A Reinforcement Learning-based Cognitive MAC Protocol

I.Kakalou[1],G.I.Papadimitriou[1], *Senior Member, IEEE,* P. Nicopolitidis[1], *Senior Member, IEEE,* P.G.Sarigiannidis[2], *Member IEEE,*andM.S.Obaidat[3], *Fellow, IEEE*

[1]*Department of Informatics, Aristotle University of Thessaloniki, Greece*
[2]*University of Western Macedonia, Kozani, Greece*
[3]*Department of Computer Science, Monmouth University, W. Long Branch NJ 07764, U.S.A*

***Abstract*—A Multi-Channel Cognitive MAC Protocol for ad-hoc cognitive networks that uses a distributed learning reinforcement scheme is proposed in this paper. The proposed protocol learns the Primary User (PU) traffic characteristics and then selects the best channel to transmit. The scheme, whichaddresses overlay cognitive networks,avoids collision with the PU nodes and manages toexceed the performance of the less adaptive statistical channel selection schemesin normal and especially bursty traffic environments. The simulation analysis results have shown that the performance of our proposed scheme outperforms that of the CREAM-MAC scheme.**

***IndexTerms*— Next Generation Networks, Cognitive, ad-hoc, MAC, Reinforcement Learning.**

I. Introduction

Cognitive Radio was introduced to answer the spectrum scarcity problem.The MAC protocol ofa Cognitive Radio network is supposed to enable the so called Secondary Users (SU) network nodes to dynamically access unused or under-utilized licensed spectrum.The licensed users also called Primary Users(PUs) can share the spectrum with the SUs whom can access the spectrum inunderlay and overlay mode. In underlay spectrum accessthe secondary user limits its transmission power below the interference temperature limit so as not to disturb licensed users' transmission. The interference temperature is a metricof the quality of the received signal and includes noise and interference of other sources signals. In the overlay spectrum access,the SUs can only access the spectrum opportunistically at the absence of the PUs. The SUs have to sense the spectrum for vacancies namely "spectrum holes" and decide whether to access the spectrum or not usually based on the PU collision probability.

In the Cognitive Radio Network(CRN), the SUs interfere with each other and this factor, i.e. aggregated interference, has to be considered and estimated. When the numberof the SUs and their traffic characteristics are not known, thenarises the problem that the common control channel answers. In ad-hoc networks that lack of a central entity, synchronization is difficult and thus has to be addressed by the protocol. Cluster-based architectures were introduced in CRNs toreduce congestion in channel access. These architectures do not share a common control channel.

This paper introduces a MAC Protocol for CR wireless ad-hoc networks for opportunistically spectrum access with a distributed learning reinforcement scheme for channel selection based on SUs observations of PU trafficand with minor computational demands and avoids collisions with the PUs and keeps SUs synchronized.

II. Related Work

A scheduling algorithm for multi-hop CR network was proposed in [1] which determines the time slot and channel for transmission by SUs presuming that each SU has a different set of available channels. A heuristic-based distributed algorithm was proposed. It consists of two phases for allocating slots in all channels. In the first one each node chooses the time slot and the channel for each link and in the second this information is propagated in the network and each node makes the necessary adjustments.

The statistical channel allocation MAC (SCA-MAC) [2] employs a control channel and aggregated data channels whilst the selection of the range of channels for transmission is an optimization problem. The protocol outperforms the random scheme. In [4] a cluster-based cognitive radio network was introduced where local traffic can be exchanged through cluster-heads and inter-cluster communication is achieved via the gateway node.There is no common control channel and the role of coordinator for nodes communication is assigned to the cluster head at the communication channel of the cluster-head. Transmission is based on superframes that include a beacon period for synchronization of the cluster heads resource allocation information.

The CREAM-MAC [5] is a protocol based on IEEE 802.11 DCF that limits each channel access time to a value of Tdmax so as to limit the interference to PUs.Protocolsthat reserve channelsfor data transmissions similar with the well-known IEEE802.11 DCF standard are DSA-MAC [6], DCRMAC[7], HC-MAC [8], DDMAC [9], SMA[10].

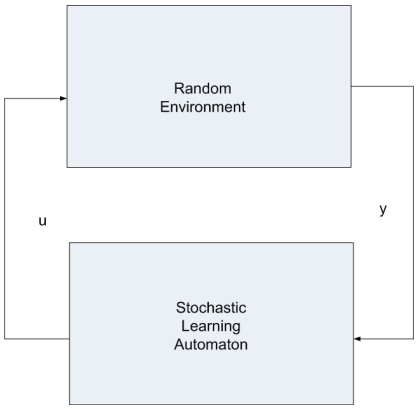

Figure 1.A stochastic learning automaton.

