Adaptive Trunk Reservation Policies in Multiservice Mobile Wireless Networks

David Garcia-Roger, M.^{a} Jose Domenech-Benlloch, Jorge Martinez-Bauset
and Vicent Pla

Departamento de Comunicaciones, Universidad Politecnica de Valencia , UPV
ETSIT Camino de Vera s/n, 46022, Valencia, Spain
{dagarro,mdoben}@doctor.upv.es,{jmartinez,vpla}@dcom.upv.es

Abstract. We propose a novel adaptive reservation scheme designed to operate in association with the well-known Multiple Guard Channel (MGC) admission control policy. The scheme adjusts the MGC configuration parameters by continuously tracking the Quality of Service (QoS) perceived by users, adapting to any mix of aggregated traffic and enforcing a differentiated treatment among services during underload and overload episodes. The performance evaluation study confirms that the QoS objective is met with an excellent precision and that it converges rapidly to new operating conditions. These features along with its simplicity make our scheme superior to previous proposals and justify that it can satisfactorily deal with the non-stationary nature of an operating network.

1 Introduction

Session Admission Control (SAC) is a key mechanism in the design and operation of multiservice mobile cellular networks that guarantee a certain degree of Quality of Service (QoS). The mobility of terminals make 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. The design of SAC 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 probabilities and forced termination probabilities.

For stationary multiservice scenarios, different SAC policies have been evaluated in [1], where it was found that trunk reservation policies like *Multiple Guard Channel* (MGC) and *Multiple Fractional Guard Channel* (MFGC) outperform those policies which stationary state probability distributions have a product-form solution. More precisely, it was found in [1] that for the scenarios studied the performance of the MFGC policy is very close to the performance of the optimal policy and that the performance of both the MGC and MFGC policies tend to the optimal as the number of resources increase beyond a few tens. In [1] the performance is evaluated by obtaining the maximum aggregated call rate that can be offered to the system, which we call the *system capacity*, while guaranteeing a given QoS objective. The QoS objective is defined in terms of upper bound for the blocking probabilities of both new session and handover requests. It was also found in [1] that the performance of trunk reservation policies is quite sensitive to errors in the setting of their configuration parameters, defining their values the action (accept/reject) that must be taken in each system state when a new session or handover request arrives.

For the class of SAC policies considered in [1] the system capacity is a function of two parameter sets: those that describe the system as a Markov process and those that specify the QoS objective. Two approaches are commonly proposed to design a SAC policy. First, consider the parameters of the first set as stationary and therefore design a static SAC policy for the worst scenario. Second, consider them as non-stationary and either estimate them periodically or use historical information of traffic patterns.

In this paper we study a novel adaptive strategy that operates in coordination with the MGC policy. Although for simplicity we only provide implementations for the scheme when operating with the MGC policy, it can be readily extended to operate with the MFGC policy. Our scheme adapts the configuration of the MGC policy according to the QoS perceived by users. The main advantage of our adaptive scheme is its ability to adapt to changes in the traffic profile and enforce a differentiated treatment among services during underload and overload episodes. In the latter case, this differentiated treatment guarantees that higher priority services will be able to meet their QoS objective possibly at the expense of lower priority services.

Recently, different SAC adaptive schemes have been proposed for mobile cellular networks. In these proposals the configuration of the SAC policy is adapted periodically according to estimates of traffic or QoS parameters. Two relevant examples of this approach in a single service scenario are [2] and [3]. A four parameter algorithm based on estimates of the blocking probability perceived by handover requests is proposed in [2] to adjust the number of guard channels. A two hour period is defined during which the system accumulates information to compute the estimates. This period is too long to capture the dynamics of operating mobile cellular networks. Besides, the value of the parameters proposed in [2] do not work properly when some traffic profiles are offered [3], (i.e. QoS objectives are not met). A two parameter probability-based adaptive algorithm, somewhat similar to that of *Random Early Detection* (RED), is proposed in [3] to overcome these shortcomings. Its main advantage is that it reduces the new requests blocking probability, once the steady state has been reached, and therefore higher resource utilization is achieved. Nevertheless, its convergence period is still of the order of hours. The scheme we propose is also probability-based like in [3] but it has a considerably lower convergence period and can be applied to single service and multiservice scenarios.

