A new CRRM Scheduling Algorithm for heterogeneous networks using Hopfield Neural Networks

N. García [(1)], J. Pérez-Romero [(2)], R. Agustí [(2)],

[(1)] Universitat Pompeu Fabra (UPF), Barcelona, Spain, E-mail : nuria.garcia@upf.edu

[(2)] Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, E-mail : [jorperez, ramon] @ tsc.upc.edu

Abstract.- **This paper proposes a novel Common Radio Resource Management (CRRM) algorithm for heterogeneous scenarios making use of the Hopfield Neural Network methodology, which provides a fast way of finding the optimum resource allocation that minimises a given energy function reflecting specific service and system constraints. The proposed algorithm is applied in a heterogeneous wireless scenario with CDMA and TDMA radio access networks to schedule the downlink transmissions of a delay-constrained service. The algorithm is evaluated by means of realistic simulations and compared with a reference scheme, revealing its ability to adapt to the specific service and traffic conditions.**

I. Introduction

The heterogeneous network concept is intended to propose a flexible and open architecture for a large variety of wireless access technologies, applications and services with different QoS demands. A typical heterogeneous scenario is constituted by several Radio Access Technologies (RATs) each having a Radio Access Network (RAN) interfacing a common Core Network (CN). RANs include cellular networks, e.g. UTRAN (UMTS Terrestrial Radio Access Network) and GERAN (GSM/EDGE Radio Access Network), as well as non-cellular systems, e.g. WLAN.

Each radio access network differs from each other by air interface technology, cell-size price, access, coverage and ownership. In spite of these basic differences, many services can be carried over any of the deployed RANs. Then, the complementary characteristics that these networks offer make possible to exploit the trunking gain resulting from the joint consideration of the different networks as a whole. Thus interworking among heterogeneous RANs leads to a better overall performance than the accumulated performances of the stand-alone systems. This challenge calls for the introduction of new radio resource management (RRM) algorithms operating from a common perspective that take into account the overall amount of resources offered by the available RANs, and therefore are referred to as CRRM (Common Radio Resource Management) algorithms [1][2].

In this scenario, whenever there is more than one radio access technology that can provide the requested services, it is needed to decide the RAN that fits better to the user constraints. In that sense, one of the key aspects to achieve an efficient usage of the available radio resources while at the same time keeping the desired QoS constraints for the different services is to execute a Dynamic Resource Allocation (DRA) that takes into account both QoS requirements and channel status for each user. DRA in a heterogeneous scenario can be considered as a combinational problem where radio resources (e.g. bandwidth in each RAN) have to be allocated to several users at each frame subject to certain restrictions in terms of Quality of Service (QoS) and of total amount of available resources. In that sense, it is known that a Hopfield Neural Network (HNN) searches for the combinational solution that minimizes an Energy Function under specific constraints [3]. In addition to that, it is claimed that real time operation at a very low time scale could be provided either by means of Hardware HNN implementations or by means of numerical algorithms. HNN have been used in different works in the literature. In [4] a HNN is proposed to manage the multirate and multiservice structure of an ATM wireless network. The convergence problems appearing in the classical HNN algorithm are overcome by using new terms in the Energy formulation. In turn, [5] shows the good HNN behavior when managing different service profiles for a user centric approach in the uplink of a CDMA system, in which each user has a range of bit rates that can be allocated depending on its service profile, defined by a minimum satisfaction bit rate. In [6] the basis of the HNN is captured for a downlink scheduling problem in a CDMA packet network where sensitive delay guarantees are to be provided to the users sharing the access.

Taking into account the above framework, in this paper the basis of the HNN is captured for a downlink scheduling problem in a CRRM scenario. The main overall goal is not necessarily to achieve the maximum throughput, but to deliver each packet without exceeding a specific time deadline using the best available RAN. The optimum solution is found by solving a nonlinear constrained optimization problem. The HNN here proposed is claimed to do that in a very low time scale of several milliseconds (the typical value for a transmission time interval in a wireless system).

The rest of the paper is organised as follows. Section II presents the problem formulation and section III provides the proposed HNN-based scheduling approach, which will be compared against a reference scheme presented in section IV. Results are presented in section V and finally conclusions and future work are summarised in section VI.