III. Primary and SUs Model and Protocol Description

A. Stochastic Automata and Reinforcement Learning

A Learning Automaton is a control mechanism that follows a predetermined sequence of operations or adapts to changes in the environment. The adaptation is the result of the learning process i.e. the permanent change in the learning automaton behavior toward a final goal as a result of the past experience. The term stochastic refers to the adaptive nature of the learning automaton. A stochastic automaton has no information of the optimal action, but selects an action randomly then the environment is observed and the action probability is updated based on the response from the environment. This procedure is repeated and the algorithm that guarantees the desired learning process is called a reinforcement scheme.

Let $w1,w2,..w_m$ be the mutually exclusive responses of the environment, Pi the probability of the occurrence of the i^{th} response, then the reinforcement scheme of the Learning automaton can be described as:

$$Pi(n+1) = \lambda Pi(n) + (1-\lambda)a_i(n) \quad n=0,1,2 \quad (1)$$

Where $Pi(n)$ is the probability of the occurrence of the i^{th} response on time n for the x_i input.

$$0 < \lambda < 1, 0 < \alpha_i(n) \leq 1$$

$$\sum_{i=1}^{m} a_i(n) = 1 \quad (2)$$

Equation (1) describes a linear reinforcement scheme and it can be easily shown that if $\alpha_i(n)=\alpha_i$ then:

$$P_i(n+1) = \lambda^n P_n(0) + (1-\lambda^n)a_i \quad (3)$$

$$\text{And} \lim_{n\to\infty} P_i(n) = a_i \quad (4)$$

The reinforcement learning can be formulated by learning automata. A stochastic learning automaton operating in a random environment is shown in Figure 1.At each step, the random environment provides a feedback of satisfactory or unsatisfactory performance to the stochastic learning automaton known asa penalty y=0with probability 1-$\pi_i$and y=1 with probability $\pi_i$respectively.

A stochastic learning automaton is a quintuple {Y,Q,U,F,G} where Y is the environment response set and it consists of only two elements i.e. If it takes 0 then is said P-Model, if it takes finitenumber values in the range [0,1] then is said Q-Model and if it takes arbitrary numbers in the range [0,1] then is said S-Model. Q is the finite set of states Q={$q1,q2,\ldots qs$}, U is the finite set of the stochastic learning automaton outputs U={$u_1,u_2,\ldots,u_m$}, F is the state transition function q(n+1)=F[y(n), q(n)] and G is the u(n)=G[q(n)]. The function G can be either stochastic or deterministic whilst function F is stochastic and this stochastic nature of learning automata makes them suitable for learning systems.

If $Imin=\min\{\pi_1,\pi_2,\ldots,\pi_m\}$ then the optimal output of the stochastic learning automaton is $u_\beta$and the reinforcement scheme is said optimal if

$$\lim_{n\to\infty} \mathbb{E}\{P_\beta(n)\} = 1 \quad (5)$$

The reinforcement scheme is said $\in -optimal$ if

$$\lim_{\lambda\to 0} \lim_{n\to\infty} \mathbb{E}\{P_\beta(n)\} = 1 \quad (6)$$

And $\in -optimality$ ensures that the operation of the stochastic learning automaton is very close to optimality, where $\lambda$is the learning rate.

B. Transmission opportunity detection

We assume a wireless network of CR-enabled nodes equipped with a radio dedicated to the common control channel, a radio for data transmission and sensors for channel sensing. The spectrum licensed to PUs consists of M channels. The nodes' transceivers can operate to any of the M channels. According to the sensing outcome of each sensing period, a state is assigned to each channel.The ON state represents that the PU transmit state and the OFF state means the SUs can opportunistically access the channel. A vector *state[i]* maps the state of each channel.

Presuming that the traffic of PUs comes in bursts, the SUs have to detect those opportunities in spectrum that are "good" i.e. they are mostly likely for successful transmission as they will not be interrupted. Due to the burstyness of its traffic, a PU which transmits during a sensing slot, most likely will transmit during the next sensing slot in order to complete its transmission.

To learn the burstyness of PUs' traffic, each SU employs a learning automata mechanism for each channel *i.* The mechanism in [12] estimates the probability *P[i,t] that* channel i will be occupied via PU transmission at time t (where t is measured in sensing slots).