Adaptive SAC mechanisms have also been studied, for example in [4–6], both in single service and multiservice scenarios, but in a context which is somewhat different to the one of this paper. There, the adjustment of the SAC policy configuration is based on estimates of the handover arrival rates derived from the current number of ongoing calls in neighboring cells and mobility patterns.

It is expected that the performance of our scheme would improve when provided with such predictive information but this is left for further study.

Our SAC adaptive scheme differs from previous proposals in: 1) it does not rely on measurement intervals to estimate the value of system parameters but tracks the QoS perceived by users and performs a continuous adaptation of the configuration parameters of the SAC policy; 2) the possibility of identifying several arrival streams as protected (with an operator defined order of priorities) and one as *best-effort*, being it useful to concentrate on it the penalty that unavoidably occurs during overloads; and 3) the high precision in the fulfillment of the QoS objective.

The remaining of the paper is structured as follows. Section 2 describes the model of the system and defines the relevant SAC policies. Section 3 illustrates the fundamentals of the adaptive scheme, introducing the policy adjustment strategy and how multiple services are handled. Section 4 describes the detailed operation of the scheme. Section 5 presents the performance evaluation of the scheme in different scenarios, both under stationary and non-stationary traffic conditions. Finally, Section 6 concludes the paper.

2 System Model and Relevant SAC Policies

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, the proposed scheme could also be deployed in non-homogeneous scenarios. In each cell a set of R different classes of users contend for C resource units, where the meaning of a unit of resource depends on the specific implementation of the radio interface. For each service, new and handover arrival requests are distinguished, which defines $2R$ arrival streams.

Abusing from the Poisson process definition, we say that for any class r, $1 \leq r \leq R$, new requests arrive according to a Poisson process with time-varying rate $\lambda_r^n(t)$ and request c_r resource units per session. The duration of a service r session is exponentially distributed with rate μ_r^s. The cell residence (dwell) time of a service r session is exponentially distributed with rate μ_r^d. Hence, the resource holding time for a service r session in a cell is exponentially distributed with rate $\mu_r = \mu_r^s + \mu_r^d$. We consider that handover requests arrive according to a Poisson process with time-varying rate $\lambda_r^h(t)$. Although our scheme does not require any relationship between $\lambda_r^h(t)$ and $\lambda_r^n(t)$, for simplicity we will suppose that $\lambda_r^h(t)$ it is a known fraction of $\lambda_r^n(t)$. We use exponential random variables for two reasons. First, for simplicity. Second, although it has been shown that the random variables of interest are not exponential, deploying them allows to obtain values of the performance parameters of interest which are good approximations. Besides, the operation of the proposed scheme is independent of the distribution of the random variables.

We denote by P_i, $1 \leq i \leq 2R$, the perceived blocking probabilities for each of the $2R$ arrival streams, by $P_r^n = P_i$ the blocking probabilities for new requests and by $P_r^h = P_{R+i}$ the handover blocking probabilities. The QoS objective is expressed as upper bounds for the blocking probabilities, denoting by B_r^n (B_r^h) the bound for new (handover) requests. Let the system state vector be $n \equiv (n_1, n_2, \dots, n_{2R-1}, n_{2R})$, where n_i is the number of sessions in progress in the cell initiated as arrival stream i requests. We denote by $c(n) = \sum_{i=1}^{2R} n_i c_i$ the number of busy resource units in state n.

The definition of the SAC policies of interest is as follows: 1) Complete-Sharing (CS). A request is admitted provided there are enough free resource units available in the system; 2) Multiple Guard Channel (MGC). One parameter is associated with each arrival stream i, $l_i \in \mathbb{N}$. When an arrival of stream i happens in state n, it is accepted if $c(n) + c_i \leq l_i$ and blocked otherwise. Therefore, l_i is the amount of resources that stream i has access to and increasing (decreasing) it reduces (augments) P_i.