II. Problem Formulation

The considered DRA problem assumes a set of N users, i=1,..., N, with their corresponding queues, located at the base station of the access network, which contain the packets pending to be transmitted in the downlink direction of a multiple-RAN system, as illustrated in Fig. 1. It is considered that non-shaped traffic is arriving to the queues so that all the incurred packet delay is introduced at the network level. It is assumed also that there are P available RANs, p=1,...,P. The scheduling algorithm operates in frames of T s and allocates a certain bit rate to each user from a set of M possible bit rates, j=1,...,M in a given RAN. Multiple transmissions of different users can be allocated in each frame for each RAN multiplexed on a TDMA or a CDMA way, depending on the specific characteristics of the RAN. It will be assumed that a user can not be simultaneously allocated in more than one RAN.

The bit rate allocation will be executed by means of an optimal mechanism based on HNN through the minimisation of a properly defined Energy Function. This Energy includes the Cost function associated to each Bit Rate usage and a proper formulation for the network restrictions, including fairness and an optimal partitioning of the total available bandwidth, which depends on the specific characteristics of the considered radio access technology. Particularly, from a general point of view, the set of bit rates allocated at a given instant in the p-th RAN must fulfil the relationship.

$$\sum_{i=1}^{N}\sum_{j=1}^{M} R_{b,i,j,p}V_{i,j,p} \le B_p \qquad (1)$$

where $R_{b,i,j,p}$ is the j-th bit rate of the i-th user in the p-th RAT and $V_{i,j,p}$ takes the value 1 if the j-th bit rate has been allocated to the i-th user in the p-th RAN and 0 otherwise. In turn, B_p is the total available bandwidth in the p-th RAN. Then, the efficient exploitation of the total available bandwidth B_p in terms of the allocated bit rate to the different users can be obtained by leading (1) to equality.

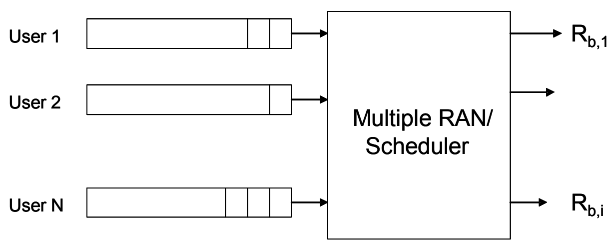

Fig. 1 Joint Packet Scheduler

With respect to the queue model, let us assume that at the beginning of the k-th frame the i-th user has L_i packets waiting for transmission, as depicted in Fig. 2. $l_{i,m}^{k}$ denotes the number of bits of the m-th packet of the i-th user in the k-th frame.

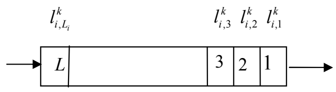

Fig. 2 Queue of the i-th user

Assuming a FIFO policy for the packets in the queue of each user, the amount of bits that should be transmitted until the complete transmission of the m-th packet of the i-th user is given by:

$$B_{i,m,left}^{k} = \sum_{n=1}^{m} l_{i,n}^{k} \qquad (2)$$

On the other hand, the delay constraint for each packet is given by $D_{max,i}$, measured as the maximum packet delay specified in the contract of each user, and measured in frames. Let $f_{i,m}^{k}$ be the elapsed time at the beginning of the k-th frame since the arrival of the m-th packet in the queue of the i-th user. Then, the maximum time-out left for transmission of this packet is given by:

$$TO_{i,m}^{k} = D_{\max,i} - f_{i,m}^{k} \qquad (3)$$

Consequently, the minimum bit rate required to guarantee the transmission in due time is given by:

$$v_{i,m}^{k} = \frac{\sum_{n=1}^{m} l_{i,n}^{k}}{TO_{i,m}^{k}} = \frac{B_{i,m,left}^{k}}{TO_{i,m}^{k}} \qquad (4)$$

We define the *Optimum Bit Rate* (OBR) for the i-th user in the k-th frame as the one that allows transmitting all the packets in due time, given by:

$$R_{b,i,opt}^{k} = \max_{m=1,\cdots,L}\left\{v_{i,m}^{k}\right\} \qquad (5)$$

If this rate was continuously allocated this would ensure that all the packets would be transmitted on due time. However, this cannot always be guaranteed for all the users because of the total bandwidth restrictions, which could lead to some packet drops if they cannot be served in due time. Consequently, a smart bandwidth allocation algorithm must ensure that the bit rate allocated to all the users is as close as possible to the optimum bit rate to minimise the packet drop ratio.

