delta l)$	$\lambda^h \cdot p_h^+ \cdot (1 - \alpha_h(\mathbf{x})) \cdot \beta(\mathbf{x})$
$(n^n, n^h, l^n - \Delta l, l^h + \Delta l)$	$\lambda^h \cdot p_h^+ \cdot (1 - \alpha_h(\mathbf{x})) \cdot (1 - \beta(\mathbf{x}))$
$(n^n - 1, n^h, l^n, l^h)$	$n^n \mu$
$(n^n, n^h - 1, l^n, l^h)$	$n^h \mu$

in Fig. 3. Although the system state vector can be defined as (n, l^n, l^h) and (n, l^h), respectively, for the sake of readability we employ a more detailed description $((n^n, n^h, l^n, l^h))$ for both cases. If the QoS objectives for s^n and s^h are expressed as $B^n = b^n/o^n$ and $B^h = b^h/o^h$, then we define $p_n^- = 1/(o^n - b^n)$, $p_n^+ = 1/b^n$, $p_h^- = 1/(o^h - b^h)$ and $p_h^+ = 1/b^h$. Note that in the system of Fig. 3 only s^h has a QoS objective defined, therefore $p^- = 1/(o^h - b^h)$ and $p^+ = 1/b^h$.

We also define the following indicator functions:

$$\alpha_n(\mathbf{x}) = \begin{cases} 1 & (n^n + n^h < C) \cap (n^n + n^h < l^n), \\ 0 & (n^n + n^h = C) \cup (n^n + n^h \geqslant l^n), \end{cases}$$

$$\alpha_h(\mathbf{x}) = \begin{cases} 1 & (n^n + n^h < C), \\ 0 & (n^n + n^h = C), \end{cases}$$

$$\beta(\mathbf{x}) = \begin{cases} 1 & l^h \leqslant C, \\ 0 & l^h > C. \end{cases}$$

3.2. Comparative performance evaluation

From now on we will refer to the scheme in [6] as ZL and to the one in [7] as WZZZ, after their authors' initials. In these schemes, as in ours, incoming handover requests are always accepted provided that there are enough free resource units available in the system.

The ZL scheme has four parameters: α_u, α_d, N and τ. It operates as follows: (i) if after a blocked handover request it is detected that $P^h \geqslant \alpha_u B^h$, then l^n is updated as $l^n \leftarrow (l^n - 1)$; (ii) if for N consecutive handover requests it is found that $P^h \leqslant \alpha_d B^h$, then l^n is updated as $l^n \leftarrow (l^n + 1)$. This scheme estimates P^h during one update period of fixed length τ. As in [7], for both the ZL and WZZZ schemes it will be assumed that $\tau \to \infty$ (i.e. the estimated P^h is the handover blocking probability experienced so far). This choice is motivated by the fact that the ZL scheme ambiguously defines how P^h is estimated. Additionally, it was also found that making $\alpha_u = \alpha_d = 1.0$ instead of the values proposed by the authors allows the ZL scheme to reach a steady state regime ($P^h = B^h$) and minimizes oscillations.

In order to reduce the number of parameters, improve the adaptability of the system to different traffic profiles, and improve the response time of the ZL scheme, a new probability-based adaptive scheme was proposed in [7]. The WZZZ scheme defines three parameters: α_u, α_d and P_{inc} (probability to increase the number of guard channels, i.e. to decrease l^n). The WZZZ scheme performs probabilistic adjustments only for each blocked handover request, producing a convergence rate slower than the one achieved by the ZL scheme. Our comparative study showed that, as with the ZL scheme, using $\alpha_u = \alpha_d = 1.0$ instead of the values proposed by the authors is better. We deployed the suggested value $P_{inc} = 0.2$, although we found that $P_{inc} = 1.0$ is more appropriate.

In the following sections, the performance of the ZL and WZZZ schemes is evaluated by simulation, as they require to estimate P^h online what precludes a CTMC-based evaluation approach. On the other hand, ours is evaluated by solving the CTMC that describes its operation. Additionally, the correctness of the CTMC model has been validated against computer simulation.

3.2.1. Stationary regime

The evaluation of the three schemes is done for the scenario described in [6,7], which is summarized as follows: $C = 50$, $B^h = 1\%$, $\lambda^h = 0.2\lambda^n$ sessions/s, and $\mu = 1/180\ \text{s}^{-1}$. We focus on a load range that allows analyzing the schemes in underload and overload conditions. An extensive comparative evaluation was conducted but only some relevant experiments are discussed here. A value for the adjustment step of $\Delta l = 1$ is assumed, unless otherwise specified. For the simulations, we used a confidence interval of $\pm 5\%$ around the mean estimate and a confidence level of 95%.

Fig. 4 shows that our scheme carries more traffic in the load region of interest, which is due to a more precise management of the guard channels. Fig. 5 shows the variation of the mean number of guard channels required to meet the QoS objective for different new session arrival rates. When the LPC is a protected class with a QoS objective $B^n = 10\%$, during underload episodes ($\lambda^n < 0.2$) the system deploys the direct adjustment mode and adapts to enforce $P^n = B^n$. Therefore it rejects more new requests than if it allowed $P^n \leqslant B^n$. On the positive side, setting l^n to a value lower than required increases the amount of resources that the HPC has access to, achieving a lower P^h as observed in Fig. 6. Furthermore, elastic traffic which is carried along with streaming traffic in modern cellular networks will benefit from that spare capacity. This improvement would be achieved while still meeting the QoS objective of both the HPC and the LPC.

As load increases, the adjustment changes from the direct to the indirect mode. In this mode the scheme adjusts l^n to meet the QoS objective of the s^h, which converts the

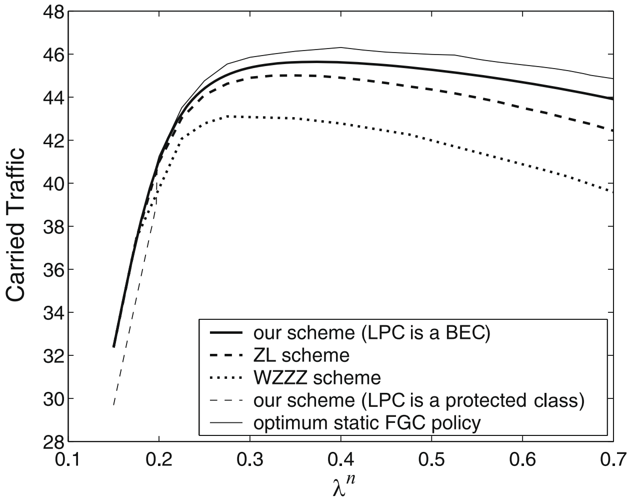

Fig. 4. Carried traffic with stationary traffic.

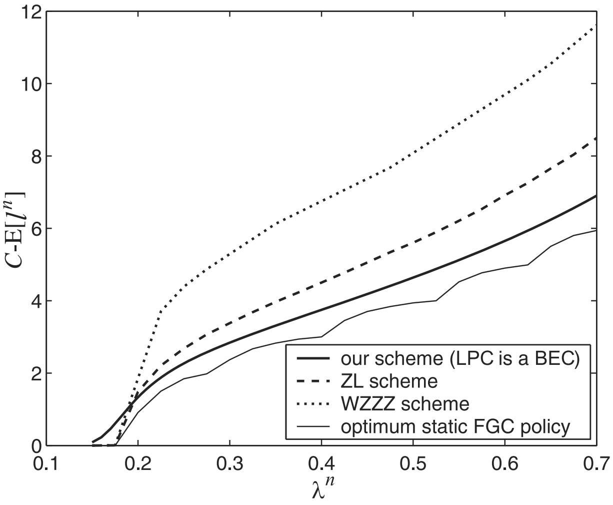

Fig. 5. Mean number of guard channels required.