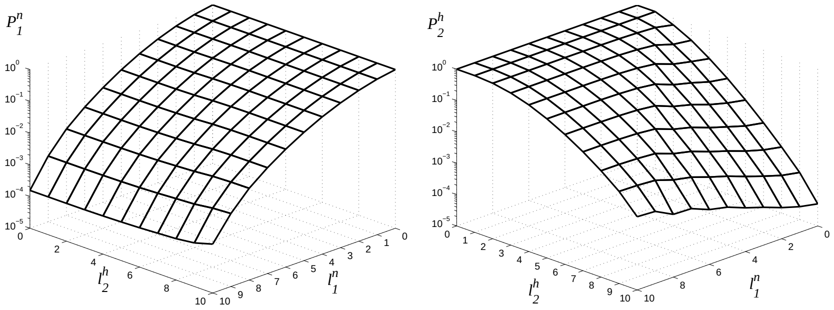
The performance evaluation of the adaptive scheme is carried out for five different scenarios (A, B, C, D and E) that are defined in Table 1, being the QoS parameters B_i expressed as percentage values. The parameters in Table 1 have been selected to explore possible trends in the numerical results, i.e., taking scenario A as a reference, scenario B represents the case where the ratio c_1/c_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. Note that the aggregated arrival rate of new requests is defined as $\lambda = \sum_{r=1}^{R} \lambda_r^n$, where $\lambda_r^n = f_i \lambda$. The system capacity is the maximum λ (λ_{max}) that can be offered to the system while meeting the QoS objective.

Table 1. Definition of the scenarios under study

	A	B	C	D	E
c_1	1	1	1	1	1
c_2	2	4	2	2	2
f_1	0.8	0.8	0.2	0.8	0.8
f_2	0.2	0.2	0.8	0.2	0.2
B_1^n%	5	5	5	1	1
B_2^n%	1	1	1	2	1
	A,B,C,D,E				
B_r^h%	$0.1 B_r^n$				
λ_r^n	$f_r \lambda$				
λ_r^h	$0.5 \lambda_r^n$				
μ_1	1				
μ_2	3				

3 Fundamentals of the Adaptive Scheme

Most of the proposed adaptive schemes deploy a reservation strategy based on *guard channels*, increasing its number when the QoS objective is not met. The extension of this heuristic to a multiservice scenario would consider that adjusting the configuration parameter l_i only affects the QoS perceived by s_i (P_i) but has no effect on the QoS perceived by the other arrival streams. As an example, Fig. 1 shows the dependency of P_1^n and P_2^h with l_1^n and l_2^h, respectively, while the other configuration parameters are kept constant at their optimum values. It has been obtained in scenario A with $C = 10$ resource units, when deploying the MGC policy and when offering an arrival rate equal to the system capacity. As shown, the correctness of the heuristic is not justified (observe P_2^h) although it might work in some cases (observe P_1^n).



(a) Arrival stream s_1^n (b) Arrival stream s_2^h

Fig. 1. Dependency of the blocking probability with the configuration parameters.

Our scheme has been designed to handle this difficulty and to fulfill two key requirements that have an impact on its performance: one is to achieve a convergence period as short as possible and the other is to enforce a certain response during underload or overload episodes. For these purposes we classify the different arrival streams into two generic categories: i) those that the operator identifies as "protected" because they must meet specific QoS objectives; ii) one *Best-Effort Stream* (BES), with no specific QoS objective.

Additionally, the operator can define priorities at its convenience in order to protect more effectively some streams than other, i.e. handover requests. If we denote the generic stream i by s_i, $1 \leq i \leq 2R$, and we assume that the order of priorities required by the operator for the different streams is $\mathbf{s}^* = (s_{\pi_1}, s_{\pi_2}, \ldots, s_{\pi_{2R}})$, then the vector $\pi^* = (\pi_1, \ldots, \pi_i, \ldots, \pi_{2R})$, $\pi_i \in \mathbb{N}, 1 \leq \pi_i \leq 2R$, is called the "prioritization order", being s_{π_1} the *Highest-Priority Stream* (HPS) and $s_{\pi_{2R}}$ the *Lowest-Priority Stream* (LPS). We study two im-

