A Dynamic Resource Allocation Algorithm for WCDMA Systems with Delay Constraints based on Hopfield Neural Networks

Daniel Calabuig[1], David Gómez-Barquero[1†], José Monserrat[1], and Nuria García [2]
[1]*Polytechnic University of Valencia (UPV) – Mobile Communications Group, Spain*
[2]*University Pompeu Fabra (UPF), Barcelona, Spain*
{dacaso, dagobar, jomondel}@iteam.upv.es

Abstract
This paper proposes an algorithm for Dynamic Resource Allocation (DRA) in uplink WCDMA systems for non-real time services. The main objectives of DRA are to ensure a minimum Quality of Service (QoS) for all users, and to maximize the utilization of the radio resources. Previous works have proved the utility of Hopfield Neural Networks (HNN) to implement DRA algorithms to maximize fairly the total allocated bandwidth among users. This paper presents an enhanced version of these algorithms by introducing a packet delay control technique, guaranteeing a better QoS.

I. Introduction

Nowadays mobile wireless communication systems are characterized by offering a wide range of services (speech, video, data download, multimedia, etc.). Services are characterized by different user profiles with distinct Quality of Service (QoS) parameters, e.g. minimum bit rate and maximum packet delay, and differ in the amount of resources required. In these increasingly complex scenarios, the need of an efficient Radio Resource Management (RRM) becomes crucial to maximize the utilization of the radio resources while assuring a minimum QoS. The Dynamic Resource Allocation (DRA) algorithm arises as a key element in the system, since it will determine the amount of radio resources allocated to each user at each instant.

The DRA problem has been widely studied in the literature, e.g. [1]-[5] and references therein. Hopfield Neural Networks (HNN) have been successfully employed to implement DRA algorithms, e.g. [3] for GSM and GPRS and [4], [5] for UMTS. The main benefit of the HNN is the high computational speed of its hardware implementation, that permits real-time running of the algorithm, what is not possible with other analytical solutions [6].

Previous works as [4] and [5] have tried to give a solution to the DRA problem by using HNN, to maximize the total allocated uplink bandwidth, and with the introduction of a cost function, they also try to obtain a proportional allocation [4] or to allocate the bit rates over the user satisfaction [5]. However, they only solve the DRA partially, not guaranteeing a global QoS (i.e. packet delay is not considered). This paper proposes a DRA HNN-based algorithm, based on [4] and [5], to consider both bit rate and delay constraints, being thus possible to provide services with controlled bit rate and delay. The delay control technique adopted is based on [7], and prioritizes users by their buffer size.

The algorithm proposed has a user-centric approach, since bit rates are not only allocated by network constraints, but users' expectations and requirements. This feature allows maximize the utilization of the radio resources as a function of the service profiles of the users [5].

The rest of the paper is organized as follows. Section II presents the WCDMA constraints that the DRA algorithm should take into account: maximum uplink system load, maximum terminal power, minimum user bit rate, and maximum packet delay. Section III presents the solution adopted for the DRA problem based on a HNN. Section IV presents the proposed DRA algorithm. Section V shows some numerical results obtained by simulation. Finally, the most important conclusions are drawn in section VI.

II. DRA Constraints

This paper assumes a WCDMA system where a set of possible bit rates are allowed. Each user is characterized by a subset of possible bit rates, defined by the type of service he is subscribed to. For the sake of simplicity, and in order to compare results with [5], a single isolated cell has been considered.

If N is the number of users demanding for resources, and M is the number of possible bit rates, the DRA algorithm shall find the optimal bit rate $R_{b,i}$, i=1,…,N, for each user satisfying the following constraints. The proposed DRA algorithm works on a frame by frame basis.

II.1. Load Constraint

In WCDMA systems the total load in the uplink is controlled to avoid exceeding a predefined value $\eta_{UL\max}$ to ensure a minimum coverage of the service area. The load parameter represents the interference experimented by the base station that comes mainly from all the code-multiplexed users in its cell. Limiting these interferences to a maximum value ensures a certain QoS since WCDMA systems are usually interference-limited.

The total load in an isolated cell can be calculated as:

$$\eta_{UL} = \sum_{i=1}^{N} \frac{1}{1+\frac{W}{\left(\frac{E_b}{N_0}\right)_i R_{b,i}}} \le \eta_{UL\max} \qquad (1)$$

where W is the total transmission bandwidth and $(E_b/N_0)_i$ is the energy per bit to spectral density noise power ratio of the i-th user.

