
Article:
Jiang, Tao, Zhao, Qiyang, Grace, David orcid.org/0000-0003-4493-7498 et al. (2 more authors) (2016) Single-State Q-learning for Self-organized Radio Resource Management in Dual-hop 5G High Capacity Density Networks. Transactions on Emerging Telecommunications Technologies. 1628–1640. ISSN 2161-3915

https://doi.org/10.1002/ett.3019

Single-State Q-learning for Self-organized Radio Resource Management in Dual-hop 5G High Capacity Density Networks

Tao Jiang, Qiyang Zhao, David Grace, Alister G. Burr and Tim Clarke
Department of Electronics, University of York, York, YO10 5DD, United Kingdom

Abstract—**In this paper, a dual-hop wireless backhaul and small cell access network has been exploited with effective spectrum sharing, to provide 1 Gb/s/km^2 ultra high capacity density for 5G ultra-dense network deployments. We develop a Single-State Q-learning (SSQL) based radio resource management algorithm for dynamic spectrum access creating a self-organized network. It intelligently utilizes the instantaneous spectrum observation information from spectrum sensing or a database, to learn long-term optimized decisions based on historical information of the system. The conventional Q learning algorithm with state-action pairs has been simplified to a stateless format and applied in a fully distributed manner on individual data file transmissions, which reduces the complexity of the learning model and improves the applicability of Q-learning algorithms to self-organized wireless networks. The results show that not only does the proposed algorithm completely remove the requirement for frequency planning, but also it improves the convergence, QoS and system capacity substantially by achieving higher link capacity on both access and backhaul networks.**

Keywords- 5G, Ultra-dense Network, Radio Resource Management, Dynamic Spectrum Access, Self-organized Network, Qlearning

I. INTRODUCTION

In future 5G radio access networks, ultra-high capacity density has been identified as a critical system requirement for mobile broadband services in dense urban areas [1]. To support this requirement, ultra-dense networks are being considered as a possible solution. The idea is to deploy a large number of small cell base stations with mobile backhaul capability, to effectively improve network spectrum efficiency and scalability. In this context, cognitive dynamic spectrum management is essential to reduce interference, maximize resource utilization and enhance system capacity [2, 3]. The network is also required to be self-organized, in order to reduce operational expenditure and improve reliability [2, 4]. Machine learning technologies have been widely studied to allow the radio nodes to intelligently select radio spectrum and mitigate interference. They learn from historical experience to deliver long-term stable performance in changing radio environments, unlike with conventional cognitive radio where decisions are made based on instantaneous measurements. However, the conventional algorithms such as Q learning require an observation of global environment states, which are not applicable to self-organized networks [5]. Thus in this paper, we developed a novel Single State Q learning (SSQL) algorithm, which operates in a fully distributed manner and directly utilizes physical environment information. The objective is to support both Dynamic Spectrum Access (DSA) and Self-organized Network (SON) configurations, and achieve 1 Gbps/km^2 capacity density for 5G ultra-dense network requirements.

A very conservative estimation indicates that in a city centre area of a typical European city with an average population density of 5000 people/km^2, the demanded throughput density is about 1 Gbps/km^2 [6]. The figure is even greater in large Asian or North American cities where the population density is much higher. It is also crucial that 5G should provide the required average data rate to all active users simultaneously within a service area. This is particularly true in highly populated urban city centre areas where the demand for wireless broadband service is the highest. The throughput density that LTE-A can provide is about 350 Mbps/km^2 [6], which may be adequate in less populated areas but insufficient for high user density areas.

A novel dual-hop hierarchical wireless system has been recently proposed in [6] to deliver a capacity density of 1 Gbps/km^2 in a cost-effective way, providing a 5G wireless access in a dense urban city area. On the other hand, the system is required to be cost effective, including a maximized utilization of the current 20 MHz 4G band and a self-organized network architecture. This ambitious goal is achieved mainly by distributing a relatively large amount of short-range, below roof-top Access Base Stations (ABSs), on streets to provide sufficient capacity density to MSs. The traffic is then aggregated at ABSs and self-backhauled to the higher level over roof-top Hub Base Station (HBS) wirelessly. By using a multi-beam directional antenna at the HBSs to provide sufficient self-backhaul capacity, and below-rooftop directional antennas at the ABSs to provide high access capacity to the MSs, this novel architecture is able to achieve a significant spectral efficiency.

