Congestion Control Strategies
in Multi-Access Networks

X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí

Dept. Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
[xavier.gelabert,jorperez,sallent,ramon]@tsc.upc.edu

Abstract— **This paper addresses the problematic of congestion control in the radio access interface when considering the allocation of voice and data services over several Radio Access Technologies (RATs). In particular, the GSM/EDGE Radio Access Network (GERAN) and the UMTS Terrestrial Radio Access Network (UTRAN) are considered for the evaluation of congestion control strategies. After a congestion situation in the radio access is detected, congestion resolution mechanisms are triggered in order to reduce the overload in the congested RAT(s). In this paper, a framework for the detection and resolution of congestion conditions in a multi-access network is presented. Moreover, three approaches intending to solve congestion situations are proposed and the evaluation of an inter-RAT handover algorithm for solving congestion events in GERAN is also presented.**

Keywords- Common Radio Resource Management (CRRM); Congestion Control; GERAN; UTRAN; Beyond 3G.

I. INTRODUCTION

The heterogeneous network concept is a very attractive notion that has been extensively addressed over the past few years [1]. It essentially proposes a flexible architecture capable of managing a large variety of wireless access technologies along with applications and services with different Quality of Service (QoS) demands and protocol stacks. A heterogeneous network may then include cellular networks like 3GPP-standardised UTRAN and GERAN along with non-cellular access networks like WLAN 802.11. Moreover, it must be also aware of the existing cellular layers, the so-called Hierarchical Cell Structure—HCS, e.g. macro, micro or pico cells covering a given area. Finally, several technologies may be available in different cells within a given RAN, such as UMTS R99 and HSDPA in UTRAN, or GPRS and EDGE in GERAN, or 802.11b and 802.11g in WLAN.

The rationale behind the heterogeneous network notion lays in the fact that new emerging technologies, e.g. HSDPA or WiMax, will have to coexist with previous and/or legacy technologies, e.g. GSM or GPRS. Then, we can take advantage of this plethora of networks by trying to exploit the trunking gain that results from the common management of all the available radio resources of all networks rather than managing those radio resources considering stand-alone networks.

Among other radio resource management (RRM) strategies, Congestion Control (CC) is the RRM function devoted to overcome potential QoS failures due to the intrinsic dynamics of the network (e.g. mobility, interference rise, traffic variability, etc.). Regardless of having a strict admission control mechanism, which may ensure some average QoS requirements at call/session establishment, if the dynamics of certain network parameters suffer from high random behaviour, the network may experience high-load/high-interference situations which in turn may degrade the QoS perceived by users. In order to account for this situation, CC strategies are designed so as to minimise the impact of these sudden changes on the network performance.

To achieve a high utilization of the scarce radio resources in multi-RAT scenarios, CC may take advantage of the common pool of resources in order to solve congestion situations. This is in line with what generically has been termed Common RRM (CRRM) [2][3].

It should be kept in mind that, throughout the paper, the term *congestion* will be used to define the congestion situations experienced at the radio interface layer due to an excessive interference, e.g. in WCDMA systems, or to the excessive radio resource sharing in e.g. FDMA/TDMA systems.

Throughout the literature it is widely accepted that three main procedures should be carried out during a congestion situation [4], namely:

a) The *Congestion Detection* (CD) monitors the network status in order to correctly identify a congestion situation by means of RAT-specific measurements.

b) The *Congestion Resolution* (CR) actuates over a set of congestion control actions (CCA) in order to reduce the load and consequently the congestion situation.

c) The *Congestion Recovery* (CRV) attempts to restore the old transmission parameters before the congestion was triggered.

Congestion control has been extensively covered in the literature in the area of fixed computer networks, e.g. [5]. Congestion control at the radio access level has also been addressed in a number of papers, e.g. [7] and [8], nevertheless considering one single RAT. To our knowledge, few efforts have been devoted on congestion detection and resolution at the radio interface comprising various RATs [9].

In this paper, we propose a generic framework for the detection and resolution of congestion situations in a

GERAN/UTRAN multi-access scenario. Moreover, three CR strategies are identified, briefly: (i) an inter-RAT or vertical handover (VHO) based mechanism; (ii) a bit-rate reduction mechanism and (iii) a user dropping mechanism. Results for the VHO-based CR will be presented for a specific study case. The paper is outlined as follows: section II describes the considered framework architecture. In section III the CD and CR mechanisms are described. In section IV, some issues regarding simulation setup and scenarios are presented. Section V presents some illustrative results and, finally, conclusions are found in section VI.