A new connection is only accepted if the load of the system with all the current users transmitting at the lowest bit rate of their subset does not exceed a certain threshold,

† Supported by a PhD scholarship from the Generalitat de Valencia (Spain)

called admission control threshold, lower than η_{ULmax}. This threshold is introduced as a security margin to allow the system to allocate higher bit rates in order to transmit packets with high delay near deadline.

II.2.Power Constraint

The power constraint takes into account the maximum power that mobile terminals can transmit, P_{Tmax}. The transmitted terminal power can be approximated by:

$$P_{T,i} = \left(\frac{E_b}{N_0}\right)_i \Bigg|_{\text{target}} \frac{R_{b,i} L_{p,i} P_N}{W(1-\eta_{UL})} \le P_{T\max} \qquad (2)$$

where $L_{p,i}$ is the path loss between the *i*-th user and the base station and P_N is the thermal noise.

When a terminal needs more power than the maximum in order to guarantee the $(E_b/N_0)_i$ target, it is out of coverage for that bit rate $R_{b,i}$. In that case, the algorithm will allocate a lower bit rate if it is available in the subset, otherwise the connection is lost. In any case, these dropped connections would become usually handovers in a real system with more than one cell.

II.3.Bit Rate Constraint

This algorithm only allocates to each user one of the bit rates predefined in the associated subset. The minimal bit rate ensured is thus the minimal bit rate in the subset.

II.4.Delay Constraint

QoS is not only defined by the minimal bit rate but also by the maximum packet delay. So far the aforementioned constraints have been considered as hard constraints, i.e. if they are exceeded they result in dropped connections. However, the delay is considered as a soft constraint since it does not result in any dropped connection. Packets exceeding their maximum delay are dropped but the connections remain on.

The delay control technique adopted is based on the Service Credit (SCr) concept proposed in [7]. The DRA algorithm intends to transmit all the traffic packets generated by the users before the maximum service delay. If the packet's queue of one user grows, it means that the bit rate allocated to that user is not high enough, and the system should reallocate a higher bit rate in detriment of other users with empty (or shorter) queues.

The SCr of each user is updated every frame following:

$$SCr_i^{(k)} = u\left(SCr_i^{(k-1)} + \frac{bits_i^{(k-1)}}{T} - R_{b,i}^{(k-1)}\right) \cdot \left(SCr_i^{(k-1)} + \frac{bits_i^{(k-1)}}{T} - R_{b,i}^{(k-1)}\right) \text{ (bits/s)} \qquad (3)$$

where $SCr_i^{(k)}$ is the credit of the *i*-th user in the *k*-th frame, $bits_i^{(k)}$ is the number of bits generated by the *i*-th user in the *k*-th frame and $R_{b,i}^{(k)}$ is the allocated bit rate for the *i*-th user in the *k*-th frame. $u(\bullet)$ is the unit step function and *T* is the frame period.

These credits do not depend on the value of the maximum delay but they only ensure that users with long queues have more priority than users with empty queues. When the delay of the transmitted frames of a user exceeds a certain threshold, the algorithm will temporally eliminate the lower bit rates of his subset to guarantee a higher transmission speed.

III. HNN-based Optimization

The DRA is a NP (Non-deterministic Polynomial time) problem that makes impossible finding an analytical solution for an elevated number of users. As mentioned in the introduction, the main reason for using a HNN is its hardware implementation speed that, taking advantage of the inherent parallelism of the network, makes possible a real-time running of the algorithm.

III.1. HNN Model

A HNN is composed by a set of interconnected neurons. Neurons will change dynamically their state until reaching an equilibrium point. Hopfield showed that an energy function *E* can represent the dynamics of the HNN, and that the problem of finding an equilibrium state of the neurons can be solved by finding a local minimum of the energy function [8], [9].

The dynamics of the HNN can be expressed as:

$$\frac{dU_i}{dt} = -\frac{U_i}{\tau} - \frac{\partial E}{\partial V_i} \qquad (4)$$

where U_i and V_i are the input and output of the *i*-th neuron, and τ is the time constant of the circuit. The relationship between the outputs and the inputs of the neurons is non-linear, and is given by the sigmoid function:

$$V_i = f(U_i) = \frac{1}{1+e^{-\alpha U_i}} \qquad (5)$$

where α is the gain of the neurons.