However, with such a dual-hop architecture, the Radio Resource Management (RRM) aspects of the system become significantly more important. First, to meet the ambitious goal of 1 Gbps/km^2, it is desirable that the limited 20MHz band is reused aggressively. As used for LTE networks, a reuse factor of 1 is necessary, where any of a HBS or ABS should have access to the entire spectrum band [7]. However, the conventional Fractional Frequency Reuse scheme proposed in LTE is based on a cellular scenario, which is not applicable to this type street coverage configuration. The idea of using

below-rooftop ABSs is to take full advantage of the 'natural' isolation of building blocks normally seen in large cities so that the radio frequency can be reused aggressively. Secondly, with a relatively larger amount of ABSs deployed on the streets, the design of the RRM becomes significantly more complex. Dedicated backhaul links, such as microwave and fiber, are normally implemented at base stations in various systems. However, it is not economically feasible in a dense small cell network, because the use of high frequency bands require line-of-sight propagation links which are difficult to achieve in urban scenarios. In band wireless backhaul has the potential to significantly improve spectrum utilization and reduce the deployment cost, which is primarily considered for the joint self-backhaul and access design. On the other hand, this makes the design, implementation, configuration, and operation of fixed RRM strategies hard to plan and optimize.

Therefore, spectrum sharing and self-organization are key technologies to reduce the complexity of the RRM design. [8] analyses the spectrum sharing schemes between different operators, and demonstrates an improvement to spectral efficiency with increasing number of users. [9] proposes radio access technology selection schemes to use spectrum resources provided from different radio access networks. [10] investigates spectrum sharing in a D2D network and the data rate uploaded to eNBs. A self-organized network requires RRM decisions to be based on local observations only, which can be potentially achieved from cognitive radio. Spectrum sensing has been extensively studied as a key technology to achieve cognitive radio. It allows the user, base station or network to observe the spectrum environment and utilize spectrum holes for data transmission. [11] proposes a distributed cooperative spectrum sensing approach using the concept of correlated equilibrium from game theory, to improve the reliability of detection. However, spectrum sensing is constrained by the cost of equipment, time and bandwidth, meaning that it cannot provide real time information for every user on every data file transmission. [12] has addressed this problem and proposes a centralized spectrum leasing algorithm to balance the cost of spectrum acquisition and QoS, though which is not applicable to a distributed SON design, because it requires well established spectrum occupancy information exchange between primary and secondary users.

Reinforcement Learning (RL) algorithms have been widely studied for cognitive radio to reduce the cost of spectrum sensing in achieving DSA. The methodology is to discover which actions yield the most reward on a trial and error basis. The implementation scenario of reinforcement learning is the Markov Decision Process (MDP), where a learning agent interacts with its environment to achieve a goal. Such a scenario is well suited to RRM, where the action of resource allocation interacts with the spectrum environment, and the goal is to achieve effective spectrum separation among adjacent users. In order to provide effective RRM in a SON architecture under the DSA scenario, a learning engine is expected to perform 1) self-adaption: identify environment changes and make subsequent decisions; 2) self-optimization: exploit and improve the best action space in specific environment state; 3) fast convergence: fix on the preferred action space and resist rapid environment changes.

Reinforcement learning based channel assignment can be generally categorized into centralised algorithms where channels are assigned at a centralized server, and distributed algorithms where spectrum decisions are made by individual users. Research work in the field has largely focused on centralised scenarios prior to the introduction of cognitive radio. The centralized Q-learning based dynamic channel assignment was originally proposed in [13], which assigns channels on a call-by-call basis by utilizing the information gained through Q learning. It has been shown that Q-learning outperforms the Fixed Channel Assignment under different traffic conditions. This work has been extended in [14], which introduces Call Admission Control (CAC) when updating the Q-values of channels.