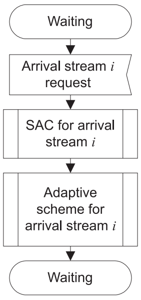

Waiting
Arrival stream i request
SAC for arrival stream i
Adaptive scheme for arrival stream i
Waiting

Fig. 2. Conceptual operation of the adaptive reservation scheme.

plementations, one in which the LPS is treated as a protected stream and one in which the LSP is the BES. For clarity in some cases we will denote by s_r^n (s_r^h) the arrival stream associated to new (handover) requests. In relation to the parameters that define the configuration of the MGC policy, we will denote by l_r^n (l_r^h) the configuration parameter associated to the arrival stream s_r^n (s_r^h) and by l_i the one associated to s_i.

3.1 Probabilistic Setting of the Configuration Parameters

A common characteristic of previous schemes like those in [2, 3] and [4–6] is that they require a time window (*update period*) at the end of which some estimates are produced. The design of this update period must trade-off the time required to adapt to new conditions for the precision of estimates. The adaptive scheme we propose overcomes this limitation. The scheme tracks the QoS perceived by each arrival stream and performs a continuous adaptation of the configuration parameters of the SAC policy.

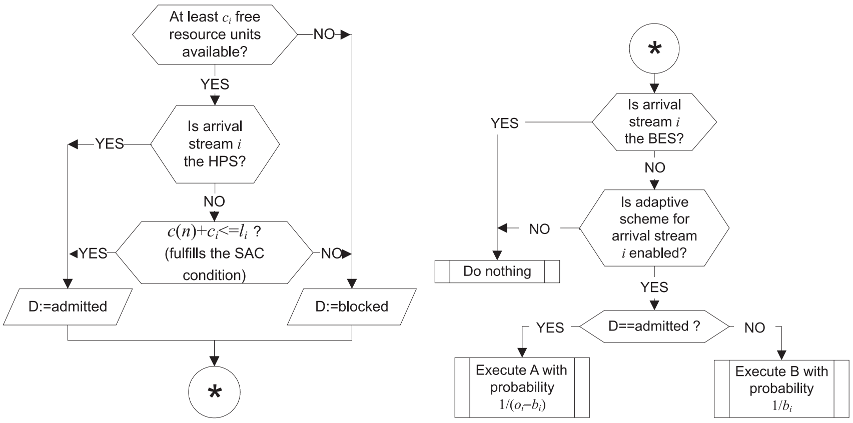
Let us assume that arrival processes are stationary and the system is in steady state. If the QoS objective for s_i can be expressed as $B_i = b_i/o_i$, where $b_i, o_i \in \mathbb{N}$, then it is expected that when $P_i = B_i$ the stream i will experience, in average, b_i rejected requests and $o_i - b_i$ admitted requests, out of o_i offered requests. It seems intuitive to think that the adaptive scheme should not change the configuration parameters of those arrival streams meeting their QoS objective. Therefore, assuming integer values for the configuration parameters, like those of the MGC policy, we propose to perform a probabilistic adjustment each time a request is processed, i.e. each time the system takes an admission or rejection decision, by adding $+1$ or -1 to l_i, when it effectively occurs.

Figure 2 shows the general operation of the proposed scheme. As seen, when a stream i request arrives, the SAC decides upon its admission or rejection and this decision is used by the adaptive scheme to adjust the configuration of the SAC policy.

4 Operation of the SAC Adaptive Scheme

Figure 3 shows the operation of the SAC subsystem and the adaptive scheme. In our proposal, two arrival streams, the HPS and the BES, receive differentiated treatment. On the one hand, a HPS request must be always admitted, if enough free resources are available. On the other hand, no specific action is required to adjust the QoS perceived by the BES, given that no QoS objective must be met.

As shown in Fig. 3(a), to admit an arrival stream i request it is first checked that at least c_i free resource units are available. Note that once this is verified, HPS requests are always admitted, while the rest of streams must also fulfill the admission condition imposed by the MGC policy. In general, the adaptive scheme is always operating (except for the BES), but meeting the QoS objective of higher priority streams could require to disable the operation of the adaptive schemes associated to lower priority streams, as explained below.