II. FRAMEWORK ARCHITECTURE

The functional model assumed by the Third Generation Partnership Project (3GPP) for CRRM operation, [1][3], considers the total amount of resources available for an operator divided into radio resource pools. Each radio resource pool consists of the resources available in a set of cells, typically under the control of a RNC (Radio Network Controller) or a BSC (Base Station Controller) in UTRAN and GERAN respectively. Two types of entities are considered for the management of these radio resource pools (see Figure 1). On one hand, the local RRM entity, which carries out the management of the resources in one radio resource pool of a certain RAN and, on the other hand, the CRRM entity, which executes the coordinated management of the resource pools controlled by different RRM entities, ensuring that the decisions of these RRM entities take also into account the resource availability in other RRM entities.

Regarding CC, local RRM entities provide cell load measurements of the cells under management of the CRRM entity. The Common Congestion Control module, within the CRRM entity (see Figure 1), will then process the information and actuate accordingly if congestion is detected. The actions to be taken after a congestion event is detected may call for local resource management in a specific RAT, e.g. by limiting the bit-rate of its users, or, on the other hand, by the coordinate management of the resources in both RATs, e.g. in the case of solving the congestion by means of VHOs.

III. CC STRATEGIES IN A GERAN/UTRAN SCENARIO

Hereon we will focus in a scenario where GERAN and UTRAN sites provide service over a same area. Due to the different medium access nature of GERAN and UTRAN systems (TDMA vs. CDMA), congestion will be detected and solved differently in each of the RATs.

The following sections tackle the congestion detection mechanisms in each of the available RATs along with congestion resolution strategies both in a common and local perspective.

A. Congestion Detection

The Congestion Detection (CD) procedures must avoid two situations: false CD and non-detected congestion. The former relates to the case when a congestion situation is detected when the air interface is actually not overloaded. The latter is concerned with congestion situations becoming unnoticed when the air interface is overloaded. In order to avoid the aforementioned problems, the CD mechanism should exhibit fast reactivity and high measurement reliability. CD in GERAN and UTRAN are described in the following.

1) Congestion Detection in GERAN

The resource allocation in EGPRS is based on the "capacity on demand" principle. An EGPRS user may transmit data using simultaneously a number of packet data channels (PDCHs). Moreover, a number of users may be multiplexed over the same PDCH. Since data and voice users in the cell share the same transport media, resources for GSM and EGPRS traffic must be managed appropriately. Several strategies may be devised for handling these types of traffic [10]. In this study we will assume that the total capacity is shared between voice and data users with pre-emptive priority for the voice service. If each cell offers a total amount of C channels for voice and data users, the number of occupied channels by voice users, C_v, and the number of occupied channels by data users, C_d, must satisfy $C_v + C_d \leq C$.

In order to account for the congestion effect of time-slot (TSL) sharing among users, we exploit the *reduction factor* (RF) presented in [11]. This parameter takes values between 0 and 1, meaning a high TSL reuse in the former, and a low TSL reuse in the latter. The RF observed after the t-th frame, RF_t, may be computed as follows:

$$RF_t = \begin{cases} 0 & \text{if } N_t = 0 \\ 1 & \text{if } 0 < N_t \leq C_{d,t} \\ \dfrac{C_{d,t}}{N_t} & \text{if } N_t > C_{d,t} \end{cases} \qquad (1)$$

with N_t the total number of assigned data TSLs and $C_{d,t}$ the number of occupied TSLs by data services at the t-th frame.

Then, fixing a reduction factor threshold RF_{CD} matched to some QoS parameter, if $0 < RF_t < RF_{CD}$ during a certain number of frames, the CR mechanism is triggered.

2) Congestion Detection in UTRAN

In UTRAN, overload situations may be detected by means of the load factor η which can be measured, for the uplink (UL) and downlink (DL), as [5]:
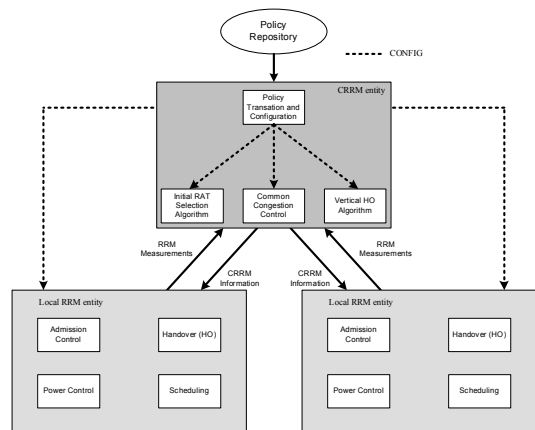