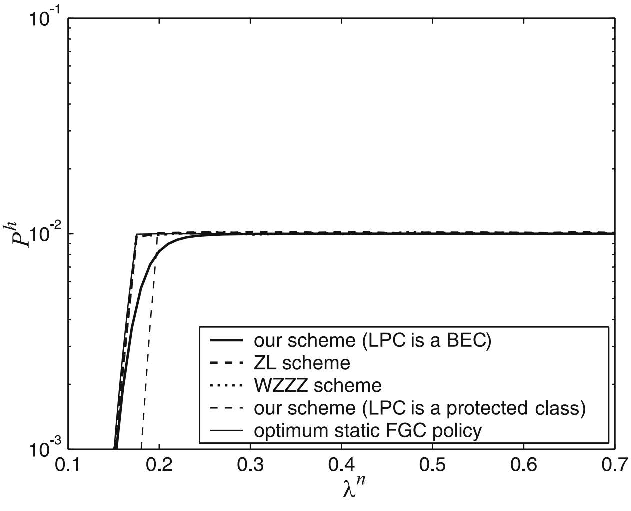

Fig. 6. Handover blocking probability with stationary traffic.

LPC from a protected class in the BEC. In summary, the capability of our scheme to operate in two different modes provides the operator with additional flexibility to manage the QoS objective.

In Figs. 4–6 the performance of the different adaptive schemes is compared to the performance of an optimum static single fractional guard channel (FGC) policy. The single FGC policy defines two configuration parameters $l^n \in \mathbb{N}$ and $q^n \in [0,1]$. Let us denote by n^n (n^h) the number of sessions in progress in the cell initiated as new (handover) requests and by $n = n^n + n^h$ the number of resource units occupied. When a new arrival happens, it is accepted with probability one if $n < l^n$, it is accepted with probability q^n if $n = l^n$ and blocked otherwise. Conversely, handover arrivals are always accepted. It has been shown that the system capacity achieved by the FGC policy is very close to the capacity achieved by an optimal policy [15,35].

Note that the evaluation has been done considering that the arrival processes are stationary. In this context, the configuration parameters of the single FGC policy have been determined by formulating the problem as a non-linear programming algorithm in which for each new session arrival rate we search for the value of the configuration parameters that maximize the carried traffic subject to the fulfillment of the QoS objective [15]. Therefore we refer to this policy as the *optimum* single FGC policy. We also refer to it as *static* because for each value of the arrival rate studied we determine the optimum configuration parameters. On the other hand, the adaptive schemes does not know the arrival rates a priori and therefore they continuously change the threshold of the single GC policy to meet the QoS objective, which limits the resource utilization that can be achieved.

3.2.2. Non-stationary regime

Unless otherwise stated, in this section we evaluate the transient regime of the proposed scheme in a scenario characterized by being the system initially in the steady state regime, i.e. empty, with $l^h = l^n = C$ and where the LPC is the BEC. We first evaluate the transient regime of P^h by applying a step-increase of traffic from $\lambda^n = 0$ to $\lambda^n = 0.333$. As observed in Fig. 7, our scheme achieves the fastest convergence rate. The ZL scheme shows a slow oscillating behavior around B^h, taking 30,000 s to reduce P^h to a $\pm 10\%$ interval around its objective ($B^h = 0.01$). Ours needs only 3400 s, about ten times less, to achieve the same operating conditions. Note that the WZZZ scheme oscillates even slower than the ZL scheme.

Note that previous scenario is unlikely to occur in real networks but provides important information of the transient regime of the system. To provide additional information, this time in a more realistic scenario, we study the transient behavior of our scheme after a step-type increase in the λ^h/λ^n ratio from 0.2 to 0.4, maintaining $\lambda^n = 0.417$ (25 sessions/min). As above, the system is in the steady state regime before the step-increase is applied. As the WZZZ scheme has not a very competitive convergence rate it has not been included in this study. Fig. 8 shows the transient behavior of P^h using our scheme when considering the LPC as a BEC and the ZL scheme with $N = 1$. Again our scheme outperforms the ZL scheme in terms of convergence rate and stability.

In summary, the ZL scheme requires tuning a series of configuration parameters for each scenario, while ours

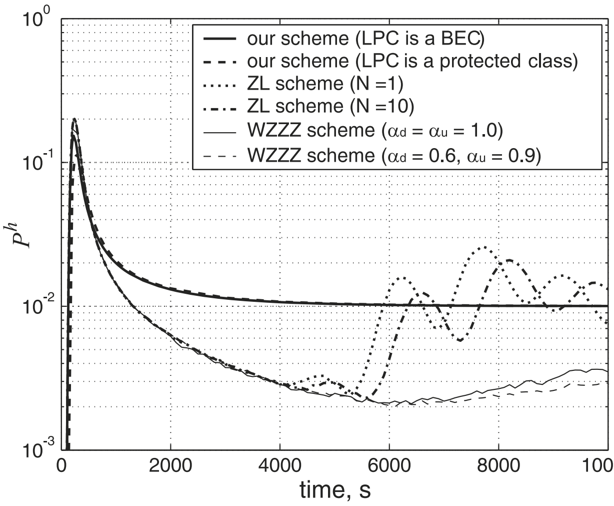

Fig. 7. Step-increase of traffic from $\lambda^n = 0$ to $\lambda^n = 0.333$.

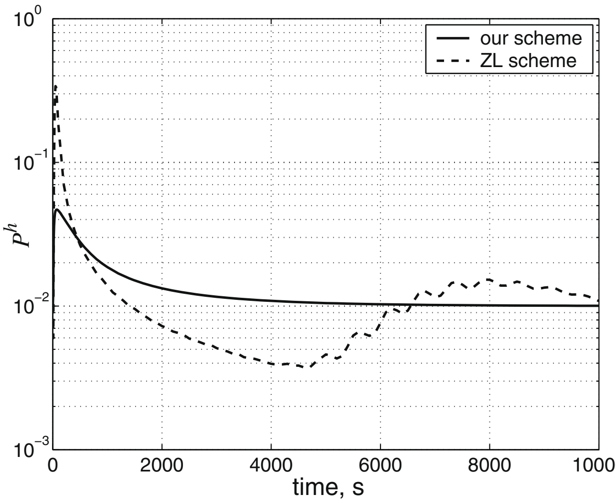

Fig. 8. Step-increase of traffic from $\lambda^h/\lambda^n = 0.2$ to $\lambda^h/\lambda^n = 0.4$, with $\lambda^n = 0.417$.

does not. Although the average resource utilization is close to the one achieved by our scheme, the transient performance of the ZL scheme is rather poor when compared to ours. Additionally, the ZL scheme was designed to support only streaming traffic in single service scenarios, while ours is designed to handle multiple streaming services, it incorporates a rate-adaptation mechanism and also handles elastic traffic.

3.3. Additional features

In this section we discuss in more detail other features of our scheme like achieving a faster convergence rate and its ability to adapt to any arrival and holding time distributions. Previous results assumed that the objective was expressed as an irreducible fraction, for example $B^h = b^h/o^h = 1/100$. Any other multiple of the irreducible fraction might have been chosen as well, i.e. $B^h = kb^h/ko^h$, $k \geqslant 1$. Fig. 9 shows that any value of k allows our scheme to eventually fulfill the QoS objective, but the convergence rate of P^h decreases as k increases. Likewise, Fig. 10 shows that the convergence rate of P^n also decreases as k increases. Therefore, $k = 1$ confirms itself as the most suitable value in terms of performance.

