A Novel Access Network Selection Scheme using Q-learning Algorithm for Cognitive Terminal

Haifeng Tan[(1)], Yizhe LI[(2)], Yami Chen[(2)], Li TAN[(2)] and Qian Li[(2)]
(1) The State Radio Monitoring Center; Beijing 100037
E-mail: tanhf@srrc.org.cn
(2) Key Laboratory of Universal Wireless Communications, Ministry of Education
Wireless Technology Innovation Institute (WTI), Beijing University of Posts and Telecommunications
Beijing, P.R.China, 100876

Abstract— **In a B3G/4G wireless communication system, the users will connect to the network using one of several available radio access technologies. In this paper, we proposed a Q-learning based algorithm for terminals' independent access network selection with the aim of improving the resource utilization and providing the best quality of service with respect to the wireless environment status, network performance and user' requirement. In particular, for the first time we introduced the concept of low-carbon as one of the evaluation indicators of wireless communication performance, in order to reduce the power consumption and achieve a balance between quality and consumption. The proposed scheme is based on the concept of cognitive network, which has been proposed recently by the motivation of complexity, heterogeneity and reliability requirements of tomorrow's network and the cognitive pilot channel used in it. The performance of the access network selection algorithm is shown in the simulation and it can be seen that this algorithm significantly reduced the blockrate and power consumption as well as increased the throughput compared with random accessing approach. In future work, we will continue to research on the effective access network seletion algorithm and try to introduce the low-carbon indicator to other aspects of the wireless communication system.**

Keywords-access network selection; Q-learning algorithm; cognitive network; low-carbon

I. INTRODUCTION

In future wireless and mobile environments it is likely that users will have access to multiple networks at the same time. Therefore there is a need to have mechanisms in place to decide which network is the most suitable for each user at each moment for every application the user requires. Meanwhile, mobile devices are now being built as multi-homed, with multiple network interfaces. And with advances in microelectronics comes the birth of multi-mode terminals, allowing multiple RANs (LTE, HSPA, WLAN, etc.) interfaces to coexist in a single terminal [1].

This paper tries to deal with such a problem that on a user call launching, how to ensure that the user is attached to the most appropriate network that yields the best performance. In the paper, the network performance is evaluated by gain and cost of the network, considering the access network's status, terminal's status and user's QoS requirements.

In some previous studies, the similar problem has been dealt with, assuming that terminals send measured information and their own status to available networks and the networks make access decisions [2]. But evolution towards next generation networks makes it necessary of a user-centric approach where the terminals have greater control on their own behaviors [3]. In literatures [4][5], some network architectures supporting terminals' access network selection and the signaling system design are given, but they are lacking of available access network selection algorithm. There are standards supporting only one specific wireless technology and related access point selection, the most notable example is 802.11 [6-8], meaning that such solutions can not be used with other technologies. Many previous works on access point selection merely focus on the maximization of throughput. However, increasing number of popular applications is sensitive with other factors such as delay and reliability. With researches making efforts to take QoS requirements of different applications into account, only performance of an already established communication is optimized, while the QoS-aware access network selection has not been considered.

Cognitive networks are motivated by the complexity, heterogeneity, and reliability requirements of the future networks, which are increasingly expected to be self-organized so as to meet user and application objectives. [9]. Furthermore, the cognitive pilot channel (CPC) [10] is introduced to such heterogeneous cognitive network as supports for the reconfiguration management of networks and user terminals.

In this paper, a Q-learning based algorithm for terminals' independent access network selection in the cognitive network is proposed. Related heterogeneous networks information is transmitted to a terminal through the CPC channel. And then, the terminal takes that into consideration together with its request of traffic type and QoS to select the best access network using Q value tables. What is more, we introduced the concept of low-carbon as one of the performance evaluation indicators for the first time. The power consumptions of heterogeneous networks for different traffics are considered as important factors when choosing the access network. The simulation results show that the appropriate choice of access network can effectively reduce the power

consumption for transmitting specific traffic. The rest of paper is organized as follows: the Q-learning algorithm and access network selection are described in Section 2. In Section 3, the performance of the algorithm is analyzed and evaluated. Finally we conclude this paper in Section 4.

