The final publication is available at www.springerlink.com

Joint Dynamic Resource Allocation for QoS Provisioning in Multi-Access and Multi-Service Wireless Systems

Daniel Calabuig, José F. Monserrat, David Martín-Sacristán and Narcís Cardona
Polytechnic University of Valencia (UPV) – iTEAM Research Institute, Spain
{dacaso, jomondel, damargan, ncardona}@iteam.upv.es

***Abstract*– This paper proposes a Joint Dynamic Resource Allocation (JDRA) algorithm that allocates simultaneously the best-suited Radio Access Technologies (RAT) and amount of resources to all the users active in a multi-access wireless system. Both distributions are performed at the same time so as to make the most of the heterogeneous network. In this scenario users can connect to several RATs but not simultaneously and, therefore, the JDRA algorithm is able to consider the required handover time in the decision making. Moreover, the algorithm guarantees the Quality of Service (QoS) provision in terms of delay and bit rate in a multi-service scenario where different users may have different QoS requirements. Such a complex optimization problem has been tackled using a Hopfield Neural Network (HNN) formulation. These neural networks have fast response times once hardware implemented, which is very significant since current and future wireless networks must rapidly adapt to changing circumstances in wireless environment and traffic. Results prove the benefits achieved by the usage of the HNN-based JDRA algorithm. Firstly, the joint decision outperforms a two-steps procedure in which, after the RAT selection, the same uni-RAT DRA algorithm is applied. Secondly, the proposed algorithm can deal with different levels of congestion and load distribution among RATs in a much better way that other reference algorithms specifically designed for multi-service scenarios.**

I. Introduction

Mobile wireless systems are in constant evolution due to the continuously evolving requirements and expectations of both users and operators. Users expect high quality communications and full access to digital contents with the same transmission capacity as wired networks, independently of the number of users active in the system. According to this user demand for wireless connectivity, new standards

have been designed and launched to the market in the last years to satisfy these increasing requirements. General Packet Radio Service (GPRS), Universal Mobile Telecommunications System (UMTS), Worldwide Interoperability for Microwave Access (WiMAX), Wireless Local Area Network (WLAN) or Bluetooth are some examples of current standardized technologies. Each Radio Access Technology (RAT) is specially suited for one type of wireless network, ranging from Wireless Wide Area Networks (WWAN) down to Wireless Personal Area Network (WPAN). In addition to the usage scenario, conventional mobile networks were devised to fulfill the specific Quality of Service (QoS) requirements of each service, whereas other technologies paid more attention to system simplicity and flexibility.

Currently it is quite common to have several independent RATs giving coverage to the same area. Moreover, users are who decide upon the technology they get connected to, either configuring the User Equipment (UE) or using different UEs for each technology. Nevertheless, users should not get involved in this type of decisions, or at least not separately, since they have not got a global view of the different RATs. Thus, the future points to a multi-RAT UE capable of getting automatically connected to the most proper RAT. This multi-access wireless system, also referred to as heterogeneous wireless system, could make the most of the individual coverage and instantaneous capacity of each technology taking into account the RAT availability, signal quality and type of service to provide the most appropriate resources for the variety of different users.

The notion of being always best connected, which was first introduced in [1], is an extension for heterogeneous systems of the notion of being always connected. Now, users not only should be able to be connected anywhere and anytime, but also they should be served with the best available connection, which can be only accomplished with the interworking of the different technologies. For that reason, the standardization bodies are doing their best to make the interworking possible. For instance, the 3GPP organization not only allows UMTS to interwork with GPRS (two 3GPP RATs) but also establishes the basis for a WLAN interworking (a non-3GPP RAT). In addition, the IEEE Standards Association is working on the 802.11u standard (scheduled for 2009), which provides WLAN with the capability of interworking with other external networks. Nowadays RAT interworking is becoming a

reality that requires more advanced mechanisms resulting in a higher resource usage and network quality.

A. RRM in Multi-access Wireless Systems

In wireless systems the concept of QoS poses several constraints to networks' management to assure an optimum distribution of the scarce radio resources among active users. In this framework, the concept of Radio Resource Management (RRM) encompasses various techniques specially designed to fulfill the negotiated QoS to the end users.

As distinguished from existing wireless networks where an independent RRM is performed by the radio network controller of each system, in a multi-access wireless network some kind of overall resource management is required to select the best RAT, dynamically allocate resources among them, control the congestion and manage handovers. The Common Radio Resource Management (CRRM) concept is widely used to refer to these tasks.

The most important functions related to radio resource management are: initial RAT and cell selection, Call or Connection Admission Control (CAC), congestion control, power control, scheduling or resource allocation, handover (HO) and vertical HO. Depending on the level of CRRM/RRM coupling these functions are handled by either a RAT-specific RRM entity or by the overall CRRM entity [2]. Anyway, it is generally agreed that the CRRM entity is, at least, responsible for the interworking of the Radio Access Networks (RANs) not only of the same RAT but also of different RATs [3].

Considering a high or tight coupling degree – the one grabbing nowadays more attention from the research community – the CRRM entity is entrusted with the management of most of the RRM functionalities, delegating only the power control and scheduling to the RRM entities. In terms of resource allocation, at each resource allocation slot the RRM entities have a number of users connected and must distribute the resources among them. The CRRM entity is responsible for the distribution of users among RATs and, consequently, for vertical HO management. The RRM entity can use any of the algorithms proposed in the literature to distribute resources in a unique RAT (see examples [2],[4]-[7]), whereas the CRRM entity chooses one RAT selection technique to reduce system

congestion while maximizing user QoS (see examples [8], [9]). Since this paper deals simultaneously with both aspects, some of these techniques are further analyzed.

Certain resource allocation techniques allocate resources to those users experiencing best channel quality [2], [4]. This kind of policy can maximize the average system throughput, but at the expense of an unfair distribution that implies relatively bad QoS for users with poor channel quality. A wide range of more sophisticated algorithms are based on the Generalized Processor Sharing (GPS) idea [5], where resources are distributed among users proportionally to some predefined weights. Within this group, Modified Largest Weighted Delay First (MLWDF) [6] and its improvement, the Cross-Layer Scheduling Algorithm (CLSA) [7], are noteworthy. Both techniques take into account some weighting coefficients that prioritize users according to their service, current QoS and perceived channel quality. For that reason, these algorithms exhibit better performance than those prioritizing users according to only one of these characteristics. Moreover, MLWDF and CLSA were designed to provide either bit rate or delay-based QoS.