III.2. Problem Formulation

The DRA problem can be formulated in terms of a 2D-HNN with $L=N \cdot M$ neurons, being *L* the number of neurons in the HNN, *N* the number of users in the system and *M* the number of possible bit rates. Users are then represented in the first dimension of the neural network (by rows), whereas the second dimension represents the set of possible bit rates (by columns). It is worth noticing that only two state neurons are taken into account: OFF or 0, and ON or 1. A neuron V_{ij} is ON if the *i*-th user has the *j*-th bit rate allocated. Note that the rest of the neurons of row *i*, corresponding to user *j*, V_{il}, $l \neq j$, must be OFF.

With these conditions, the absolute energy function minimum occurs in one of the 2*L* corners (since each neuron has two different states, $V_{ij} \in \{0,1\}$) of the *L*-dimensional hypercube. After solving (4) numerically by the Euler technique and reaching a stable state, each neuron is set to ON if V_{ij} is greater or equal than 0.5, or to OFF if V_{ij} is lower than 0.5.

The energy function proposed in this paper is based on the proposal made in [4] and [5], with a modification on the second term to introduce the load of the system instead of

the total allocated bandwidth. The energy function formulation is:

$$E = \frac{\mu_1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} C_{ij}V_{ij} + \frac{\beta^{\zeta}\mu_2}{2}\left|1-\frac{\eta_{UL}}{\eta_{UL\max}}\right| + \\ + \frac{\mu_3}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{ij}V_{ij} + \frac{\mu_4}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}V_{ij}\left(1-V_{ij}\right) + \\ + \frac{\mu_5}{2}\sum_{i=1}^{N}\left(1-\sum_{j=1}^{M}V_{ij}\right)^2 \qquad (6)$$

The first term of the energy function, weighted by μ_1, introduces the cost function C_{ij}. This function takes into account the prioritization made by the Service Credits of (3) as follows:

$$C_{ij} = \frac{u\left(\Gamma_{ij}\right)\Gamma_{ij}}{\max\limits_{\substack{\forall x\in\{1\cdots N\}\\ \forall y\in\{1\cdots M\}}}\left\{u\left(\Gamma_{xy}\right)\Gamma_{xy}\right\}} \qquad (7)$$

$$\Gamma_{ij} = SCr_i^{(k)} + \frac{bits_i^{(k)}}{T} - R_{b,i}^{(j)} \qquad (8)$$

here $R_{b,i}^{(j)}$ represents the j-th bit rate of the i-th user. The term $u\left(\Gamma_{ij}\right)\Gamma_{ij}$ is the Service Credit of the next frame $(k+1)$ if the j-th bit rate is allocated to the i-th user at the end of the algorithm. The value of the cost function increases with the credits, increasing thus the energy function as well. The dynamics of the HNN will tend to minimize the cost function value and hence to minimize the number of credits. That is, it will increase bit rates of the users that are not high enough to transmit all the packets they generate.

The μ_2 term forces the system to allocate the maximum resources that can be utilized, making the total load tend to the maximum.

The β^{ζ} factor, with:

$$\zeta = u\left(\frac{\eta_{UL}}{\eta_{UL\max}} - 1\right) \qquad (9)$$

is introduced to penalize the situations where $\eta_{UL} > \eta_{UL\max}$. The load expression in (1) has been modified to introduce the voltage of the neurons as follows:

$$\eta_{UL} = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{V_{ij}}{1+\dfrac{W}{\left(\dfrac{E_b}{N_0}\right)_i R_{b,i}^{(j)}}} \le \eta_{UL\max} \qquad (10)$$

On the other hand, the μ_3 term uses the ψ matrix. This matrix represents the bit rate subset of each user, where $\psi_{ij}=0$ if the j-th bit rate is usable by the i-th user and $\psi_{ij}=1$ otherwise. The μ_3 term penalizes the energy function if the bit rate allocated to a user is not in his subset. As stated before in section II.3, if one user exceeds a certain service delay threshold, the ψ matrix is temporarily changed to suppress the minimum bit rate of that user.

Finally, the last two terms ensure a rapid convergence to correct and stables states of neurons. The μ_4 term forces the minimum to be in the corners of the hypercube, and the μ_5 term forces users to have only one bit rate allocated.