III. HNN-BASED SCHEDULING MODEL

The proposed scheduling algorithm tries to ensure that the total bandwidth is efficiently used while at the same time it takes into account the time-out requirements of the different packets to ensure an optimal allocation. As stated in the previous section, the problem consists in selecting, from the predefined set of bit rates j=1,...,M at the p-th RAN, the bit rate $R_{b,i,j,p}^{k}$ to be allocated in the k-th frame to the i-th user, as well as the appropriate p-th RAN. A user can only be allocated in one RAN. Then, a dynamic cost function $C_{i,j,p}^{k}$ is defined for each possible bit rate as:

$$C^k_{i,j,p} = \begin{cases} 1 & \text{if} \quad R^k_{b,i,j,p} < R^k_{b,i,p,opt}(1-a) \\ \dfrac{R^k_{b,i,j,p}\ TO^k_i - B^k_{i,pleft}(1-a)}{B^k_{i,pleft}(1+a) - B^k_{i,pleft}(1-a)} & \text{if} \quad R^k_{b,i,opt}(1-a) \le R^k_{b,i,j,p} \le R^k_{b,i,opt}(1+a) \\ 0 & \text{if} \quad R^k_{b,i,p,opt}(1+a) < R^k_{b,i,j,p} \end{cases} \quad (6)$$

where TO^k_i and $B^k_{i,left}$ are the time-out and the number of bits to be transmitted, respectively, corresponding to the most restrictive packet in the queue of the *i*-th user, i.e. the one having the maximum value of $v^k_{i,m}$ as defined in (4). In turn, *a* is a parameter to be set.

The above DRA problem subject to the mentioned restrictions can be formulated in terms of a three-dimensional L=N×M×P neural network. In such a 3D HNN each neuron is modelled as a nonlinear device with a sigmoid monotonically increasing function defined by the logistic function:

$$V_{i,j,p} = f(U_{i,j,p}) = \frac{1}{1+e^{-\alpha U_{i,j,p}}} \quad (7)$$

where $U_{i,j,p}$ and $V_{i,j,p}$ are the input and output, respectively, of the *i,j,p*-th neuron, and α is the corresponding gain of the amplifier of the neuron. The optimization process of the HNN is carried out on a frame basis and relies on minimizing an energy function E through the convergence of the expression:

$$\frac{dU_{i,j,p}}{dt} = -\frac{U_{i,j,p}}{\tau} - \frac{\partial E}{\partial V_{i,j,p}} \quad (8)$$

which characterizes the HNN behaviour [7]. The minima of the Energy occur at the 3L corners inside the L-dimensional hypercube defined on $V_{i,j,p} \in \{0,1\}$ [8]. By solving numerically (8) and after reaching a stable state each neuron is either ON (i.e. 1 if $V_{i,j,p}$ is greater or equal than 0.5) or OFF (i.e. 0, if $V_{i,j,p}$ is lower than 0.5).

The numerical iterative solution of (8) is obtained following the Euler technique as:

$$U_{ijp}(n+1) = U_{ijp}(n) + \Delta\left[-\frac{U_{ijp}(n)}{\tau} - \frac{\partial E}{\partial V_{ijp}}\right] \quad (9)$$

The neuron's voltage is updated at the *n*-th iteration with the setting $\Delta=10^{-4}$. After reaching a stable state each neuron is either ON or OFF.

The iterative numerical solution in (9) converges when the iterations *n* and *n*-1 satisfy

$$\left\| V^n - V^{n-1} \right\|_2 < \varepsilon \quad (10)$$

being $\left\| \ \right\|_2$ the Euclidean Norm, provided that only one element of each row of the V matrix is greater than 0.5. We have set $\varepsilon = 5\cdot 10^{-6}$. If all these conditions are fulfilled we decide the process converges and the matrix $V_{i,j,p}$ provides us the $R_{b,i,j,p}$ values.

It is worth mentioning that, due to the constraints in the considered DRA problem, the convergence is not mathematically guaranteed. Keeping this in mind, and from a practical point of view, a maximum number of iterations $N_{max}=1000$ is considered in the iterative solution of (9). If this maximum is reached it is assumed that the procedure has not converged and the $R_{b,i,j,p}$ values are decided to be the same obtained in the last frame. In any case, it should be noted that, under this strategy, only a 0.005% of the frames simulated in the results presented in the next Section failed to converge.