Regarding RAT selection, Pérez-Romero *et al*. proposed different policies [8] to sort, in order of preference, the list of candidate RATs the user could get connected to. As an example, they proposed selecting GPRS for voice and indoor users and UMTS for web and outdoor users whenever possible. This policy is motivated by the good performance of UMTS for high data rate users as compared with GPRS and by the bad behavior of UMTS for indoor users. Nevertheless, this policy does not consider quality expectations of users. With this aim, Pérez-Romero *et al*. defined a fittingness factor that reflects the degree of adequacy of each RAT to each user [9]. The fittingness factor takes into account two concepts: (a) the capabilities of the RAT and the UE, i.e. if a RAT can provide the service the user is asking for and if the UE can connect to the RAT, and (b) also the suitability of each possible connection in terms of channel quality and bit rate. Although both algorithms only consider one RAN per RAT, [9] can also choose the best-suited RAN inside the selected RAT. Conversely, [8] needs an additional process for RAN selection once the destination RAT is known.

So far, Unlicensed Mobile Access (UMA) is the only mechanism being implemented to dynamically select the best-suited RAT. UMA is the commercial solution of the 3GPP standard called Generic Access

Network (GAN) [10], [11]. In the UMA solution, dual UEs can migrate from High-Speed Downlink Packet Access (HSDPA) network to a WLAN Access Point (AP) and vice versa. Thus, anytime a UE finds an AP, it tries to get a WLAN connection since this technology is supposed to provide much better throughput capacity than HSDPA. This philosophy is in some way similar to the policies proposed in [8].

B. Scope of this work

In addition to the abovementioned CRRM/RRM interaction degree, in the very high interaction degree the CRRM entity decides also on the resource allocation. This fact makes possible to jointly distribute both resources among users and users among RATs to make the most of the available technologies. Of course, in this paradigm users can be connected to more than one type of RAT but not simultaneously. In this context, there is not any joint scheduling solution proposed in the literature, beyond our preliminary works of [12] and [13]. The complexity of this optimization problem makes it difficult not only to find a solution but also to define an algorithm capable of performing the search. For that reason, most research groups are just focusing on RAT selection techniques. Nevertheless, this article proves that a joint scheduler can outperform the combination of an optimum RAT selection technique and uni-RAT scheduler.

To this aim, this article proposes an extremely efficient Joint Dynamic Resource Allocation (Joint DRA, JDRA) algorithm. The complexity of the optimization problem requires the usage of advanced techniques to find, at least, a sub-optimum solution. Many types of algorithms have been proposed in the literature to solve such huge optimization problems, like genetic algorithms, game theory, linear programming or Hopfield Neural Networks (HNNs). Within this group, HNNs have been identified as fast hardware optimizers that can obtain a valid solution in few microseconds [14]. This fast response is a consequence of the simplicity of each individual neuron and their parallel interworking. Therefore, problems that are more complex need more neurons, i.e. more hardware, but maintain the fast response of simpler problems. This feature makes HNNs be the best candidates for sub-optimal and real-time schedulers.

HNNs have been widely used in a variety of scientific domains [15]-[17]. The first study that introduced an HNN-based algorithm in a wireless system was presented by Del Re et al. in [18]. The research work carried out by Lázaro and Girma in [19] was built on this algorithm. They proposed the usage of HNNs for the dynamic distribution of frequency channels over the cells of a GSM system together with a guard channel technique for handovers. Ahn and Ramakrishna [14] were the first authors to use HNNs for solving the DRA problem. In the main, their algorithm aimed at maximizing the allocated resources and obtaining a fair distribution among users. This seminal work was extended in our previous work of [20], where not only resources were maximized, but also delay was minimized in order to improve the QoS support. Nevertheless, this work is only focused on delay and it only distributes resources in a unique RAN.

To sum up, the previous works have shown the utility of HNNs to allocate dynamically resources to users. Thus, starting from the original work of Ahn and Ramakrishna [14], this paper aims at proposing a new HNN-based algorithm to jointly distribute the set of resources of RATs among users and users among the RATs available in a heterogeneous wireless system. This algorithm will take simultaneously into account the kind of service and specific QoS requirements of each user and the resource availability and characteristics of each RAT. These are the main contributions of this paper and the main differences with respect to [14] and [20].

The remaining of this paper is organized as follows: Section II reviews the fundamentals of HNNs. In Section III the HNN energy formulation proposed for the JDRA problem is presented. The simulation scenario is described in Section IV whereas Section V copes with the analysis of the simulation results. Finally, the main conclusions are drawn in Section VI.

II. Fundamentals of Hopfield Neural Networks

HNNs are a type of recurrent neural networks completely characterized by an energy function E that describes their dynamics – i.e. the time evolution of the neuron inputs and outputs. The main contributions of Hopfield are summarized in [15] and [21]. Hopfield considered a network with interconnected neurons using resistors. He defined $T_{ij} = 1/R_{ij}$ as the interconnection weight and R_{ij}

the resistance between the i-th neuron input and the j-th neuron output. In addition, each neuron has also an input current I_i. Let us define U_i as the input voltages, V_i as the output voltages and g_i as the gain function of neuron i, i.e. $V_i = g_i(U_i)$. Neuron outputs take on any value between 0 and 1. g_i is usually defined as the sigmoidal, step or linear functions. The sigmoidal function is similar to the response of operational amplifiers that form the neurons in the original Hopfield model. The step function is a first approximation of the sigmoid but HNN with such gain function have very bad outcomes [22]. A better approximation is the linear function, which has been used in this work. Moreover, the proportionality constant has been reduced to the unit for simplicity reasons. Thus:

$$V_i = g_i(U_i) = U_i. \tag{1}$$

Note that U_i belongs to the interval [0,1] too. The time evolution of the i-th neuron is [21]:

$$C_i \frac{dU_i}{dt} + \frac{U_i}{R_i} = \sum_{j=1}^{N} T_{ij} V_i + I_i, \tag{2}$$

where R_i and C_i are, respectively, the equivalent resistance and capacitance of the i-th neuron input and N is the number of neurons in the network. Incorporating R_i and C_i into T_{ij} and I_i and using (1), (2) can be rewritten as:

$$\frac{dU_i}{dt} = \sum_{j=1}^{N} T_{ij} V_i + I_i. \tag{3}$$

Let consider the following energy function:

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} V_i V_j - \sum_{i=1}^{N} I_i V_i. \tag{4}$$

Hence, if $T_{ij} = T_{ji}$:

$$\frac{dU_i}{dt} = \frac{dV_i}{dt} = -\frac{\partial E}{\partial V_i}. \tag{5}$$

(5) can be approximated with computer simulations using the Euler forward method. Consequently, defining a time step Δt, neuron outputs are updated each Δt seconds following this procedure:

$$V_i(t + \Delta t) = V_i(t) - \Delta t \frac{\partial E}{\partial V_i}(t). \tag{6}$$

A noteworthy conclusion that can be extracted from (5) or (6) is that neurons evolve towards states with less energy. Moreover, the equilibrium point, i.e. when $V_i(t + \Delta t) = V_i(t)\ \forall i$, is reached at a local minimum of E. As a consequence, E must be defined to be minimum at the desired solutions and then the HNN will find and stop at those solutions.