After each sensing period, this probability is updated according to the following scheme:

- If channel $i$is occupied by a PU, then P *[i,t]*is increased:$P[i,t+1] = P[i,t] + L*(1-P[i,t])$ (7)
- If channel i is occupied by PU, then *P[i,t]*is decreased:$P[i,t+1] = P[i,t] - L*(P[i,t] - \alpha)$(8)

It holds that $L,\alpha \in (0,1)$ and $P[i,t] \in (\alpha,1) \ \forall$ *i,t.* L is a parameter that governs the speed of the automaton(7)(8) convergence.. The lower the value of L, the more accurate the estimation made by the automaton; a fact that comes at the expense of convergence speed. Parameter "α" prevents the probabilities *P[i,t]* from taking values in the neighborhood of zero and thus increases the adaptivity of the automaton.

After estimating the channel usage made by the PUs, the protocol has to identify transmission opportunities for the secondary ones. Thus, the probability that a SU at channel I can transmit isthen equal to $S[i,t] = 1-P[i,t]$.We update (7) and (8) with S[i, t+1]=1-P[i, t+1] and S[i]=1-P[i]. We have the following linear reinforcement scheme for the stochastic learning automaton of P-Model:

- If channel *i*is not occupied by a PU, then S *[i,t]*is increased:*S[i,t+1] =S[i,t] (1-L)+L*(1-a) (9)*

- If channel *i* is occupied by PU, then *S[i,t]*is decreased:*S[i,t+1] =S[i,t] (1-L) (10)*

The probabilities*S[i,t+1]* are updated at the end of each sensing slot.As (1-α) is constant according to (1),(3),(4) and due to the use of very smallvalue(i.e. 10^{-4}) for parameter α mentioned above(7)(8)(9), during the burst of the licensed user at channel i, the S*[i,t]* will approach zero whilst during a spectrum hole on the channel, it will approach*(1-α).*For very small values of α and according to (5):

$$\lim_{t\to\infty} \mathbb{E}\{S[i,t]\} = \mathbb{E}\{\lim_{t\to\infty} S[i,t]\} = 1 (11)$$

Thus the channel selection learning reinforcement scheme is optimal for very small values of *α*.An opportunity discover function provides the SU with the best selection metric i.e. the first channel in S*[i,t]* that is also idle according to *state[i]* . The algorithm decreases the selection probability of channel *i*, if it is occupied by a PU during the last sensing slot and increases the selection probability of channel*i*, if the channel is not occupied by a PU. As PU traffic comes in bursts the selection probability of channel *i* keeps decreasing. On the other hand, as the spectrum holes are contiguous, the selection probability will increase when the PU is not transmitting on channel *i*.

This paper introduces an architecture where the SUs form clusters of nodes which experience the same PUs presence and local traffic can be exchanged within the cluster whilst remote traffic is achieved by the inter-cluster communication.A common control channel answers the problem of aggregated interference uncertainty. The protocol introduced in this paper employs a common control channel for resource reservation per cluster in order to limit the aggregated interference of SUs and to handle the hidden terminal and multi-channel hidden terminal problems. The common control channelis dynamically selected asthe most reliable and with the best selection metric.

The proposed protocol relies on the synchronization achieved within the cluster by the use of the distributed scheme as the sender and receiver experience the presence of the same PUs and they share the same channel selection probabilities. As there are no collisions between SUs and PUs, transmissions are identifiedand SUs are synchronized. For spectrum sensing, energy detection and cyclostationary feature detection can beemployed.

C. Spectrum Sensing Inaccuracy

If spectrum sensing is not accurate thenthe selection metric S[i,t+1] is modified by the term:

$$S'[i,t+1] = S[i,t+1] + L(1-\alpha)(1-L)^{N+1} (12)$$
$$S'[i,t+1] = S[i,t+1] - L(1-a)(1-L)^{N+1} (13)$$

The equations above stand for PU misdetection and false alarm cases. In PU misdetection the selection metric is updated wrongly according to equation (9) for *N* slotsin the sequence –as long as faulty sensing occurs- and then as the system seems to overcome misdetection, the selection update is done according to equation (10). Thus, we have an increase