II. SCENARIO AND ALGORITHM

Large amount of the future wireless communication is very likely to happen a scenario of multi-network coverage. In this paper, we suppose that the LTE, WLAN and HSPA networks cover the same region and users' terminals have the capability of discovering and accessing any of these three networks with the help of CPC. In this heterogeneous network environment, the traffics, networks, and users' terminals will present various requirements and should be described and calculated in a unified parameters system. So the parameter congregation that describes the access networks selection should contain the generality of different functional entity mentioned above, moreover, it ought to be easy accessed as well as shield low-level detail and reduce computational complexity. In this paper parameters for access selection algorithm include: the various access networks' current coverage, load, types of service and average power consumption of unit bandwidth.

Here we used Q learning method to achieve dynamic adjustment of network configuration, so as to realize the optimization of overall system performance. Q learning algorithm is derived from the classic standard reinforcement learning algorithm [11] ~ [13], which is developed from animal learning theory. There is no need of prior knowledge during reinforcement learning process; the agent can choose action autonomously through continuous interaction with the environment to acquire knowledge. The reinforcement learning algorithms have been widely used in the field of robot behavior learning [14]. According to reinforcement learning the environment does not tell the agent how to generate the correct action, but through evaluating the actions aroused and produce reinforcement signal. The agent learns using the information provided by the external environment and their own experiences, through this method, the agent acquires knowledge in constant action-assessment (return) process and improves their own program of action to adapt to the environment. The goal of reinforcement learning system is to keep on learning, dynamicly adjust behavior strategy, and maximize the reinforcement signal. The basic reinforcement learning mechanism is: when the agent makes an action decision based on the current state of the environment, the environment gives the agent a feedback: return of benefit or negative returns, namely the reinforcement signal. The basic principle of reinforcement learning algorithm is: if the agent selects an action and is given positive return then the trend that the agent chooses this action will be strengthened; the other hand, the trend that the agent chooses this action will be weakened.

Reinforcement learning aims to acquire a behavior strategy by learning, in order to make the agent get the greatest reward after actions select. Reinforcement learning needs to define an objective function to indicate what the excellent action is in the long term. Usually the objective function is represented by the value of state or state-action value. The three main function forms are:

$$V^{\pi}(s_t)=\sum_{i=0}^{\infty}\gamma^{i}r_{i+1},\quad 0<\gamma\le 1 \tag{2-1}$$

$$V^{\pi}(s_t)=\sum_{i=0}^{h}r_i \tag{2-2}$$

$$V^{\pi}(s_t)=\lim_{h\to\infty}(\frac{1}{h}\sum_{i=0}^{h}r_i) \tag{2-3}$$

Where $\gamma\in(0,1)$ is the discount factor, r_t is the return after the agent changed from s_t to s_{t+1}, it can be positive, negative or zero. (2-1) is the infinite discount model; the agent takes return from infinite steps in future into return and cumulates it with some form of discount in the value function. (2-2) is the finite model, the agent only takes the sum of next h steps' return into account. (2-3) is the average return model; the agent takes the long-term average return. Obviously, if the objectives function can be determined, then the optimal behavioral strategies can be determined by virtue of (2-4):

$$\pi^{*}=\arg\max_{\pi}V^{\pi}(s),\quad \forall s\in S \tag{2-4}$$

the main feature and difficulty of reinforcement learning is that only the return value can be used to find the optimal strategy, and it is different from supervised learning algorithm in that it is an online learning algorithm in which the prior training sequence is not necessary.

Provided that the environment is a finite-state discrete Markov process, reinforcement learning system can select the action in a limited action set. The state transfers to s_{t+1} with the probability $Prob[s=s_{t+1}|s_t,a_t] = P[s_t,a_t,s_{t+1}]$ after the environment accepts the action and gives the agent return r_t.