According to [4] a valid expression for the energy function E follows as:

$$\begin{aligned} E = &\frac{\mu_1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{p=1}^{P} C^k_{i,j,p} V_{i,j,p} + \sum_{p=1}^{P}\frac{\eta^{\varsigma_p}\mu_2}{2}\left|1-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{R_{b,i,j,p}}{B_p}V_{i,j,p}\right| \\ &+\frac{\mu_3}{2}\sum_{p=1}^{P}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{i,j,p}V_{i,j,p} + \frac{\mu_4}{2}\sum_{p=1}^{P}\sum_{i=1}^{N}\sum_{j=1}^{M}V_{i,j,p}\left(1-V_{i,j,p}\right) \\ &+\frac{\mu_5}{2}\sum_{i=1}^{N}\left(1-\sum_{p=1}^{P}\sum_{j=1}^{M}V_{i,j,p}\right)^2 \end{aligned} \quad (11)$$

The minimisation of this function provides the allocation of the *i*-th user given by $V_{i,j,p}$, which takes the value 1 for the *j*-th bit rate (i.e. $R_{b,i,j,p}$) and *p*-th RAN and 0 otherwise. The first summand in (11) contains the cost function defined in (6). In turn, the second summand penalizes the undesired situations in which the total allocated bandwidth in a given RAN is higher than B_p, and, at the same time, it drives expression (1) to the equality, so that resources are efficiently used. In this term, the following exponent is defined:

$$\varsigma_p = u\left(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{R_{b,i,j,p}}{B_p}V_{i,j,p} - 1\right) \quad (12)$$

where u(.) is the unit step function.

The third summand in (11) simply penalizes non allowed bit rates because of e.g. the user profile. That is, for the bit rates $R_{b,i,j,p}$ of the *i*-th user included in its profile, $\psi_{i,j,p}=0$, while for the bit rates not included in the profile $\psi_{i,j,p}=1$. The fourth term forces convergence towards $V_{i,j,p} \in \{0,1\}$ and the last term forces the physical condition that only one bit rate $R_{b,i,j,p}$ and RAN are possible for the *i*-th connection.

On the other hand, one problem to face up in wireless networks is the ping pong effect. This is a typical effect in Horizontal Handover procedures coming up when, due to the pilot signal fluctuations, the user point of attachment is transferred repeatedly from the old base station to the new one and vice versa. Such non desired intermediate handovers cause a signalling and hence a resource waste and must be avoided as much as possible. When CRRM is considered, a Vertical Handover is said to occur when the user is moved from one to another RAN. Then similar signalling expense could be demanded if the RAN change assignment is too often executed. Furthermore, it could happen that this RAN handover requires a minimum execution time due to implementation constraints. Then, because of these reasons we will define T_{VH} as the minimum time a user has to be anchored to a given RAN once it is assigned. The obtained final scheduling performances obtained when this time is considered are expected to be degraded with respect to the ideal theoretical case T_{VH}=0 (in which the RAN can be changed from frame to frame) and the purpose here will be to assess this degradation for different T_{VH} values. In practice, we can use the $\psi_{i,j,p}$ value in the third summand of (11) to manage this T_{VH} constraint. Particularly, provided that the *i*-th user is currently allocated in the *p*-th RAN, then we will set $\psi_{i,j,r}$ =1 for any other *r*-th RAN ($r \neq p$) during T_{VH} seconds after the initial *p*-th RAN assignment. In this way, the algorithm will keep the *i*-th user allocated in the same RAN during T_{VH}.

IV. Reference Scheduling Scheme (RSS)

For comparison purposes, a reference scheduling algorithm has been considered. It operates under the same assumptions regarding the previous constraint delay and T_{VH} setting but, instead of making the optimisation procedure, it simply allocates to each user its optimum bit rate OBR as defined in (5). The algorithm operates then in the following steps in the k-th frame:

Step 1.- Order the users in increasing value of $R^k_{b,i,opt}$.

Step 2.- Allocate the RAT having the lower number of connected users at the allocation time and keep it during T_{VH} for a given user.

Step 3.- Allocate sequentially to each user the first available bit rate higher or equal than $R^k_{b,i,opt}$ (i.e. OBR) until having exhausted the total bandwidth B_T.

