Operator's RAT Selection Policies Based on the Fittingness Factor Concept

O. Sallent, J. Pérez-Romero
Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
[sallent, jorperez] @ tsc.upc.edu

R. Ljung, P. Karlsson
Telia Sonera
Malmö, Sweden
[Rickard.M.Ljung, Peter.C.Karlsson] @ teliasonera.com

A. Barbaresi
Telecom Italia
Torino, Italy
andrea.barbaresi@telecomitalia.it

Abstract.— **This paper considers a heterogeneous radio access network scenario where different Radio Access Technologies (RATs) are jointly managed. The RAT selection algorithm is based on the fittingness factor concept, which is a generic representation of the different factors that may influence the RAT selection decision. The flexibility provided by the fittingness factor characterization is illustrated in this paper with two different examples of operator's policies for RAT selection, which are implemented by means of different settings of this factor.**

Keywords- Heterogeneous networks, Beyond 3G, UMTS, GERAN, Common Radio Resource Management

I. Introduction

The heterogeneous characteristics that different radio access networks offer in beyond 3G scenarios make possible to exploit the trunking gain resulting from the joint consideration of the different networks as a whole. That is, the additional dimensions introduced by the multiplicity of RATs available provide further flexibility in radio resources management and, consequently, overall improvements may follow with the use of the so-called Common Radio Resource Management (CRRM) strategies [1]-[3]. The scenario heterogeneity is also present from the customer side, because users may access to the requested services by means of a variety of terminals with different capabilities (e.g. single or multi-mode) and different market segments can be identified (e.g. business or consumer users) with their corresponding QoS levels. Then, selecting the proper RAT and cell is a complex problem due to the number of variables involved, as reflected in Figure 1 where some possible inputs are depicted. Furthermore, some of these variables may vary dynamically, which makes the process even more difficult to handle automatically.

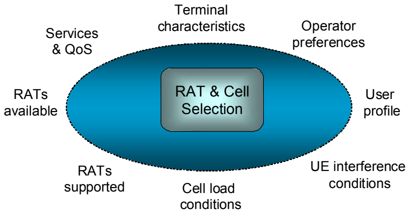

Figure 1.- Factors influencing the RAT and Cell selection.

In this context, CRRM in general and RAT selection algorithms in particular have recently received a lot of attention, clearly acknowledging the key role that these strategies will have for a full realization of Beyond 3G (B3G) scenarios. Research efforts have been oriented either to propose and assess the performance of heuristic algorithms [4]-[7] or to identify architectural and functional aspects for CRRM support [1][8][9].

With all the above, it would be of prime importance to devise a generic framework that takes all these diverse aspects into account and come up with suitable RAT selection principles under any possible circumstance that may arise in a practical implementation. In this respect, in [10] a novel metric was introduced, named *fittingness factor*, which reflects the suitability of selecting each available RAT depending on the multi-dimensional context as well as on the terminal and network capabilities. Then, by abstracting all these multiple aspects into a single and generalized metric, a more clear representation of the dynamic reality can be obtained. Based on the work in [10], this paper presents, on the one hand, a modification to introduce a network-centric component in the fittingness factor definition, and, on the other hand, it addresses the implementation of different operator policies using the framework provided by the algorithm.

The fittingness factor will provide a flexible framework for RAT selection by enabling the implementation of different operator's policies with respect to the management principles to follow. For example, the network operator may target achieving as high as possible capacity, which thereby gives the operator opportunities to reduce the capex and opex due to reduced network deployment requirements. However, the operator may also prefer to utilise the RAT selection possibilities within the heterogeneous network in an advanced manner e.g. in order to handle different service level agreements where the operator is engaged to provide an improved QoS to a given market segment at the expense of a suboptimal capacity.

Furthermore, the mentioned fittingness factor can be particularly useful when considering the foreseen evolution of wireless mobile networks towards IP-based architectures accompanied by more and more decentralized management concepts. In this respect, such a single metric can facilitate the implementation of regional or even cell-by-cell RRM strategies by reducing signalling exchanges. In that sense, the proposed methodology could be extended to cope with LTE (Long Term Evolution) procedures, like the cell re-selection, assuming that the network can broadcast signalling information to the users for load-balancing purposes [11]. Similarly, the proposed framework integrates in a natural way concepts being under discussion in LTE like the handling of UE capabilities [12].