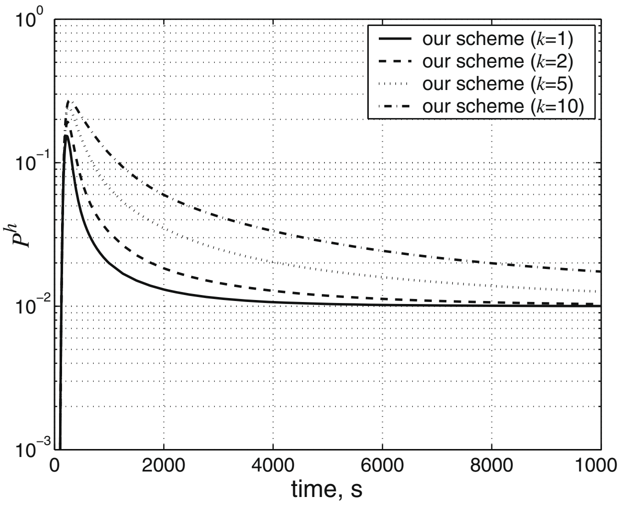

Fig. 9. Dependence of the convergence rate of P^h with k.

Up to now, it had been assumed an adjustment step value of $\Delta l = 1$, but Fig. 11 shows that as Δl increases, the convergence rate of the scheme increases as well. As observed in Fig. 12, there is a trade-off between the convergence rate of P^h and the penalty experienced by P^n, but step values below or equal to $\Delta l = 5$ guarantee a remarkable gain in convergence rate and a minor impact on the value observed for P^n. As seen, the scheme is able to adapt to the new operating conditions in less than 700 s. It is clear that the convergence time would become shorter as the offered load increases, because adjustment events (i.e. arrivals) will occur at a higher rate. Finally, Fig. 13 shows that the distribution of the resource holding time has only a minor impact on convergence rate of the

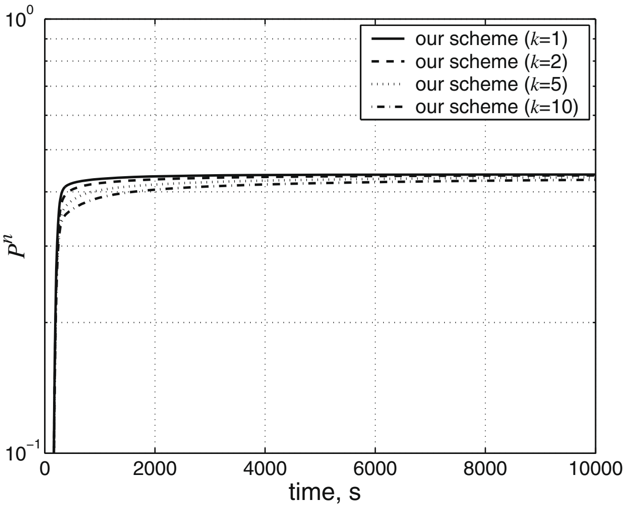

Fig. 10. Convergence rate dependence of P^n with k.

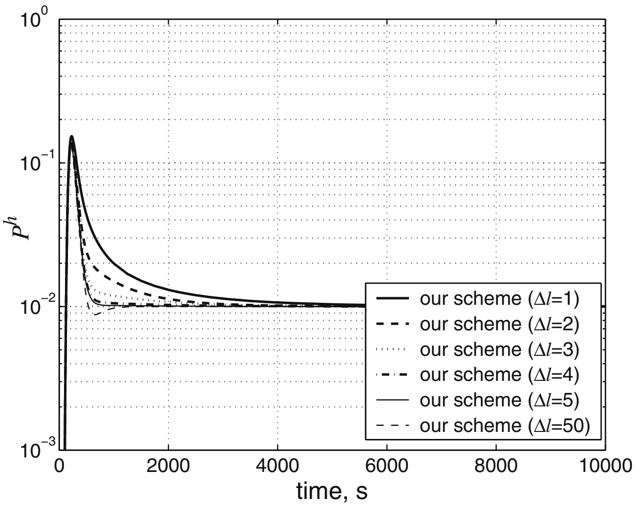

Fig. 11. Convergence rate dependence of P^h with Δl.

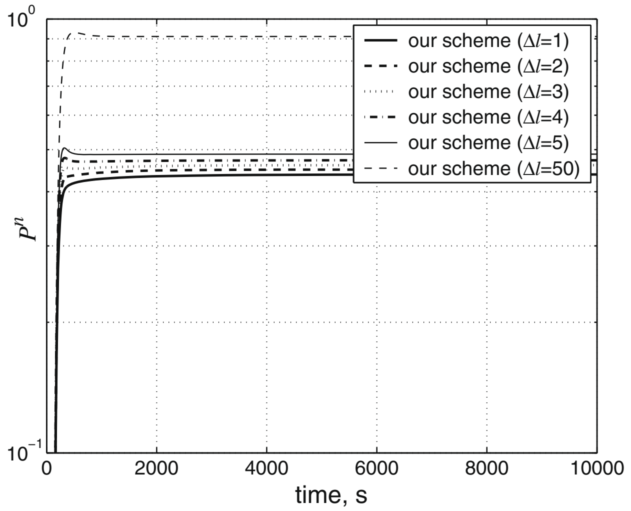

Fig. 12. Convergence rate dependence of P^n with Δl.

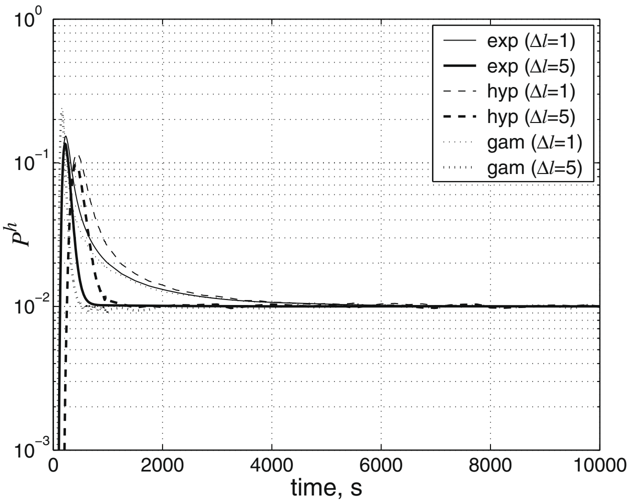

Fig. 13. Convergence rate dependence of P^h with the coefficient of variation of the distribution of the resource holding time.

scheme. The gamma distribution has a pdf given by $f(x) = (\lambda e^{-\lambda x}(\lambda x)^{\alpha-1})/\Gamma(\alpha), x \geqslant 0, \alpha > 0, \lambda > 0, \Gamma(\alpha) = \int_0^\infty e^{-y}y^{\alpha-1}\,dy$, with mean α/λ and variance α/λ^2. We deployed $\alpha = 2.9370$ and $1/\lambda = 61.2868$ to obtain a mean of 180 and a coefficient of variation of $0.5835 \approx 1/\sqrt{3}$. The hyper-exponential distribution has a pdf $f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}$, $x \geqslant 0$, $p_2 = 1 - p_1$, with mean $p_1/\lambda_1 + p_2/\lambda_2$ and variance $2p_1/\lambda_1^2 + 2p_2/\lambda_2^2 - (p_1/\lambda_1 + p_2/\lambda_2)^2$. We deployed $1/\lambda_1 = 357.1429$, $1/\lambda_2 = 2.8571$ and $p_1 = p_2 = 0.5$ to obtain a mean of 180 and a coefficient of variation of $1.7138 \approx \sqrt{3}$.