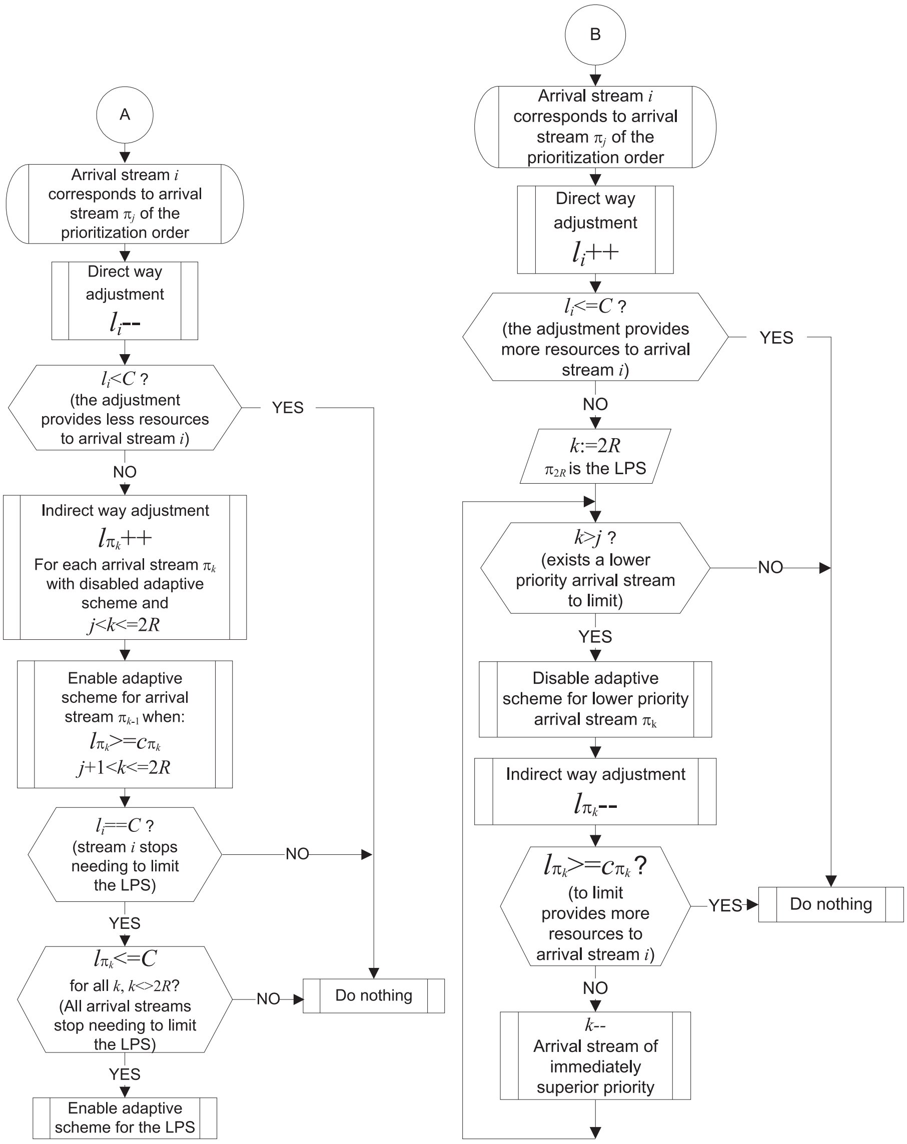

At least c_i free resource units available?
NO
YES
Is arrival stream i the HPS?
YES
NO
$c(n)+c_i<=l_i$? (fulfills the SAC condition)
YES
NO
D:=admitted
D:=blocked
*
*
Is arrival stream i the BES?
YES
NO
Is adaptive scheme for arrival stream i enabled?
NO
Do nothing
YES
D==admitted ?
YES
NO
Execute A with probability $1/(o_i-b_i)$
Execute B with probability $1/b_i$

(a) Description of the *SAC for arrival stream i* block in Fig. 2.

(b) Description of the *Adaptive scheme for arrival stream i* block in Fig. 2.

Fig. 3. Operation of SAC policy and adaptive scheme.