In this context, this paper is organised as follows. Section II develops the fittingness factor concept. Section III describes the corresponding RAT selection algorithm. The algorithm is evaluated under the simulation model described in Section IV and results are provided in Section V. Finally, Section VI summarises the conclusions.

II. Fittingness Factor Concept and Formulation

In order to cope with the multi-dimensional heterogeneity reflected in Figure 1, two main aspects are identified in the RAT selection problem:

1) Capabilities. A user-to-RAT association may not be possible for limitations in e.g. the user terminal capabilities (single-mode terminal) or the type of services supported by the RAT (e.g. videocall is not supported in 2G).

2) Suitability. A user-to-RAT association may or may not be suitable depending on the matching between the user requirements in terms of QoS and the capabilities offered by the RAT (e.g. a business user may require bit rate capabilities feasible on HSDPA and not on GPRS or these capabilities can be realised in one technology or another depending on the RAT occupancy, etc.). There is a number of considerations, which can be split at two different levels:

a) Macroscopic. Radio considerations at cell level such as load level or, equivalently, amount of radio resources available.

b) Microscopic. Radio considerations at local level (i.e. user position) such as path loss, inter-cell interference level. This component will be relevant for the user-to-RAT association when the amount of radio resources required for providing the user with the required QoS significantly depends on the local conditions where the user is (e.g. power level required in WCDMA downlink).

The above concepts can be captured in a novel measurement, the so-called *fittingness factor*, which reflects the degree of adequacy of a given RAT to a given user. The fittingness factor is defined with respect to each cell of the *j*-th RAT for each *i*-th user, who belongs to the *p*-th customer profile requesting the *s*-th service, as follows:

$$\psi_{i,p,s,j} = C_{i,p,s,j} \times Q_{i,p,s,j} \times \delta(\eta_{NF}) \tag{1}$$

In the following subsections the different terms in (1) are further detailed.

A. Capabilities

The first term in (1) reflects the hard constraints posed by the capabilities of either the terminal or the technology, and therefore is defined as:

$$C_{i,p,s,j} = T_{i,p,j} \times S_{s,j} \tag{2}$$

where the terms depending on the terminal and RAT capabilities are defined, respectively, as $T_{i,p,j}$=1 if the *i*-th user's terminal supports the *j*-th RAT and 0 otherwise, and $S_{s,j}$=1 if the *s*-th service is supported by the *j*-th RAT and 0 otherwise.

B. User-centric suitability

The term $Q_{i,p,s,j}$ reflects the suitability of the *j*-th RAT to support the *s*-th service requested by the *i*-th user with the *p*-th customer profile. This terms accounts for a user-specific suitability which depends on the bit rate that can be allocated to the user depending on the existing load and the path loss experienced by the user. The function $Q_{i,p,s,j}$ can be defined empirically or analytically. In particular, in [10] some analytical expressions for the computation of this function for voice, videocall and interactive services in GERAN (GSM/EDGE Radio Access Network) and UTRAN (UMTS Terrestrial Radio Access Network) RATs were presented.

C. Network-centric suitability

The term $\delta(\eta_{NF})$ intends to capture the suitability from an overall RAT perspective, then to provide further flexibility on the fittingness factor definition. Let define the non-flexible load η_{NF} in one RAT as the total load coming from non-flexible traffic, which is the traffic that can only be served through one specific RAT and therefore it does not provide flexibility to CRRM. For instance in UTRAN η_{NF} could be the total load from videocall users assuming they cannot be served through GERAN. Similarly, in case of GERAN, η_{NF} could be the load from mono-mode terminals, which cannot be served through UTRAN. $\delta(\eta_{NF})$ is a function that reduces the fittingness factor of flexible traffic depending on the amount of non-flexible load. The idea is that if there is a high amount of non-flexible load in a given RAT, this RAT is made less attractive for flexible load, thus leaving room to non-flexible users.

