A Cellular Neural Network and Utility-Based Radio Resource Scheduler for Multimedia CDMA Communication Systems

Scott Shen, Chung-Ju Chang, *Fellow, IEEE*, and Li-Chun Wang, *Senior Member, IEEE*

***Abstract*—The paper proposes a cellular neural network and utility (CNNU)-based radio resource scheduler for multimedia CDMA communication systems supporting differentiated quality-of-service (QoS). Here, we define a relevant utility function for each connection, which is its radio resource function weighted by a QoS requirement deviation function and a fairness compensation function. We also propose cellular neural networks (CNN) to design the utility-based radio resource scheduler according to the Lyapunov method to solve the constrained optimization problem. The CNN is powerful for complicated optimization problems and has been proved that it can rapidly converge to a desired equilibrium; the utility-based scheduling algorithm can efficiently utilize the radio resource for system, keep the QoS requirements of connections guaranteed, and provide the weighted fairness for connections. Therefore, the CNNU-based scheduler, which determines a radio resource assignment vector for all connections by maximizing an overall system utility, can achieve high system throughput and keep the performance measures of all connections to meet their QoS requirements. Simulation results show that the CNNU-based scheduler attains the average system throughput greater than the EXP [9] and the HOLPRO [5] scheduling schemes by an amount of 23% and 33%, respectively, in the QoS guaranteed region.**

***Index Terms*—Cellular neural networks (CNN), fairness, quality of service (QoS), radio resource, scheduling, utility function.**

I. Introduction

In future wireless networks, heterogeneous and customized services with diverse traffic characteristics and QoS requirements are expected to be provided via a number of air interfaces. Also, multimedia applications are commonly accepted as enabling services, which are categorized into several classes [1]. To meet various traffic characteristics and QoS requirements of these potential applications, a sophisticated scheduling algorithm plays an essential role so that the system resource allocation is optimal, while retaining pre-defined QoS requirements and fairness among them.

Several studies on radio resource scheduling and allocation among connections in wireless networks with consideration of physical layer processing, power control range, and link conditions have been carried out [2]-[3]. Bhargharvan, Lu, and Nandagopal [4] developed a framework to achieve long-term fairness in wireless networks.

There are schemes considering either delay bound or minimum rate as its QoS requirements. For those schemes considering delay requirements, Varsou and Poor [5] proposed a head-of-line pseudo-probability (HOLPRO) scheduling algorithm, based on an earliest deadline first (EDF) concept, adapted to wireless environments. They also proposed a simple analysis for the performance of the generalized powered earliest deadline first (PEDF) and the HOLPRO scheduling schemes [6]. Stolyar and Ramanan studied a throughput-optimal scheduling algorithm for delay bounded system [7]; a variational scheduling algorithm for rate guarantee was also investigated. For non-real-time interactive connections, the rate guarantee is desirable. Kam and Siu considered the minimal rate guarantee with fairness in their proposed scheme [8]. Moreover, some schemes considered joint scheduling criteria to deal with complicated needs for systems. Shakkottai and Stolyar [9] considered both link quality and QoS requirements as the criteria and derived an exponential rule (EXP) scheduling scheme via fluid Markovian techniques.

Many of these scheduling algorithms above, [2]-[3], [7]-[10], were formulated in utility-based approaches. Generally, the utility function is defined as the benefit from receiving an amount of service for each connection so that the overall utility is maximized, in addition to fulfilling the design constraints such as QoS requirements and fairness.

The utility-based scheduling algorithm over radio channels is usually formulated as a complicated optimization problem with real-time requirement. To solve the complicated constrained optimization problem, many intelligent techniques have been applied successfully, for example, genetic algorithm, feed-forward neural network, and generalized Hopfield neural networks (HNN) [11], [12]. Among those, the class of generalized HNN has been mostly adopted for real-time tasks. However, the stability and spurious response problems make the HNN ineffective in practical applications. A special type of HNN, named *cellular neural network* (CNN), has been proposed [13] and has been proved that it can rapidly converge to a desired equilibrium on vertex along the prescribed trajectories by applying a proper learning or design procedure [14], [15]. It can converge to an unique equilibrium as long as the energy function corresponding to the architecture of CNN is properly designed, while HNN may converge to a

Manuscript received November 7, 2007; revised July 28, 2008 and April 17, 2009; accepted July 30, 2009. The associate editor coordinating the review of this paper and approving it for publication was V. K. Bhargava.

This work was supported by the National Science Council of Taiwan, ROC, under contract number NSC 97-2221-E-009-098-MY3, and the Ministry of Education (MOE) under ATU Program 98w959.

C.-J. Chang is with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan (e-mail: cjchang@mail.nctu.edu.tw).

Digital Object Identifier 10.1109/TWC.2009.071242

local minimum. Also, the CNN has digital outputs which is a vertex located in its state space, but the HNN has analog outputs and may converge to spurious response. Moreover, the CNN has an architecture that all cells (neurons) are with the same structure, i.e. the same inter-connection weights and bias current. Accomplished by locally recurrent inter-connection, it has much fewer number of inter-connection than HNN and more suitable for VLSI implementation. The CNN has been applied in telecom switching systems to optimize the system throughput constrained on given QoS requirements [16].

In this paper, we propose a CNN and utility (CNNU)-based radio resource scheduler, which adopts CNN to solve the complicated optimization problem of the utility-based scheduling algorithm, for downlinks of multimedia CDMA communication systems. This paper extends the work in [17] to include how the CNN architecture is modified for the optimization problem of the radio resource scheduling and what basic assumptions are needed to make.

The proposed CNNU-based scheduler contains a utility function (UF) preprocessor, a radio-resource range (RR) decision maker, and a CNN processor. A relevant utility function of each connection is designed in the UF preprocessor. It jointly considers radio resource efficiency, diverse QoS requirements, and fairness. It is a radio resource function weighted by both its QoS requirement deviation function and its fairness compensation function. The UF preprocessor also generates a matrix of normalized utility functions of all connections. The RR decision maker determines a matrix showing the upper limit of radio resource assignment for each connection. The CNN processor receives the matrix of normalized utility functions and the matrix of upper limits of radio resource assignment vector as inputs. It determines an *optimal normalized radio resource assignment vector* for connections in multimedia CDMA cellular systems by minimizing the system cost function, which is in terms of the overall system utility function under system constraints of maximum transmission power, minimum spreading factor, and remaining queue length. The architecture of the CNN processor is constructed by a Lyapunov method [18], [19], via mapping the system cost function to a proper energy function. It is designed in a two-layered configuration, which consists of a decision layer and an output layer, to reduce the number of inter-connections in the CNN. It can be shown that the stability exists and the stable equilibriums locate in the desired state space. The CNN is powerful for complicated optimization problems.

Simulation results show that CNNU-based scheduler outperforms the EXP [9] (HOLPRO [5]) scheduling scheme in the average throughput by an amount of 15% (29%). Moreover, it also has higher maximum achievable throughput in the QoS guarantee region by 23% (33%) than the EXP (HOLPRO) scheme. The CNNU-based scheduler can allocate radio resources to different types of connections more adequately to achieve higher capacity and keep various QoS requirements fulfilled to a similar extent. Therefore, the CNNU-based scheduler is efficient and effective for multimedia CDMA cellular networks.

The rest of the paper is organized as follows. Section II presents the model of a considered multimedia CDMA cellular system. Section III proposes a relevant utility function. In section IV, a utility-based scheduling problem is formulated. In section V, the design of a CNNU-based scheduler is discussed. Simulation results to examine the performance of the CNNU-based scheduler are illustrated in section VI. Finally, the paper is concluded in section VII.

II. System Model

Assume that the multimedia CDMA cellular system has N real-time (RT) and non-real-time (NRT) active connections (users) in the downlink transmissions with chip rate W. The RT connections transmit on dedicated channels, while the NRT connections transmit on shared channels. For every active connection using either dedicated or shared channels, a fixed number of code channels with their corresponding spreading factors is given in the connection setup phase. A minimum spreading factor SF_i is therefore associated with the assigned code channels for connection i. The system radio resource is here defined to be the transmission power. It is limited by a maximum power budget denoted by P^*_{max} and scheduled to all connections every frame time period T_f.

For a downlink connection i, there are four QoS requirements defined in the packet level, such as BER^*_i, or the call level, such as delay bound D^*_i, packet dropping ratio $P^*_{D,i}$, and minimum transmission rate $R^*_{m,i}$. For RT connections, hard delay bound D^*_i exists and $P^*_{D,i}$ can be larger than zero; while for NRT connections, no explicit delay bound is imposed, but $R^*_{m,i} > 0$ should be satisfied for interactive connections and $R^*_{m,i} = 0$ be set for best effort connections.