III. JDRA Problem Formulation

The most relevant contribution of this paper is the formulation of the HNN energy function that satisfies all the constraint established by the JDRA problem in a multi-access and multi-service wireless system. This section gives a full account of all the terms included in this energy function and the procedure followed to come up with such formulation.

Let I be the number of users demanding resources at a specific resource allocation time. All users will be distributed among K RANs that may belong to different radio technologies or not. The resources of all RANs are divided into minimum resource quanta, e.g. time slots in GPRS or 256 Spreading Factors (SFs) in UMTS. Due to this division, for the k-th RAN only one finite set exists, $\Re_k$, with all the feasible resource quantities that may be allocated to one user. For example, in GPRS only 0 up to 8 time slots can be allocated to one user, hence the finite set is $\Re_k = \{0,1,2,3,4,5,6,7,8\}$. Let J_k be the number of elements of $\Re_k$. The optimization problem consists in finding the best combination of RAN and quantity of resources that must be allocated to each user in order to satisfy its QoS requirements and system constraints.

A. QoS Provision

Depending on the type of service, quality requirements can be concreted differently, for instance in terms of maximum packet delay or minimum bit rate. In order to have a common definition for QoS requirements, the concept of minimum target bit rate firstly introduced in [20] is extended in this paper. This minimum bit rate has to fulfill the user-specific QoS requirements. Therefore, it must be calculated independently for each user and type of QoS.

QoS based on minimum bit rate: if the i-th user requires a minimum instantaneous bit rate of $\underline{R_i}$, then the minimum target bit rate for that user is:

$$R_{\min,i} = \underline{R_i}. \tag{7}$$

Nevertheless, this tight condition is usually relaxed for actual services. For example, data transfer services using File Transfer Protocol (FTP) require not an instantaneous minimum bit rate $\underline{R_i}$ but an average one, $\bar{R}_i$. In this case $R_{\min,i}$ can be calculated using a leaky-bucket approach. $\bar{R}_i \cdot \Delta t_a$ tokens are generated each resource allocation time, being Δt_a the resource allocation period. On the other hand, $R_i(t_a) \cdot \Delta t_a$ tokens are spent each period, where $R_i(t_a)$ is the bit rate allocated to the i-th user at time t_a. Note that time is now represented with t_a to avoid confusion with time evolution of HNN. Each available token can be understood as a bit that must be transmitted to reach the average minimum bit rate $\bar{R}_i$. The quantity of bits the system owes the i-th user is calculated as follows:

$$b_{\text{owed},i}(t_a + \Delta t_a) = \begin{cases} 0, & b_i(t_a) = 0, \\ b_{\text{owed},i}(t_a) + \left(\bar{R}_i - R_i(t_a)\right)\Delta t_a, & \text{otherwise}, \end{cases} \tag{8}$$

where $b_i(t_a)$ is the quantity of bits stored in the buffer of the i-th user at time t_a. The definition of (8) assumes that $b_{\text{owed},i}$ is reset at the beginning of every data burst. Thus, $\bar{R}_i$ is actually the minimum average bit rate per burst. The objective is that the quantity of owed bits be at most 0 at the burst end. If so, the average bit rate allocated to the user would be greater or equal to $\bar{R}_i$. $R_{\min,i}$ is the minimum bit rate that accomplishes this objective. Thus, if at time t_a the user is served with $R_{\min,i}$ until the burst end, then the burst would last $b_i(t_a)/R_{\min,i}$ additional seconds and:

$$b_{\text{owed},i}\left(t_a + \frac{b_i(t_a)}{R_{\min,i}}\right) = 0 = b_{\text{owed},i}(t_a) + \left(\bar{R}_i - R_{\min,i}\right)\frac{b_i(t_a)}{R_{\min,i}}. \tag{9}$$

Finally, from (9):

$$R_{\min,i} = \frac{b_i(t_a)\bar{R}_i}{b_i(t_a) - b_{\text{owed},i}(t_a)}. \tag{10}$$

It is worth noting that (10) makes sense only if $b_i(t_a) > b_{\text{owed},i}(t_a)$. If the quantity of owed bits is greater than those available for transmission then it would not be possible to achieve the objective bit rate $\bar{R}_i$. In that case, the allocated bit rate should be the greatest one in order to approach $\bar{R}_i$ as possible. Consequently, $R_{\min,i}$ can be defined as $R_{\min,i} = \infty$ for $b_i(t_a) \leq b_{\text{owed},i}(t_a)$.

QoS based on maximum delay: if the service of the i-th user is delay-sensitive and packets must be transmitted before a certain maximum delay $D_{\max,i}$, then $R_{\min,i}$ can be obtained following [20]:

$$R_{\min,i} = \begin{cases} \max\limits_{p=1\dots P_i} \dfrac{\sum_{s=1}^{p} \beta_{s,i}}{D_{\max,i} - D_{p,i}}, & D_{\max,i} > \max\limits_{p} D_{p,i}\,, \\ \infty, & D_{\max,i} \leq \max\limits_{p} D_{p,i}\,, \end{cases} \tag{11}$$

being P_i the number of packets in the buffer, $\beta_{s,i}$ the size in bits of the s-th packet and $D_{p,i}$ the delay of the p-th packet of user i. Refer to [20] for more details about eq. (11).

B. Resources to bit rate mapping

Each RAN may have a different amount of resources available for distribution. Besides, the type of resource can be highly different from one RAT to another. Therefore, it is quite important to calculate the quantity of resources that each user requires by converting $R_{min,i}$ to amount of resources or vice versa. Let us define $Q_k(c_i)$ as the effective bit rate that the i-th user is capable of achieving with a resource unit (r.u.) of the k-th RAN. c_i is the channel Signal to Noise and Interference Ratio (SNIR), which measures the received signal quality. Q_k can be understood as a kind of Look Up Table (LUT).

Supposing an optimal link adaptation, Q_k can be obtained as:

$$Q_k(c_i) = \max_{s=1\dots S_k} \frac{L_s}{L_s + C_s} Br_{sk}\big(1 - Er_{sk}(c_i)\big), \tag{12}$$

where S_k is the number of Transmission Modes (TMs), i.e. modulation and coding scheme pairs, Br_{sk} and Er_{sk} are respectively the bit rate per r.u. and error rate of the s-th TM of the k-th RAT, L_s is the payload length and C_s is the header length.