A proposed empirical definition of this function would be as:

$$\delta(\eta_{NF}) = \begin{cases} 1 & \text{if } \eta < 1-\min(\eta_{NF},D) \text{ OR traffic is non-flexible} \\ \left(\dfrac{1-\eta}{\min(\eta_{NF},D)}\right)^2 & \text{if } \eta > 1-\min(\eta_{NF},D) \text{ AND traffic is flexible} \end{cases} \tag{3}$$

where η is the average normalised load and *D* (0≤*D*≤1) is a parameter of the function.

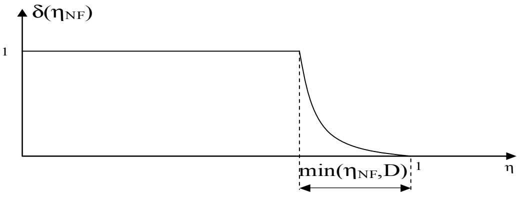

Figure 2.- Network-centric suitability function.

Figure 2 shows the shape of this function when applied for flexible traffic (notice that for non-flexible traffic the function will always be 1) as a function of the total normalised load η and the non-flexible load η_{NF}, which should be averaged over a sufficient amount of time. According to this function, whenever the average normalised load η exceeds the threshold 1-min(η_{NF},D) the fittingness factor for flexible traffic will be reduced, in order that this traffic is preferably handed over to other RATs. Consequently, the parameter min(η_{NF},*D*) is the room that the algorithm tries to keep in the specific RAT for non-flexible load, and it depends on the amount of non-flexible load, in the sense that, if there have not been non-flexible users during a certain amount of time (i.e. η_{NF}=0 after averaging during some period) it is not necessary to reserve room for this type of traffic. Consequently, by controlling the value of the parameter *D* different operator policies can be enforced. Hence, the variation of the parameter *D* is an important parameter for the RAT selection procedure. Different examples of the impact of *D* parameter settings and the impact on the RAT utilisation will be further evaluated in the sections below.

III. RAT Selection Algorithm

Based on the above framework, the proposed RAT selection algorithm for the *i*-th user of the *p*-th profile requesting a given *s*-th service would be as follows:

A. Initial RAT selection

Step 1.- Measure the fittingness factor for each candidate cell k_j of the j-th RAT. Since the measurement should be done separately for uplink and downlink, both measurements can be weighted to obtain a unique measurement:

$$\psi_{i,p,s,j}(k_j)=\alpha_{p,s}\psi^{UL}_{i,p,s,j}(k_j)+\left(1-\alpha_{p,s}\right)\psi^{DL}_{i,p,s,j}(k_j) \quad (4)$$

Here, the weight factor is $\alpha_{p,s}$, depending in general on the specific service and profile, in the sense that for very asymmetric services $\alpha_{p,s}$ should be close to 0 so that the downlink is basically considered in the computation of the total fittingness factor (alternatively close to 1 if the uplink is the most important link). For symmetric services a proper setting could be $\alpha_{p,s}$=0.5, to give the same importance to both links.

Step 2.- Select the RAT J having the cell with the highest fittingness factor among all the candidate cells:

$$J=\max_{j}\left(\max_{k_j}\Psi_{i,p,s,j}\left(k_j\right)\right) \quad (5)$$

Step 3.- Try admission in the RAT J.

Step 4.- If admission is not possible, try with the next RAT in decreasing order with respect to the fittingness factor, provided that its fittingness factor is higher than 0. If no other RATs with fittingness factor higher than 0 exist, block the service request.

In case that two or more RATs have the same value of the fittingness factor, then a decision can be taken based on other criteria (e.g. select the less loaded RAT).

B. Vertical handover

Similarly, the proposed criterion to execute a vertical handover algorithm based on the fittingness factor would be as follows, assuming that the terminal is connected to the RAT denoted as "servingRAT" and the cell denoted as "servingCell".