Assume that the linkgain $\zeta_i(t)$ and the interference $\mathcal{I}_i(t)$ for connection i at the frame time t can be measured at the user side and perfectly signaled to the base station. Here $\zeta_i(t)$ consists of the mean path loss, long-term fading, and short-term fading. It is given by

$$\zeta_i(t) = d_i^{-4} \cdot 10^{\frac{\zeta_i^L(t)}{10}} \cdot \zeta_i^S(t), \tag{1}$$

where d_i is the distance between the user i and its base station, $\zeta_i^L(t)$ is the log-normal shadowing component, and $\zeta_i^S(t)$ is the Rayleigh-fading component. $\mathcal{I}_i(t)$ is given by

$$\mathcal{I}_i(t) = \left[(1-\alpha) P^*_{max} \cdot \zeta_i(t) + \sum_{b \in B_a} P^*_{max} \cdot \zeta_{i,b}(t) + N_0 W \right], \tag{2}$$

where α is the orthogonality factor for downlink, B_a is the set of adjacent base stations for connection i, $\zeta_{i,b}(t)$ is the link gain from base station b to connection i, and N_0 is the spectrum noise power density. The adaptive QAM modulation is adopted, and the modulation order M_{κ_i} with index κ_i for connection i would be determined according to the link gain quality and interference. The traffic source of connection i generates packets, and packets are queued in its individual buffer and the buffer size is infinite. Models for these traffic source are assumed to be on-off for RT connections, Pareto for NRT interactive connections, and batch Poisson with truncated geometrical batch size for NRT best effort connections.

III. The Utility Function

We define the utility function of each connection as the throughput achievable by the connection according to its link

condition weighted by both the deviation of QoS measurement from its QoS requirement and the amount of fairness compensation required by this connection. Mathematically, the utility function for connection i at frame time t, denoted by $\mathcal{U}_i(t)$, is given by

$$\mathcal{U}_i(t) = \mathcal{R}_i(t) \cdot \mathcal{A}_i(t) \cdot \mathcal{F}_i(t), \tag{3}$$

where $\mathcal{R}_i(t)$ is the radio resource function of connection i, $\mathcal{A}_i(t)$ is its QoS requirement deviation function, and $\mathcal{F}_i(t)$ is its fairness compensation function. Unlike those utility functions in existing works which consider only subset of above factors, this proposed utility function takes link condition, adaptive modulation, differentiated QoS requirements, and fairness into account. Thus the scheduling algorithm based on the proposed utility function can address more complicated situations.

A. Radio Resource Function $\mathcal{R}_i(t)$

The radio resource function $\mathcal{R}_i(t)$ is to denote the maximum achievable transmission rate the connection i can achieve. For connection i with modulation order M_{κ_i} of the adaptive QAM modulation scheme and the corresponding $(E_b/N_0)^*_{\kappa_i}$ to satisfy its BER_i^* requirement, the following inequality should be held,

$$\frac{W}{R_{s,i}(t)} \cdot \frac{c_i(t) \cdot P^*_{max} \cdot \zeta_i(t)}{\mathcal{I}_i(t)} \geq \left(\frac{E_b}{N_o}\right)^*_{\kappa_i}, \tag{4}$$

where $R_{s,i}(t)$ is its symbol rate and $c_i(t)$ is the normalized radio resource (power) assignment to connection i at frame time t. Since the channel state is assumed to be known and remain constant during a frame time, the $(E_b/N_0)^*_{\kappa_i}$ in (4) is given by [20]

$$(E_b/N_0)^*_{\kappa_i} = \frac{-(M_{\kappa_i} - 1) \cdot \ln\{5BER_i^*\}}{1.5}. \tag{5}$$

The BER_i^* in (5) can be obtained by

$$BER_i^* = 0.2 \int_{\gamma} \exp\left\{\frac{-1.5\gamma_i}{M_{\kappa_i} - 1}\right\} f_{\gamma_i}(\gamma) d\gamma, \tag{6}$$

where γ_i is the instantaneous $(E_b/N_0)_{\kappa_i}$ received by connection i and $f_{\gamma_i}(\gamma)$ is the pdf of γ_i [20]. Denote $R^*_{s,i}(t)$ the *maximum achievable symbol rate* that $(E_b/N_0)^*_{\kappa_i}$ can be fulfilled at $c_i(t) = 1$. Clearly, $R^*_{s,i}(t) = \frac{W}{(E_b/N_0)^*_{\kappa_i}} \cdot \frac{P^*_{max} \cdot \zeta_i(t)}{\mathcal{I}_i(t)}$. However, the $R^*_{s,i}(t)$ is further limited by $\frac{W}{SF_i}$ for a given minimum spreading factor SF_i of the allocated code channel. Thus the $R^*_{s,i}(t)$ can be obtained by

$$R^*_{s,i}(t) = \min\left\{\frac{W}{(E_b/N_0)^*_{\kappa_i}} \cdot \frac{P^*_{max} \cdot \zeta_i(t)}{\mathcal{I}_i(t)}, \frac{W}{SF_i}\right\}. \tag{7}$$

According to (7), the most efficient modulation order M_{κ_i} can be selected by the following inequality,

$$M_{\kappa_i} \leq \frac{1.5 \cdot SF_i \cdot P^*_{max} \cdot \zeta_i(t)}{\mathcal{I}_i(t) \cdot (-\ln\{5BER_i^*\})} + 1 \leq M_{(\kappa_i+1)}, \tag{8}$$

where M_{κ_i+1} denotes the next modulation order higher than M_{κ_i}. Since the information bit of one symbol is $\log_2 M_{\kappa_i}$, the radio resource function of connection i, $\mathcal{R}_i(t)$, can be consequently obtained by (9).

B. The QoS Requirement Deviation Function $\mathcal{A}_i(t)$

The QoS requirement deviation function $\mathcal{A}_i(t)$ is used to indicate the extent of deviation of connection i from its call-level QoS requirements. The larger the extent of deviation from the QoS requirements is, the more resource the connection be allocated. For RT connection i, a hard delay bound D_i^* is imposed on each packet. The QoS requirement of its packet dropping ratio due to excess delay must be $Prob\{D_i(t) > D_i^*\} < P^*_{D,i}$, where $D_i(t)$ is the waiting time delay of head-of-line packet at time t. For NRT interactive connection i, the QoS requirement is that its average transmission rate, $\mathbf{E}[r_i(t)]$, cannot be less than the minimum transmission rate, $R^*_{m,i}$, i.e. $\mathbf{E}[r_i(t)] \geq R^*_{m,i}$, where $\mathbf{E}[\cdot]$ indicates the expectation. As for NRT best-effort connection i, since no call level QoS requirements are guaranteed, $R^*_{m,i}$ is set to be 0.

The proposed *Modified Largest Weighted Delay First* (M-LWDF) algorithm in [21] suggests that an exponential rule [9] be the form with throughput optimal for the above call level QoS requirement constraints. Therefore, the *QoS requirement deviation function* $\mathcal{A}_i(t)$ is defined as (10).

In (10), $\overline{D}(t) = \frac{1}{N}\sum_i \left[\left(\frac{-\log(P^*_{D,i})}{D_i^*}\right) \cdot D_i(t)\right]$ is the average weighted delay, $\hat{L}_i(t) = \min\left\{\hat{L}_i(t-1) + \left(\frac{R^*_{m,i} - r_i(t)}{R^*_{m,i}}\right), L_{max,i}\right\}$ is the number of buckets, $L_{max,i}$ is the maximum buffer size of buckets for connection i, and $\overline{L}(t) = \frac{1}{N}\sum_i \hat{L}_i(t)$. In (10), for RT connections, the delay of connection i is weighted by the log-scale packet dropping ratio and the inverse of the delay bound requirements [9]. If the weighted delay is more than the average weighted delay of all connections, $\mathcal{A}_i(t)$ will be exponentially increased, and more resource will be scheduled. On the other hand, if the weighted delay is less than the average weighted delay, $\mathcal{A}_i(t)$ will be dramatically decayed, and less resource will be allocated. Similarly, for NRT interactive connections, if the accumulated difference of the guaranteed minimum transmission rate and the assigned rate is greater than the average value, more resource will be assigned. As to the NRT best-effort connections, this function is simply assigned to be 1. This is because none of the QoS requirements are imposed on NRT best-effort connections, and the constant 1 would take no effect on the utility function.

C. The Fairness Compensation Function $\mathcal{F}_i(t)$

The fairness compensation function is to ensure that RT connections using dedicated channels can attain a relative high priority over NRT connections using shared channels. It is also how the radio resource shared by all NRT best-effort connections without any QoS requirements is assigned according to a predefined target weighting factor. With the predefined target weighting factors w_i and w_k for NRT best-effort connections i and k, respectively, the radio resources are here expected to be allocated to them so that their average assigned transmission rates, $\mathbf{E}[r_i(t)]$ and $\mathbf{E}[r_k(t)]$, can be achieved according to the ratio: $\frac{\mathbf{E}[r_i(t)]}{\mathbf{E}[r_k(t)]} = \frac{w_i}{w_k}$ [4].