With the rapid development of Cognitive Radio, distributed RL algorithms have received more attention recently. A secondary cognitive radio system model based on IEEE 802.22 standard is considered in [15], where distributed Q-learning based techniques are applied to learn how to control the transmit power in order to reduce the aggregated interference at Primary Users (PUs) receivers. The system state is defined jointly by the aggregated interference at PUs, the approximate distance between the Secondary Users (SUs), the PU protection contour, and the transmit power level at SUs. A theoretical study of a simple 2 SU x 2 channel case has been carried out in [16]. No PU is assumed in the paper and spectrum sensing is ignored. The system state is defined by the availability of channels, which in practice would require system level information. A fixed set of rewards are also assumed throughout the learning process. Multi-agent reinforcement learning for cognitive radio has been studied in a more realistic scenario in [17]. A Q-learning based joint channel and power allocation scheme has been proposed. The state of the system has been defined by using the transmit power level and the channel utilization information of all users. In [18, 19], the authors studied a multi-agent reinforcement learning algorithm in a Carrier Sense Multiple Access (CSMA) based system. It is assumed that the Q-values are updated after every packet transmission. The learning model in their work requires the location information of entities at the system level in order to define the system states.

However, most of the existing work only concerns the access link of the network, and also in most cases system level information is required to define the system state S in the learning model. The state-action formulation is an important process in any RL model, including Q-learning. Properly defined system state-action pairs are often of fundamental importance to any learning based system.

The wireless communication system is a multi-server multi-user queueing system [20], where the physical state is usually referred to the number of resources (channels) that a base station provides to the users. In this context, global state information is required when applying conventional multi-

state Q learning for RRM [13]. This requires a centralized network controller or a fully coordinated control plane protocol, which is highly complex in the dual-hop architecture with joint access and backhaul design, and is not applicable to SON. More importantly, the convergence time of centralized Q learning highly relies on the size and format of the network topology. It takes a significantly long trial-and-error process for the network to obtain optimized solutions on every distributed node, which is inefficient in large ultra-dense networks. Furthermore, it is difficult to define the states in Q learning that match the physical states in the network. Thus in this paper, we introduce a stateless Q learning algorithm where the state formulation is no longer required, and the actions can be modelled as channels assigned to data file transmissions in the network. As a result, the SSQL algorithm can effectively model the physical behaviour of the wireless system in a fully distributed manner without the observation of global states, which reduces the complexity of the system control plane architecture.

In summary, a system state formulation is required in all the algorithms reviewed above, which in practice is not straightforward. This is particularly true in a fully distributed network where only local observation is available. Thus in our previous work [21], a simple reinforcement learning algorithm without state formulation has been developed for a simplified transceiver pair system. This has been further extended in [22] for the application on a multi-hop backhaul network. However, none of them has been investigated in the context of a joint access and backhaul system.

The Single-State Q-Learning (SSQL) based RRM algorithm proposed here is designed to utilize not only the instantaneous spectrum sensing measurements but also the historical information of the system. Such information has not been sufficiently utilized in normal wireless systems to facilitate the optimization process which could prove wasteful. This work also aims to evaluate and compare the capacity density performance of the dual-hop beyond next generation mobile broadband system with different RRM strategies, including the conventional frequency planning algorithm and cognitive RRM with sensing information only.