To be able to guarantee that the QoS objective is always met, particularly during overloads episodes or changes in the load profile (i.e. new f_i), the probabilistic adjustment described in Section 3.1 requires additional mechanisms. Two ways are possible to change the policy configuration when the QoS objective for stream i is not met. The direct way is to increase the configuration parameter l_i, but its maximum value is C, i.e. when $l_i = C$ full access to the resources is provided to stream i and setting $l_i > C$ does not provide additional benefits. In these cases, an indirect way to help stream i is to limit the access to resources of lower priority streams by reducing their associated configuration parameters.


A
Arrival stream i corresponds to arrival stream π_j of the prioritization order
Direct way adjustment
l_i--
$l_i<C$?
(the adjustment provides less resources to arrival stream i)
YES
NO
Indirect way adjustment
l_{π_k}++
For each arrival stream π_k with disabled adaptive scheme and
$j<k<=2R$
Enable adaptive scheme for arrival stream π_{k-1} when:
$l_{\pi_k}>=c_{\pi_k}$
$j+1<k<=2R$
$l_i==C$?
(stream i stops needing to limit the LPS)
NO
YES
$l_{\pi_k}<=C$
for all k, $k<>2R$?
(All arrival streams stop needing to limit the LPS)
NO
Do nothing
YES
Enable adaptive scheme for the LPS
B
Arrival stream i corresponds to arrival stream π_j of the prioritization order
Direct way adjustment
l_i++
$l_i<=C$?
(the adjustment provides more resources to arrival stream i)
YES
NO
$k:=2R$
π_{2R} is the LPS
$k>j$?
(exists a lower priority arrival stream to limit)
NO
YES
Disable adaptive scheme for lower priority arrival stream π_k
Indirect way adjustment
l_{π_k}--
$l_{\pi_k}>=c_{\pi_k}$?
(to limit provides more resources to arrival stream i)
YES
Do nothing
NO
k--
Arrival stream of immediately superior priority

(a) Adjustment algorithm after an admission decision.

(b) Adjustment algorithm after a rejection decision.

Fig. 4. The adaptive algorithm.

As shown in Fig. 4(b), upon a rejection the adaptive scheme uses first the direct way and after exhausted it resorts to the indirect way, in which case the adaptive schemes of the lower priority streams must be conveniently disabled. Figure 4(a) shows the reverse procedure. Note that when stream π_k is allowed to access the resources, then the adaptive scheme of the π_{k-1} stream is enabled. When the LPS is the BES then its adaptive scheme is never enabled. Note also that we allow the values of the l_i parameters to go above C and below zero as a means to remember past adjustments.

The scheme described in this paper is a generalization of the one proposed in [8] because it incorporates two notable features. First, it provides the operator with full flexibility to define any prioritization order for the arrival streams and for selecting one of the two implementations proposed. Second, the penalty induced on the lower priority streams increases progressively to guarantee that the QoS objective of the higher priority streams is met.

5 Performance Evaluation

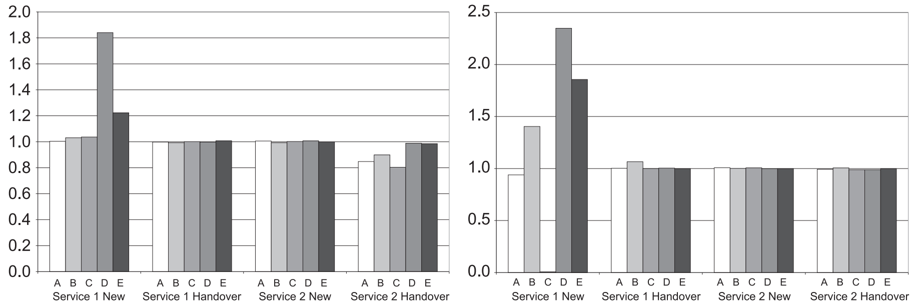
The performance evaluation has been carried out using Möbius$^{\text{TM}}$ [7], which is a software tool that supports *Stochastic Activity Networks* (SANs). Möbius$^{\text{TM}}$ allows to simulate the SANs that model the type of systems of interest in our study, and under certain conditions, even to numerically solve the associated continuous-time Markov chains.