Step 4.- Once the bandwidth is exhausted, allocate a bit rate =0 kb/s (i.e. no transmission) to the remaining users.

V. Results

This section presents some illustrative results obtained with the proposed allocation strategy. Two RANs have been selected for illustration purposes, namely a CDMA-based and a TDMA-based RAN. The wireless network scenario consists of one isolated circular cell with radius 0.5 km where the two RANs are co-sited. The propagation and mobility models defined in [9] are considered. In particular, the path loss at distance d is given by L(dB)= 128.1 + 37.6 $\log_{10}$(d), with a log-normal distributed shadowing with deviation 10 dB. Users are randomly distributed. The noise power is P_N=-102 dBm.

The model for the downlink of CDMA is described in [6] and considers a single isolated cell transmitting at maximum power P_{Tmax}, i.e. assuming that the base station operates at the capacity limit, and that background noise is negligible with respect to intracell interference. Then, assuming the same (Eb/No) requirement for all the users, it is shown in [6] that the power constraint can be translated into a bit rate constraint given by:

$$\sum_{i=1}^{N} R_{b,i} \leq \frac{\beta W}{\rho\left(\frac{E_b}{N_o}\right)} = B_{CDMA} \tag{13}$$

where W is the transmission bandwidth W=3.84 Mchips/s, the orthogonality factor is ρ =0.4 and (1-β) is the fraction of total power devoted to common control channels, with β=0.95. The maximum power available at the base station is 43 dBm. Furthermore, the Eb/No target is 5 dB which leads to an available bandwidth B_{CDMA}= 2.88Mb/s. The available set of possible bit rates is given by: {384 kb/s, 256 kb/s, 128 kb/s, 64 kb/s, 32 kb/s and 16 kb/s}. With respect to TDMA-based RAN the EDGE system [10] is taken for illustrative purposes with a total bandwidth of B_{TDMA}=1920 kb/s corresponding to 4 carriers with MCS-9. Assuming a multislot capability of 8 slots (i.e. 8 slots per frame can be allocated to the same user terminal) and MCS-9 the available bit rates for a given mobile are approximated by 60*j* kb/s where *j*=1,..,8.

The selected parameters appearing in the formulation of the HNN are μ_1=1000, μ_2=4000, μ_3=8000, μ_4=800, μ_5=6000, τ=1, α=1.0 and η=10. These values have been set based on [4]. The frame period is set to 10 ms and the parameter *a* in the cost function (6) takes the value 0.1. An interactive service has been considered for simulation purposes and a www traffic model has been assumed [9] as representative. Specifically, www sessions are composed by an average of 5 pages, with an average time between pages of 30s. In each page, the average number of packets is 25 with an average time between packets of 0.0277s. The packet length follows a Pareto with cut-off distribution with parameters α=1.1, k=81.5 and m=6000 bytes. The average time between www sessions is 0.1s (i.e. it is assumed that a user is continuously generating sessions).

Two user classes, namely Class 1 and Class 2, have been considered, as representative of two different user profiles, with maximum allowed delays of 120 ms and 60ms, respectively. The proposed framework allows them to be jointly managed. Fig. 3 shows the average packet loss ratio (i.e. the ratio of packets dropped because of exceeding the deadline) versus the number of users for different T_{VH} settings for class 1 users. It can be observed that the performance degrades when increasing the value of T_{VH}. Fig. 4 shows the comparison between the proposed HNN strategy and the reference scheme (RSS) in terms of average packet dropping ratio, revealing the improvement attained when a HNN is used. The same trend is observed in Fig. 5, which compares the Cumulative Function Distribution (CDF) of the delay for the two classes of users. Fig. 6 plots the histogram of bit rate assignment in the CDMA RANs, with HNN and RSS. Notice that, for a high number of users, RSS assigns the higher bandwidths at the extent of leaving without service (i.e. bit rate=0) most of the users. On the contrary, HNN tries to serve all the users so as all of them receive some no null bit rate, thus offering better performances. Similar results would apply to TDMA

VI. Conclusions

In this paper a novel HNN based scheduling procedure for a constrained delay service has been proposed for the common radio resource management of multiple RANs. The methodology incorporates in a native way user differentiation profiles as well as delay constraints in the RAN switching. Results have been shown for a scenario with CDMA and TDMA RANs that compare favourably with a reference scheduling procedure. The algorithm allows also taking into consideration timing constraints to avoid continuous changes of RAN for a given user, and it has been shown the impact of this constraint in the obtained performance.