The novel contribution of this work lies mainly in four aspects: (1) The SSQL algorithm is applied jointly to both the access and self-backhaul network. In other words, we consider the joint RRM design of access and self-backhaul links based on SSQL which has not been addressed before. (2) Unlike most of the previous work, this paper adapts a SSQL model where the issue of state-action pair formulation is less significant. Hence the complexity of the Q-learning model is reduced, and because the system state formulation is no longer required, the adaptability of such a model to wireless communication systems is significantly improved; (3) The traditional Q-learning model is modified to take the physical measurement of the wireless system as rewards, e.g. data rate or Signal-to-Noise plus Interference Ratio (SINR). Most of the existing literature, on the contrary, uses sets of predetermined reward values which are not directly linked to the physical state of the wireless system. By properly linking the learning model to the wireless system, the learning model is tailored to utilize as much system information as possible so that the benefit of applying learning is maximized. (4) For the first time, the capacity of the dual-hop beyond next generation high capacity density mobile broadband system is evaluated on a large scale at the system level with different RRM algorithms, i.e. fixed frequency planning, cognitive RRM, and SSQL based RRM.

The rest of the paper is organized as follows: Section II introduces the dual-hop ultra-dense network model. Section III specifies the detail of distributed SSQL learning model. In Section IV, system-level simulation is conducted and the performance of the proposed algorithm is compared with the fixed frequency planning and cognitive RRM algorithm. Finally the conclusion is given in Section V.

II. DUAL-HOP ULTRA-DENSE 5G ARCHITECTURE

2.1. System Model

A novel dual-hop hierarchical architecture has been proposed as an ultra-dense network solution for 5G [6]. In order to enable a cost-efficient way of delivering high capacity density in the service area, the system is composed of an access network and a self-backhaul network. The key elements of the novel architecture are:

- Hub Base Station (HBS): an entity that is connected to the operator's backhaul network, which can be co-located with conventional Macrocell BSs. A multi-beam directional antenna is deployed over roof-top at HBSs, providing a high capacity self-backhaul link to the access network entities.
- Access Base Station (ABS): a low-cost entity that provides the access to the Mobile Subscribers (MS). A large number of ABSs will be mounted below roof-top on electricity poles, traffic lights, traffic signs, etc, which establishes a dense small cell network. The ABSs have several single-beam directional antennas. This includes one above-rooftop backhaul antenna and two to four below roof-top access antennas which point in opposite directions to provide spatial diversity to MSs on streets.

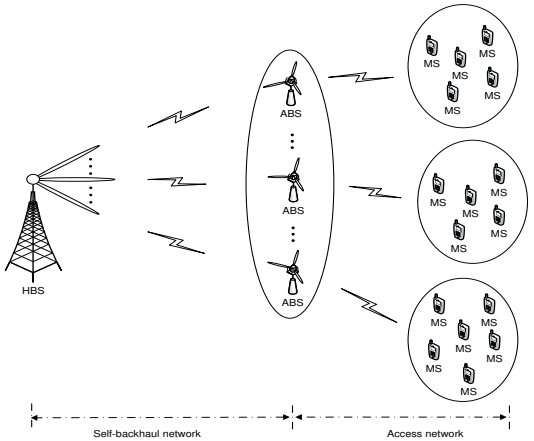

Fig. 1. Dual-hop High Capacity Density Network Model

A large number of these low-cost ABSs are deployed along the streets, providing sufficient capacity to MSs in the Access Network. The aggregated traffic at the ABSs is then transmitted to the associated HBSs wirelessly via the dedicated directional antennas. Therefore, by carefully designing advanced RRM techniques, the dual-hop beyond next generation mobile network has the potential to meet the much desired 1 Gbps/km^2 capacity density.

We consider a wireless network with M HBSs. Each HBS has L beams, serving N ABSs. Each ABS provides service to K MSs. Then the HBS transmit power can be denoted as $P^H = \begin{pmatrix} p_1^{H,1} & \cdots & p_1^{H,L} \\ \vdots & \ddots & \vdots \\ p_M^{H,1} & \cdots & p_M^{H,L} \end{pmatrix}$, where $p_m^{H,l}$ is the transmit power of beam l of HBS m. The ABS transmit power can be expressed as$P^A = \begin{pmatrix} p_1^{A,1} & \cdots & p_N^{A,1} \\ \vdots & \ddots & \vdots \\ p_1^{A,M} & \cdots & p_N^{A,M} \end{pmatrix}$, where $p_n^{A,m}$ denotes the transmit power of ABS n, which is associated with HBS m.