Step 1.- For each candidate cell and RAT, monitor the corresponding fittingness factor $\psi_{i,p,s,j}(k_j)$. Measures should be averaged during a period *T*.

Step 2.- If the condition $\psi_{i,p,s,j}(k_j) > \psi_{i,p,s,servingRAT}(servingCell)+\Delta_{VHO}$ holds during a period T_{VHO} then a vertical handover to RAT *j* and cell k_j should be triggered, if there are available resources in the cell.

IV. Scenario And Simulation Model

The scenario considered in this paper is built upon a legacy GERAN network, where UMTS has been later on deployed following a co-sited approach, taking as a reference the scenarios defined in the AROMA project [13]. 7 omnidirectional cells for GERAN and 7 for UTRAN are considered. The distance between cell sites is 2km. In case of GERAN, the 7 cells operate with different carrier frequencies. The main parameters of the User Equipment (UE) and the Base Station (BS) are summarised in Table I. All terminals have multi-mode capabilities. The urban macrocell propagation model in [14] is considered for both systems, corresponding to L(dB)=128.1+37.6log(d(km)) with an additional shadowing with standard deviation 10 dB. The mobility model in [15] is considered with speed 3 km/h.

Table I UTRAN BS and UE parameters

BS parameters	**UTRAN**	**GERAN**
Maximum transmitted power	43 dBm	43 dBm
Thermal noise	-104 dBm	-117 dBm
Common Control Channels Power	33 dBm	43 dBm
Maximum DL power per user	41 dBm	N/A
Number of carriers	1	3
UE parameters	**UTRAN**	**GERAN**
Maximum transmitted power	21 dBm	33 dBm
Minimum transmitted power	-44 dBm	0 dBm
Thermal noise	-100 dBm	-113 dBm
DL Orthogonality factor	0.4	N/A
Multislot class (UL, DL, UL+DL)	N/A	2,3,4

Voice, videocall and interactive web browsing services are considered. Voice and video calls are generated according to a Poisson process with an average call rate of 10 calls/h/user and exponentially distributed call duration with an average of 180 s. In UTRAN, the Radio Access Bearer (RAB) for voice users is the 12.2 kb/s speech defined in [16], while for videocall users the bit rate is 64 kb/s. In turn, GERAN does not support the videocall service and voice users are allocated to a TCH-FS (traffic channel full-rate speech), i.e. one time slot in each frame. Interactive users follow the www browsing model given in [15], with 5 pages per session and an average reading time between pages of 20s. In the uplink, there is an average of 25 packets per page, an interarrival packet time 0.05s and an average packet size of 366 bytes. In turn, in the downlink there are 50 packets per page on average, the interarrival packet time is 0.01s and the average packet size is 392 bytes. The average time between user sessions is 30s. It is assumed that half of the interactive users belong to the consumer profile and half to the business profile. WWW browsing service is provided in UTRAN by means of dedicated channels (DCH) using the transport channel type switching procedure. The considered RAB assumes a maximum bit rate of 64 kb/s in the uplink and 128 kb/s in the downlink for consumer users and 384 kb/s for business users [16]. In turn, in GERAN, the www service is provided through a PDCH (Packet Data Channel) with a round robin scheduling algorithm to allocate transmissions to users sharing the same time slot. The algorithm allocates three times more resources to business users than to consumer users in order to have the same bit rate relation than in UTRAN. A link adaptation mechanism operating in periods of 1s is used to select, for each user, the highest modulation and coding scheme (MCS) that ensures the specific sensitivity.

In the UTRAN admission control procedure three conditions are checked [3], namely the uplink load factor after user acceptance should be below 1, the downlink transmitted power below 42 dBm and there must be downlink channel code sequences available. The active set size is 1 with a replacement hysteresis of 3 dB. The time to trigger an horizontal handover is 0.64 s. The target BLER is 1% for voice and videophone and 10% for interactive. In GERAN, voice users are accepted provided that there are available time slots, while interactive users are always accepted at session initiation in idle state. Voice users have precedence over www users, so that slots occupied by www users are allocated to incoming voice users when there are not other free slots. All slots are reversible except the slot 0 of the carrier transmitting the broadcast channel. For www users, a maximum of 8 simultaneous TBFs per uplink slot and 32 TBFs per downlink slot are allowed. Horizontal handover is triggered when the received power is below -100 dBm during 3 samples.