The *fairness compensation function* for connection i till time t, $\mathcal{F}_i(t)$, is defined by (11). The β_i in (11) is the

$$\mathcal{R}_i(t) = \log_2 M_{\kappa_i} \cdot R^*_{s,i}(t) = \frac{1.5W \cdot \log_2 M_{\kappa_i}}{(M_{\kappa_i} - 1) \cdot (-\ln\{5BER^*_i\})} \cdot \frac{P^*_{max} \cdot \zeta_i(t)}{\mathcal{I}_i(t)} \tag{9}$$

$$\mathcal{A}_i(t) = \begin{cases} \exp\left\{ \dfrac{\frac{-\log_{(P^*_{D,i})}}{D^*_i} \cdot D_i(t) - \overline{D}(t)}{1 + \left[\overline{D}(t)\right]^{1/2}} \right\}, & \text{if } i \in \{\text{RT connections}\} \\ \exp\left\{ \dfrac{\hat{L}_i(t) - \overline{L}(t)}{1 + \left[\overline{L}(t)\right]^{1/2}} \right\}, & \text{if } i \in \{\text{NRT interactive connections}\} \\ 1, & \text{if } i \in \{\text{NRT best-effort connections}\} \end{cases} \tag{10}$$

$$\mathcal{F}_i(t) = \begin{cases} \beta_i, & \text{if } i \in \{\text{RT connections}\} \\ \beta_0, & \text{if } i \in \{\text{NRT interactive connections}\} \\ max\left\{(w_i - \overline{w}_i(t)), 1\right\}, & \text{if } i \in \{\text{NRT best-effort connections}\} \end{cases} \tag{11}$$

priority bias for RT connections, β_0 is the basic reference value set for NRT interactive connections, and $\overline{w}_i(t)$ is the moving-average of $r_i(t)$. The β_i is usually larger than the β_0 for differentiation. For RT connections and NRT interactive connections, only priority bias is set and the weighted fairness is not considered due to their QoS-driven nature. For NRT best-effort connections, $(w_i - \overline{w}_i(t))$ indicates the unfairness of connection i. The more the extent of the unfairness of connection i is, the larger the $\mathcal{F}_i(t)$ will be; then more resource will be scheduled to connection i, and the $(w_i - \overline{w}_i(t))$ will be smaller afterwards. In the stationary situation, the unfairness of all NRT best-effort connections should be almost the same via the linear feedback control.

The target weighting factor w_i of NRT connection i is defined as its target average transmission rate. It is considered to be a function of its equivalent traffic source rate s^*_i [22], mean link gain $\overline{\zeta}_i$, mean interference level $\overline{\mathcal{I}}_i$, and its guaranteed minimum transmission rate, $R^*_{m,i}$. Also, w_i is proportional to its mean maximum transmission rate, $\frac{P^*_{max} \cdot \overline{\zeta}_i}{\overline{\mathcal{I}}_i \cdot (\frac{E_b}{N_0})^*_i}$, and its normalized effective bandwidth, $\frac{s^*_i}{\sum_k s^*_k}$. Therefore, we have define w_i as

$$w_i = \max\left\{ \frac{P^*_{max} \cdot \overline{\zeta}_i}{\overline{\mathcal{I}}_i \cdot (\frac{E_b}{N_0})^*_i} \cdot \frac{s^*_i}{\sum_k s^*_k},\ R^*_{m,i} \right\}, \tag{12}$$

where $(\frac{E_b}{N_0})^*_i$ is the required E_b/N_0 to achieve BER^*_i of connection i using the least-order modulation scheme. The s^*_i can be given according to the *effective bandwidth method* proposed in [22], [23].

The lower bound $R^*_{m,i}$ for w_i is to avoid the starvation problem of the connection i in bad link condition. Note that $R^*_{m,i} = 0$ for the best-effort connection, and the target weighting factor of the best-effort connection is usually less than that of the interactive connection.

The priority bias β_i for RT connection i is a relative margin for $\zeta_i(t)$ over the link gains of NRT connections, and is a function of its transmission suspension threshold ζ^*_i, the average of the mean link gains of all NRT connections $\overline{\zeta}_{NRT}$, and the E_b/N_0 requirements. We have defined β_i as

$$\beta_i = \left(\frac{\overline{\zeta}_{NRT}}{\zeta^*_i} \cdot \frac{(\frac{E_b}{N_0})^*_i}{(\frac{E_b}{N_0})^*_{NRT}} \right) \cdot \beta_0. \tag{13}$$

This β_i makes the product $\mathcal{R}_i(t) \cdot \mathcal{F}_i(t)$ of the RT connection i greater than the product of NRT connections till $\zeta_i(t) \geq \zeta^*_i$. Therefore, RT connections can share the radio resource with relatively higher priority than NRT connections via the setting of priority bias β_i. Moreover, β_0 is designed to protect the NRT interactive connections against the NRT best-effort connections capturing the radio resource in the overloaded situation. Here, β_0 is defined as $\beta_0 = \overline{L}(t)$, where $\overline{L}(t)$ is the averaged bucket size for NRT interactive connections. The higher the bucket size of NRT interactive connections is, the less important the weighted fairness of NRT best-effort connections would be.

IV. Problem Formulation

The utility-based scheduling problem is formulated as a constrained optimization problem given by

$$\vec{c}^{\,*}(t) \quad = \quad \arg\max_{\vec{c}\,(t)} \left\{ \sum_i^N c_i(t) \cdot \mathcal{U}_i(t) \right\},$$

subject to constraints:

$$\Psi_1 = \left\{ \vec{c}(t) : \sum_n c_n(t) \leq 1 \right\},$$
$$\Psi_2 = \left\{ \vec{c}(t) : c_i(t) \leq \min\left\{ \frac{W \cdot \log_2 M_{\kappa_i}}{SF_i \cdot \mathcal{R}_i(t)}, \frac{Q_i(t)/T_f}{\mathcal{R}_i(t)} \right\}, \forall i \right\}, \tag{14}$$

where $\vec{c}(t)$ is the normalized radio resource (power) assignment vector determined at the t-th frame and $\vec{c}(t) = (c_1(t), \ldots, c_i(t), \ldots, c_N(t))$, $\sum_i^N c_i(t) \cdot \mathcal{U}_i(t)$ is the overall system utility function, and Ψ_1 and Ψ_2 are two constraints on $\vec{c}(t)$. The constraint Ψ_1 is because of the system transmission power limited by a maximum power budget P^*_{max}. Notice that the assigned transmission rate to connection i at the t-th frame, denoted by $r_i(t)$, is determined according to the $c_i(t)$ and the modulation order M_{κ_i}; and the $r_i(t)$ is further limited by the minimum spreading factor SF_i and the waiting queue length $Q_i(t)$. The constraint Ψ_2 indicates that no further utility can be gained if $r_i(t)$ exceeds the supported rate which is the rate when the power ratio $c_i(t)$ equals $(\frac{W \cdot \log_2 M_{\kappa_i}}{SF_i \cdot \mathcal{R}_i(t)})$, or the necessary rate to transmit all remaining packets in $Q_i(t)$ at the t-th frame, which equals $(\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)})$. The optimal transmission rate for connection i at the t-th frame, denoted by $r^*_i(t)$, is then equal to $c^*_i(t) \cdot \mathcal{R}_i(t)$.

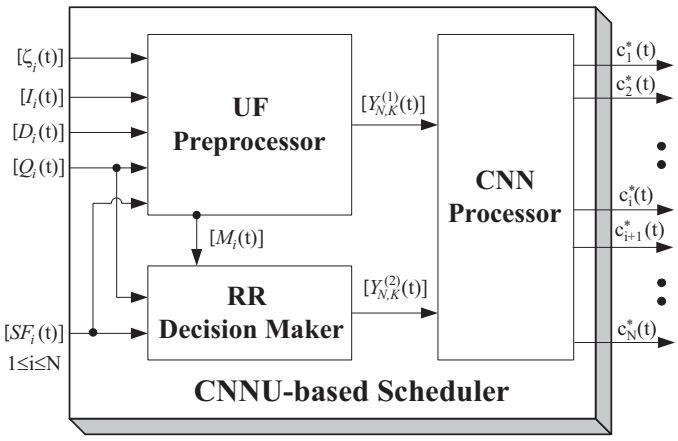


Fig. 1. The block diagram of CNNU-based scheduler.

V. The CNNU-based Scheduler

Figure 1 shows the block diagram of the CNN and utility (CNNU)-based scheduler. It contains a *utility function* (*UF*) *preprocessor*, a *radio-resource range* (*RR*) *decision maker*, and a *CNN processor*. The proposed CNNU-based scheduler takes the link information ($\zeta_i(t)$), interference ($\mathcal{I}_i(t)$), delay ($D_i(t)$), queue length ($Q_i(t)$), and spreading factor (SF_i), $1 \le i \le N$, as inputs at the t-th frame, and finally outputs an optimal normalized radio resource (power) assignment vector $\vec{c}^{\,*}(t) = (c_i^*(t), \ldots, c_N^*(t))$, where $c_i^*(t)$, $1 \le i \le N$, is expressed by K bits. The quantization of the radio resource assignment vector into K bits is due to the nature of the binary expression of each neuron for the CNN processor. The higher the quantization level is, the higher the precision will be, but the more the convergence time of the CNN processor would be taken. Note that the $\vec{c}^{\,*}(t)$ can be for the current frame or the next frame, depending on the convergence time of the CNN processor.