The purpose of reinforcement learning is to find an optimal strategy, so that the total expected return will be the largest. Under strategy π, the value of state s_t is:

$$V^{\pi}(s_t)=r(\pi(s_t))+\gamma\sum_{s_{t+1}\in S}p[s_t,a_t,s_{t+1}]V^{\pi}(s_{t+1}) \tag{2-5}$$

Dynamic programming theory ensures that there is at least one strategy which achieves that:

$$V^{\pi^{*}}(s_t)=\max_{a\in A}\{r(\pi(s_t))+\gamma\sum_{s_{t+1}\in S}p[s_t,a_t,s_{t+1}]V^{\pi}(s_{t+1})\} \tag{2-6}$$

The idea of Q-learning is not to estimate the environment model, but rather optimize an Q function that can be iterative calculated directly. The Q function implements the action a_t under the state s_t, and then cumulates the reinforcement value according to the discount of the optimal action sequence executed later:

$$Q(s_t,a_t)=r_t+\gamma\max_{a\in A}\{Q(s_{t+1},a_t)\mid a_t\in A\} \tag{2-7}$$

Q-learning can be shown to converge under certain conditions and implemented through various neural networks. Each network corresponds to the Q value of a movement: $Q(s_t,a_t)$.

According to the definition of Q function,

$$Q(s_{t+1},a_t)=r_t+\gamma\max_{a\in A}\{Q(s_{t+1},a)\} \quad (2\text{-}8)$$

Only under premise of being the optimal strategy is the above formula established. During the learning phase the two sides in function (2-8) are not equal, resulting in the error signal as follows:

$$\Delta Q(s_t,a)=r_t+\gamma\max_{a'\in A}Q(s_{t+1},a)-Q(s_t,a) \quad (2\text{-}9)$$

Where $Q(s_{t+1},a_t)$ represents the Q value corresponding to the next state, and Q iterative learning rule is shown below:

$$Q(s_{t+1},a)=\begin{cases}Q(s_t,a)+\alpha\Delta Q(s_t,a), & x=x_t,\ a=a_t\\ Q(s_t,a), & \text{others}\end{cases} \quad (2\text{-}10)$$

Specific implementation is to establish a two-dimensional Q value table in each terminal. One dimension is used to represent the index of all the possible status, the other to represent the index of all possible actions. Each cell in the Q value table stores the current Q value q(s, a) corresponding to certain action a, in the given status s. The action selection algorithm is : the terminal selects an action with a certain probability from the action congregation based on Q values; the bigger the Q value, the higher probability the corresponding action is chosen. There are two types of action selection algorithm: one is based on the experiences of action selection already got, while the other is to continuously try new actions, both of which have their own advantages and disadvantages. Depending on learning experience can make the algorithm process achieve convergence quickly, however, with risk of falling into local optimum; trying new action space will be more extensive and comprehensive experience to achieve optimize performance, but at the expense of more learning time [15]. In this paper we chose the action selection algorithm used which is based on Boltzmann distribution.

Algorithm flows:

(1) Each terminal initializes all the data of its Q value table to 0.

(2) When access request arrives, the terminal makes construction of new state s, according to the session condition and network load information. Four elements (C, G, P, L) constitute s, among which $C\in\{0,1,...,8\}$ represents which access networks cover the terminal, $G=[g_1,...,g_i,...,g_k]$, $g\in\{1,2,...,G\}$ represents the type of traffic that requests admission. $P=[p_1,...,p_i,...,p_k]$, $p_i\in\{0,1,2\}$ represents the number of traffic initiated redirection. $L=[l_1,...,l_i,...,l_k]$, $l_i\in\{0,1,...,10\}$ represents the current load levels of all access networks. After obtaining the new state *s*, all its corresponding Q values in the currently used Q value table are to be calculated.