Given that in real operation it is not expected that the conditions change in a step-like way, we believe that our scheme handles satisfactorily the non-stationarity of real networks.

3.4. Performance evaluation in a multiservice scenario

The performance evaluation is carried out for five different scenarios $\{A, B, C, D, E\}$ that are defined in Table 3, being the QoS parameters B_i expressed as percentage values. The parameters in Table 3 have been selected to explore possible trends in the numerical results, i.e., taking scenario A as a reference, scenario B represents the case where the ratio d_1/d_2 is smaller, scenario C where f_1/f_2 is smaller, scenario D where B_1/B_2 is smaller and scenario E where B_1 and B_2 are equal.

When deploying the MGC policy without the adaptive scheme, the system capacities for the five scenarios defined in Table 3, $\{A, B, C, D, E\}$, with $C = 10$ are $\lambda_{max} = \{1.89, 0.40, 1.52, 1.97, 1.74\}$, respectively. Refer to García et al. [15] for details on how to determine the system capacity. For all scenarios defined in Table 3 we assume the following prioritization order $(s_2^h, s_1^h, s_2^n, s_1^n)$. We evaluate two implementations that differ in the treatment of the LPC (s_1^n), one in which it is a protected class and one in which it is the BEC.

3.4.1. Stationary regime

For the two implementations of the adaptive scheme, Table 4 shows the ratio P_i/B_i for the four arrival classes in the five scenarios considered. In all cases, an aggregated load equal to the system capacity (λ_{max}) is offered. Note that the adjustment is more precise when the LPC is the BEC.

Figs. 14 and 15 show the variation of P_i with the relative offered load $((\lambda - \lambda_{max})/\lambda_{max})$ in scenario C with $C = 10$ resource units. Note that the adaptive scheme tries to enforce $P_i = B_i$ when possible for the protected classes, and therefore during underload episodes the system is rejecting more requests than strictly required. Nevertheless, some classes (BEC and/or HPC) benefit from this extra capacity. When the LPC is a protected class (Fig. 14) it does not benefit from the capacity surplus during underload episodes and it is the first to be penalized during overload episodes. On the other hand, when the LPC is the BEC (Fig. 15) it benefits during underload episodes and, as before, it is the first to be penalized during overload episodes. In both implementations, note that s_2^n is also penalized when keeping on reducing l_1^n (bellow zero) would be ineffective to meet the QoS objective of higher priority classes.

Table 3
Definition of the scenarios under study.

	d_1	d_2	f_1	f_2	$B_1^n(\%)$	$B_2^n(\%)$	$B_r^h(\%)$	λ_r^n	λ_r^h	μ_1	μ_2
A	1	2	0.8	0.2	5	1	$0.1B_r^n$	$f_r\lambda$	$0.5\lambda_r^n$	1	3
B	1	4	0.8	0.2	5	1					
C	1	2	0.2	0.8	5	1					
D	1	2	0.8	0.2	1	2					
E	1	2	0.8	0.2	1	1					

Table 4
P_i/B_i when deploying the MGC policy and a stationary load equal to λ_{max}.

P_i/B_i	Scenario				
	A	B	C	D	E
(a) LPC is a protected class					
Class 1N	1.004	1.030	1.036	1.841	1.223
Class 2N	0.998	0.992	1.001	0.998	1.007
Class 1H	1.006	0.992	1.002	1.007	0.999
Class 2H	0.848	0.899	0.803	0.988	0.985
(b) LPC is the best-effort class					
Class 1N	0.938	1.404	0.007	2.348	1.857
Class 2N	1.003	1.065	1.000	1.004	0.999
Class 1H	1.007	1.001	1.007	0.999	0.999
Class 2H	0.993	1.006	0.988	0.989	0.999

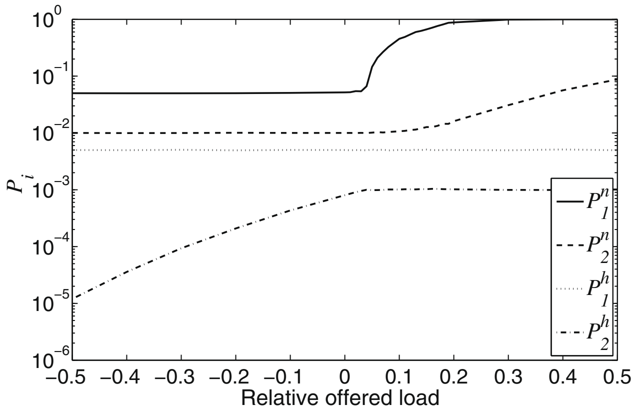

Fig. 14. Variation of P_i with the relative offered load in stationary conditions when the LPC is a protected class.

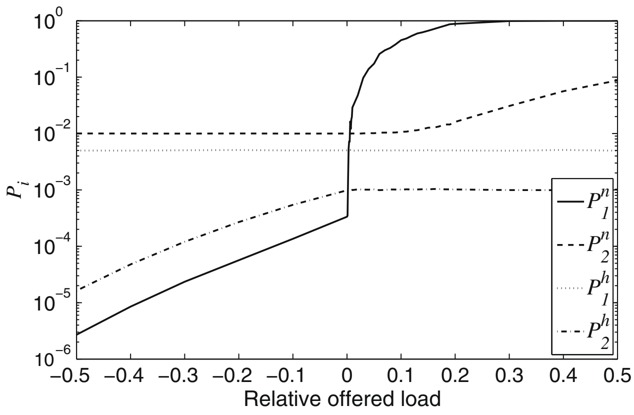

Fig. 15. Variation of P_i with the relative offered load in stationary conditions when the LPC is the BEC.

In Fig. 16 the resource utilization factor $E[c(\mathbf{n})]/C$ of the adaptive scheme in scenario A is compared to the one of an optimum static MFGC policy, whose performance is close to the performance of an optimal policy [15]. As for the single FGC, the configuration parameters for the MFGC policy have been determined by formulating the problem as a non-linear programming algorithm in which for each λ we search for the values of the configuration parameters that maximize the carried traffic subject to the fulfillment of the QoS objective. The term "adapt. MGC" refers to the adaptive scheme when the LPC is a protected class, while "adapt. MGC–BEC" refers to the adaptive scheme when the LPC is the BEC. Note that for $\lambda = \lambda_{max}$ the utilization achieved by the optimum static MFGC policy is only 2% higher than the utilization achieved by the adaptive scheme. Note that both implementations of the adaptive scheme behave identically in overload ($\lambda > \lambda_{max}$).

3.4.2. Non-stationary regime

We study the transient regime after a step-type traffic increase from $0.66\lambda_{max}$ to λ_{max} is applied to the system in scenario A when the LPC is a protected class. Before the step-increase is applied the system is in the steady state regime. Fig. 17 shows that the scheme converges rapidly and in an oscillation-free manner to the new operating conditions. As mentioned before, note that the convergence rate increases with the offered load, as adjustment events (i.e. arrivals) occur at a higher rate.