Once the required $R_{min,i}$ and the current SNIR perceived by the user are known, the amount of r.u. to be reserved to the user can be easily obtained dividing $R_{min,i}$ by $Q_k(c_i)$.

C. HNN for JDRA

Previous works – see [14] and [20] – used 2-dimensional (2D) HNNs for solving the DRA problem. This work uses the natural evolution of these networks – first proposed in [12] –introducing the different RANs in a third dimension. Therefore, neurons are organized in a 3D grid where if the neuron at

position (i, j, k) is active this represents the allocation of the j-th resource quantity in the k-th RAN, the j-th element of $\Re_k$, to the user i. At the equilibrium the rest of neurons of user i must be inactive.

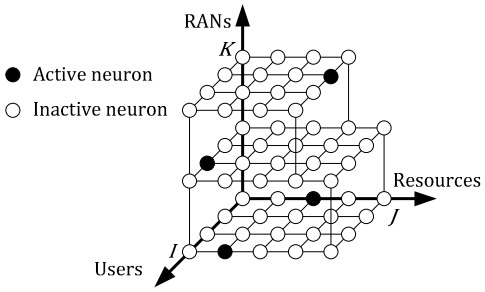

Figure 1. Hopfield Neural Network and equilibrium example. 4 users ask for resources in 3 RANs. The first two RANs have 5 different resource quantities whereas the last RAN has only 4.

Figure 1 shows an example with 4 active users demanding resources in 3 different RANs.

D. Energy function

Let us define the following objective function that the HNN will minimize:

$$E_1 = -\frac{\mu_1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k} B_{ijk}V_{ijk} - \frac{\mu_2}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\frac{Q_k(c_i)\rho_{jk}}{\max_{lmn}(Q_n(c_l)\rho_{mn})}\,V_{ijk}, \tag{13}$$

where B_{ijk} is the benefit – see Section III.E for more details – the i-th user receives in terms of QoS from the allocation of the j-th resource quantity of the k-th RAN, ρ_{jk} is the j-th resource quantity of the k-th RAN and μ_1 and μ_2 weight each term. As explained before, $Q_k(c_i)\rho_{jk}$ is the effective bit rate transmitted to user i for a given ρ_{jk}, whereas $\max_{lmn}(Q_n(c_l)\rho_{mn})$ is constant in a resource allocation period regardless the value of i, j and k and aims at normalizing the cost of the second term. Minimizing (13) will determine the resources allocated to each user pursuing two objectives, first maximizing the benefit from users' perspective and second maximizing the total throughput of the heterogeneous system. Note that this maximization is possible due to the negative sign of both terms.

Nevertheless, additional constraints have to be taken into account. First and most importantly, RAN resources are finite and, hence, the total amount of allocated resources must be controlled. To this aim a new term was added to (13):

$$E_2 = E_1 + \frac{\mu_3}{2} \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \xi_{ijk} V_{ijk}, \tag{14}$$

where $\xi_{ijk} = 0$ if the k-th RAN has enough resources to allocate the j-th resource quantity to the i-th user and $\xi_{ijk} = 1$ otherwise – see Section III.F for further details. – In other words, this term penalizes the allocations that imply exceeding the maximum available resources in any RAN.

Additionally, some resource quantities may not be allowed to some users. Let us define ψ_{ijk} as a permission table where $\psi_{ijk} = 1$ if the j-th resource quantity of the k-th RAN should be prohibited to the i-th user and $\psi_{ijk} = 0$ otherwise.

Then, the following objective function takes this effect into account:

$$E_3 = E_2 + \frac{\mu_4}{2} \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \psi_{ijk} V_{ijk}. \tag{15}$$

Thanks to this term, the heterogeneous system can define different user priority sets – Gold, Silver and Bronze users – limiting the maximum allowed bit rate according to the user quota. Moreover, when a user starts a vertical handover procedure the connection must migrate from the original RAN to the new serving one. This process takes some non-negligible time. During RAN changes users cannot consume any resource for data transmission. Therefore, the permission table can be modified in accordance with this wasted time so that the user in a vertical handover is not capable of using some of the highest resource quantities. The amount of resource quantities prohibited will depend on the RAN reconfiguration time.

Finally, some additional terms must be introduced to ensure a rapid convergence to correct and stable neuron states. Neuron outputs V_{ijk} must be 0 or 1 at the equilibrium and furthermore, only one neuron must be active, i.e. $V_{ijk} = 1$, for each user. These two constraints can be introduced in the energy function with the two terms proposed in [14] resulting finally:

$$E = -\frac{\mu_1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k} B_{ijk}V_{ijk} - \frac{\mu_2}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\frac{Q_k(c_i)\rho_{jk}}{\max_{lmn}(Q_n(c_l)\rho_{mn})}V_{ijk} + \frac{\mu_3}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\xi_{ijk}V_{ijk} + \frac{\mu_4}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\psi_{ijk}V_{ijk} + \frac{\mu_5}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k}V_{ijk}(1-V_{ijk}) + \frac{\mu_6}{2}\sum_{i=1}^{I}\left(1-\sum_{k=1}^{K}\sum_{j=1}^{J_k}V_{ijk}\right)^2. \quad (16)$$

This energy function has been used in this article to solve the JDRA problem. The weighting coefficients μ_1 to μ_6 must be carefully selected. The same rationale stated in [20] has been followed giving as a result the coefficient values summarized in Table I.

Table I. HNN weighting coefficients.

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
1500	500	2500	16000	15	7000

E. Benefit function

Once QoS is homogenized in terms of minimum bit rate $R_{min,i}$, benefit that users perceive depends on two main factors. First, the higher the bit rate the higher the benefit. Moreover, the lower the quality the user is perceiving, i.e. higher requirements of $R_{min,i}$, the higher the achievable benefit. This policy will increase the fairness among users. Because of the second factor, two users with different qualities have different maximum benefits. To achieve that, users are weighted inversely to their quality.