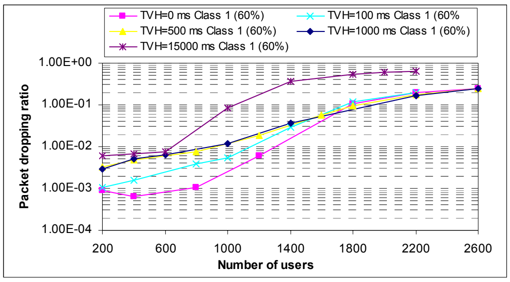

Fig. 3 Average packet dropping for different values of T_{VH}

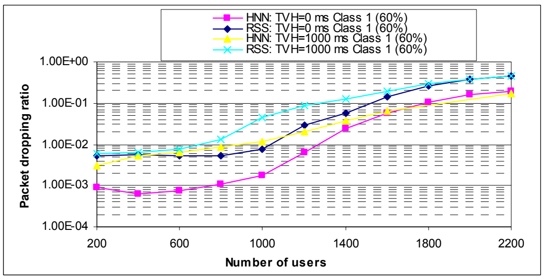

Fig. 4 Comparison with the Reference Scheme RSS

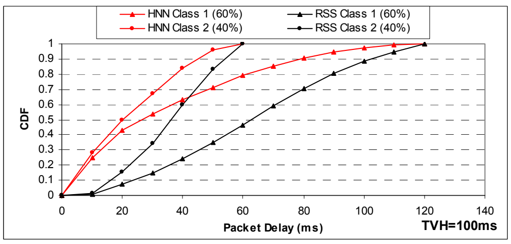

Fig. 5 Cumulative Distribution Function of the packet delay for a situation with 1200 users (60% class 1 and 40% class 2)

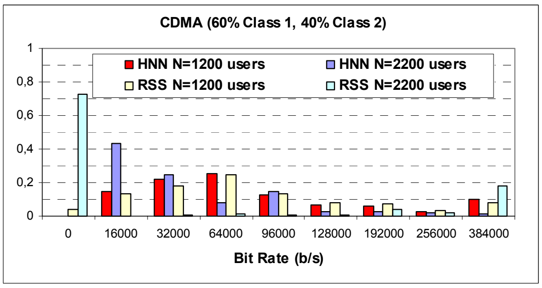

Fig. 6 Histogram of the utilisation of the different bit rates in CDMA with HNN and RSS for a traffic mix with 60% of class 1 users and 40% of class 2 users

Acknowledgements

This work is partially funded by the European Network of Excellence NEWCOM (Contract number 507325), and by the COSMOS grant (ref. TEC2004-00518, Spanish Ministry of Science and Education and European Regional Development Fund).

References

[1] 3GPP TR 25.881 v5.0.0 “Improvement of RRM across RNS and RNS/BSS”

[2] A. Tölli, P. Hakalin, H. Holma, “Performance Evaluation of Common Radio Resource Management (CRRM)”, ICC Conference, Vol. 5, April, 2002, pp. 3429-3433.

[3] Shigeo Abe, *Neural Networks and Fuzzy Systems. Theory and Applications*, Kluwer Academic Publishers, 1997

[4] C.W.Ahan, R.S.Ramakrishna, “QoS Provisioning Dynamic Connection-Admission Control for Multimedia Wireless Networks Using Hopfield Neural Networks”, IEEE Transactions on Vehicular Technology, Vol 53, Jan 2004

[5] N. García , R. Agustí, J. Pérez-Romero, “A User-Centric Approach for Dynamic Resource Allocation in CDMA systems based on Hopfield Neural Networks”, IST Summit Dresden June 2005

[6] N. García, J. Pérez-Romero, R.Agustí, “A novel scheduling algorithm for delay-oriented services based on Hopfield Neural Networks methodology”, IEEE WCNC Conference, Las Vegas, USA, 2006.

[7] Haykin, *Neural Networks. A Comprehensive Foundation*, New York: Mac Millan,1994

[8] J.J. Hopfield , “Networks with graded response have collective computational properties like those of two state neurons”, Proc. Nat. Acad. Sci., vol 81, May 1984

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

[10] T. Halonen, J. Romero, J. Melero, *GSM,GPRS and EDGE Performance*, John Wiley & Sons, 2002.