The BuNGee architecture uses Frequency-division Duplexing (FDD), where the uplinks and downlinks are allocated a pair of fully separated and equal band assignments [6]. Within each band, the access and backhaul links on the same direction may interfere with each other. However in the FDD context, the uplink-downlink interference can be effectively eliminated by sufficient spectrum band separation. FDD remains one of the preferred methods of duplexing by operators for mobile deployments today, owing to its general ease of use. Moreover, future 5G communication systems are likely to have a large amount of data traffic on both uplinks and downlinks, particularly in dense city areas. As a result, the uplinks and downlinks will exhibit similar system performance from the RRM perspective, as they are allocated equal and separated spectrum. In order to not be unduly repetitive, we focus solely on the downlink performance in this paper. If we assume that there are U channels available in total, and each channel is divided into R OFDM subchannels.

2.2. System Architecture

As we discussed previously, the primary objective of the dual-hop wireless system is to provide sufficient capacity density to the dense city centre area. Therefore, the simulator is based on a grid like system illustrated in Fig. 3, where the building heights are 6m, and the ABS antennas are located below rooftop height on the street lamps. The building block size is 75*75m, and the street width is 15m. Such layouts are seen in many modern cities today. This was originally modelled on the Diagonal district in Barcelona in Spain and has also been adopted as basis for modelling these dual hop architectures in the ETSI report [6]. The simulator characterises the performance in 3 spatial dimensions, to model the benefits of isolation caused by buildings in this scenario, and over roof-top transmissions of the self-backhaul links.

The HBSs are placed above roof-top at the centre of each square cell covering 25 buildings. A directional antenna with 20 beams is placed at the HBS to provide high-capacity self-backhaul links. The beams of the HBS antenna point towards different ABSs in the cell. The ABSs are located below rooftop at the horizontal and vertical streets cross. Each ABS has two single-beam directional antennas pointing in two opposing directions either North-South (N-S) or East-West (E-W) along the streets. The corner ABSs have 4 beams covering the N-S and E-W neighbouring streets. Dedicated backhaul antennas at the ABSs are directed in the direction of the largest power ray towards the centre of the cell as suggested in [23].

A number of channel models have been used to calculate the path loss in the simulation, which effectively models the real environment, including random effects such as variable attenuation due to shadowing. The ray-tracing based channel model introduced in [6] is used to estimate the path loss between entities with the self-backhaul network. WINNER II provides a comprehensive set of channel models that are capable of covering the propagation environment of the access network of the system. In this simulation, WINNER II B1 is used to calculate the pass loss between ABS and MS that is located outside of a building block. The path loss between the ABS and MS inside a building block is estimated using WINNER II B4 [24].

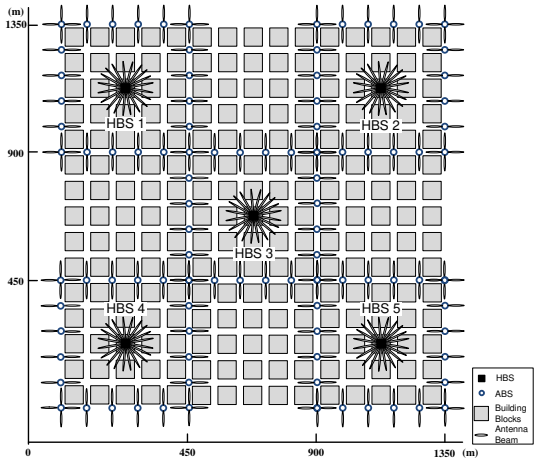

Fig. 2. Square Cell Deployment Scenario: A 5 HBS case

2.2. Self-Backhaul network

The SINR for ABS n is (signal transmitted from HBS $m$in channel u and subchannel r):