For the five scenarios defined in Table 1, $\{A, B, C, D, E\}$, with $C = 10$ and with no adaptive scheme, the system capacity when deploying Complete Sharing is $\{1.54, 0.37, 1.37, 1.74, 1.54\}$, while when deploying the MGC policy is $\{1.89, 0.40, 1.52, 1.97, 1.74\}$. Refer to [1] for details on how to determine the system capacity. For all scenarios defined in Table 1 we assume the following prioritization order $\mathbf{s}^* = (s_2^h, s_1^h, s_2^n, s_1^n)$. We evaluate by simulation two implementations that differ in the treatment of the LPS (s_1^n), one in which it is a protected stream and one in which it is the BES.

5.1 Performance under Stationary Traffic

Figure 5(a) and (b) show the ratio P_i/B_i for the four arrival streams in the five scenarios considered and for the two implementations of the adaptive scheme. In all cases, an aggregated calling rate equal to the system capacity (λ_{max}) is offered.

Figure 6 provides additional information on the variation of performance for scenario C with $C = 10$ resource units. When the LPS is a protected stream (Fig. 6(a)) 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 LPS is the BES (Fig. 6(b)) 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 penalizing the LPS would be ineffective. Note also that during underload episodes

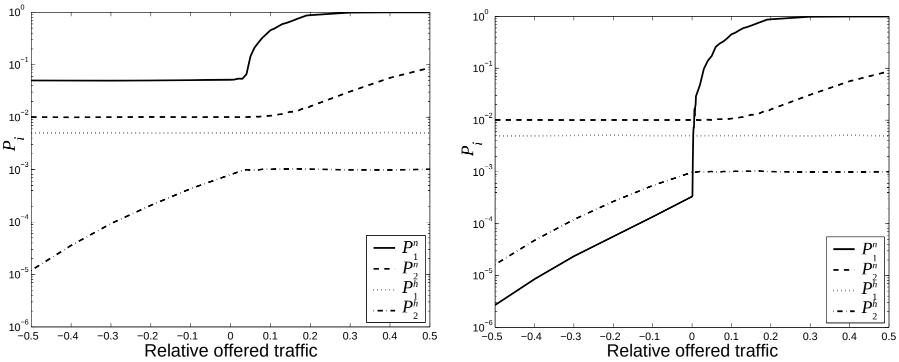


(a) Implementing the LPS as a protected stream.

(b) Implementing the LPS as the BES.

Fig. 5. P_i/B_i for a system with a stationary load equal to λ_{max}.

$P_i = B_i$ is held for protected streams and therefore the system is rejecting more requests than required, but some streams (HPS and BES) benefit from this extra capacity.



(a) Implementing the LPS as a protected stream.

(b) Implementing the LPS as the BES.

Fig. 6. P_i as a function of $(\lambda - \lambda_{max})/\lambda_{max}$ in stationary conditions.

5.2 Performance under Non-Stationary Traffic

In this section 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 LPS is a

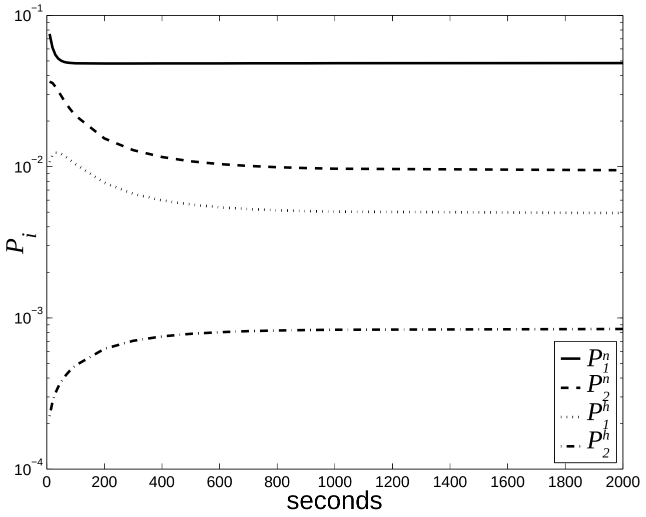


Fig. 7. Transient behavior of the blocking probabilities.

protected stream. Before the step increase is applied the system is in the steady state regime.