Concerning the fittingness factor evaluation for the vertical handover algorithm, the measurements are averaged in periods of T=1s. The hysteresis margin is Δ_{VHO}=0.1 and T_{VHO}=3s. Furthermore, $\alpha_{p,s}$=0.5 for all services giving equal importance to uplink and downlink in the fittingness factor computation. In turn, the non-flexible load is averaged in periods of 5s and the parameter D will be varied in the simulations.

The simulations consider a total of 100 voice users in the scenario, 200 www users (50% consumer and 50% business) and a variable number of videocall users.

V. Results

In order to illustrate the flexibility of the presented RAT selection framework, in this section, two possible examples of operator's policies are considered:

Policy 1) Let us consider that the operator is interested in promoting 3G services particularly to the business segment through enhanced Internet experience. In this case, providing highly satisfactory QoE (Quality of Experience) for interactive business users is a key target.

Policy 2) Let us consider that the operator is interested in promoting 3G services to mass market through video calls in order to differentiate from legacy 2G services. In this case, providing highly satisfactory videocall service is a key target.

For comparison purposes, a load-balancing RAT selection strategy, where the target is to achieve a load distribution among RATs as evenly as possible, is considered [7]. In this strategy, the users are allocated to the less loaded RAT, provided that the terminal and the service support it (e.g. videocall users are only allocated in UTRAN regardless the load in GERAN).

A. Fittingness factor evolution and service distribution

In order to illustrate the evolution of the fittingness factor under different load conditions, Figure 3 plots the value of average fittingness factor measured in UTRAN for interactive business users for different values of the parameter D in function (3). Notice that, by increasing the value of D the fittingness factor is significantly reduced. This is because with higher D values the algorithm gives higher preference to non-flexible load (videocall in the considered case, which can only by served through UTRAN) and thus UTRAN is made less attractive for flexible load (such as interactive business).

The evolution of the fittingness factor shown in Figure 3 at the end impacts over the traffic distribution between the two RATs, as it is observed in Figure 4 and Figure 5, which plot the percentage of traffic served through GERAN for interactive business and consumer users, respectively. For comparison purposes, the load balancing strategy is also presented. It can be observed that, with the load balancing strategy, for low conversational loads (i.e. 100 videocall users) around half of the interactive traffic is served through GERAN and half through UTRAN. In turn, when increasing the number of videocall users, which are only served through UTRAN, it is more likely that the interactive users find a higher load in UTRAN than in GERAN and therefore most of the interactive traffic is served through GERAN, which applies for both consumer and business users. On the contrary, the behaviour for the fittingness factor based algorithm is quite different. In particular, for the case D=0, most of the interactive business traffic is served through UTRAN (notice in Figure 4 that only for high videocall loads a small fraction around 7% of traffic is served through GERAN). In turn, by increasing the value of the parameter D it is possible to direct more traffic to GERAN when the videocall load increases. In turn, concerning consumer users, there is in general a higher amount of traffic served through GERAN even with the D=0 case (although in general it is lower than with the load balancing case). Similarly, the increase in the parameter D turns into the fact that the traffic served through GERAN is even higher. In the case of voice traffic, results not shown here for the sake of brevity reveal that similar behaviour as for the interactive users is observed. In particular, when increasing the value of D more voice traffic is served through GERAN.

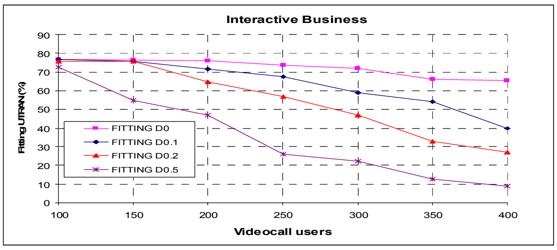

Figure 3.- Fittingness factor of UTRAN for interactive business users.