The *UF preprocessor* first calculates the utility function $\mathcal{U}_i(t)$ given in (3), $1 \le i \le N$. Then it normalizes $\mathcal{U}_i(t)$ by a *compression function* $(1 - e^{-\sigma \mathcal{U}_i(t)})$, expresses $(1 - e^{-\sigma \mathcal{U}_i(t)})$ to be an $1 \times K$ vector given by (15), and finally constructs an $N \times K$ input matrix $\left[Y_{i,k}^{(1)}\right]$ for the CNN processor, where $Y_{i,k}^{(1)} = (1 - e^{-\sigma \mathcal{U}_i(t)}) \cdot 2^{-k}$. Notice that the σ is a constant related to the slope and the linear region of the compression function $(1 - e^{-\sigma \mathcal{U}_i(t)})$, which normalizes $U_i(t) \in [0, \infty)$ into the range of $[0, 1)$. A good compression function is the one with broad linear range so that the individual utility function is normalized linearly within a reasonable range. This normalization is to implement the first constraint in (14). *The UF preprocessor* also determines a vector of modulation order $[M_{\kappa_i}]$ for all connections according to (8) and outputs to the *RR decision maker*. The *RR decision maker* determines the upper limit for the radio resource assignment for every connection i and expresses it by a $1 \times K$ vector which is the upper limit multiplied by the bit-weighted vector $(2^{-1}, \ldots, 2^{-k}, \ldots, 2^{-K})$, $1 \le i \le N$. Then the *RR decision maker* constructs the second input matrix $\left[Y_{i,k}^{(2)}\right]$, $1 \le i \le N$, $1 \le k \le K$, where the element $Y_{i,k}^{(2)}$ is given by $\left(\min\left\{\frac{W \cdot \log_2 M_{\kappa_i}}{SF_i \cdot \mathcal{R}_i(t)}, \frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right) \cdot 2^{-k}$ to implement the second constraint in (14). The *CNN processor* receives the two input matrices, $\left[Y_{i,k}^{(1)}\right]$ and $\left[Y_{i,k}^{(2)}\right]$, and computes the optimal radio resource assignment vector $\vec{c}^{\,*}(t)$ for the t-th frame. During the computation process, denote by τ the instantaneous time index of the CNN and by $\vec{c}(t, \tau)$ the instantaneous radio resource assignment vector at time τ during the frame t. For each $c_i(t, \tau)$, $1 \le i \le N$, it is represented by K bits, $X_{i,k}(\tau)$, $1 \le k \le K$, and $c_i(t, \tau)$ can be expressed by

$$c_i(t, \tau) \cong \sum_{k=1}^{K} X_{i,k}(\tau) \cdot 2^{-k}. \tag{16}$$

When the CNN processor arrives at an equilibrium, the output will converge to the optimal radio resource assignment vector, ie., $\lim_{\tau \to \infty} \vec{c}(t, \tau) = \vec{c}^{\,*}(t)$.

In the following, the design of the CNN processor for the CNNU-based scheduler is described. Characteristics of the original CNN proposed in [13] are first briefed. Then, we define a cost function associated with the constrained optimization problem formulated in (14). Subsequently, we design an energy function based on the cost function. According to the trajectory of the energy function, we can construct the architecture of the desired CNN processor to fit the defined scheduling problem. Notice that it can be proven that the designed CNN processor can be with good stability and convergence [23], based on the principle of Lyapunov and the stability theorem of CNN [18], [19].

A. Preliminaries for Cellular Neural Networks

Consider a neural network with $N \times K$ neurons arranged in a rectangular array, where neuron (i, k) is denoted by $z_{i,k}$. The output of $z_{i,k}$ at time τ, denoted by $X_{i,k}(\tau)$, can be expressed by $X_{i,k}(\tau) = f\left(X_{i,k}^{(s)}(\tau)\right)$, where $f(x) = \frac{1}{2}\left[|x| - |x-1|\right] + \frac{1}{2}$ is an activation function of $z_{i,k}$ and $X_{i,k}^{(s)}(\tau)$ is the state variable of $z_{i,k}$ at time τ. The $X_{i,k}^{(s)}(\tau)$ consists of recurrent inputs, external inputs, and a bias current. Each neuron $z_{i,k}$ connects with all other neurons within its neighborhood, denoted by $Z_n(i, k)$. The area of $Z_n(i, k)$ is determined according to the design of the neural network. Generally, the dynamics of the CNN at time τ is represented by [13], [14], as (17), where ν in (17) is a time constant for all neurons, $A_{i,k;j,m}$ is the recurrent inter-connection weight from neuron $z_{j,m}$ to $z_{i,k}$, $B_{i,k;j,m}$ is the control weight of external input from $z_{j,m}$ to $z_{i,k}$, $Y_{j,m}$ is the external input to the neuron $z_{j,m}$, and $V_{i,k}$ is the bias current to $z_{i,k}$, which is usually a fixed value V. It is worth mentioning that $A_{i,k;i,k} > 1/\nu$ is held so that the neuron $z_{i,k}$ will eventually enter into a saturation region [13]. Also, the inter-connection weights are assumed to be symmetric, that is, $A_{i,k;j,m} = A_{j,m;i,k}$, thus the CNN is stable [13].

An energy function at time τ which decreases along the trajectories of (17), denoted by $\mathcal{E}(\tau)$, is generally expressed by [13], as (18). At the stable state, outputs of neurons will arrive at an equilibrium with the minimum energy function. If the energy function is properly designed and acts as a cost function, such an optimization problem can be solved via the Lyapunov method [15]-[19]. By the Lyapunov method, the CNN can be designed with a set of prescribed trajectories.

$$\left((1-e^{-\sigma\mathcal{U}_i(t)})\cdot 2^{-1},\ldots,(1-e^{-\sigma\mathcal{U}_i(t)})\cdot 2^{-k},\ldots,(1-e^{-\sigma\mathcal{U}_i(t)})\cdot 2^{-K}\right) \tag{15}$$

$$\begin{aligned}\frac{dX_{i,k}^{(s)}(\tau)}{d\tau} &= -\frac{X_{i,k}^{(s)}(\tau)}{\nu}+A_{i,k;i,k}\cdot X_{i,k}(\tau)+B_{i,k;i,k}\cdot Y_{i,k}\\ &\quad+\sum_{z_{j,m}\in Z_n(i,k)}A_{i,k;j,m}\cdot X_{j,m}(\tau)+\sum_{z_{j,m}\in Z_n(i,k)}B_{i,k;j,m}\cdot Y_{j,m}+V_{i,k}\end{aligned} \tag{17}$$

$$\begin{aligned}\mathcal{E}(\tau) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}A_{i,k;i,k}X_{i,k}^2(\tau)-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{m=1}^{K}A_{i,k;j,m}X_{j,m}(\tau)X_{i,k}(\tau)\\ &\quad-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{m=1}^{K}B_{i,k;j,m}Y_{j,m}X_{i,k}(\tau)-\sum_{i=1}^{N}\sum_{k=1}^{K}V_{i,k}\cdot X_{i,k}(\tau)\end{aligned} \tag{18}$$

The trajectories are described by the gradient of the Lyapunov function $\mathcal{E}(\tau)$ which is the energy of the CNN network at time τ. With an appropriate energy function designed according to the cost function, the minimization of the cost can be achieved along the designed trajectories. In the mean time, it can be proved that the architecture of the designed CNN can be related with the energy function by

$$\frac{dX_{i,k}^{(s)}(\tau)}{d\tau}=-\frac{X_{i,k}^{(s)}(\tau)}{\nu}-\frac{\partial\mathcal{E}(\tau)}{\partial X_{i,k}(\tau)}. \tag{19}$$

Using (19), the desired system parameters of inter-connection weights, control weights, and bias currents can be found from the trajectories of energy function.