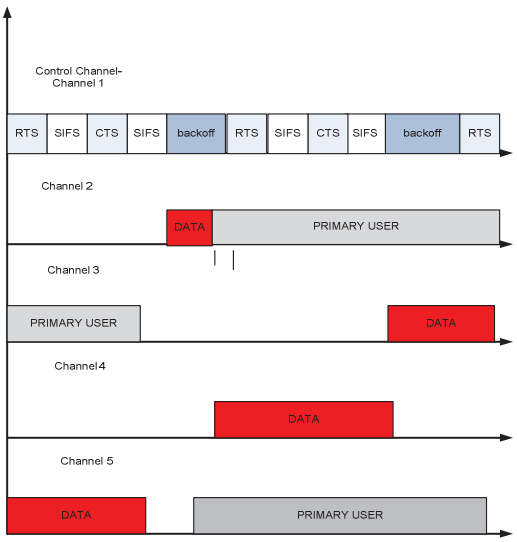

Figure 2.Example of the proposed protocol operation.

inthe term as appearing in equation (12). On the other hand, false alarm causes a selection metric update for N sensing slots in a sequence according to equation (10) and then the system overcomes false alarm by persistent updating the selection metric according to equation (9). Hence, we have again a decrease in the value of the selection metric by the term appearing at equation (13).

The limit of the error term is zero:

$$\lim_{N\to\infty} L(1-\alpha)(1-L)^{N+1} = 0 (14)$$

Thus for higher values of Lthe error term reaches very small values around 10^{-4} in a few slots.The proposed reinforcement learning scheme can operate in inaccurate sensing conditions as the system overcomes inaccurate sensing easily when the value of *L* is high.

D.SUschannel access

The control packets are namely ReadyToSend/ClearToSend/ACKnowledgement. The ACKnowlegdement is returned to the sender at the successful completion of data transfer on the data channel.In case of unsuccessful handshake binary exponential back-off is invoked. The SU that wants to transmit and finds the control channel busy backs off for a period of 2^{CW} **Min_Waiting_Period* where *CW* is the contention window and *Min_Waiting_Period* is the minimal backoff period.

The whole handshake procedure for data channel reservation is shown in Fig. 2 for each user that wins the contention. If data transmission is interrupted by the PU presence then data transmission continues in the first available channel provided by the opportunity discover function. On successful data reception an ACK message is returned to the sender via the data channel.

The RTS/CTS messages between the nodes of the same cluster do not include the available channels on each node as they experience the presence of the same PUs.Thenodes remain synchronized with the learning automata and so the RTS/CTS messages hold other

information useful for cluster operations e.g.for cluster orientation.

Cluster operations are the necessary functions taking places in the control channel and they are responsible for the cluster orientation, neighborhood orientation, the voting process for the selection of the control channel. When the quality of the control channel drops or PU presence occurs and remains so for at least period Tc, a voting process based on both the quality of the signal and the learning reinforcement scheme metric will determine the next control channel of the cluster. Upon collapse of CC the cluster e.g. due to jamming, nodes will mitigate to the next channel with the better metric of all channels.

The cluster synchronization is necessary when a node entering the cluster or experiencing synchronization problems e.g. at the limits of the cluster, thus wants to synchronize itself with the rest of the cluster, they will be supported by the cluster via the control messages. The nodes will demonstrate their cluster orientation information within their RTS/CTS messages.Each cluster operation is related to a certain field update with the appropriate information in the control messages.

Neighboring clusters neither can share the same control channel nor allow data transmission on the control channels of their neighboring clusters. They cannot share the same control channel in order to allow nodes belonging to different clusters to communicate avoiding overlapping. Furthermore, data transmission with no collisions on the non-control channels is guaranteed by the control messages whilst the communication on the control channel can be guaranteed if the nodes of the neighboring clusters do not interfere with the control channel operation i.e. no data transmission can take place at a channel that operates as control channel of a neighboring cluster..

SUs on different clusters communicate asynchronously-i.e. they have to agree upon data transmission channel- at the receiver's Control Channel.

IV. Performance Evaluation

We used simulation analysis in order to evaluate the performance of the proposed protocol. The OMNET package was used. As our main goal is to show the superiority of learning PUs' traffic characteristicsfor identifying transmission opportunities, we compared the performance of the proposed scheme with that of CREAM-MAC [5], which identifies transmission opportunities via gathering statistics on licensed channels usage by PUs. The simulation parameters of CREAM-MAC were identical with those of [5]. The parameters used in our simulations are: transmission rate of 2Mbps per channel, 30 SUs, 30 channels and PUs activity which follows the exponential distribution. The proposed protocol is compared to CREAM-MAC for a maximum interference period Tdmax=10msecin terms of aggregate throughput (Fig. 3,4,5), access delay (Fig.7) and collision with the PU ratio (Fig.8). The data packet length is assumed to be 20,000 bits and *Min_Waiting_Period* is assumed to be 10ms. A one cluster architecture was simulated for the proposed protocol.