(3) The terminal chooses an access network with certain probability according to the Q value table. If the chosen network has enough bandwidth to accept then the terminal calculates return according to the outcome of the choice, or else the terminal tries to access to another network. The return can be calculated as: :

$$r_t(s,a)=\zeta(g)\tau^2/\lambda p \quad (2\text{-}11)$$

τ is the traffic sustained duration, p is the terminal's energy consumption, λ is the weight of energy consumption, $\zeta(g)$ is the traffic utility function, $\zeta(g)=\beta*d*\partial$, β is the breakage factor of redirection's delay, d is the average data rate and in particular, for the purpose of minimizing the terminal's power consumption, we set the gain parameter ∂ to adaptively adjust the traffic utility function according to different coverage situations. Simplicity, here the voice and data are of no distinction.

(4) According to the return and previous round Q values, the terminal updates the Q value related to the status s and chosen action a as:

$$Q_k(x,a)=R(x,a)+\gamma\sum_y P_{xy}[\pi(x)]V^{\pi}(y) \quad (2\text{-}12)$$

and

$$Q_k(x,a)=R(x,a)+\gamma\sum_y P_{xy}[\pi(x)]V^{\pi}(y) \quad (2\text{-}13)$$

in which: $R(x,a)=E\{r\mid x,a\}$; $V^{\pi}=\max_{b\in A}Q^{\pi}(y,b)$, $V_{k-1}(y_k)\equiv\max_b\{Q_{k-1}(y,b)\}$, π is the scheme, α is learning factor.

III. PERFORMANCE ANALYSIS

The reconfigurable system in the simulation is composed of three heterogeneous networks: LTE, HSPA and WLAN, each deploying one cell, with coverage of red, green and yellow area respectively, as shown in Fig. 1. There are the two cells configured respectively as LTE and HSPA with partial overlapping area. And there is a WLAN access point in the LTE cell, the HSPA cell also has overlapping coverage with WLAN. The LTE base station is located at coordinates (0, 0), its wireless coverage area radius is 1000 meters and cell capacity is 10Mbps; HSPA base station is located at coordinates (900,150), its wireless coverage area radius is 600 meters, cell capacity is 7.2Mbps; WLAN base station locates at coordinates (250,250), its wireless coverage area radius is 200 meters, cell capacity is 2Mbps. All users are uniformly distributed in the area covered by at least one base station. In the simulation, the mobility of user is not considered in order to focus on the performance of the access selection scheme, although the scheme can also be applied to moving user terminals.

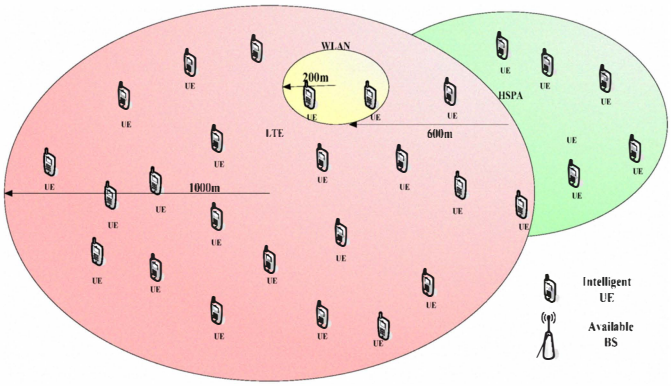

Fig.1. The Multi Access Networks Scenario

All terminals are reconfigurable and have a mean session arrival rate of μ following the Poisson distribution and different values of μ are used to prove the performance of the proposed scheme under different traffic load, as shown in Table 2. The multi-service traffic is considered in the simulation including voice and data sessions. And the session duration follows the negative exponential distribution with mean value of 1/μ. All the relevant simulation parameters are shown in table 1.

Table 1 Simulation parameters configuration

	LTE	HSPA	WLAN
Cell capacity (Mbps)	**10**	**7.2**	**2**
Coverage scope (m)	**1000**	**600**	**200**
Session arrival rate μ (calls/s)	1/6, 1/5, 1/4, 1/3, 1/2, 2/3		
Session duration $1/\mu$ (s)	120		
Traffic bandwidth (kbps)	32		
Time discount factor γ	0.9		
Learning factor α	0.1		
Simulation iterations	10,000		

Table 2 lists the corresponding traffic gain parameters ∂ of different coverage situations.