B. Cost Function of CNN Processor

The cost function of the proposed CNN processor to achieve the optimal resource allocation at time τ during frame t, denoted by $\mathcal{H}(t,\tau)$, consists of a cost function for the utility function, denoted by $\mathcal{H}_u(t,\tau)$, in conjunction with cost functions for system constraints Ψ_1 and Ψ_2 given in (14), denoted by $\mathcal{H}_{\Psi_1}(t,\tau)$ and $\mathcal{H}_{\Psi_2}(t,\tau)$, respectively. The $\mathcal{H}(t,\tau)$ has the form of

$$\mathcal{H}(t,\tau) = \mathcal{H}_u(t,\tau)+\mathcal{H}_{\Psi_1}(t,\tau)+\mathcal{H}_{\Psi_2}(t,\tau). \tag{20}$$

The $\mathcal{H}_u(t,\tau)$ is defined to be the difference between an overall normalized utility function and its maximum, where the overall normalized utility function is defined as $\sum_{i=1}^{N}c_i(t,\tau)\cdot(1-e^{-\sigma\mathcal{U}_i(t)})$. When $\sum_{i=1}^{N}c_i(t,\tau)\le 1$, $\sum_{i=1}^{N}c_i(t,\tau)\cdot(1-e^{-\sigma\mathcal{U}_i(t)})$ is bounded by 1. Thus the $\mathcal{H}_u(t,\tau)$ is given by

$$\mathcal{H}_u(t,\tau)=\eta_0\left[1-\sum_{i=1}^{N}c_i(t,\tau)\left(1-e^{-\sigma\mathcal{U}_i(t)}\right)\right], \tag{21}$$

where η_0 is the coefficient of $\mathcal{H}_u(t,\tau)$.

The $\mathcal{H}_{\Psi_1}(t,\tau)$ is defined as

$$\mathcal{H}_{\Psi_1}(t,\tau)=\eta_1\left[\sum_{i=1}^{N}c_i(t,\tau)-1\right]^2, \tag{22}$$

where $\eta_1 = \eta_1^+\cdot u\left(\sum_{i=1}^{N}c_i(t,\tau)-1\right)+\eta_1^-\cdot\left(1-u\left(\sum_{i=1}^{N}c_i(t,\tau)-1\right)\right)$, $u(\cdot)$ is the unit-step function, η_1^+ is the slope constant for the cost increment when the total radio resource is greater than the maximum, and η_1^- is the slope constant for the cost increment otherwise. The ranges of η_1^+ and η_1^- are further investigated in the next section to ensure the stability and the desired output pattern of the *CNN processor*.

The $\mathcal{H}_{\Psi_2}(t,\tau)$ is defined to be proportional to the difference $\left[c_i(t,\tau)-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\ \frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right]$ if $c_i(t,\tau)>\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}$; otherwise, no cost will be incurred because the radio resource will be allocated to other connections for efficiency. It is given by (23), where η_2 is the coefficient of $\mathcal{H}_{\Psi_2}(t,\tau)$; $(a)^+=a$ if $a\ge 0$, $(a)^+=0$ if $a<0$.

C. The Architecture of CNN Processor

According to the cost function $\mathcal{H}(t,\tau)$ at time t, the energy function $\mathcal{E}(t,\tau)$ can be designed for the CNN processor in the CNNU-based scheduler. However, some modifications on $\mathcal{H}(t,\tau)$ in (20) should be made for $\mathcal{E}(t,\tau)$ to ensure the correctness of the desired output and the stability of the CNN processor. The $\mathcal{E}(t,\tau)$ is given by (24), where η_3 is a constant for additional auxiliary terms. The first three terms of (24) are convex function for the $\mathcal{E}(t,\tau)$, which are transformed from the three concave functions of the $\mathcal{H}(t,\tau)$, respectively. However, several constants are removed to simplify the architecture of CNN with the equilibrium unchanged. The first item differs from (21) in that the scalar 1 is ignored and the remaining term $\sum_{i=1}^{N}(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})\cdot(1-e^{-\sigma\cdot\mathcal{U}_i(t)})$ is bounded above by 1 and has the same minimum as in the cost function. For the second and the third terms, the quadratic forms in (22) and (23) are replaced by convex functions which merely contain state variable $X_{i,k}(\tau)$ without any scalar. The local minimums would be the same; the resulting energy at any equilibrium would be shifted by a constant value, compared to the cost in (20), and independent of the inputs and the output pattern. The last term of (24) is an auxiliary factor to ensure the convergence of the whole CNN to one of its vertex.

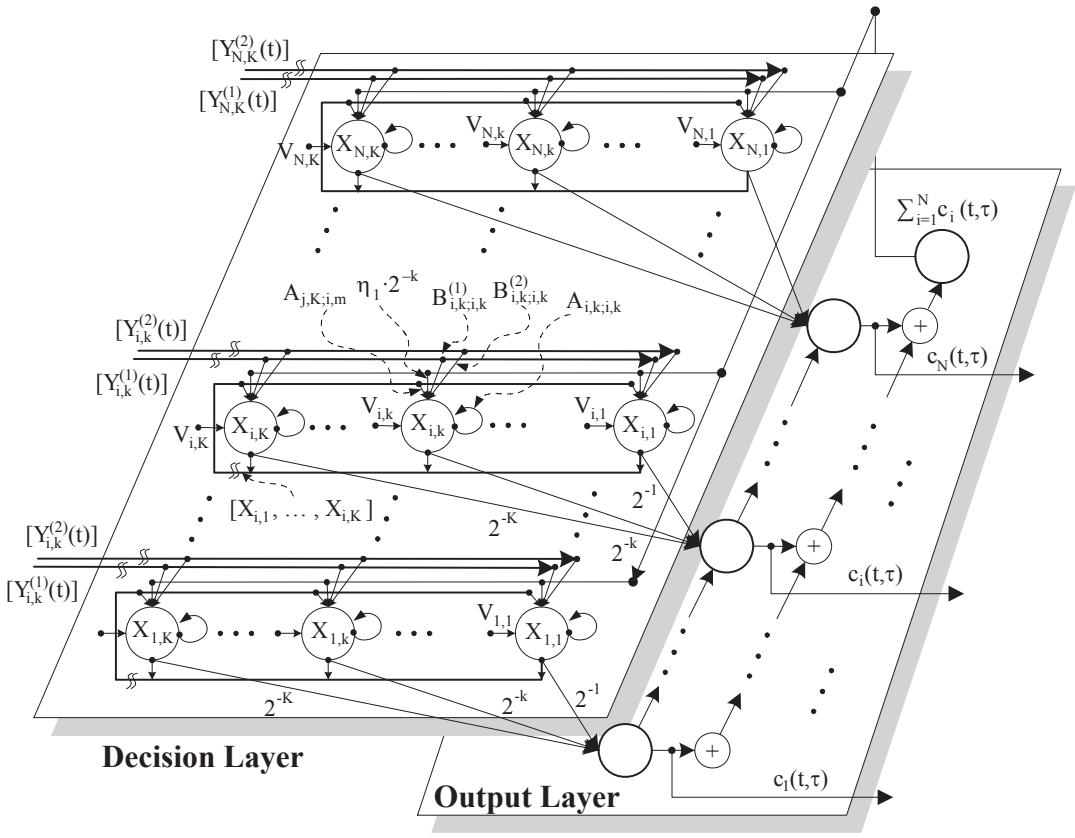


Fig. 2. The two-layer structure of CNN processor.

$$\mathcal{H}_{\Psi_2}(t,\tau)=\eta_2\left[\sum_{i=1}^{N}\left(\left(c_i(t,\tau)-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)^{+}\right)^2\right] \tag{23}$$

The necessity of this auxiliary term is shown in Lemma 2 in [23]. When the coefficient η_3 is properly selected as shown in Lemma 6 in [23], the energy function due to this auxiliary term will approach zero only when every state variable output approaches either one or zero, which is one of the CNN vertex.

By (19) and (24), the dynamics of each neuron in the proposed CNN processor for the CNNU-based scheduler can be expressed by (25), where ν_k is modified to 2^k to retain the stability and desired output pattern of the designed CNN processor. From (17) and (25), we can determine the inter-connection structure of the CNN processor. The complexity of the CNN has an order of $O(N^2)$.

In order to reduce the complexity of CNN processor, we propose a two-layer structure for the CNN processor, which the complexity has an order of $O(NK)$. Note that the K is bounded by the precision and $K \ll N$. To re-arrange (25) and replace $\sum_{k=1}^{K} X_{i,k}(t)$ by $c_i(t,\tau)$ and in contrast to (17), the first decision layer, $\left[z_{i,k}^1\right]$, is with state variable output $X_{i,k}(\tau)$, which is determined by regarding the term of $c_i(t,\tau)$ as a constant. The second output layer, $\left[z_{i,k}^2\right]$, is with state variable output $c_i(t,\tau)$, which is determined by regarding the term of $\sum_{k=1}^{K} X_{i,k}(t)$ as a constant. The decision layer consists of $N\times K$ neurons; the output layer is with an $(N+1)\times 1$ array, where the output of the first neuron is the summation of all the others. The inter-connections between the neurons of decision layer and those of output layer are defined by

- *For the first decision layer to the second output layer, the connection weight between $X_{i,k}(\tau)$ and $c_j(t,\tau)$ is 2^{-k}, $\forall k$ if $j=i$; zero if $j\neq i$.*
- *For the second output layer feedback to the first decision layer, only the first neuron output is connected to the $X_{i,k}(\tau)$ of the decision layer with the inter-connection weight $\eta_1\cdot 2^{-k}$ for $\forall i$.*

Fig. 2 shows the two-layer structure of the CNN processor. The recurrent inter-connection weights and the external control weights for the first decision layer can be determined by (26), where $B_{i,k;i;k}^{(1)}$ and $B_{i,k;i;k}^{(2)}$ are the external control weights for the first and the second external inputs, $Y_{i,k}^{(1)}=\left(1-e^{-\sigma\mathcal{U}_i(t)}\right)\cdot 2^{-k}$ and $Y_{i,k}^{(2)}=\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\cdot 2^{-k}$, respectively, and $\delta_{x,y}=1$ if $x=y$; $\delta_{x,y}=0$ otherwise. For the second output layer, there are no external inputs, and only the recurrent inter-connection weights exist. The inter-connection weight between $c_i(t,\tau)$ and $c_j(t,\tau)$ is given by $\delta_{0,j}$ with $i=0$.