As compared with the benefit function proposed in [20], this article proposes some modifications so as to include the second factor by adding some weights to the users. Let us define C_{ijk} as the benefit defined in [20], then:

$$C_{ijk} = \frac{S\left(Q_k(c_i)\rho_{jk}, s_i, r_i\right) - S(0, s_i, r_i)}{S(R_{\max}, s_i, r_i) - S(0, s_i, r_i)}, \quad (17)$$

$$S(x, s, r) = \frac{1}{1 + e^{-s(x+r)}}, \quad (18)$$

$$s_i = \begin{cases} \dfrac{2\ln(9)}{R_{\min,i}}, & R_{\min,i} \le R_{\max,i}, \\ \dfrac{2\ln(9)\,R_{\min,i}}{\left(R_{\max,i}\right)^2}, & R_{\min,i} > R_{\max,i}, \end{cases} \tag{19}$$

$$r_i = \begin{cases} -\dfrac{R_{\min,i}}{2}, & R_{\min,i} \le R_{\max,i}, \\ -R_{\max,i} + \dfrac{\left(R_{\max,i}\right)^2}{2R_{\min,i}}, & R_{\min,i} > R_{\max,i}, \end{cases} \tag{20}$$

$$R_{\max} = \max_{ijk}\left(Q_k(c_i)\rho_{jk}\right), R_{\max,i} = \max_{jk}\left(Q_k(c_i)\rho_{jk}\right). \tag{21}$$

The main properties of the previous definition are that C_{ijk} is bounded by 0 and 1 and is increasing with a high increment near $R_{\min,i}$ [20]. Thanks to the bounds C_{ijk} does not increase uncontrollably and because of the increment the HNN will tend to allocate resource quantities that satisfy the target $R_{\min,i}$. Finally, the benefit used in this article is defined as:

$$B_{ijk} = \frac{\min\left(R_{\max}, R_{\min,i}\right)}{\min\left(R_{\max}, \max\limits_{i} R_{\min,i}\right)} C_{ijk}. \tag{22}$$

B_{ijk} preserves the two properties of C_{ijk} and additionally introduces weights for each user favoring those users with higher needs.

F. Resource saturation control

The saturation control mechanism used in this article is similar to that proposed in [20]. ξ_{ijk} is used as an indicator of which resource allocations may be supported by the RANs. The ξ_{ijk} indicator is calculated for each user i assuming that the rest of users, $l \neq i$, maintain the resource allocation of the current neuron outputs. Thus:

$$\xi_{ijk} = \begin{cases} 1, & \rho_{jk} + \sum\limits_{\substack{l=1 \\ l\neq i}}^{I} \sum\limits_{m=1}^{J_k} \rho_{mk} V_{lmk} > \rho_{\max,k}, \\ 0, & \rho_{jk} + \sum\limits_{\substack{l=1 \\ l\neq i}}^{I} \sum\limits_{m=1}^{J_k} \rho_{mk} V_{lmk} \le \rho_{\max,k}, \end{cases} \tag{23}$$

where $\rho_{\max,k}$ is the maximum quantity of resources available in the k-th RAN. If $\xi_{ijk} = 1$ then ρ_{jk} cannot be supported for the i-th user with the current resource distribution. In that case, ξ_{ijk} increases

the energy function and, consequently, the HNN tends to decrease V_{ijk}, what finally means not allocating ρ_{jk} to the i-th user.

IV. Simulation Environment

A. Technologies used in the simulations

The proposed JDRA algorithm was tested in an artificial environment by means of computer simulations. The simulated heterogeneous network comprised two RATs: HSDPA and 802.11e WLAN.

802.11e WLANs use convolutional codes to protect data from errors. Furthermore, the Packet Error Rate (PER) depends not only on the channel quality but also on the payload length. Assuming a Viterbi decoding at the receiver, the PER of the s-th TM is [24]:

$$\mathrm{PER}_s(L_s, c_i) = 1 - \left(1 - P_s^u(c_i)\right)^{L_s}, \tag{24}$$

where P_s^u is the bit error probability of the s-th TM. The optimum payload length that maximizes the throughput for each TM is [24]:

$$L_s^*(c_i) = -\frac{C_s}{2} + \frac{1}{2}\sqrt{C_s^2 - \frac{4C_s}{\ln\left(1 - P_s^u(c_i)\right)}}. \tag{25}$$

Finally, from (12) and (25), function Q_k is for WLAN:

$$Q_k(c_i) = \max_{s=1\ldots S_k} \frac{L_s^*(c_i)}{L_s^*(c_i) + C_s} Br_{sk}\left(1 - P_s^u(c_i)\right)^{L_s^*(c_i)}. \tag{26}$$

The available resources in WLAN are slots of channel occupancy which are collision free thanks to the use of the HCF (Hybrid Coordinator Function) Controlled Channel Access (HCCA) mechanism.

HSDPA uses turbo codes instead of convolutional codes to protect data from errors. The Block Error Rate (BLER) depends also on the block size and on the channel quality. Nevertheless, each TM has a fixed block size and, hence, the BLER for a specific TM is only a function of the channel quality. HSDPA has a wide range of possible TMs, from which 30 have been defined in the standard as Channel Quality Indicators (CQIs). Only these 30 TMs are used in this work. The BLER of the s-th CQI can be approximated as [25]:

$$Er_s(c_i) = \left(10^{\frac{2(c_i - 1.03s + 17.3)}{\sqrt{3} - \log(c_i)}} + 1 \right)^{-\frac{1}{0.7}}. \quad (27)$$

Users are supposed to be time multiplexed. Thus, the 15 available codes are always allocated to a unique user each 2 ms. This assumption implies that the actual BLER differs from the one obtained in [25], since BLER is a function of the SNIR per code. If more codes are allocated then more SNIR is needed to maintain the same SNIR per code. Therefore, if 15 codes are allocated (27) has to be modified to:

$$Er_s(c_i) = \left(10^{\frac{2\left(c_i - 10\log\left(\frac{15}{N_s}\right) - 1.03s + 17.3\right)}{\sqrt{3} - \log(c_i)}} + 1 \right)^{-\frac{1}{0.7}}, \quad (28)$$

where N_s is the number of codes of the s-th CQI, shown in Table II. Moreover, if all the codes are allocated to a unique user, bit rates of the considered TMs also differ from the standard. Note that the new bit rates are:

$$Br_s = Br_s^* \frac{15}{N_s}, \quad (29)$$

where Br_s^* is the s-th CQI bit rate, shown also in Table II. Finally function Q_k is for HSDPA:

$$Q_k(c_i) = \max_{s=1\ldots30} Br_s\left(1 - Er_s(c_i)\right). \quad (30)$$

The DRA algorithm can allocate all codes to any user every 2 ms. Being one resource element one period of 2ms, the quantity of available resources is 500 periods of 2 ms each second.

Table II. Number of codes and bit rate of each CQI.