III.3. Dynamics of Hopfield Neural Networks

Each frame the HNN algorithm schedules the active connections, taking into account incoming users or ending communications. The HNN algorithm starts with an initial state that is exactly the last resource allocation. Moreover all new and accepted connections are supposed to transmit in this initial state with their smallest bit rates. Thanks to these conditions it is possible to obtain the solution with the minimum changes.

The numerical Euler's technique to solve (4) in a 2D-HNN is:

$$U_{ij}\left(t+\Delta t\right) = U_{ij}\left(t\right) + \Delta t\left\{-U_{ij}\left(t\right) - \frac{\partial E}{\partial V_{ij}}\right\} \qquad (11)$$

where Δt is the time interval over which output voltages of neurons are observed and updated. The gradient of the energy function can be calculated as:

$$\frac{\partial E}{\partial V_{ij}} = \frac{\mu_1}{2}C_{ij} - \frac{(-\beta)^{\zeta}\mu_2}{2\eta_{UL\max}}\frac{1}{1+\dfrac{W}{\left(\dfrac{E_b}{N_0}\right)_i R_{b,i}^{(j)}}} + \\ + \frac{\mu_3}{2}\psi_{ij} + \frac{\mu_4}{2}\left(1-2V_{ij}\right) - \mu_5\left(1-\sum_{l=1}^{M}V_{il}\right) \qquad (12)$$

The equilibrium state is reached when the output of the neurons V_{ij} changes less than a tolerance. All the outputs are calculated each frame using (5) and the solution of (11).

IV. DRA HNN-based Algorithm

The DRA proposed algorithm consists on three parts: the admission control algorithm, the delay state check and the HNN optimization.

The admission control algorithm decides, at the beginning of each frame, which new connections are admitted in the system. A new connection is only accepted if the load of the system, computed assuming that all the existing users and the new one are transmitting at their lowest satisfactory bit rate, does not exceed the admission control threshold. The lowest satisfactory bit rate is set in the user profile as a fraction of the maximum service bit rate. In case that satisfactory bit rate was eliminated from the subset for one user, then it is set to his new minimum bit rate. For all new connections their service credits are initialized to zero.

Once the admission control algorithm ends, the DRA algorithm checks the delay status of all users. The algorithm increases the minimal bit rate of users that exceed their maximum delay threshold, and deletes the packets that exceed the maximum allowed delay. The new minimal bit rate is calculated to ensure the transmission of the entire buffer in time, but it must never exceed the user's maximal bit rate. With these restrictions, the new minimal bit rate for the i-th user can be expressed as:

$$R_{b\min,i} = \min\left\{\max_j\left(R_{b,i}^{(j)}\right), \sum_{b=1}^{L_{B,i}}\frac{B_i(b)}{t_i(b)}\right\} \qquad (13)$$

where $L_{s,i}$ is the buffer length, $B_i(b)$ is the number of bits stored in the b-th position of the buffer and $t_i(b)$ is the deadline for the data in the b-th position.

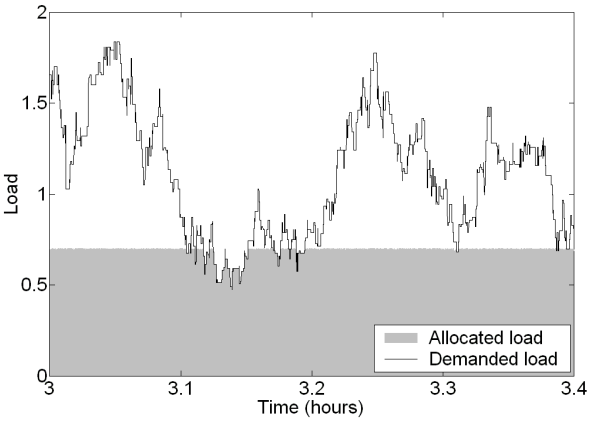

Figure 1: Comparison of the demanded and allocated load in a simulation time interval.

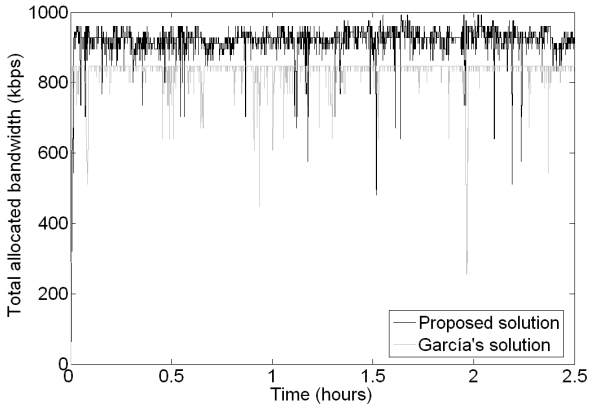

Figure 2: Total allocated bit rate of the proposed solution and the García's solution [5].