4. Rate-adaptation policy

The RA policy takes advantage of the adaptivity of multimedia applications to limit the blocking probability of

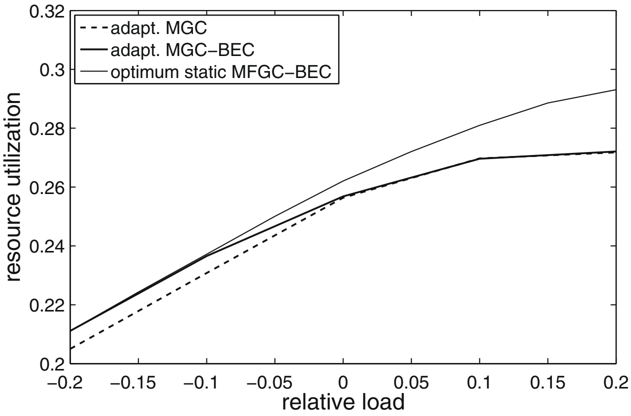

Fig. 16. Resource utilization factor.

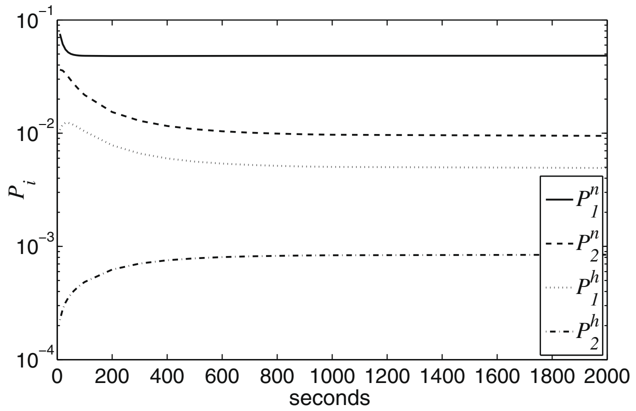

Fig. 17. Transient behavior of blocking probabilities.

new and handover requests even during overload episodes, while minimizing the frequency of rate adaptations. The operation of the RA policy differs considerably from the operation of common policies based on the full sharing principle where, in overload, the rate of ongoing sessions is adapted downwards with each new or handover request that arrives to the system and upwards after each departure. The proposed RA policy makes use of the early congestion detection mechanism provided by the threshold adaptation scheme to request rate adaptations only when sessions join a cell. That is, rate adaptations are never exerted on ongoing sessions but are requested at session initiation or when sessions are handed over from a neighboring cell, minimizing in this manner the adaptation frequency and the signaling load.

Its operation is based on the same principles described for the adaptive AC scheme. Recall that when any arrival class s_i sets its threshold above C, this is an indication that it requires more resources to maintain its blocking probability objective. Then, an *indirect* way to help s_i is to degrade arriving sessions to a lower rate, making in this manner available the required additional resources. We assume that it is more disturbing for a subscriber to have a new or handover session request dropped than having its rate adjusted downwards, producing a degraded but still acceptable playback experience. Therefore, when an arrival class is experiencing congestion then, lower, equal and higher priority classes (in this order) can be degraded to control the overload episode. The prioritization order for degradation can be freely defined by the operator, but for simplicity we assume the same one defined for the AC scheme. If the complete degradation of all classes is not sufficient, then the *indirect* mechanism of the AC scheme is activated to guarantee that higher priority classes will be able to meet their QoS objective, possibly at the expense of lower priority ones.

We denote by $\{d_{r1},\ldots,d_{rM_r}\}$ the acceptable rate values for service-r sessions, ordered in decreasing value. Likewise, $\{c_{i1},\ldots,c_{iM_i}\}$ are the acceptable rate values for sessions of the arrival class s_i, where $d_{rm}=c_{rm}=c_{(r+R)m}$, $1\leqslant r\leqslant R$ and $1\leqslant m\leqslant M_r$. We denote by $(\rho_{1m_1},\ldots,\rho_{2Rm_{2R}})$ the current rate vector of the RA policy, being ρ_{im_i} the rate at which arriving sessions of s_i must operate. At a given operation point of the policy the RA vector must fulfill the following condition, $\exists i:\forall j>i\ \rho_{jm_j}=c_{jM_j}$ (lowest rate), $\forall k<i\ \rho_{km_k}=c_{k1}$ (highest rate) and ρ_{im_i} takes decreasing (increasing) values in $\{c_{i1},\ldots,c_{iM_i}\}$ as congestion increases (decreases), $1\leqslant i,j,k\leqslant 2R$. In other words, as congestion increases the RA policy requests rate degradations first to the lowest-priority arrival class and, if congestion persists, then to the other classes in inverse order of priority.

In order to achieve a graceful degradation and at the same time avoiding to issue more degradation requests than required, we define degradation probabilities associated to the RA vector $(p_{1n_1},\ldots,p_{2Rn_{2R}})$, $1\leqslant n_i\leqslant N_i$, $p_{in_i}=n_i/N_i$. Then, for a rate vector $(\rho_{1m_1},\ldots,\rho_{2Rm_{2R}})$, an arriving session of s_i is requested to operate at rate ρ_{im_i} with probability p_{in_i} and at the rate immediately lower $\rho_{i(m_i+1)}$ with probability $1-p_{in_i}$. When $\rho_{im_i}=c_{iM_i}$ (lowest rate) then, arriving sessions of s_i are requested to operate at rate c_{iM_i} with probability 1, while arriving sessions of $s_j, j=i-1$, are requested to operate at rate c_{j1} with probability p_{jn_j} and at rate c_{j2} with probability $1-p_{jn_j}$. As congestion increases, the degradation probability p_{in_i} takes decreasing values in $\{1,(N_i-1)/N_i,\ldots,1/N_i\}$. Allowing different values of N_i for each arrival class, makes it possible for the operator to configure with more precision the degradation degree. We define the state of the RA policy by two $2R$-tuples, the rate and the degradation probability vectors.

As an example, assume that $2R=4$, $M_i=2$ and $N_i=2\ \forall i$. As congestion increases, the evolution of the RA policy state would be: $\{(c_{11},c_{21},c_{31},c_{41}),\ (1,1,1,1)\}$, $\{(c_{11},c_{21},c_{31},c_{41}),\ (1,1,1,0.5)\}$, $\{(c_{11},c_{21},c_{31},c_{42}),\ (1,1,1,1)\}$, $\{(c_{11},c_{21},c_{31},c_{42}),\ (1,1,0.5,1)\}$, $\{(c_{11},c_{21},c_{32},c_{42}),(1,1,1,1)\}$, and so on. As observed, when congestion increases, first the fraction of arrivals that are requested to degrade their rate increases and, if congestion persists, then the rate is set one step lower. Note that we can perceive the states of the RA policy as ordered in terms of degradation degree, like in the example. It is clear that the evolution of the RA policy state must depend on the evolution of congestion, the more congested is the system the more severe is the degradation degree. We have associated the state of the RA policy to the value of the thresholds maintained by the AC policy in the following way. The RA policy starts degrading arrivals when the threshold of any arrival class is above C, while no rate adaptation occurs when the value of the thresholds is equal to C or lower. When the threshold of any class whose value is above C increases (decreases) by one, then the RA policy moves one step upwards (downwards) the degradation degree.