Figure 1 Common Congestion Control in the CRRM framework.

$$\eta_{UL} = 1 - \frac{P_N}{I_{total}} \quad (2) \qquad \eta_{DL} = \frac{P_{total}}{P_{max}} \quad (3)$$

with P_N the background thermal noise, I_{total} the total received wideband power, P_{total} the total transmission power and P_{max} the maximum Node-B transmission power. Then, the criterion to decide whether we have entered a congestion situation consists in checking if $\eta_{\{UL,DL\}} \geq \eta_{CD}$ during a certain percentage of frames within a period of time.

B. Congestion Resolution Strategies

In simple terms, whenever the total sum of resource demands by the users in a particular RAT is higher than a given amount of available resources we may encounter a congestion situation, i.e.

$$\sum \text{Demands}_{(C_v, C_d, P_{total}, R_b, ...)} > \text{Resources}_{(C, \eta_{max}, P_{max})} \quad (4)$$

This simple definition of congestion enables the characterization of two alternative schemes to solve a congestion situation [5]:

a) *Resource Creation Schemes (RCS)*: Such schemes may increase the capacity of existing resources by reconfiguration of network parameters or by exploiting available resources in other cells or networks (e.g. via handover procedures). In other words, the goal in this case is to increase the right-hand side of (4).

b) *Demand Reduction Schemes (DRS)*: DRS schemes try to reduce the demand to the level of available resources, that is, to reduce left-hand side of (4). Such strategies usually entail service degradation so as to reduce the load in the system, e.g. by means of transmission rate reduction or by user dropping in the worse cases.

In this paper, the proposed CR algorithms operate, on a user-by-user basis, over a set of prioritized list of users. After CC actions have been taken over a given user i, the algorithm checks if the congestion situation has been solved, i.e. if $\eta^{new} \leq \eta_{CD}$ in UTRAN, or if $RF^{new} \geq RF_{CD}$ in GERAN. Where η^{new} and RF^{new} can be expressed as:

$$\begin{aligned} \eta^{new} &= \eta^{old} - \Delta\eta_i \\ RF^{new} &= RF^{old} + \Delta RF_i \end{aligned} \quad (5)$$

with η^{old} and RF^{old} the measured CD metrics before the CR was performed for UTRAN and GERAN respectively. $\Delta\eta_i$ and ΔRF_i account for the contribution of user i on the load factor and the RF respectively.

In the following subsections, three mechanisms to overcome congestion situations are described.

1) Vertical Handover Congestion Resolution (VHO-CR)

This strategy intends to alleviate congestion by means of performing a VHO (inter-RAT handover) over a set of prioritized users in the congested RAT/Cell. If congestion is detected in a UTRAN cell, a successful VHO attempt of user i from UTRAN to GERAN will contribute to decrease the uplink load factor in an amount which can be estimated as [5]:

$$\Delta\eta_i = (1 + f_{UL}) \left[\frac{W/R_{b,i}}{(E_b/N_0)_i} + 1 \right]^{-1} \quad (6)$$

Where f_{UL} is the inter-to-intra-cell interference ratio, W the chip-rate, $R_{b,i}$ the i-th user bit rate and $(E_b/N_0)_i$ the target bit-energy-to-noise-density requirement for user i. Similarly, for the DL, a VHO of user i from UTRAN to GERAN will decrease the total downlink power a quantity $\Delta P_{T,i}$.

On the other hand, given a congestion situation in GERAN, we intend to increase the RF by directing users to UTRAN. In this way, resources can be re-allocated and a new RF, RF^{new}, measured by means of (1).

Note that a VHO will not be allowed if the target cell/RAT is also congested or the addition of this user forces it to fall into a congestion state.

The VHO procedure involves also a base station (BS) selection. We assume that the BS with best signal strength is selected in GERAN while in UTRAN the BS with higher E_c/I_0 is chosen.

2) Bit-rate Reduction Congestion Resolution (BRR-CR)

This scheme aims to lessen congestion by reducing the transmission rate demands of data users being served in the congested cell/RAT. In this way, however, the QoS in terms of throughput perceived by the users affected by the reduction can be significantly degraded.

According to the bit-rate reduction pace to be carried out over a given user, two BRR strategies may be considered:

a) Maximum BRR (MAX-BRR): Applying the maximum allowable transmission rate reduction on a given user.

b) Minimum BRR (MIN-BRR): In this case, the reduction on each user is the minimum allowable reduction.