The range of coefficients η_0, η_1 (η_1^+, η_1^-), η_2, and η_3 must

$$\begin{aligned}\mathcal{E}(t,\tau) &= -\eta_0\left[\sum_{i=1}^{N}(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})\cdot(1-e^{-\sigma\cdot\mathcal{U}_i(t)})\right]\\ &+\eta_1\left[\left(\sum_{i=1}^{N}\frac{1}{2}(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})-1\right)\cdot\left(\sum_{i=1}^{N}(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})\right)\right]\\ &+\eta_2\left[\sum_{i=1}^{N}\left(\left(\frac{1}{2}(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)^{+}\cdot\right.\right.\\ &\left.\left.(\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k})\right)\right]+\eta_3\left[\sum_{i=1}^{N}\left(\sum_{k=1}^{K}X_{i,k}(\tau)\left(1-X_{i,k}(\tau)\right)\cdot 2^{-k}\right)\right]\end{aligned}\tag{24}$$

$$\begin{aligned}\frac{dX_{i,k}^{(s)}(\tau)}{d\tau} &= -\frac{X_{i,k}^{(s)}(\tau)}{\nu_k}+\eta_0\left(1-e^{-\sigma\mathcal{U}_i(t)}\right)\cdot 2^{-k}-\eta_1\left(\sum_{i=1}^{N}\sum_{k=1}^{K}X_{i,k}(\tau)\cdot 2^{-k}-1\right)\cdot 2^{-k}\\ &-\eta_2\left(\sum_{m=1}^{K}X_{i,m}(\tau)\cdot 2^{-m}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)^{+}\cdot 2^{-k}\\ &-\eta_3\left(1-2X_{i,k}(\tau)\right)\cdot 2^{-k}\\ &= -\frac{X_{i,k}^{(s)}(\tau)}{\nu_k}+\left[2\eta_3\cdot 2^{-k}-\eta_1\cdot 2^{-2k}-\eta_2\cdot u\left(\sum_{k=1}^{K}X_{i,k}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\right.\right.\right.\\ &\left.\left.\left.\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)\cdot 2^{-2k}\right]\cdot X_{i,k}(\tau)+\eta_0\left(1-e^{-\sigma\mathcal{U}_i(t)}\right)\cdot 2^{-k}\\ &+\eta_2\cdot u\left(\sum_{m=1}^{K}X_{i,m}(\tau)\cdot 2^{-m}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)\cdot\\ &\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\cdot 2^{-k}-\left(\sum_{j=1,j\neq i}^{N}\sum_{m=1}^{K}\eta_1\cdot 2^{-(m+k)}\cdot X_{j,m}(\tau)\right)\\ &-\left(\sum_{m=1,m\neq k}^{K}\left[\eta_1-\eta_2\cdot u\left(\sum_{m=1}^{K}X_{i,m}(\tau)\cdot 2^{-m}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\right.\right.\right.\right.\\ &\left.\left.\left.\left.\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)\right]\cdot 2^{-(m+k)}\cdot X_{i,m}(\tau)\right)+\eta_1\cdot 2^{-k}-\eta_3\cdot 2^{-k}\end{aligned}\tag{25}$$

$$\begin{cases}A_{i,k;i;k}=-\eta_1\cdot 2^{-2k}-\eta_2\cdot u\left(\sum_{k=1}^{K}X_{i,k}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)\cdot 2^{-2k}+2\eta_3,\\ B_{i,k;i;k}^{(1)}=\eta_0,\\ B_{i,k;i;k}^{(2)}=\eta_2\cdot u\left(\sum_{k=1}^{K}X_{i,k}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right),\\ A_{i,k;j;m}=-\eta_2\cdot u\left(\sum_{k=1}^{K}X_{i,k}-\min\left\{\frac{W\cdot\log_2 M_{\kappa_i}}{SF_i\cdot\mathcal{R}_i(t)},\frac{Q_i(t)/T_f}{\mathcal{R}_i(t)}\right\}\right)\cdot\delta_{i,j}\cdot 2^{-(k+m)},\\ V_{i,k}=\eta_1\cdot 2^{-k}-\eta_3.\end{cases}\tag{26}$$

be properly selected to ensure the stability and the desired response. For a tolerant error level ε, which is the maximum difference between stable output $\lim_{\tau\to\infty}\vec{c}(t,\tau)$ and the optimum $\vec{c}^{\,*}(t)$, the ranges of these coefficients are obtained as follows [23]: $0<\eta_0<\eta_3$, $\eta_1^+>2^K$, $\eta_1^-\geq\frac{\eta_0\cdot 2^{-3}}{\varepsilon}$, $\eta_2\geq 2^K$, $\eta_3>\frac{1}{2}+\frac{\eta_1^-}{2}$. We have proved in [23] that with a matrix of given utility function and a matrix of radio resource assignment ratio upper limits, the proposed architecture of CNN processor will converge to the neighborhood of the optimal pattern $\vec{c}^{\,*}(t)$ within the difference ε, with the ranges of these coefficients. If $\varepsilon\leq 2^{-K}$, the CNN can converge to $\vec{c}^{\,*}(t)$. Note that the complexity of inter-connections in the two-layer CNN processor is proportional to $[3N\times K+N]$, which is linear with respect to the number.

VI. Simulation Results and Discussion

In simulations, a scenario with five types of services in three classes is assumed. Type-1 service is a real-time class of traffic with a peak rate of 15 kbps, an activity factor of 0.57, $P_D^*=0.05$, $D^*=40$ ms, and $BER^*=10^{-3}$; type-2 (type-3) service is a non-real-time interactive class of traffic with Pareto process [24] of which the mean rate is 8 kbps (12 kbps), $R_{m,i}^*$=7.2 kbps ($R_{m,i}^*$=11 kbps), and $BER^*=10^{-5}$ ($BER^*=10^{-5}$); and type-4 (type-5) service is a non-real-time best effort class of traffic in batch Poisson distribution with a mean rate of 6 kbps (15 kbps) and a mean batch size of 1.2 kbits (1.2 kbits), and $BER^*=10^{-5}$. The proportion in the number of connections from type-1 to type-5 is kept at 1:1:1:1:1. Also, three modulation schemes, QPSK, 16QAM, and 64QAM, are available for transmission as long as the BER requirement can be fulfilled and the remaining queue is enough.

We compare the proposed CNNU-based radio resource scheduler with the EXP scheduling scheme [9] and the HOLPRO scheme [5]. The EXP scheduling scheme can be directly extended for RT and NRT connections. However, the HOLPRO scheme is only suitable for RT connections. As for NRT connections, the pseudoprobability is obtained according to the fraction between the queue length of the connection and the summed queue length among all NRT connections. Also, the RT connections have absolutely higher priority than the NRT connections.

The performance measures are such as the average system throughput, the average packet dropping ratio of RT connections, $\overline{P}_D$, the average transmission rate of NRT interactive connections, $\overline{R}_m$, the ratio of RT connections in which their packet dropping ratio requirement is not guaranteed, ϕ_{P_D}, the ratio of NRT interactive connections in which their minimum transmission rate requirement is not guaranteed, ϕ_{R_m}, and the fairness variance index of NRT connections, F_v.

The F_v is defined for measuring the variance of fairness to share the radio resource among all NRT connections. It is given by

$$F_v=\frac{1}{N_{NRT}}\sum_{i}^{N_{NRT}}\left|\frac{\mathbf{E}\left[r_i(t)\right]}{\sum_j^{N_{NRT}}\mathbf{E}\left[r_j(t)\right]}-\frac{w_i}{\sum_j^{N_{NRT}}w_j}\right|^2, \quad (27)$$

where N_{NRT} is the number of NRT connections. The fairness variance index shows the variance of the normalized radio resource allocated and the normalized proportion of resource desired to share.