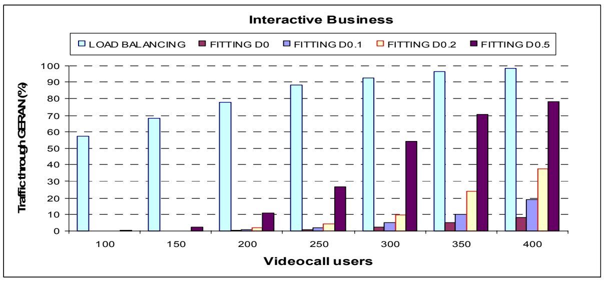

Figure 4.- Fraction of traffic served in GERAN for interactive business users.

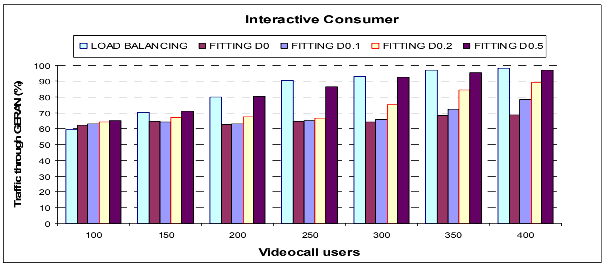

Figure 5.- Fraction of traffic served in GERAN for interactive consumer users.

B. Operator's policies implementation

In the following some illustrative results are presented in order to show how the two considered policies can be implemented by means of a proper setting of the parameter D in the fittingness factor-based CRRM algorithm. As illustrative Key Performance Indicators, the average downlink packet delay for interactive users (shown in Figure 6 and Figure 7 for business and consumer users, respectively) and the throughput for videocall users (shown in Figure 8) are considered.

Figure 6 shows that the strategy based on the fittingness factor algorithm achieves the best performance for interactive business users with D=0.1-0.2. In these cases, remarkable improvements compared to the load balancing case are observed. In turns, for D=0.5, an excessive amount of interactive traffic is moved to GERAN, which finally turns into a delay increase compared to D=0.1-0.2. Nevertheless, even in this case the performance is better than with the load balancing case. On the other hand, notice that the setting D=0 offers a worse performance than D=0.1-0.2 for interactive business users. The reason is that by moving a certain fraction of the interactive traffic to a less loaded RAT like GERAN in case that the videocall load in UTRAN is high, can still be beneficial to interactive users than keeping them in UTRAN. Figure 7 reveals similar conclusions for the consumer users, although in this case the differences are smaller.

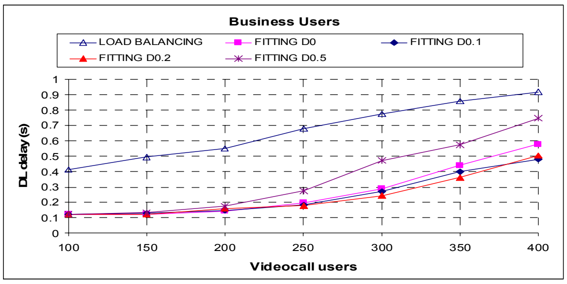

Figure 6.- DL packet delay of interactive business users

Figure 8 shows that setting D to low values is not beneficial for videocalls, since in this case an excessive amount of interactive users are served through UTRAN, thus leaving less room for videocall connections. In this case, a load balancing strategy, in which, for high videocall loads, eventually most of the flexible traffic will be served through GERAN, turns out to be beneficial for videocall users. However, a similar performance for videocall users can also be achieved with D=0.5, which provides the additional advantage that the delay of interactive users is lower than with load balancing.

Bearing all these results in mind, Policy 1, intending to provide enhanced Internet experience for business users, would be implemented with D=0.1-0.2. In turn, Policy 2, intending to promote videocalls, would be enforced with D=0.5.