CQI	N_s	Br_s^* (kb/s)	CQI	N_s	Br_s^* (kb/s)	CQI	N_s	Br_s^* (kb/s)
1	1	68.5	11	3	741.5	21	5	3277.0
2	1	86.5	12	3	871.0	22	5	3584.0
3	1	116.5	13	4	1139.5	23	6	4859.5
4	1	158.5	14	4	1291.5	24	7	5709.0
5	1	188.5	15	5	1659.5	25	10	7205.5
6	1	230.5	16	5	1782.5	26	13	8774.0
7	2	325.0	17	5	2094.5	27	15	10877.0
8	2	396.0	18	5	2332.0	28	15	11685.0
9	2	465.5	19	5	2643.5	29	15	12111.0
10	3	631.0	20	5	2943.5	30	15	12779.0

B. Reference algorithms

The HNN-based algorithm proposed in this paper was compared with different combinations of RAT selection policies and uni-RAT DRA algorithms. The UMA solution and Maximum Bit Rate (MBR) policy were used for RAT selection. As explained in the introduction, UMA terminals select WLAN when they are in the coverage area of an AP. With the MBR policy, the user connects to the RAN that could transmit with the highest bit rate given the current channel quality of the user. Once users are distributed among the available RANs using UMA or MBR policy, a DRA algorithm was applied to perform scheduling and allocate resources inside each RAN. As DRA algorithms, MLWDF, CLSA and the proposed HNN for only one technology were selected. This choice was motivated because MLWDF and CLSA take simultaneously into account channel quality, QoS satisfaction and type of service of the user. Besides, it was important to make a fair comparison including other algorithms that were also able to cope with a composite of bit rate and delay-based services.

Finally, combining all the possibilities, six different reference algorithms were defined: UMA-MLWDF, UMA-CLSA, UMA-HNN, MBR-MLWDF, MBR-CLSA and MBR-HNN.

C. Scenario

The simulation scenario comprised 7 cells with the cell under study in the center. Each cell had 2 RANs, one HSDPA and another WLAN, being both, the HSDPA base station and the WLAN access point, at the cell center. Users were time multiplexed in both technologies. The studied services were web browsing and FTP data downloading, whose traffic models were extracted from [23]. Two different user classes were defined for each service. The QoS requirements for web users were a maximum delay of 30 s and 60 s, whereas FTP users expect a minimum average bit rate per burst of 150 kb/s or 50 kb/s depending on the service class. $R_{\min,i}$ was obtained from (10) and (11) for FTP and web users respectively. Moreover, MLWDF and CLSA algorithms can use different weights for each service. For these simulations, these were extracted from [7], i.e. 1 for web and 0.8 for FTP. Regarding mobility modeling, two different areas were considered. Some users could move all around the cell with a random speed uniformly distributed between [0,50] km/h. The remaining users just moved within a

hot spot located at the cell center with a random speed uniformly distributed between [0,3] km/h. The cell and hot spot radii were 500 m and 50 m respectively.

The maximum transmitted power was set to 43 dBm for the HSDPA base station and 20 dBm for the WLAN AP. Interfering cells were supposed to transmit half the maximum power being, therefore, half loaded. Noise power at the receiver was set to -102 dBm and -95 dBm for HSDPA and WLAN respectively, as a consequence of the different bandwidth. The path losses expressed in decibels for the i-th user in both technologies were:

$$L_{\mathrm{HSPDA},i} = 137.4 + 35.2\log\left(d_{\mathrm{HSDPA},i}\right), \tag{31}$$

$$L_{\mathrm{WLAN},i} = 135 + 45\log\left(d_{\mathrm{WLAN},i}\right), \tag{32}$$

where $d_{\mathrm{HSDPA},i}$ and $d_{\mathrm{WLAN},i}$ are the distances in km from the i-th user to the base station and the AP.

The proposed HNN-based algorithm was run every simulation second. Therefore, RAN changes may only occur every second at most. Besides, users were supposed to spend 0.5 s completing a vertical handover procedure. For the reference algorithms the RAT selection procedures were run also every second, whereas DRA algorithms were run every 0.1 s for computational simplicity reasons. In order to have access to the same set of solutions, the sets of resource quantities were reduced to $\Re_k \equiv \{0, 0.1M_k, 0.2M_k, \ldots, 1M_k\}$ for both RATs, where M_k is the quantity of available resources in one second.

V. Simulation Results

The results assessment has been divided into two parts. The first part is focused on the improvement achieved by a joint scheduling. Consequently, it compares the proposed HNN-based algorithm with the UMA-HNN and MBR-HNN algorithms. Next, the HNN algorithm is evaluated against the rest of reference algorithms previously defined.

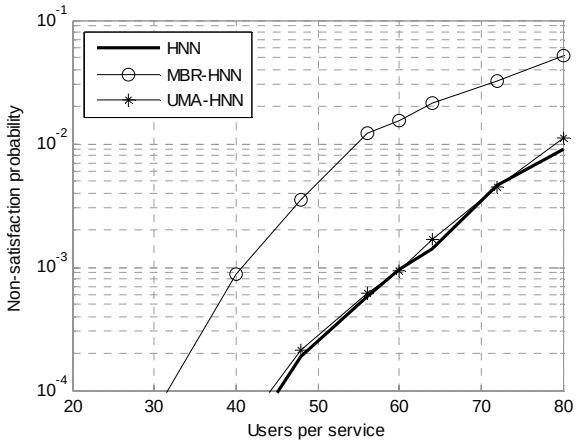

Figure 2. Probability of non-satisfaction with an increasing number of users.

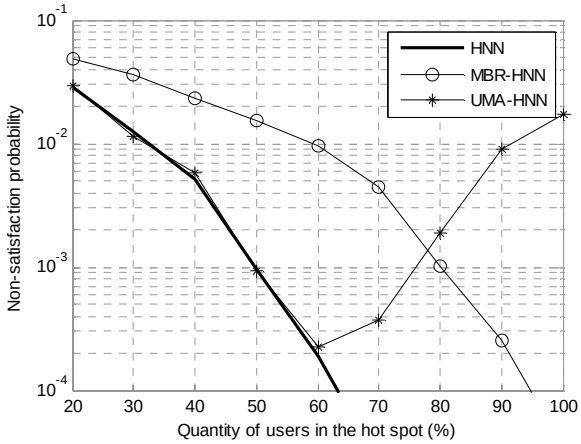

Figure 3. Probability of non-satisfaction for different load ratios in the hot spot.