In all the experiments the Learning Automata-based Multi-Channel Cognitive MAC outperforms the CREAM-MAC.In the first two experiments (see Fig.2, and Fig. 3) the aggregated throughput is studied for different PU burst lengths when the bursts are generated every 2s and 0.1s. The proposed

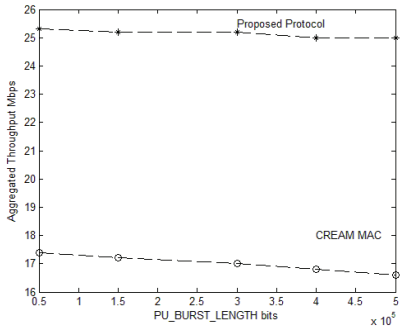

Figure 3.The aggregated throughput versus the PU burst length for burst generation every 2s.

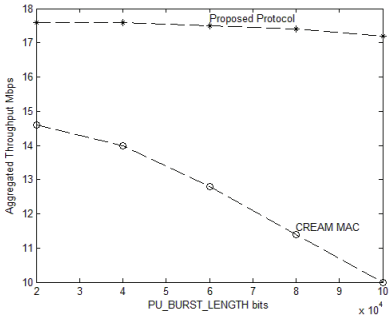

Figure 4. The aggregated throughput versus PU burst length for burst generation every 0.1s.

protocol seems to be less affected by the changes in PU burst length as it utilizes the most of the spectrum holes for every PU burst generation rate that learns, something that does not hold for CREAM-MAC as its performance drops significantly by the PU burst length.

In the third experiment(Fig. 5) the aggregated throughput is studied for different PU burst arrival rateswhen the PU burst length is 100,000 bits.The fourth experiment (Fig. 6) follows the results of the third experiment and studies the SU channel utilization for different PU burst arrival rates and PU burst length equal to 100,000 bits. As the PU burst generation rateincreases, the difference in performance of the two protocols also increases. This is due to the increase of the collisions with the licensed users that CREAM-MAC cannot predict; whereas these areavoided by our proposed protocol.

The access delay was studied for different PU burst arrival rates (Fig. 7) with PU burst length of 100,000 bits. As the PU burst mean arrival rate increases, the variance of the PU traffic distribution decreases. Then the mean arrival rate of the statistics becomes better metric for the PU traffic and thus the performance of CREAM-MAC increases in terms of access delay.

In the last experiment the burst length equals 100,000 bits (Fig. 8). According to the results, the proposed protocol predicts and avoids completely collisions with the PUs.

The main reason behind the superiority of the proposed protocol compared to CREAM-MAC, is that the former learns the PUs' traffic-PU traffic distribution characteristics and burst As the Learning Reinforcement Scheme is optimal there is no collision with the PU packets.The statistics metrics seem inefficient as they do not avoid collision with the PU and they do not utilize sufficiently

the spectrum holes; the CREAM-MAC does not differentiate

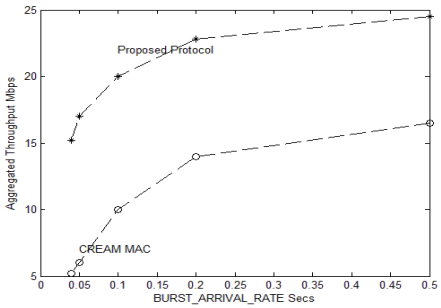

Figure 5: The aggregated throughput versus the burst arrival rate.

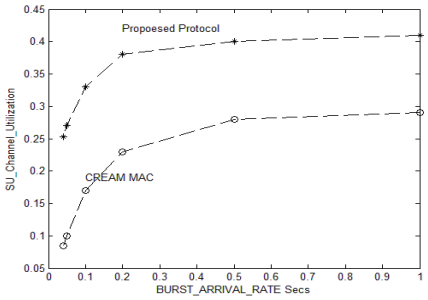

Figure 6. The SU channel utilization versus burst arrival rate.

its response to long andshort spectrum holes, i.e. in the forme

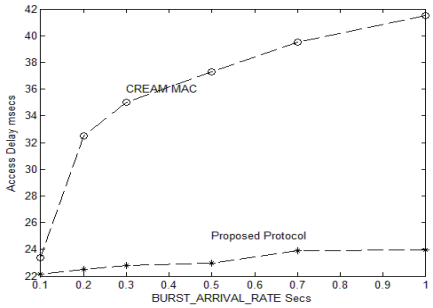

Figure 7: The access delay versus the burst arrival rate.