The performance evaluation is carried out by simulation for the same five different scenarios defined in Table 3, although only the results for scenario B with $C=50$ are shown. Scenario B has been slightly modified to handle rate degradation. For simplicity, now $d_1=2$ instead of $d_1=1$ as it is shown in Table 3. The system capacity for the new scenario B is $\lambda_{max}=10.3$. Recall that this is the maximum arrival rate that can be offered to a system that deploys the MGC policy (with the RA policy and the adaptive AC policy both disabled) while still meeting the blocking objectives. The acceptable rate values for service 1 sessions are $\{d_{11}=2,d_{12}=1\}$, while for service 2 are

$\{d_{21} = 4, d_{22} = 2\}$, then $M_i = 2\ \forall i$. When the RA policy is operating, we assume the following prioritization order $(s_2^h, s_1^h, s_2^n, s_1^n)$. Then, when any arrival class is experiencing congestion it will request rate adaptation to arriving sessions first of s_1^n, next to s_2^n and so on. The degradation probabilities associated to all arrival classes have been made identical, i.e. they take decreasing values in $\{1, 0.5\}$ i.e. $N_i = 2\ \forall i$. We only show results for the implementation in which the LPC (s_1^n) is a protected class. In the simulations, we used a confidence interval of $\pm 5\%$ around the sample mean and a confidence level of 95%.

Fig. 18 shows the variation of the fraction of the total traffic carried by each arrival class that is carried by non-degraded and degraded sessions (those with a cross symbol). Recall that the relative offered load was defined as $(\lambda - \lambda_{max})/\lambda_{max}$. As observed, when the system experiences congestion, the LPC is the first class that is requested to degrade its rate. If the congestion persists, then the rest of classes are requested to degrade their rates in inverse order of priority. Fig. 19 shows the variation of blocking probabilities perceived by each arrival class. Note that a system that enforces the RA policy is able to manage a much higher arrival rate while still meeting the blocking objectives than a system that does not. The maximum traffic that can be carried by the system while still meeting the blocking objectives and the aggregated arrival rate at which this happens are: $(27.75, \lambda = 10.30)$ in a system without RA policy and $(32.96, \lambda = 23.69)$ in a system with RA policy. This is clearly due to the fact that the system requests rate adaptation to incoming sessions as soon as it perceives congestion and therefore accepts more requests. Note also that, while possible, the blocking objectives are met with high precision. That is, when the arrival rate increases above $\lambda = 23.69$ the AC scheme resorts to the indirect adjustment mechanism and the blocking objective of the LPC is no longer achieved.

From the operational point of view, the RA scheme runs independently at each cell, and therefore it has only a limited visibility of the complete life of sessions. Our scheme allows to couple the operation of the RA policy with the operation of a management entity that keeps track of the history of sessions. This might be required in high mobility scenarios where handover rate is high, like in geographical areas serviced by picocells. In this sense, the RA mechanism can be seen as an entity that provides *rate-adaptation advices* that may be followed or not by the system. Note that not following the rate-adaptation advices would force the RA policy to move toward more severe degradation, likely affecting to higher priority arrivals, or even it might trigger the activation of the indirect adaptation way of the AC scheme, increasing in this manner the losses of low priority arrivals. In conclusion, the RA scheme designed provides operators with a great flexibility in configuring their mobile services, allowing to offer differentiated perceptual QoS to different services.

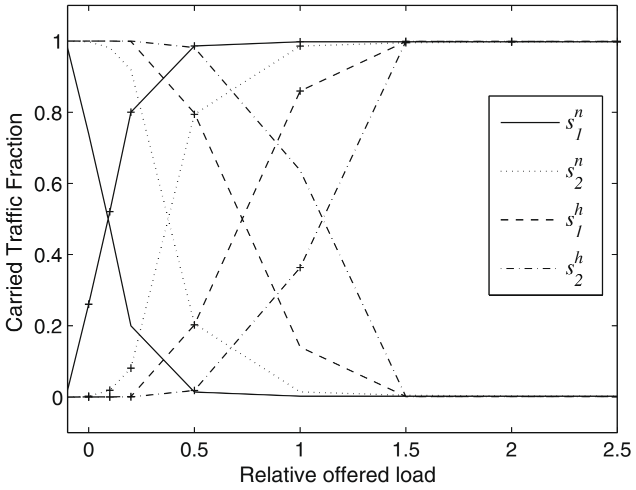

Fig. 18. Fraction of the carried traffic of each class associated to non-degraded and degraded sessions.

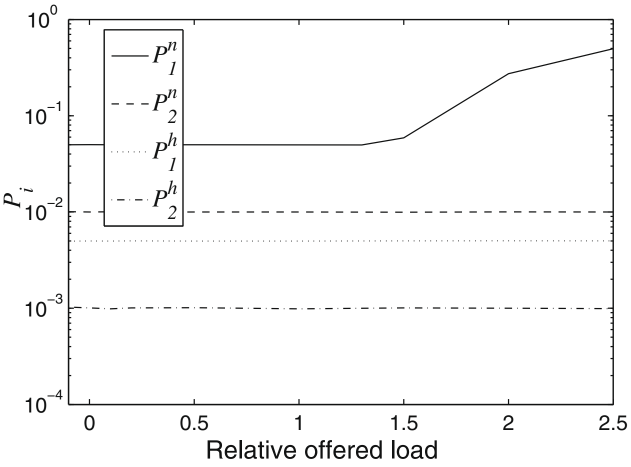

Fig. 19. Variation of P_i with the relative offered load.

5. Adaptive scheme for elastic flows

Like in other studies of the same nature, we focus on the flow level and ignore the detailed mechanisms operating at the packet level [5]. Since our focus is the radio interface at the access network, we assume that each elastic flow is rate limited either by terminal capabilities or because it is bottlenecked at the radio link, i.e. it will receive its fair share of the radio link bandwidth up to a maximum which has a common value for all terminals. Note that if different rate limits are possible for different terminals, or multiple abandonment probability objectives are defined, the scheme can handle this scenario by defining multiple elastic flow types. For the sake of mathematical tractability we assume that the flow size (given in bytes) is exponentially distributed. While it is commonly accepted that the statistical distribution of Internet document sizes shows a greater variability than the exponential distribution, in the light of the results in [5] the numerical results obtained by using an exponential document size can be considered as a lower bound of performance.

We consider the same system model described in Section 3.1, adding a service with elastic demands as follows. We denote by s_e^n (s_e^h), the arrival class associated to new (handover) requests of elastic flows. Their requests arrive according to a Poisson processes with rate λ_e^n (λ_e^h). For an elastic session, its cell residence (dwell) time is exponentially distributed with rate μ_e^d. If we denote by d_e the maximum number of resource units an elastic flow uses, and by n_e the number of elastic flows in the system, then we define the flow service rate as μ_e^s when $n_e d_e \leqslant (C - c(\boldsymbol{n}))$ and $\mu_e^s(C - c(\boldsymbol{n}))/(n_e d_e)$ when $n_e d_e > (C - c(\boldsymbol{n}))$, where $c(\boldsymbol{n})$ is the number of resource units occupied by streaming sessions.

To model the behavior of users we consider the impatience time as an independent exponentially distributed random variable. We assume that the impatience rate μ_I is equal to zero when enough free resources are available for the elastic traffic (i.e. $n_e d_e \leqslant (C - c(\boldsymbol{n}))$). Otherwise, the impatience rate is made inversely proportional to the share of resources allocated to each elastic flow. Thus, we define $\mu_I = K(n_e d_e/(C - c(\boldsymbol{n})))$ when $n_e d_e > (C - c(\boldsymbol{n}))$, where K is a constant. We denote by BA the QoS objective expressed as an upper bound for the abandonment probability, i.e. the ratio between unsuccessfully completed flows and accepted flows, and by PA the actual perceived abandonment probability.