A. Joint scheduling improvement

For this first study, the number of users per service and class ranged from 20, for the least loaded case, up to 80 for the most loaded case. Moreover, half the users were located in the hot spot whereas the other half moved around the entire cell. Figure 2 shows the mean probability of non-satisfaction or outage probability, i.e. the probability of not fulfilling the expected QoS, for the three algorithms studied in this section. The improvement of the JDRA is highly noticeable with respect to the MBR policy, whereas the difference with UMA policy is quite negligible, being the HNN-based JDRA algorithm slightly better. The good performance observed with the UMA policy is due to the fact that

the WLAN RAN is much less loaded than HSDPA. Consequently, this policy is always the best since it unloads HSDPA from users as soon as they are in the coverage area of the WLAN. Nevertheless, in a different scenario the UMA policy may not be the best one. Figure 3 shows the probability of non-satisfaction with a fixed number of users per service and class – 60 – distributed with different ratios between the hot spot and the entire cell. It can be observed that the MBR policy outperforms UMA when more than 80% of users are in the hot spot. Thus, if the WLAN coverage area is highly loaded it is not optimum to allocate all users to WLAN but distribute them among the overlapping RANs. It is worth highlighting that the proposed algorithm is the best one in all cases and, moreover, it extends the good behavior outlined by the UMA policy with low loaded WLAN to more saturated scenarios.

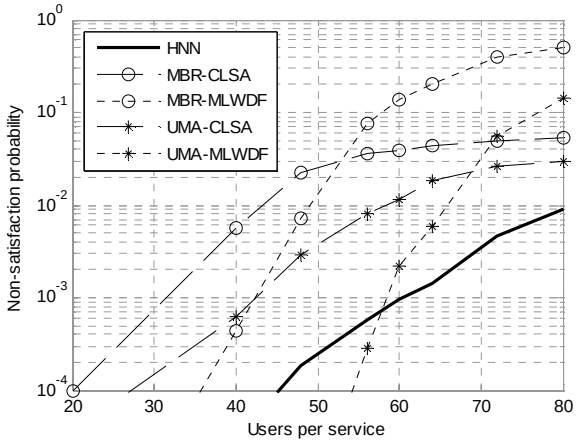

Figure 4. Probability of non-satisfaction with an increasing number of users.

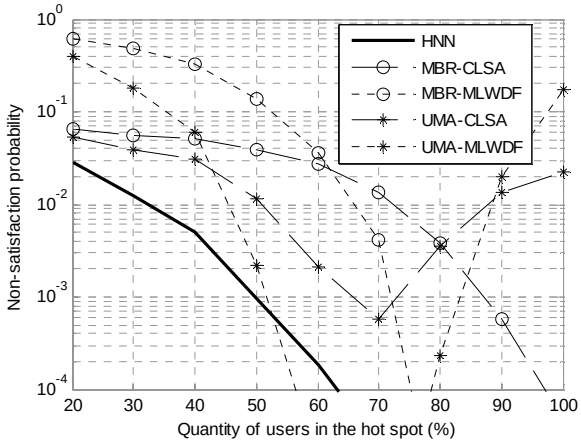

Figure 5. Probability of non-satisfaction for different load in the hot spot.

B. JDRA algorithm evaluation

Now, the same scenarios – 20 to 80 users per service and class equally split up into the hot spot and the entire cell and 60 users per service and class with different users distribution ratios between the hot spot and the entire cell – were simulated with the rest of algorithms. Non-satisfaction results are depicted in Figures 4 and 5. Performance of UMA and MBR policies are equivalent to that previously explained. If at least 80% of users are in the hot spot MBR outperforms UMA policy. Concerning the proposed JDRA algorithm, it is the best one in almost all situations. Only the UMA-MLWDF mix behaves better when RANs are sufficiently low loaded. Since HNNs find suboptimal solutions, they are not able to approach the optimum as UMA-MLWDF does in those cases. Nevertheless, the difference is not of much relevance taking into account that in comparison the proposed algorithm can reduce the non-satisfaction probability from 13% to 0.9% with 80 users per service and class. Besides, the behavior of UMA-MLWDF is highly dependent on the relative distribution between RATs and, therefore, it cannot be considered as a reliable algorithm for whatever heterogeneous system.

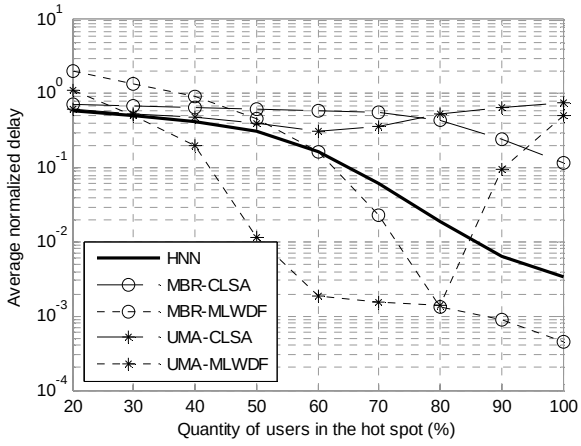

Figure 6. Average delay normalized by maximum delay.

For delay-based services, the average delay normalized by the maximum delay is depicted in Figure 6, considering 60 users per service and class. From this graph, the UMA-MLWDF could be considered a better approach than the one proposed in this paper, particularly if more than the 30% of users are in the hot spot. Nevertheless, although UMA-MLWDF reduces the average delay about 30 times at 50%

and about 75 times at 60% this fact is not reflected in the non-satisfaction probability (recall Figure 5), where UMA-MLWDF doubles the non-satisfaction at 50% and is only 10 times lower at 60%. This difference in the average delay and non-satisfaction probability stresses that resources are not being allocated to the correct users, i.e. those that need them more. Moreover, note that again the behavior of UMA-MLWDF is highly degraded when the WLAN network gets overloaded. In that case the normalized delay for UMA-MLWDF can be up to 170 times greater than the HNN-based algorithm. On the other hand, the MBR-MLWDF approach shares also this behavior, although in this case the non-satisfaction probability is much worse than that achieved with HNN. The main reason for this difference is that MLWDF is not able to distinguish between different user classes and, hence, the non-satisfaction probability of high priority classes increases despite of the average delay decreasing.

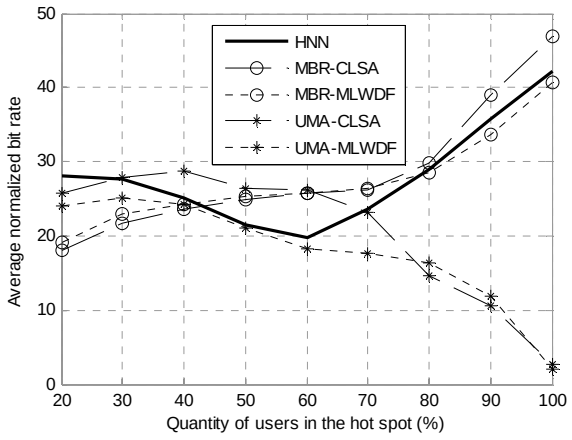

Figure 7. Average bit rate normalized by minimum bit rate.