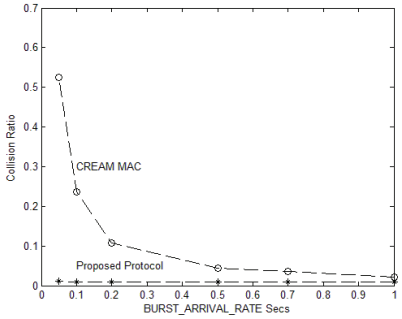

Figure 8. The collision ratio versus the burst arrival rate

case it bypasses opportunities and in the latter it does not avoid collision. The statistics do not overtake opportunities on the channels with the worse statistics when they experience longer spectrum vacancies due to PU-traffic distribution variance. The statistics do not respond promptly to the network conditions.

V. Conclusion

Most of the proposed MAC protocols reported in the literature utilize statistics on spectrum usage in licensed channels so as to identify transmission opportunities over these channels for SUs. However, using these statistics does not capture the fact that, nowadays, most traffic is of bursty nature; something that is not exploited in the identification of transmission opportunities. This paper proposes a Multi-Channel Cognitive MAC Protocol for ad-hoc cognitive networks that uses a reinforcement learning scheme forchannel selection and access based on observations of PUs traffic. The protocol learns the bursty nature of PUs' traffic to exceed the performance of the less adaptive statistical channel selection. The proposed scheme is shown to have better performance compared to the transmission opportunities via a statistics mechanism and avoids collision with the PUs. In the future work QoS will be considered in accordance with the proposed scheme.

References

[1] M.Thoppian, S.Venkatesan, R.Prakash, and R.Chandrasekaran, "MAC-layer scheduling in cognitive radio based multi-hop wireless networks", Proceedings of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM), 2006.

[2] A-C.Hsu, D.S.L. Weit, C.C.J. Kuo, " A cognitive MAC protocol using statistical channel allocation for wireless ad-hoc networks", Proceedings of Wireless Communications and Networking Conference(WCNC) March 2007, pp.105-110.

[3] C.Doerr, M.Neufeld, J.Fifield, T.Weingart, D.Sicker, D.Grunwald, "Multi-MAC-an adaptive MAC framework for dynamic radio networking", Proceedings of IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks(DySPAN), November 2005, pp.548-555.

[4] T.Chen, H.Zhang, G.M.Maggio, I.Chlamtac, "CogMesh: acluster-based cognitive radio network", Proceedings of IEEE International Symposium in Dynamic Spectrum Access Networks, April 2007, pp.168-178.

[5] H.Sui, X.Zhang, "CREAM-MAC: An Efficient Cognitive Radio EnAbledMultiChannel MAC Protocol for Wireless Networks", Proceedings of IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks, 2008,pp.1-8

[6] S.L.Wu, CY Lin, Y.C.Tseng, J.P.Sheu, "A new multi-channel MAC protocol with on-demand channel assignment for multi-hop mobile ad hoc networks", IEEE DySPAN, Maryland, USA,2005, pp. 203–213

[7] S.J.Yoo, H.Nan, T.I.Hyon, "DCR-MAC: distributed cognitive radio MAC protocol for wireless ad hoc networks". WirelCommun Mobile Comput. 9(5), 2009,pp.631–653

[8]J.Jia, Q.Zhang, X.Shen, HC-MAC: "A Hardware-Constrained Cognitive MACfor Efficient Spectrum Management", IEEE J Sel Areas Commun. 26(1),2008,pp.106–117

9]H.A.B.Salameh, M.M.Krunz, O.Younis, "Cooperative adaptive spectrum sharingin cognitive radio networks", IEEE/ACM Trans Netw. 18(4),2010,pp. 1181–1194

[10] X.Wang, A.Wong, P.H.Ho, "Stochastic Medium Access for Cognitive Radio AdHoc Networks", IEEE J Sel Areas Commun. 29(4),2011, pp. 770–783

[11]B.F.Lo, "A survey of common control channel design in cognitive radio networks",Elsevier Physical Communication, Vol 4, 1, 2011, pp. 26-39

[12]G.PapadimitriouandA.S.Pomportis, "On the use of learning automata in medium access control of single-hop lightwave Networks" Elsevier Computer Communications,vol.23,2000,pp.783-792