After a BRR is performed on a given user, congestion metrics are checked in order to see whether the congestion has been solved or not. If so, the CR process is ended, otherwise, we perform BRR on the following user in the prioritized list.

C. User Dropping Congestion Resolution (DROP-CR)

This CR strategy reduces the overload of the system by selectively dropping users in the congested cell/RAT. By doing so, load factor is reduced in UTRAN and RF is increased in the same fashion than in the VHO-CR scheme. However, the DROP-CR presents the highest negative impact on users' perceived QoS and should, therefore, be only used if other strategies fail to solve congestion.

D. User Prioritization Considerations

As mentioned earlier, congestion resolution mechanisms are applied on a number of users in order to reduce the overload of the system. How to select these users can be based on a number of criteria, among those we propose:

1) Service-class prioritization

Users are ordered based on the expected QoS requirements from high to low priority. Then, congestion may be resolved by acting over those users with low priority. An example of this ordering may be split users in Real Time (RT) and Non Real Time (NRT) service demands. Because RT services are

more stringent in QoS demands, the CR algorithm may start dropping NRT users in order to solve the congestion situation.

2) User-type prioritization

Premium users are expected to receive a preferential treatment in terms of perceived QoS. Therefore, consumer users will be the first users to get downgraded in order to mitigate an overload situation.

3) Capacity-consumption prioritization

Different service users consume different amounts of resources. For example, a voice user in GSM consumes a whole slot for its transmission in both directions, while a GPRS data user may share the same timeslot with other users thus contributing to lower resource consumption. Bearing this in mind, the congestion control algorithm may first de-allocate those users with the highest resource consumption in order to decongest the network.

IV. SIMULATION SETUP

In order to illustrate the performance of the congestion control strategies within the CRRM framework we consider a study case where congestion is detected in GERAN and the strategy VHO-CR tries to solve the congestion. For such purpose, a system-level simulator based on snapshots was devised. The scenario considers 7 co-located GERAN/UTRAN sites with equal coverage over an area of 4.5 km by 4.5 km and with cell radius of 1km. The urban macrocell propagation model is assumed and omnidirectional antennas are considered in both systems. A mix of voice and traffic users is considered, and it is assumed that all terminals have multi-mode capabilities.

In GERAN, voice users are allocated to full-rate channels, i.e. one timeslot in each frame, which offers a bit-rate per user of 12.2 kbps both in the UL and DL. In UTRAN, the Radio Access Bearer (RAB) for voice users is the 12.2 kbps speech defined in [12], considering a dedicated channel (DCH) with spreading factor (SF) 64 in the UL and 128 in the downlink.

Interactive (web browsing) users in GERAN are allocated assuming multislot capabilities up to 2 UL slots and 3 DL slots, with maximum number of UL+DL slots equal to 4. The considered Modulation and Coding scheme (MCS) is considered to be MCS-7 [11], which offers a bit-rate of 44.8 kbps per time-slot. In UTRAN, the RAB for interactive users assumes a maximum bit-rate of 64 kbps in the UL (corresponding to a minimum SF of 16) and 128 kbps in the DL (with a SF of 16) [12].

Admission control procedures for voice and interactive users in UTRAN consider checking the UL load factor ($\eta_{UL,\max} = 1$), the downlink transmitted power ($P_{DL,\max} = 42\text{dBm}$) and the availability of OVSF codes at the BS [5]. In GERAN, voice users are accepted provided there are free available time slots. Otherwise, they make use of voice priority by reducing the slot requirements of ongoing data users, or by dropping data users if necessary. Data users are accepted given that there are free timeslots and that the maximum number of users sharing the same slot is at most 8 for the UL and 32 for the DL.

Users are distributed over the aforementioned area in a non-homogeneous way considering a 1km radius circular hot-spot around the central cell. This hot-spot "captures" 25% of the users offered to the whole simulation scenario.

Regarding user allocation in each RAT, i.e. GERAN and UTRAN, a service-based RAT selection policy presented in [13] is used. In particular, voice traffic is directed to GERAN and interactive traffic is directed to UTRAN provided capacity is available in each of the RATs. Otherwise, users attempt admission in the opposite RAT. If finally the admission is not possible, users are blocked.

We will assume that congestion is detected in GERAN using the DL RF defined in (1) and assume that congestion is detected when RF falls below the $RF_{CD} = 0.2$ threshold. The users on which to perform VHO are chosen randomly over the users being served in the congested RAT/cell.