The AC policy for the elastic service defines only one threshold $l_e \in \mathbb{N}$, associated with s_e^n. When there are n_e ongoing elastic flows, a new request is accepted if $n_e < l_e$ and blocked otherwise. Handover requests of elastic flows are always accepted. The adaptive scheme for elastic flows follows a similar approach to that described in Section 2. When the QoS objective for elastic flows can be expressed as $BA = a/b$, where $a, b \in \mathbb{N}$, then we propose to perform a probabilistic adjustment in the following way: (i) each time an elastic flow abandons due to impatience, do $\{l_e \leftarrow (l_e - 1)\}$ with probability $1/a$; (ii) each time an elastic flows completes its service successfully, i.e. either it finishes or it hands over to another cell, do $\{l_e \leftarrow (l_e + 1)\}$ with probability $1/(b - a)$. A methodology to infer TCP flow interruption has been proposed in [36].

We evaluate the performance of the scheme by simulation. We consider scenario A with $C = 10$, where the LPC is the BEC and where streaming traffic and elastic flows compete for network resources. The streaming traffic offers a constant load equal to the system capacity ($\lambda = \lambda_{max} = 1.89$). To avoid starvation the system reserves 1 resource unit for elastic traffic. The values of the other parameters that model the elastic traffic are: $\mu_e^s = 2.0$, $\mu_e^d = 2.0$, $K = 0.4$, $\lambda_e^h = 0.5\lambda_e^n$, with a QoS objective of $BA = 0.1$. Fig. 20 shows that the adaptive scheme guarantees the QoS objective. Without the adaptive scheme the abandonment probability increases as the elastic arrival rate increases. This is due to the fact that less resources are available per elastic flow as more elastic flows are accepted in the system, which consequently increases the abandonment rate. Finally, note that high abandonment probabilities bring as a consequence an inefficient use of system resources because resources assigned to flows that are not completed are totally wasted. The performance seen by streaming flows is not shown here as it is unaffected by the existence of elastic traffic.

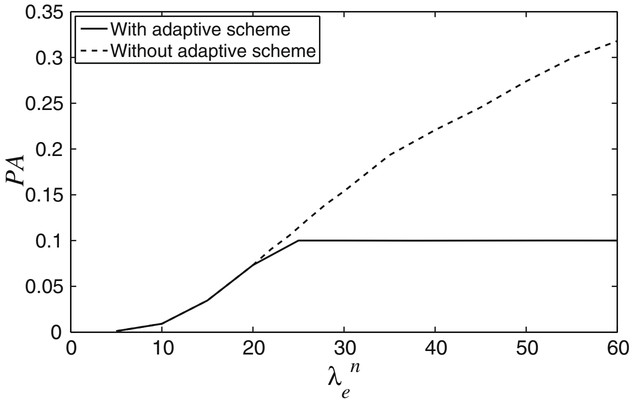

Fig. 20. Abandonment probability of elastic flows with and without adaptive scheme.

6. Conclusions

We developed a novel adaptive reservation scheme that can adapt to non-stationary traffic both in fixed and variable capacity systems. The operation of our scheme is based on simple balance equations which hold for any arrival process and holding time distribution. Our proposal has four relevant features. First, its capability to handle in an integrated way streaming and elastic traffic. Second, its ability to continuously guarantee the QoS objective even in overload by initially requesting rate adaptation at session initiation or when handovers occur, and when congestion persists by rejecting low priority requests. Third, its remarkable fast and oscillation-free transient response. And fourth, its implementation simplicity.

We provide two implementations of the scheme. First, when the lower-priority class (LPC) has a QoS objective defined, which obviously must be met when possible. Second, when the LPC is treated as a best-effort class (BEC) and therefore obtains an unpredictable QoS, which tends to be *good* during underload episodes but is *rather poor* as soon as the system enters the overload region.

We evaluated the performance of the scheme when handling multiple streaming services and showed that the QoS objective is met with an excellent precision. We also showed that it achieves an oscillation-free convergence which duration is much shorter than the one displayed by previous proposals. This confirms that our scheme can handle satisfactorily the non-stationarity of a real operating network. We also evaluated the performance of the scheme when incorporating the RA policy. Results exhibit a graceful degradation as congestion increases and a considerable increase in arrival rate that can be handled, i.e. more streaming sessions can be accepted by the system while at the same time guaranteeing the QoS objective. Finally, we evaluated the performance of the scheme when handling elastic flows in a scenario with streaming background traffic. We showed that the scheme is able to guarantee an upper bound for the abandonment probability of elastic flows.

As mobile terminals integrate positioning systems to provide location services, mobile network operators can exploit the new functionality to predict the occurrence of handovers and improve in this way the performance of the network. Then, future work will include extending the adaptive scheme so that the adjustment of the thresholds would be based not only on decisions of the AC subsystem but also based on predictive information regarding the movement of mobile terminals, both in the cell and in its neighborhood.

Table A.1
Notation used.

Symbol	Meaning
C	Number of resource units
R	Number of streaming services
r	Streaming service index
i	Index of streaming arrival class
s_i	The *i*th streaming arrival class
s_r^n s_r^h	The arrival class associated to the new and handover requests of streaming service r
λ_r^n λ_r^n	New and handover arrival rates for streaming service r
λ_{max}	System capacity perceived by streaming services
f_r	Penetration of streaming service r
μ_r	Resource holding time of streaming service r
d_r c_i	Resource units demanded by service r requests and arrival class s_i requests
P_i	Blocking probability perceived by the *i*th arrival class
P_r^n P_r^h	Blocking probabilities perceived by service r new and handover classes
B_i	Target blocking probability for the *i*th arrival class
B_r^n B_r^h	Target blocking probability for service r new and handover classes
n_r	Number of sessions in progress in the cell initiated as service r requests
$\boldsymbol{n}$	Vector of ongoing streaming sessions
$c(\boldsymbol{n})$	Number of resource units occupied by streaming sessions

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments and suggestions as they helped to improve the paper.

This work has been supported by the Spanish Ministry of Education and Science (30%) and by the European Union (FEDER 70%) under projects TSI2007-66869-C02-02 and TIN2008-06739-C04-02/TSI. It is also supported by the "Cátedra Telefónica de Internet y Banda Ancha" (e-BA) of the Universidad Politecnica de Valencia and under contract AP-2004-3332.

Appendix A. Notation

See Table A.1.

References

[1] S.C. Borst, D. Mitra, Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic, IEEE Journal on Selected Areas in Communications 16 (5) (1998) 668–678.

[2] C.-T. Chou, K.G. Shin, Analysis of combined adaptive bandwidth allocation and admission control in wireless networks, in: Proceedings of INFOCOM 2002.

[3] W.-S. Soh, H.S. Kim, A predictive bandwidth reservation scheme using mobile positioning and road topology information, IEEE/ACM Transactions on Networking 14 (5) (2006) 1078–1091.

[4] T. Bonald, J. Roberts, Scheduling network traffic, SIGMETRICS Performance Evaluation Review 34 (4) (2007) 29–35.

[5] T. Bonald, J. Roberts, Congestion at flow level and the impact of user behaviour, Computer Networks 42 (4) (2003) 521–536.

[6] Y. Zhang, D. Liu, An adaptive algorithm for call admission control in wireless networks, in: Proceedings of the IEEE Global Communications Conference (GLOBECOM), 2001, pp. 3628–3632.

[7] X.-P. Wang, J.-L. Zheng, W. Zeng, G.-D. Zhang, A probability-based adaptive algorithm for call admission control in wireless network, in: Proceedings of the International Conference on Computer Networks and Mobile Computing (ICCNMC), 2003, pp. 197–204.

[8] O. Yu, V. Leung, Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN, IEEE Journal on Selected Areas in Communications 15 (7) (1997) 1208–1224.