$$\gamma_{n,u,r}^{m,l} = \frac{p_m^{H,l} g_{u,r}^{B,m,l,n}}{\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{L} p_i^{H,j} g_{u,r}^{B,i,j,n} + \sum_{i=1,i\neq l}^{L} p_m^{H,l} g_{u,r}^{B,m,i,n} + \sigma^2} \quad (1)$$

where $g_{u,r}^{B,m,l,n}$ is the gain of the wireless link from the lth beam of HBS m to ABS n. $\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{L} p_i^{H,j} g_{u,r}^{B,i,j,n}$ is the interference from other HBSs to ABS n. $\sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{u,r}^{B,m,i,n}$is the interference from the other beams of HBS m, using the same channel u and subchannel r.$\sigma^2$is the noise power.

The link gain $g_{u,r}^{i,j}$ between two entities *i,j* is obtained by:

$$g_{u,r}^{i,j} = \frac{G_i(\theta_i)G_j(\emptyset_j)}{PL(d_{ij})} \quad (2)$$

Where $PL(.)$ is path loss, $G_i(x_i)$ is antenna gain of entity i on horizontal and vertical directions, which can be obtained by:

$$G(x) = D(x)E_a \quad (3)$$

E_a is the antenna efficiency and $D(x)$ is the directivity of the antenna. $D(x)$ is obtained by [25]

$$D(x) = D_{max}(\cos x)^{\mu} \quad (4)$$

and

$$D_{max} = \frac{32\log 2}{\theta_{3dB}^2 + \emptyset_{3dB}^2} \quad (5)$$

where θ_{3dB} and $\emptyset_{3dB}$ are 3dB beamwidth on horizontal and vertical directions, respectively.

The directional antenna is an important part of the network. By using directional antennas at the base stations, the interference between the self-backhaul links and the interference between the ABS beams are kept to a minimum. Therefore, in order to clearly identify the impact of the directional antenna, equation (1) can be rewritten as:

$$\gamma_{n,u,r}^{m,l} = \frac{p_m^{H,l} D_{max}^H(\cos\theta_{m,l})^{\mu_H} E_a^H D_{max}^A(\cos\theta_n)^{\mu_A} E_a^A}{PL(d_{m,n})} \Big/ \left\{ \sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} \frac{p_i^{H,j} D_{max}^H(\cos\theta_{i,j})^{\mu_H} E_a^H D_{max}^A(\cos\theta_n)^{\mu_A} E_a^A}{PL(d_{i,n})} + \sum_{i=1,i\neq l}^{L} \frac{p_m^{H,i} D_{max}^H(\cos\theta_{m,i})^{\mu_H} E_a^H D_{max}^A(\cos\theta_n)^{\mu_A} E_a^A}{PL(d_{m,n})} + \sigma^2 \right\} \quad (6)$$

2.3. Access Network

When carrying out the interference study of the dual-hop wireless system, two categories of interference need to be considered: the interference within each hop of the network, and the interference between the access network and the self-backhaul network. In the case of the downlink, the potential interference at the MS not only comes from entities within the access network, but also the interference from the self-backhaul network. The SINR at MS k is a little more complicated to calculate than the SINR at the ABSs. The SINR for MS k is (signal transmitted from ABS n (associated with HBS m) in channel u and subchannel r):

$$\gamma_{k,q,r}^{n,m} = p_n^{A,m} g_{u,r}^{A,n,k} \Big/ \left\{ \sum_{i=1,i\neq m}^{M} \sum_{j=1}^{N} p_j^{A,i} g_{u,r}^{A,j,k} + \sum_{i=1,i\neq n}^{N} p_i^{A,m} g_{u,r}^{A,i,k} + \sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} p_i^{H,j} g_{u,r}^{B,i,j,k} + \sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{u,r}^{B,m,i,k} + \sigma^2 \right\} \quad (7)$$

where $g_{u,r}^{A,n,k}$ is the link gain between ABS n and MS k. $\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{N} p_j^{A,i} g_{u,r}^{A,j,k}$ is the interference from all the ABSs in other cells that are using the same frequency. $\sum_{i=1,i\neq n}^{N} p_i^{A,m} g_{u,r}^{A,i,k}$ is the interference from other ABSs in the same cell. $\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{L} p_i^{H,j} g_{u,r}^{B,i,j,k}$ is the interference from the HBS beams in other cells and $\sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{u,r}^{B,m,i,k}$ is the interference from the HBS beam within the same cell. σ^2 is the noise power.