Finally, Figure 7 shows the average bit rate normalized by the minimum target bit rate for 60 users per service and class. The previous effect is not significant and all techniques have very similar behavior until 70% of users in the hot spot. From that point, the UMA policy starts allocating less bit rate to users than the rest of techniques due to the saturation of WLAN.

VI. Conclusions

This paper has presented a JDRA algorithm that jointly distributes users of diverse types of service and classes among RANs of different technologies and the RANs resources among users.

HNN have been used to solve this complex problem. The neuron parallel interconnection of these networks have made the definition of the algorithm easier as shown in the rationale followed in section III.D. Moreover and most importantly, hardware implementations of HNNs have very fast response times what makes feasible a real-time functioning of the algorithm.

The JDRA algorithm has shown a significant reduction in the non-satisfaction probability of users as compared with other RAT selection techniques. Moreover, it approaches the UMA policy when optimum. The proposed algorithm is also better than other DRA techniques proposed in the literature unless for low loaded networks. Despite of the sub-optimum nature of HNN solutions, in those cases the MLWDF algorithm is near optimum, which justifies the slight differences. Nevertheless, the region where MLWDF outperforms the proposed algorithm is below a threshold of non-satisfaction probability of 0.05%. Therefore, it is preferred the use of the HNN-based JDRA algorithm instead of MLWDF since it reduces the non-satisfaction probability from 13% to 0.9% for other cases

References

[1] E. Gustafsson and A. Jonsson, “Always best connected,” IEEE Wireless Communications, vol. 10, no. 1, pp. 49-55, 2003.

[2] J. Pérez-Romero, O. Sallent, R. Agustí and M. A. Díaz-Guerra, “Radio resource management strategies in UMTS,” Ed. John Wiley & Sons, 2005.

[3] 3GPP TR 25.891 v0.3.0 “Improvement of RRM across RNS and RNS/BSS (Post Rel-5),” 2003.

[4] C. Mihailescu, X. Lagrange, and P. Godlewski, “Performance evaluation of a dynamic resource allocation algorithm for UMTS-TDD systems,” 51st IEEE Vehicular Technology Conference (VTC), Tokyo. Spring 2000.

[5] A. K. Parekh and R. G. Gallager, “A generalized processor sharing approach to flow control in integrated services networks: the single-node case,” IEE/ACM Transactions on Networking, vol. 1, no. 3, pp. 344-357, 1993.

[6] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting and R. Vijayakumar, “Providing quality of service over a shared wireless link,” IEEE Communications Magazine, vol. 39, no. 2, pp. 150-154, 2001.

[7] Q. Liu, X. Wang and G. B. Giannakis, “A cross-layer scheduling algorithm with QoS support in wireless networks,” IEEE Transactions on Vehicular Technology, vol. 55, no. 3, pp. 839-847, 2006.

[8] J. Pérez-Romero, O. Sallent and R. Agustí, “Policy-based initial RAT selection algorithms in heterogeneous networks,” 7th International Workshop on Mobile and Wireless Communications Network (MWCN), Marrakesh. Fall 2005.

[9] J. Pérez-Romero, O. Sallent and R. Agustí, “A novel metric for context-aware RAT selection in wireless multi-access systems,” IEEE International Conference on Communications (ICC), Glasgow. Spring 2007.

[10] 3GPP TS 43.318 V8.2.0, “Generic Access Network (GAN); stage 2.”

[11] 3GPP TS 44.318 V8.3.0, “Generic Access Network (GAN); mobile GAN interface layer 3 specification.”

[12] D. Gómez-Barquero, D. Calabuig, J. Monserrat, N. García and J. Pérez-Romero, “Hopfield neural network – based approach for joint dynamic resource allocation in heterogeneous wireless networks,” 64th IEEE Vehicular Technology Conference (VTC), Montreal. Fall 2006.

[13] D. Calabuig, J. Monserrat, D. Martín-Sacristán and N. Cardona, “Joint dynamic resource allocation for coupled heterogeneous wireless networks based on Hopfield neural networks,” 67th IEEE Vehicular Technology Conference (VTC), Singapore. Spring 2008.

[14] C. W. Ahn 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, 2004.

[15] J.J. Hopfield and D.W. Tank, "'Neural' computation of decisions in optimization problems," Biological Cybernetics, vol. 52, no. 3, pp. 141-152, 1985.

[16] K. C. Tan, H. Tang and S. S. Ge, "On parameter settings of Hopfield networks applied to traveling salesman problems," IEEE Transactions on Circuits and Systems, vol. 52, no. 5, pp. 994-1002, 2005.

[17] T.-N. Le and C.-K. Pham, "A new N-parallel updating method of the Hopfield type neural network for N-queens problem," IEEE International Joint Conference on Neural Networks (IJCNN), Montreal. Spring 2005.

[18] E. Del Re, R. Fantacci and L. Ronga, "A dynamic channel allocation technique based on Hopfield neural networks," IEEE Transactions on Vehicular Technology, vol. 45, No. 1, pp. 26-32, 1996.

[19] O. Lázaro and D. Girma, "A Hopfield neural-network-based dynamic channel allocation with handoff channel reservation control", IEEE Transactions on Vehicular Technology, vol. 49, no. 5, pp. 1578-1587, 2000.

[20] D. Calabuig, J. F. Monserrat, D. Gómez-Barquero and N. Cardona, "A delay-centric dynamic resource allocation algorithm for wireless communication systems based on HNN," IEEE Transactions on Vehicular Technology, vol. 57, No. 6, pp. 3653-3665, 2008.

[21] J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," Proceedings of the National Academy of Sciences of the United States of America, vol. 81, no. 10, pp. 3088-3092, 1984.

[22] G. Joya, M. A. Atencia and F. Sandoval, "Hopfield neural networks for optimization: study of the different dynamics," Neurocomputing, vol. 43, no. 1, pp. 219-237, 2002.

[23] 3GPP2-TSGC5, "HTTP and FTP Traffic Model for 1xEV-DV Simulations."

[24] S. Choudhury and J. D. Gibson, "Payload length and rate adaptation for throughput optimization in wireless LANs," IEEE Vehicular Technology Conference (VTC), Melbourne. Spring 2006.

[25] F. Brouwer, I. de Bruin, J. C. Silva, N. Souto, F. Cercas and A. Correia, "Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS," IEEE International Symposium Spread Spectrum Techniques and Applications (ISSSTA), Sydney. Fall 2004.