At this point it must be ensured that the HNN algorithm could find a feasible solution, i.e. that the total load with all the users transmitting with their minimum bit rate is below the maximum allowed. Otherwise users are dropped following a predefined priority scheme.

Next, the HNN algorithm is called and the new resource allocation is obtained. However, after the HNN an additional function is included to perform a local search with a greedy algorithm, since there is always a possibility that the network finds a local optimum located near the global optimum. The greedy algorithm tries to increase or decrease the bit rates of connections with a priority scheme that favors small changes in bandwidth. Note that it does not compromise on computation time because the number of combinations needed for finding a better (or optimal) solution is very small [4].

Finally the SCr for all the users are updated.

V. Results

V.1. Simulation Scenario

This section presents some results obtained with the proposed DRA algorithm. The scenario adopted is the same used in [5], to be able to compare results. The scenario consists in one isolated cell with radius 1 km. The path loss for the *i*-th user is calculated using:

$$L_{p,i}(\text{dB}) = 128.1 + 37.6\log_{10}(d_i) \qquad (14)$$

where d_i is the distance in km between the *i*-th user and the center cell. Users are on the move with a random speed uniformly distributed between 0 and 60 km/h. The thermal noise power level is -102 dBm. The total transmission bandwidth, *W*, is 3.84 Mchips/s. The maximum load factor, $\eta_{UL\max}$, is set to 0.7. The allowed bit rates are {256 kb/s, 128 kb/s, 64 kb/s, 32 kb/s, 16 kb/s} and the E_b/N_0 ratio for all the bit rates is 5 dB [5]. The load and delay thresholds are set to the 50% and 80% of the maximum respectively.

Table 1: Delay statistics.

Bit rate class	Delay class	Average delay	% of packets with exceeded delay
*SR*1	0.5 s	0.046 s	0.0043 %
	1.5 s	0.094 s	0 %
*SR*2	0.5 s	0.041 s	0.0036 %
	1.5 s	0.087 s	0 %
*SR*3	0.5 s	0.027 s	0.0002 %
	1.5 s	0.033 s	0 %

Table 2: García's solution [5] delay statistics.

Bit rate class	Delay class	Average delay	% of packets with exceeded delay
*SR*1	0.5 s	2.011 s	9.564 %
	1.5 s	2.139 s	8.993 %
*SR*2	0.5 s	0.034 s	1.128 %
	1.5 s	0.060 s	0.613 %
*SR*3	0.5 s	0.049 s	0.018 %
	1.5 s	0.051 s	0 %

The parameters of the HNN network considered are the following:

$$\mu_1 = 1000 \quad \mu_2 = 5000 \quad \mu_3 = 8000 \quad \mu_4 = 100$$
$$\mu_5 = 6000 \quad \beta = 10 \quad \Delta t = 10^{-4} \quad \alpha = 1$$

User's sessions have a Poisson arrival distribution with an arrival average rate of 0.2 calls/s, and an exponential duration distribution with mean 120 s. The frame period is equal to 0.1 s.

User's packets are generated each frame with a random size. The packet size probability density function is set to:

$$\left(\frac{12R_{b\text{av,i}}}{R_{b\max,i}^3} - \frac{6}{R_{b\max,i}^2}\right)\frac{B_i}{T} + \frac{4}{R_{b\max,i}} - \frac{6R_{b\text{av,i}}}{R_{b\max,i}^2} \qquad (15)$$

where $T \cdot R_{b\text{av},i}$ is the average packet size of the *i*-th, $R_{b\max,i}$ is the maximum bit rate of the *i*-th user and B_i is the packet size. The average packet size is set to the 40% of the maximum, i.e. $R_{b\text{av},i}=0.4R_{b\max,i}$.