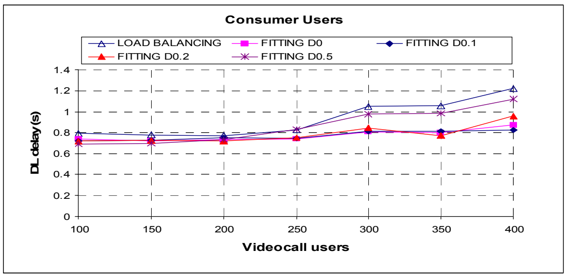

Figure 7.- DL packet delay of interactive consumer users

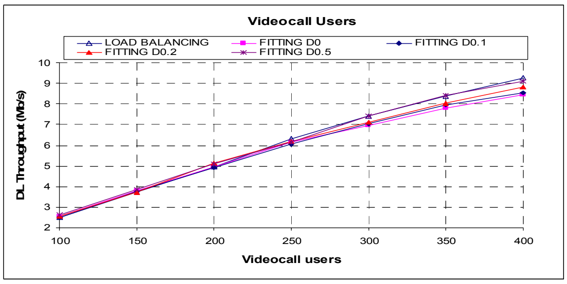

Figure 8.- Throughput of videocall users

VI. CONCLUSIONS

This paper has presented a novel strategy based on the fittingness factor concept for RAT selection in heterogeneous wireless networks. By integrating both user and network-centric suitability concepts, it provides a flexible framework enabling the implementation of different operator's policies by controlling a specific parameter of the algorithm.

ACKNOWLEDGEMENTS

This work has been performed in the framework of the project IST-AROMA (http:// www.aroma-ist.upc.edu), which is funded by the European Commission.

REFERENCES

[1] 3GPP TR 25.881 v5.0.0 "Improvement of RRM across RNS and RNS/BSS (Release 5)"

[2] 3GPP TR 25.891 v6.0.0 "Improvement of RRM across RNS and RNS/BSS (Release 6)"

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

[4] A. Tölli, P. Hakalin, "Adaptive load balancing between multiple cell layers", IEEE VTC Fall, Vol. 3, Sept. 2002, pp.1691 – 1695.

[5] A. Pillekeit, F. Derakhshan, E. Jugl, A. Mitschele-Thiel, "Force-based load balancing in co-located UMTS/GSM networks", VTC 2004-Fall. 2004 IEEE 60th Vol. 6, 26-29 Sept. 2004 pp. 4402 – 4406.

[6] S. Lincke-Salecker, "The Benefits of Load Sharing when Dimensioning Networks", Proceedings of the 37th Annual Simulation Symposium (ANSS'04), April, 2004.

[7] X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "On the suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks", Wireless Personal Multimedia Communications Symposium (WPMC'05), Aalborg, Denmark, September, 2005.

[8] W. Zhuang, Y. Gan, K. Loh, K. Chua, "Policy-Based QoS Management Architecture in an Integrated UMTS and WLAN Environment", IEEE Communications Magazine, November, 2003, pp. 118-125.

[9] J. Pérez-Romero et al. "Common Radio Resource Management: Functional Models and Implementation Requirements", IEEE PIMRC Conference, Berlin, September, 2005.

[10] J. Pérez-Romero, O.Sallent, R.Agustí, "A Novel Metric for Context-Aware RAT Selection in Wireless Multi-Access Systems", ICC conference, Glasgow, 2007.

[11] 3GPP R2-070050, "On Intra-LTE Cell Reselection Methods", 3GPP TSG-RAN WG2 Meeting #56-bis, January, 2007.

[12] 3GPP R2-070066, "Framework for UE capability handling in LTE", 3GPP TSG-RAN WG2 Meeting #56-bis, January, 2007.

[13] R. M. Ljung et al., "D05 Target Scenarios specification: vision at project stage 1", Deliverable of the AROMA project, April, 2006. Available at http:// www.aroma-ist.upc.edu

[14] 3GPP TR 25.942 "Radio Frequency (RF) system scenarios"

[15] UMTS 30.03 v3.2.0 TR 101 112 "Selection procedures for the choice of radio transmission technologies of the UMTS", ETSI, April, 1998.

[16] 3GPP TS 34.108 "Common Test Environments for User Equipment (UE); conformance testing"