V. RESULTS

Figure 2 shows the congestion detection probability (CDP), measured as the ratio between the number of congestion events and the number of simulation events (snapshots), at the GERAN central cell. For 50 interactive users and the range of voice users, we notice that hardly any congestion situation is detected. This is because UTRAN can handle interactive users and only in very few cases interactive users are directed to GERAN, thus rarely causing congestion. If we increase the number of interactive users up to 100, the CDP gets quite noticeable. In this case, interactive users are directed from UTRAN to GERAN causing higher timeslot sharing and thus congestion. If we increase the number of voice users, due to voice pre-emption, interactive users are forced to share even more their resources. However, if we keep raising the number of voice users up to 70, the CDP falls due to interactive users getting blocked caused by voice user pre-emption priority. This effect is present for 150 interactive users, where CDP decreases as the number of voice users rises.

Figure 3 shows the congestion resolution probability (CRP) when VHO-CR is used to solve congestion situations in GERAN central cell. The CRP is measured as the ratio between the number of successful congestion events solved by VHO-CR in GERAN central cell and the total number of congestion detections in GERAN central cell. For 50 interactive users we have seen (Figure 2) that congestion situations happen very rarely. If they do happen, GERAN is able to solve them with ease by performing VHO to UTRAN.
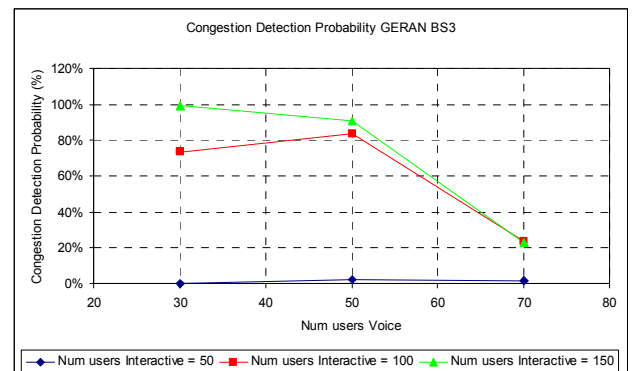

Figure 2 Congestion Detection Probability in GERAN central cell for several service mixes.

If the number of voice and interactive users is increased, congestion situations are more severe and thus they are harder to solve, because more VHO are needed.

Figure 4 shows the average DL reduction factor in the GERAN central cell measured before the congestion was solved (full bullets) and after the congestion was solved (empty bullets). Also the reduction factor threshold is plotted to assess the congestion resolution. Note that, in order to obtain relevant results, the average measurements are computed conditioned that the congestion is solved. Clearly, RF before congestion is solved lies below the threshold, and RF after congestion is solved lies over the threshold. The RF in the UL is less restrictive than in the DL since multislot class assigns fewer slots in the UL as compared to the DL.

Finally, Figure 5 illustrates the average DL throughput per interactive user before the congestion was solved and after the congestion was solved (again, measurements are conditioned to the congestion resolution success). Results are consistent with the fact that congestion control in GERAN aim to lessen the number of data users sharing the same slot. In this way, throughput per user is improved as we can see in Figure 5. It is important to remark that, for this case study, measurements revealed no significant degradation in terms of outage probability in UTRAN. So, in this case study, GERAN congestion is solved without degradation of users in UTRAN.
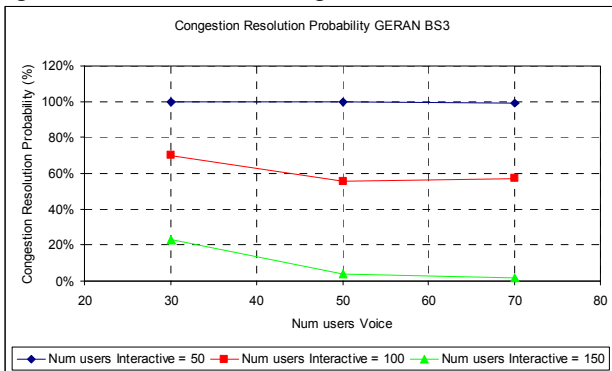