Table 2. Traffic Utility Function Gain Parameters Corresponding to Different Coverage Situations

Area	LTE	HSPA	WLAN
1	**1**	**0**	**0**
2	**0**	**1**	**0**
3	**0**	**0**	**1**
4	**2**	**1**	**0**
5	**1**	**0**	**2**
6	**0**	**1**	**2**
7	**1**	**1**	**2**

The area 3 and 6 respectively represent coverage situations that the area covered by only WLAN and both WLAN and HSPA, there are not existing areas like that, but in order to ensure the integrity, here we provide the corresponding parameters. And the reasons for selecting the gain parameters are that: such as area 7, the terminals are covered by LTE, HSPA, and WLAN, because the WLAN coverage is small as well as provides more resources and consumes less power, greater gain will be acquired if accessing WLAN. So the WLAN's gain parameter is 2 times as the other two. Other parameters are also based on similar considerations.

In the simulation, we compared the performance of proposed access network selection method and the random access method. It can be seen from Fig. 2 that after access network selection, the terminals using Q-learning algorithm adaptively tend to access the network with more available bandwidth resource and a high session quality, resulting in less block rate than the random access method for different traffic arrival rates.

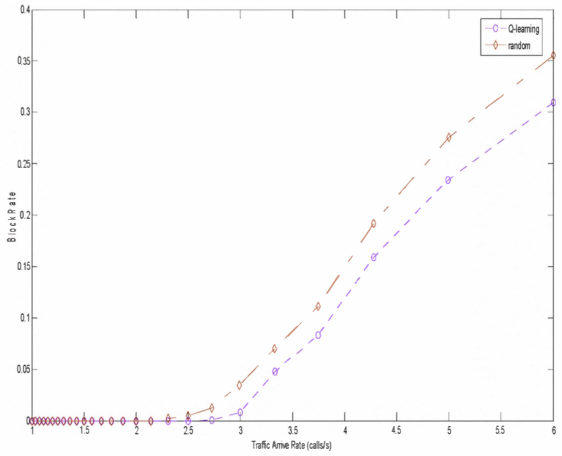

Fig.2. Blocking Rate for Different Traffic Arrival Rate

Fig. 3 shows that the throughput of terminals using Q-learning access network selection algorithm is higher than that deploying random access. This is because the terminals adaptively choose to access the network with higher average data rate as well as the reduction of the block rate.

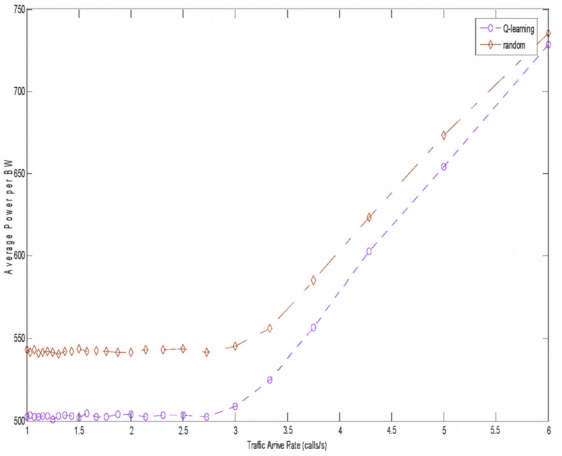

Fig.3. Throughput for Different Traffic Arrival Rate

The results shown in Fig.4 give an demonstration of the performance in the proposed low-carbon target. Through Q-

learning access network selection algorithm, the terminals take the trade-offs between quality of traffic and power consumption into account while accessing the sessions with satisfying QoS.