[9] P. Ramanathan, K.M. Sivalingam, P. Agrawal, S. Kishore, Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks, IEEE Journal on Selected Areas in Communications 17 (7) (1999) 1270–1283.

[10] O. Yu, S. Khanvilkar, Dynamic adaptive QoS provisioning over GPRS wireless mobile links, in: Proceedings of the IEEE International Conference on Communications (ICC), vol. 2, 2002, pp. 1100–1104.

[11] W.S. Jeon, D.G. Jeong, Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink, IEEE Transactions on Vehicular Technology 50 (1) (2001) 59–66.

[12] Y. Wei, C. Lina, F. Rena, R. Raadb, E. Dutkiewiczb, Dynamic handoff scheme in differentiated QoS wireless multimedia networks, Computer Communications 27 (10) (2004) 1001–1011.

[13] L. Huang, S. Kumar, C.-C.J. Kuo, Adaptive resource allocation for multimedia QoS management in wireless networks, IEEE Transactions on Vehicular Technology 53 (2) (2003) 547–558.

[14] D. Garcia-Roger, Mª Jose Domenech-Benlloch, J. Martinez-Bauset, V. Pla, Adaptive admission control scheme for multiservice mobile cellular networks, in: Proceedings of the First Conference on Next Generation Internet Networks (NGI2005), 2005, pp. 18–20.

[15] D. García, J. Martínez, V. Pla, Admission control policies in multiservice cellular networks: optimum configuration and sensitivity, Lecture Notes in Computer Science 3427 (2005) 121–135.

[16] B. Girod, Psychovisual aspects of image communications, Signal Processing 28 (3) (1992) 239–251.

[17] Y. Xiao, C.L.P. Chen, B. Wang, Bandwidth degradation QoS provisioning for adaptive multimedia in wireless/mobile networks, Computer Communications 25 (13) (2002) 1153–1161.

[18] N. Argiriou, L. Georgiadis, A framework for providing user level quality of service guarantees in multi-class rate adaptive systems, Journal of Network Systems Management 16 (4) (2008) 375–397.

[19] F.A. Cruz-Pérez, L. Ortigoza-Guerrero, Flexible resource allocation strategies for class-based QoS provisioning in mobile networks, IEEE Transactions on Vehicular Technology 53 (3) (2004) 805–819.

[20] G. Schembra, A resource management strategy for multimedia adaptive-rate traffic in a wireless network with TDMA access, IEEE Transactions on Wireless Communications 4 (1) (2005) 65–78.

[21] T. Kwon, Y. Choi, C. Bisdikian, M. Naghshineh, QoS provisioning in wireless/mobile multimedia networks using an adaptive framework, Wireless Networks 9 (1) (2003) 51–59.

[22] T. Kwon, Y. Choi, S.K. Das, Bandwidth adaptation algorithms for adaptive multimedia services in mobile cellular networks, Wireless Personal Communications 22 (3) (2002) 337–357.

[23] W. Li, X. Chao, Call admission control for an adaptive heterogeneous multimedia mobile network, IEEE Transactions on Wireless Communications 6 (2) (2007) 515–525.

[24] J.M. Gimenez-Guzman, J. Martinez-Bauset, V. Pla, A reinforcement learning approach for admission control in mobile multimedia networks with predictive information, IEICE Transactions on Communications E90-B (7) (2007) 1663–1673.

[25] F.R. Yu, V.W.S. Wong, V.C.M. Leung, A new QoS provisioning method for adaptive multimedia in wireless networks, IEEE Transactions on Vehicular Technology 57 (3) (2008) 1899–1909.

[26] 3GPP, QoS Concept and Architecture, 3GPP TS 23.107 V.4.4.0, 2002.

[27] IEEE Std 802.16, Air interface for fixed broadband wireless access systems, IEEE Std 802.16-2004, October 2004.

[28] S. Choi, K.G. Shin, Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks, IEEE Transactions on Parallel and Distributed Systems 13 (9) (2002) 882–897.

[29] S.K. Biswas, B. Sengupta, Call admissibility for multirate traffic in wireless ATM networks, in: Proceedings of INFOCOM, vol. 2, 1997, pp. 649–657.

[30] J.S. Evans, D. Everitt, Effective bandwidth-based admission control for multiservice CDMA cellular networks, IEEE Transactions on Vehicular Technology 48 (1) (1999) 36–46.

[31] J.Z.I. Koo, A. Furuskar, K. Kim, Erlang capacity of multiaccess systems with service-based access selection, IEEE Communications Letters 8 (11) (2004) 662–664.

[32] P.V. Orlik, S.S. Rappaport, On the handoff arrival process in cellular communications, Wireless Networks Journal 7 (2) (2001) 147–157.

[33] B. Jabbari, Teletraffic aspects of evolving and next-generation wireless communication networks, IEEE Personal Communications 3 (6) (1996) 4–9.
[34] F. Khan, D. Zeghlache, Effect of cell residence time distribution on the performance of cellular mobile networks, in: Proceeding of the 47th IEEE Vehicular Technology Conference (VTC'97), vol. 2, 1997, pp. 949–953.
[35] R. Ramjee, R. Nagarajan, D. Towsley, On optimal call admission control in cellular networks, Wireless Networks Journal 3 (1) (1997) 29–41.
[36] D. Rossi, M. Mellia, C. Casetti, User patience and the web: a hands-on investigation, in: Proceedings of the IEEE Global Communications Conference (GLOBECOM), vol. 7, 2003, pp. 4163–4168.

Jorge Martinez-Bauset holds a Ph.D. from the Universidad Politecnica de Valencia (UPV), Spain, and currently is an associate professor at the same university. From 1987 to 1991 he was with QPSX Communications in Perth (Western Australia), working with the team that designed the first IEEE 802.6 MAN. He has been with the Dep. of Communications of the UPV since 1991. He is a member of the Euro-NF Network of Excellence. He was recipient of the 1997s Alcatel Spain best Ph.D. Thesis award in access networks. His research interests are in the area of performance evaluation and traffic control for multiservice networks.

David Garcia-Roger, received the M.S. degree in telecommunication engineering (2001) and the Ph.D. degree in telecommunication (2007) from the Universidad Politecnica de Valencia, Spain. Currently, he belongs to the Departamento de Telemática or the Universidad Carlos III of Madrid. His professional interests are in the area of telematic engineering, particularly quality of service in mobile cellular networks and optical networks.

M.ª Jose Domenech-Benlloch graduated in 2002 in Telecommunications Engineering by the Universidad Politecnica de Valencia (Spain) and currently is a Ph.D. student in the same university. She belongs to the Broadband Networks Interconnection Group (GIRBA) from the Applications of Advanced Information and Communication Technologies institute (ITACA). Her Ph.D. Thesis is supported by Spanish government under the FPU program. Her research interests are in the area of telecommunication networks engineering.

Vicent Pla, received his M.Sc. and Ph.D. in Telecommunication Engineering from Universidad Politecnica de Valencia (UPV), Spain. He received the 1999 third national award for academic excellence in Telecommunication Engineering degree, and the 2006 Universidad Politécnica de Valencia Ph.D. award. Currently, he is an associate professor at the UPV in the Department of Communications. In 1999 he was a visiting Ph.D. student in the Multimedia and Networking Lab at the University of Pennsylvania, and a visiting scholar in the Department of Communications and Networking at the Helsinki University of Technology in 2006. He is a member of the Euro-NF Network of Excellence. His research interests lie primarily in the areas of teletraffic and performance evaluation of communication networks. In these areas he has published several papers in refereed journals and conference proceedings.