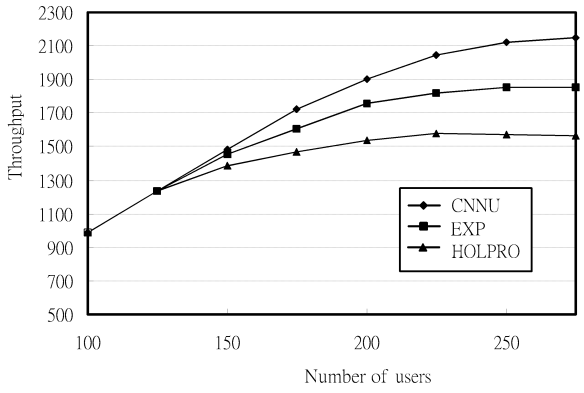


Fig. 3. The average system throughput.

At first, we measure the stability and the convergence of the CNN processor during the simulation. The measure of stability of the CNN processor is assumed to be the percentage of which the output of each neuron $z_{i,k}^1$ takes the value from $\{0,1\}$ and no chaotic state occurs; the measure of convergence of the CNN processor is assumed to be the percentage of which the output of neurons $z_{i,k}^2$ is within the system constraints Ψ_1 and Ψ_2. It is found that the CNN processor can always be stable without chaotic trajectory. Also, the output $\vec{c}_i(t)$ can mostly converge to a reasonable set of stable output of the CNN processor in a percentage of about 99.9%. On the other hand, the convergence speed of CNN processor would be in few miniseconds range [15], [18] since the CNN has an architecture that all neurons are in the same structure, which makes the CNN be suitable for VLSI implementation.

Fig. 3 shows the average system throughput. It can be found that the CNNU-based scheduler can always have a higher system capacity than the EXP and HOLPRO scheduling schemes in all traffic load conditions. It achieves the improvement of system throughput over the EXP (HOLPRO) scheduling scheme by more than 9% (19%) as the number of connections is greater than 200, and by higher than 15% (25%) as the number of connections increases up to 250. This is because the CNNU-based scheduler is with the radio resource function that makes CNN processor adapt to the link variation and allocate radio resource in an efficient way. Both RT and NRT connections with relatively worse link conditions have lower probability to be scheduled as long as their QoS requirements can be achieved in a long term sense. The CNNU-based scheduler is with the fairness compensation function that makes the NRT connections share the radio resource according to the location dependent fairness, achieving a higher radio resource efficiency. Also, the CNN processor can determine an optimal radio resource assignment vector in the sense that the allocation of downlink power by CNNU-based scheduler is the most efficient one, with given utilities and upper limits of the radio resource assignment. Additionally, beyond the point of 250 (200) connections, the throughput of the EXP

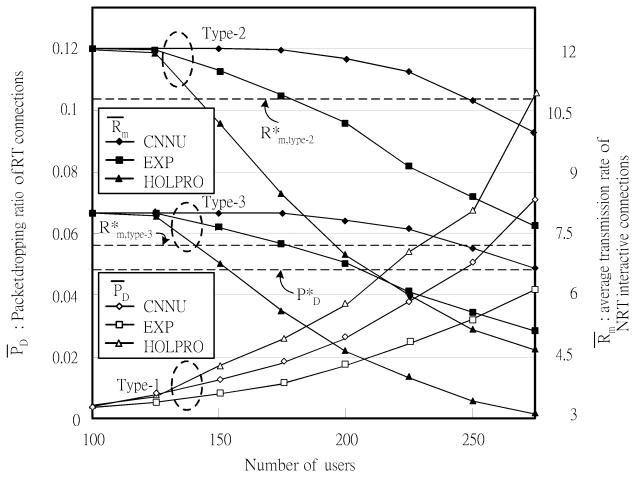


Fig. 4. QoS performance measures of packet dropping ratio $\overline{P}_D$ and average transmission rate $\overline{R}_m$.

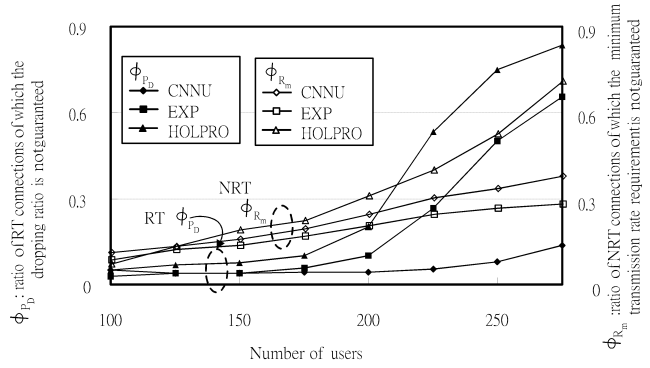


Fig. 5. The ratio ϕ_{P_D} for RT connections and the ratio ϕ_{R_m} for NRT interactive connections.

(HOLPRO) scheme is almost saturated, while the throughput of the CNNU-based scheduler continues to grow up but with a slightly lowering slope. It is because the CNNU-based scheduler can achieve the utilization of multiuser diversity gain better than the EXP scheduling scheme. The HOLPRO scheme always has the lowest throughput among these three schemes due to the lack of link condition information.

Fig. 4 depicts the performance measures of the average packet dropping ratio of type-1 RT connections $\overline{P}_D$ and the average transmission rate of type-2 and type-3 NRT interaction connections $\overline{R}_m$. It can be found that the $\overline{P}_D$ of the CNNU-based scheduler is larger than that of the EXP scheme but lower than that of HOLPRO scheme, and it violates the P_D^* requirement as the number of users is about 250; on the other hand, all the $\overline{R}_m$ of type-2 and type-3 connections of the CNNU-based scheduler are greater than those of the EXP and HOLPRO schemes, and the EXP (HOLPRO) scheme violates the R_m^* requirements as the number of users is about 170 (150). These indicate that the QoS guaranteed region by the CNNU-based scheduler can be up to 250 connections, while those by the EXP and HOLPRO schemes are only to 175 and 150 connections, respectively. The QoS guaranteed region achieved by the CNNU-based scheduler is larger than those achieved by the EXP and HOLPRO schemes. This is because the CNNU-based scheduler is designed with the QoS deviation function together with the priority bias that can balance the extent of deviation of every performance measure from the QoS requirement. The worse the QoS performance measure is, the more the radio resource will be scheduled. Besides, since the CNNU-based scheduler has higher throughput performance, the more number of connections can be served in the QoS guaranteed region. Moreover, if we define the maximum throughput achievable in QoS guaranteed region to be the average system throughput, the CNNU-based scheduler can have the average system throughput equal to 2.083 Mbps at 235 connections, while the EXP and the HOLPRO schemes can have the average system throughput equal to 1.6 Mbps at 175 connections and 1.388 Mbps at 150 connections, respectively. The former attains the average system throughput greater than the latter by an amount of 23% and 33%, respectively.

Fig. 5 shows the ratio of RT connections of which the packet dropping ratio requirement is not guaranteed, ϕ_{P_D}, and the ratio of NRT interactive connections of which the minimum transmission rate requirement is not guaranteed, ϕ_{R_m}. It can be seen that the total ratio of connections with QoS requirements un-guaranteed for the CNNU-based scheduler is about 0.0435, while those for the EXP and HOLPRO schemes are greater than 0.18 and 0.365, respectively, in heavily loaded situations as the number of connections is greater than 225. The total ratio of connections with QoS requirements un-guaranteed is defined as $\frac{1}{3}\phi_{P_D} + \frac{2}{3}\phi_{R_M}$, which is weighted by the number of RT and NRT interactive connections.

The CNNU-based scheduler can achieve the total ratio of QoS requirements un-guaranteed connections in all traffic types lower than the EXP and HOLPRO scheme. The reason is that the CNNU-based scheduler can balance the allocation of radio resources among traffic types and avoid allocating excess radio resource to connections with bad link condition, while the EXP and HOLPRO schemes prefer RT connections and overprotects them so that the QoS guaranteed region is reduced. Note that the ratios of ϕ_{P_D} and ϕ_{R_m} are greater than zero at any traffic load condition due to the existence of connections with very bad link quality. These results imply that the CNNU-based scheduler will not guarantee all the QoS requirements all the time, and a properly designed call admission control is required to reject the connections with very bad link quality in terms of the current traffic load conditions.