Six types of service have been considered. Each service is characterized by a subset of allowed bit rates and a maximum delay. Specifically, three different subsets of bit rates are considered: $SR1 \equiv$ {256 kb/s, 128 kb/s, 64 kb/s, 32 kb/s, 16 kb/s}, $SR2 \equiv$ {128 kb/s, 64 kb/s, 32 kb/s, 16 kb/s} and $SR3 \equiv$ {32 kb/s, 16 kb/s}, and two maximum delays for each subset: 0.5 s and 1.5 s. The minimum satisfactory bit rate is set to the half of the maximum service bit rate.

V.2. Simulation Results

Figure 1 shows the total load demanded by users, assuming they ask for their maximum bit rate, and the real system load after scheduling. It can be seen that the allocated load approaches always the maximum allowed (i.e. 0.7), except when the load demanded is below that threshold. In that case the allocated and demanded loads are identical. In the 98.42 % of the simulation time the demanded load is over the threshold of 0.7. The average allocated load is 0.696, very closed to the maximum allowed value.

Tables 1 and 2 show the delay statistics for the DRA algorithm proposed in this paper and the one presented in [5]. The solution of [5] tries to allocate the maximum possible transmission bandwidth, obtaining also a good approximation to the maximum possible. However this algorithm works with a bit rate constraint instead of a load constraint what forces to make the following approximation:

$$\frac{W}{\left(\frac{E_b}{N_0}\right)_i R_{b,i}} >> 1$$

With this assumption a constant maximum system bit rate can be obtained from (1). This approximation makes the total allocated bandwidth to lie below the maximum reachable. Figure 2 shows the difference between the proposed algorithm, which do not assume the approximation, and the García's solution in terms of total allocated bit rate.

Another problem of the García's solution [5] is the excessively high delay for the service class *SR1*, what makes impossible a good communication. With the improvements added, i.e. the Service Credits (3)(7)(8) and the temporal suppression of low bit rates, this results can be practically corrected.

VI. Conclusions

This paper has presented a Dynamic Resource Allocation uplink approach for multi-service wireless WCDMA networks, where different user profiles, characterized by a subset of possible data rates and a maximum packet delay, exist. A Hopfield Neural Network has been proposed to maximize the resource allocation, while trying to guarantee a maximum service delay. The proposed algorithm is based on previous works where only the rate allocation problem is addressed. Results obtained show that a great delay performance improvement can be achieved by introducing a delay control technique that prioritizes users according their packet queue length, while keeping an excellent rate allocation performance.

Acknowledgments

This work has been carried out in the framework of the European Network of Excellence NEWCOM (Network of Excellence in Wireless Communications).

References

[1] C. Mihailescu, X. Lagrange, Ph. Godlewski, "Performance evaluation of a dynamic resource allocation algorithm for UMTS-TDD systems". IEEE Vehicular Technology Conference (VTC) Fall, Boston, USA, 2000.

[2] L. Forkel, T. Kriengchaiyapruk, B. Wegmann, E. Schulz, "Dynamic Channel Allocation in UMTS Terrestrial Radio Access TDD Systems". IEEE Vehicular Technology Conference (VTC) Spring, Boston, USA, 2001.

[3] O. Lázaro, "Dynamic radio resource management algorithms and traffic models for emerging mobile communication systems", Ph.D. dissertation, University of Strathclyde, Glasgow, Scotland, 2001.

[4] C. W. Ahan and R. S. Ramakrishna, "QoS provisioning dynamic connection-admission control for multimedia wireless networks using Hopfield Neural Networks", *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 106-117, Jan. 2004.

[5] N. García, R. Agustí, and J. Pérez-Romero, "A user-centric approach for dynamic resource allocation in CDMA systems based on Hopfield Neural Networks". IST Summit, Dresden, Germany, 2005.

[6] D. Abramson, K. Smith, P. Logothetis, and D. Duke, "FPGA-based implementation of a Hopfield Neural Network for solving constraint satisfaction problems", 24th Euromicro conference, Vaesteraas, Sweden, 1998.

[7] L. Almajano and J. Pérez-Romero, "Packet scheduling algorithms for interactive and streaming services under QoS guarantee in a CDMA system". IEEE Vehicular Technology Conference (VTC) Fall, Vancouver, Canada, 2002.

[8] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proc. Natl. Acad. Sci.*, vol. 79, pp. 2554-2558, April 1982.

[9] J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons", *Proc. Natl. Acad. Sci.*, vol. 81, pp. 3088-3092, May 1984.

[10] H. Holma and A. Toskala, "*WCDMA for UMTS: radio access for third generation mobile communications*", John Wiley & Sons, 2001.