Figure 3 Congestion Resolution Probability when VHO-CR is applied to solve congestion in GERAN central cell for several service mixes.
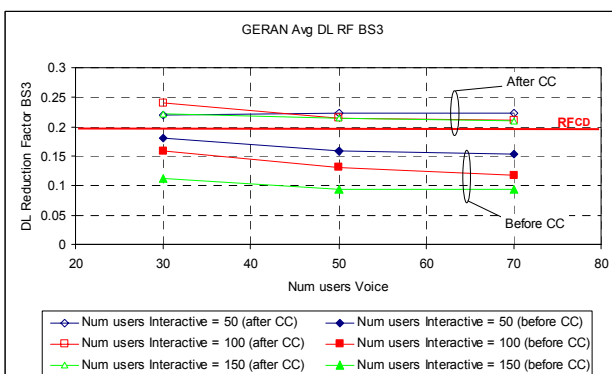

Figure 4 Average Reduction Factor (RF) for the DL before and after applying VHO-CR in GERAN central cell for several service mixes.
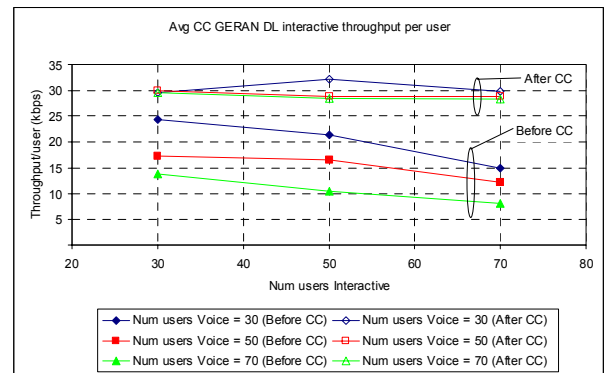

Figure 5 Average DL Throughput per interactive user before and after applying VHO-CR in GERAN central cell for several service mixes.

VI. CONCLUDING REMARKS

This paper has presented a framework for the evaluation and resolution of congestion events in a heterogeneous network comprising GERAN and UTRAN RATs. The mechanisms for congestion detection in both systems have been presented and some strategies aiming to solve congestion have also been addressed. In particular, a VHO procedure has been evaluated for congestion resolution in GERAN. Simulation results revealed that, under certain scenarios, it is possible to solve a congestion situation with minor impact in the QoS of users both in the congested cell/RAT and the destination cell/RAT.

REFERENCES

[1] H. Honkasalo, K. Pehkonen, M.T. Niemi, A.T. Leino, "WCDMA and WLAN for 3G and beyond," *Wireless Communications, IEEE [see also IEEE Personal Communications]* , vol.9, no.2pp. 14- 18, April 2002.

[2] 3GPP TR 25.881 v5.0.0, "Improvement of RRM across RNS and RNS/BSS". http://www.3gpp.org/.

[3] 3GPP TR 25.891 v0.3.0, "Improvement of RRM across RNS and RNS/BSS (Post Rel-5) (Release 6)". http://www.3gpp.org/.

[4] 3GPP 25.922 v6.0.1, "Radio Resource Management Strategies (release 6)". http://www.3gpp.org/.

[5] R. Jain, "Congestion control in computer networks: issues and trends" *Network, IEEE*, vol. 4, no. 3, pp. 24-30, 1990.

[6] J. Pérez-Romero, O. Sallent, R. Agustí, and M. Díaz-Guerra, *Radio Resource Management Strategies in UMTS* , Wiley, 2005.

[7] J. Pérez-Romero, O. Sallent, R. Agustí, "A Novel Approach for a Multicell Load Control in W-CDMA", *5th International Conference on 3G Mobile Communication Technologies*, London, UK, October, 2004.

[8] W. Rave, T. Kohler, J. Voigt, G. Fettweis; "Evaluation of Load Control Strategies in an UTRA/FDD Network", *IEEE 53rd Vehicular Technology Conference Spring*, Rhodes, Greece, May 2001, pp. 2710-2714.

[9] F. Malavasi, M. Breveglieri, L. Vignali, P. Leaves, J. Huschke, "Traffic control algorithms for a multi access network scenario comprising GPRS and UMTS", VTC Spring, Vol. 1, pp. 145-149, 2003.

[10] M. Ermel, et al., "Analytical comparison of different GPRS introduction strategies" in *Proc. of the 3rd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems* Boston, MA, US, 2000, pp. 3-10.

[11] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE performance: evolution towards 3G/UMTS* , Wiley, 2002

[12] 3GPP TS 34.108 "Common Test Environments for User Equipment (UE); conformance testing" http://www.3gpp.org/.

[13] J. Pérez-Romero, O. Sallent, R. Agustí "Policy-based Initial RAT Selection algorithms in Heterogeneous Networks", *7th International Conference on Mobile and Wireless Communications Networks* (MWCN), Marrakesh, 2005.