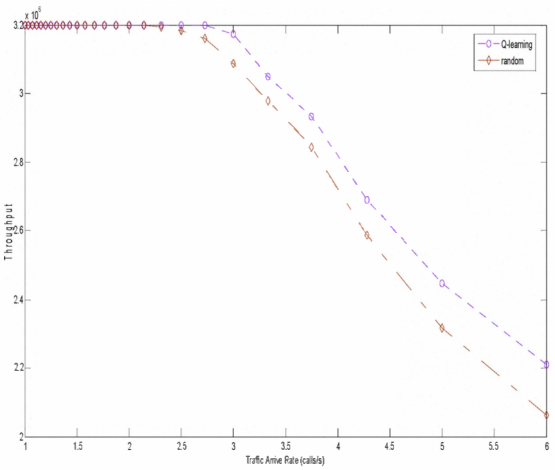

Fig.4. Average Power Consumption per BW for Different Traffic Arrival Rate

IV. CONCLUSION

In this paper, we analyzed the issue that how can the users connect to the network using the best access technology in the future heterogeneous wireless network environment. A Q-learning based algorithm is proposed for the terminals' autonomous access network selection, which takes the network status, user preference as well as QoS requirement of different applications into fully account. At the same time, we introduced the low-carbon indicator to optimize power consumption while providing satisfactory service. The simulation results show the effectiveness of the proposed algorithm.

V. ACKNOWLEDGMENT

This work was sponsored by National Key Technology R&D Program of China (2009ZX03007-004), Key Project of National Natural Science Foundation of China (60632030) and National Basic Research Program of China (973 Program) (2009CB320406).

REFERENCES

[1] Baldo N, Zorzi M, Cognitive Network Access using Fuzzy Decision Making, IEEE International Conference on Communications (ICC 2007), JUN 24-28, 2007 Glasgow, SCOTLAND.

[2] Al-Gizawi T., Peppas K., Axiotis D.I. et al., "Interoperability Criteria, Mechanisms, and Evaluation of System Performance for Transparently Interoperating WLAN and UMTS-HSDPA Networks," IEEE Network, vol. 19 (4), Aug. 2005, pp. 66–72.

[3] Gustafsson E, Jonsson A. Always best connected. IEEE Wireless Commun Mag 2003; 10(1):49–55.

[4] Nguyen-Vuong, Quoc-Thinh, Agoulmine, Nazim, Ghamri-Doudane, Yacine. Terminal-controlled mobility management in heterogenous wireless networks. IEEE Commun Mag 2007;45(4):122–9.

[5] Inoue M，Mahmud K，Murakami H．MIRAI：a solution to seamless access in heterogeneous wireless networks[C]．Anchorage USA：ICC 03．May 2003．PP．1033—1037.

[6] K. Sundaresan and K. Papagiannaki, "The need for cross-layer information in access point selection," in Internet Measurement Conference, Rio De Janeiro, Brazil, Oct 2006.

[7] A. Nicholson, et al., "Improved access point selection," in Int'l Conference on Mobile Systems, Applications and Traffics, Uppsala, Sweden, June 2006.

[8] V. Mhatre and K. Papagiannaki, "Using smart triggers for improved user performance in 802.11 wireless networks," in Int'l Conference on Mobile Systems, Applications and Traffics, Uppsala, Sweden, June 2006.

[9] R. W. Thomas, D. H. Friend, L. A. DaSilva, and A. B. MacKenzie, "Cognitive Networks: Adaptation and Learning to Achieve End-toEnd Performance Objectives," IEEE Communications Magazine, vol. 44, no. 12, pp. 51–57, 2006.

[10] Pascal Codier, Didier Bourse, David Grandblaise, Klaus Moessner, Jijun Luo, Clemens Kloeck, et. al.: Cognitive Pilot Channel, Proccedings of WWRF15, Paris, 8.- 9.12.2005 CPC,WWRF.

[11] Litao Liang, "The key technologies of self-management in heterogeneous reconfigurable networks." PhD Thesis, BUPT.

[12] Kaelbling L P., "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, 1996, 4: 237~285.

[13] Watkins C. J. C. H., "Learning from Delayed Rewards," Ph.D. Dissertation, Cambridge University, Cambridge, U.K., 1989.

[14] Singh S., "Agents and Reinforcement Learning," *Miller freeman publish Inc*, San Mateo, CA, USA, 1997

[15] Senouci S.-M., Pujolle G., "Dynamic Channel Assignment in Cellular Networks: A Reinforcement Learning Solution", in Proc. 10th Intl. Conf. on Telecommun., vol.1, 2003, pp. 302–309.