Figure 7 shows the transient behavior of the blocking probabilities. As observed, the convergence period is lower than 1000 s., which is 10 to 100 times lower than in previous proposals [2, 3]. Note that the convergence period will be even shorter when the offered load is above the system capacity thanks to the increase in the probabilistic-adjustment actions rate, which is an additional advantage of the scheme. Additional mechanisms have been developed that allow to trade-off convergence speed for precision in the fulfillment of the QoS objective, but will not be discussed due to paper length limitations.

6 Conclusions

We developed a novel adaptive reservation scheme that operates in coordination with the Multiple Guard Channel policy but that can be readily extended to operate with the Multiple Fractional Guard Channel policy. Three relevant features of our proposal are: its capability to handle multiple services, its ability to continuously track and adjust the QoS perceived by users and its simplicity. We provide two implementations of the scheme. First, when the LPS has a QoS objective defined, which obviously must be met when possible. Second, when the LPS is treated as a best-effort stream and therefore obtains an unpredictable QoS, which tends to be “good” during underload episodes but is “quite bad” as soon as the system enters the overload region.

The performance evaluation shows that the QoS objective is met with an excellent precision and that the convergence period, being around 1000 s., is 10 to 100 times shorter than in previous proposals. This confirms that our scheme can handle satisfactorily the non-stationarity of a real network.

Future work will include the evaluation of the scheme when operating with other SAC policies, for example those for which the stationary probability distribution has a product-form solution. Another interesting extension would be to base the adjustment of the configuration parameters not only on the decisions of the SAC subsystem but also on predictive information, like movement prediction.

Acknowledgments

This work has been supported by the Spanish Ministry of Education and Science (30%) and by the EU (FEDER 70%) under projects TIC2003-08272, TEC2004-06437-C05-01 and under contract AP-2004-3332, and by the Generalitat Valenciana under contract CTB/PRB/2002/267.

References

1. D. García, J. Martínez and V. Pla,“Admission Control Policies in Multiservice Cellular Networks: Optimum Configuration and Sensitivity,” Wireless Systems and Mobility in Next Generation Internet, Gabriele Kotsis and Otto Spaniol (eds.), Lecture Notes in Computer Science, vol. 3427, pp.121-135, Springer-Verlag 2005.
2. Y. Zhang, D. Liu, “An adaptive algorithm for call admission control in wireless networks”, Proceedings of the IEEE Global Communications Conference (GLOBECOM), pp. 3628-3632, San Antonio, (USA), Nov. 2001.
3. X.-P. Wang, J.-L. Zheng, W. Zeng, G.-D. Zhang, “A probability-based adaptive algorithm for call admission control in wireless network”, Proceedings of the International Conference on Computer Networks and Mobile Computing (ICCNMC), pp. 197-204, Shanghai, (China), 20-23 Oct. 2003.
4. O. Yu, V. Leung, “Adaptive Resource Allocation for prioritized call admission over an ATM-based Wireless PCN”, IEEE Journal on Selected Areas in Communications, pp. 1208-1224, vol. 15, Sept. 1997.
5. P. Ramanathan, K. M. Sivalingam, P. Agrawal, S. Kishore, “Dynamic Resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks”, Journal on Selected Areas in Communications, pp. 1270-1283, vol. 17, Jul. 1999.
6. O. Yu, S. Khanvilkar, “Dynamic adaptive QoS provisioning over GPRS wireless mobile links”, Proceedings of the IEEE International Conference on Communications (ICC), pp. 1100-1104, vol. 2, New York, (USA), 28 Apr.- 2 May 2002.
7. Performability Engineering Research Group (PERFORM), Möbius^{TM}. User Manual. Version 1.6.0: http://www.perform.csl.uiuc.edu/mobius/manual/Mobius Manual_160.pdf.
8. D. Garcia-Roger, M^{a} Jose Domenech-Benlloch, J. Martinez-Bauset, V. Pla, “Adaptive Admission Control Scheme for Multiservice Mobile Cellular Networks”, Proceedings of the 1st Conference on Next Generation Internet Networks (NGI2005), Roma, (Italy), 18-20 Apr. 2005.