Fig. 6 shows the fairness variance index of NRT connections. It can be found that the fairness variance indexes of the CNNU-based scheduler retains within 1 in almost all simulation cases, and grows up slowly as the number of connections increases; the fairness variance indexes of the EXP and HOLPRO schemes, on the other hand, increase with slightly higher slope than that of the CNNU-based scheduler. This is because the fairness compensation function of the CNNU-based scheduler considers the location dependent information and aims to share the radio resource fairly as long as the minimum rate is guaranteed, while the design

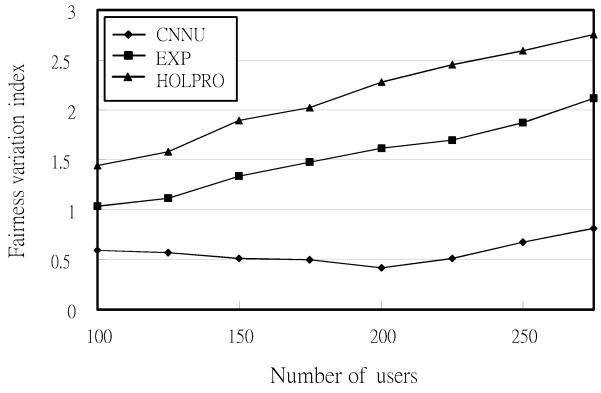


Fig. 6. The fairness variation index for NRT connections.

of the EXP and the HOLPRO schemes ignore the location dependent information to allocate rate fairly to all connections. The fairness compensation function, considering the location dependent information, also facilitates the higher capacity for the CNNU-based scheduler shown in Fig. 3.

VII. Concluding Remarks

This paper presents a cellular neural network and utility (CNNU)-based radio resource scheduler, which jointly considers its radio resource efficiency, diverse QoS requirements, and fairness, to schedule the radio resource for downlink connections in multimedia CDMA communication systems. The utility function is defined to be the radio resource function properly weighted by the QoS requirement deviation function and the fairness compensation function. Also, the cellular neural network (CNN) is successfully manipulated to be a two-layer structure to solve the constrained optimization problem defined for the radio resource scheduling in a real-time fashion. Simulation results show that the CNNU-based scheduler can efficiently allocate the radio resource to achieve higher throughput than the conventional EXP and HOLPRO scheduling schemes. It can also effectively support differentiated QoS requirements for connections with various traffic characteristics. Moreover, the CNNU-based scheduler can enlarge the QoS guaranteed region under the complicated QoS requirements. The CNNU-based radio resource scheduler is effective for multimedia CDMA communciation systems with diverse QoS requirements when both dedicated and shared channels are adopted.

Acknowledgment

The authors would like to give many thanks to those anonymous reviewers for their kind helps to improve the presentation of the paper.

References

[1] 3rd Generation Partnership Project (June 2001), QoS Concept and Architecture, *3GPP TS 23.107*. Online: http://www.3gpp.org.

[2] Y. M. Lu and R. W. Brodersen, "Integrating power control, error correction coding, and scheduling for a CDMA downlink system," *IEEE J. Select. Areas Commun.*, vol. 17, no. 5, pp. 978-989, May 1999.

[3] V. K. N. Lau and Y. K. Kwok, "On generalized optimal scheduling of high data-rate bursts in CDMA systems," *IEEE Trans. Commun.*, vol. 51, no. 2, pp. 261-266, Feb. 2003.

[4] V. Bharghavan, S. W. Lu, and T. Nandagopal, "Fair queuing in wireless networks: issues and approaches," *IEEE Personal Commun.*, vol. 6, no. 1, pp. 44-53, Feb. 1999.

[5] A. C. Varsou and H. V. Poor, "HOLPRO: a new rate scheduling algorithm for the downlink of CDMA networks," in *Proc. IEEE VTC 2000*, pp. 948-954.

[6] ———, "Waiting time analysis for the generalized PEDF and HOLPRO algorithms in a system with heterogeneous traffic," in *Proc. IEEE VTC 2001*, pp. 2152-2156.

[7] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: large deviations and optimality," *Annal Appl. Prob.*, vol. 11, no. 1, pp. 1-48, 2001.

[8] A. C. Kam, T. Minn, and K. Y. Siu, "Supporting rate guarantee and fair access for bursty data traffic in W-CDMA," *IEEE J. Select. Areas Commun.*, vol. 19, no. 11, pp. 2121-2130, 2001.

[9] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," *Bell Lab Reports*, 2000.

[10] C. X. Li, X. D. Wang, and D. Reynolds, "Rate control and fairness scheduling for downlink utility-based power control systems," in *Proc. IEEE GLOBECOM 2004*, pp. 465-469.

[11] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: MacMillan, 1994.

[12] S. Abe, "Global convergence and suppression of spurious states of Hopfield neural networks," *IEEE Trans. Circuits Syst.*, vol. 40, no. 4, pp. 246-257, 1993.

[13] L. O. Chua and L. Yang, "Cellular neural networks: theory," *IEEE Trans. Circuits Syst.*, vol. 35, no. 10, pp. 1257-1272, Oct. 1988.

[14] "Cellular neural networks: applications," *IEEE Trans. Circuits Syst.*, vol. 35, no. 10, pp. 1273-1290, Oct. 1988.

[15] J. A. Nossek, "Design and learning with cellular neural networks," in *Proc. IEEE CNNA-1994,* Rome, pp. 137-146, Dec. 1994.

[16] R. Fantacci, M. Forti, M. Marini, and L. Pancani, "Cellular neural network approach to a class of communication problems," *IEEE Trans. Circuits Syst.*, vol. 46, no. 12, pp. 1457-1467, Dec. 1999.

[17] S. Shen and C. J. Chang, "A cellular neural network and utility-based scheduler for multimedia CDMA cellular networks," in *Proc. IEEE IWCMC 2005,* pp. 768-773.

[18] W. E. Lillo, M. H. Loh, S. Hui, and S. H. Zak, "On sovling constrained optimization problems with neural networks: a penalty method approach," *IEEE Trans. Neural Net.,* vol. 4, no. 6, pp. 931-940, Nov. 1993.

[19] C. C. Chiu, C. Y. Maa, and M. A. Shanblatt, "Energy function analysis of dynamic programming neural networks," *IEEE Trans. Neural Net.*, vol. 2, no. 4, pp. 418-426, July 1991.

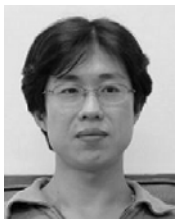
[20] A. J. Goldsmith and S. G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218-1230, Oct. 1997.

[21] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150-154, Feb. 2001.

[22] C. S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.,* vol. 13, no. 6, pp. 1091-1100, Aug. 1995.

[23] S. Shen, "The connection admission control and radio resource allocations for multimedia WCDMA Networks," doctoral dissertation, National Chiao Tung University, 2005.

[24] Universal Mobile Telecommunications System (UMTS), Selection procedures for the choice of radio transmission technologies of the UMTS, UMTS 30.03, version 3.2.0, 1998.



Scott Shen (S'99) received the B.S. degree in electronics engineering from National Tsing Hua University, Hsing-Chu, Taiwain, in 1996, the M.A. and Ph.D degree in communication engineering from National Chiao Tung University, Hsing-Chu, Taiwain, in 1998 and 2005. His interest area includes wireless networks, mobile communications, high-speed networks, communications protocol design, and network performance evaluation. He is currently focusing the research areas on resource management for cellular networks.

Chung-Ju Chang was born in Taiwan, ROC, in August 1950. He received the B.E. and M.E. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1972 and 1976, respectively, and the Ph. D degree in electrical engineering from National Taiwan University, Taiwan, in 1985. From 1976 to 1988, he was with Telecommunication Laboratories, Directorate General of Telecommunications, Ministry of Communications, Taiwan, as a Design Engineer, Supervisor, Project Manager, and then Division Director. He also acted as a Science and Technical Advisor for the Minister of the Ministry of Communications from 1987 to 1989. In 1988, he joined the Faculty of the Department of Electrical Engineering, College of Electrical and Computer Engineering, National Chiao Tung University, as an Associate Professor. He has been a Professor since 1993. He was Director of the Institute of Communication Engineering from August 1993 to July 1995, Chairman of Department of Communication Engineering from August 1999 to July 2001, and Dean of the Research and Development Office from August 2002 to July 2004. Also, he was an Advisor for the Ministry of Education to promote the education of communication science and technologies for colleges and universities in Taiwan during 1995–1999. He is acting as a Committee Member of the Telecommunication Deliberate Body, Taiwan. Moreover, he serves as Editor for IEEE COMMUNICATIONS MAGAZINE and Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. His research interests include performance evaluation, radio resources management for wireless communication networks, and traffic control for broadband networks. Dr. Chang is a member of the Chinese Institute of Engineers (CIE).

Li-Chun Wang (S'92-M'96-SM'06) received the B.S. degree in electrical engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1986, the M.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1988, and the M.Sc. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1995 and 1996, respectively. From 1990 to 1992, he was with Chunghwa Telecom, Taoyuan, Taiwan. In 1995, he was with Northern Telecom, Richardson, TX. From 1996 to 2000, he was a Senior Technical Staff Member with the Wireless Communications Research Department, AT&T Laboratories. In August 2000, he became an Associate Professor with the Department of Electrical Engineering, National Chiao Tung University, where he has been a Full Professor since August 2005. His research interests include cellular architectures, radio network resource management, and cross-layer optimization for cooperative and cognitive wireless networks. He is the holder of three U.S. patents and has three patents pending. Dr. Wang is a co-recipient of the Jack Neubauer Best Paper Award from the IEEE Vehicular Technology Society in 1997.