Similarly, if we rewrite equation (7) the same way as we did for equation (1) to consider the impact of directional antennas, the following equation can be obtained:

$$\gamma_{k,u,r}^{n,m} = \frac{p_n^{A,m} D_{max}^A(\cos\theta_n)^{\mu_A} E_a^A G^M}{PL(d_{m,n,k})} \Big/ \left\{ \sum_{i=1,i\neq m}^{M} \sum_{j=1}^{N} \frac{p_j^{A,i} D_{max}^A(\cos\theta_{i,j})^{\mu_A} E_a^A G^M}{PL(d_{i,j,k})} + \sum_{i=1,i\neq n}^{N} \frac{p_i^{A,m} D_{max}^A(\cos\theta_{m,i})^{\mu_A} E_a^A G^M}{PL(d_{m,i,k})} + \sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} \frac{p_i^{H,j} D_{max}^H(\cos\theta_{i,j})^{\mu_H} E_a^H G^M}{PL(d_{i,k})} + \sum_{i=1,i\neq n}^{N} \frac{p_i^{A,m} p_m^{H,i} D_{max}^H(\cos\theta_{m,i})^{\mu_H} E_a^H G^M}{PL(d_{m,k})} + \sigma^2 \right\} \quad (8)$$

where G^M is the antenna gain of mobile handset antenna. G^M is a fixed value regardless of θ since we assume omnidirectional antenna at the MS end.

Adaptive modulation is assumed and the truncated Shannon bound (TSB) is considered in this work to represent the achievable data rates in practice given an Adaptive Modulation and Coding (AMC) codeset [26]. The achievable data rate of the self-backhaul links is given by:

$$C_{n,u,r}^{m,l} = \begin{cases} 0 & if\ \gamma_{n,u,r}^{m,l} \le SINR_{min} \\ \alpha B_r \log_2(1+\gamma_{n,u,r}^{m,l}) & if\ SINR_{min} \le \gamma_{n,u,r}^{m,l} \le SINR_{max} \\ B_r C_{max} & if\ SINR_{max} \le \gamma_{n,u,r}^{m,l} \end{cases} \quad (9)$$

where α is an attenuation factor, $SINR_{min}$ denotes the minimum SINR of the AMC codeset, $SINR_{max}$ is the maximum SINR of the codeset, C_{max} is the maximum achievable capacity and B_r is the bandwidth. A set of parameter values customized to the dual-hop high capacity density system is used where $\alpha = 0.65$, $SINR_{min} = 1.8$ dB, $SINR_{max} = 40$ dB and $C_{max} = 8.6$ bps/Hz[26]. Similarly, the capacity of the access links can be denoted as:

$$C_{k,u,r}^{n,m} = \begin{cases} 0 & if\ \gamma_{k,u,r}^{n,m} \le SINR_{min} \\ \alpha B_r \log_2(1+\gamma_{k,u,r}^{n,m}) & if\ SINR_{min} \le \gamma_{k,u,r}^{n,m} \le SINR_{max} \\ B_r C_{max} & if\ SINR_{max} \le \gamma_{k,u,r}^{n,m} \end{cases} \quad (10)$$

Therefore, the end-to-end link (HBS to MS) capacity is obtained by:

$$C_{m,k} = min(C_{n,u,r}^{m,l}, C_{k,u,r}^{n,m}) \quad (11)$$

III. SINGLE-STATE Q-LEARNING FOR RADIO RESOURCE MANAGEMENT OF SELF-BACKHAUL AND ACCESS NETWORK

As indicated previously in Section II, delivering sufficient capacity density is the key aim of the dual-hop mobile broadband system. This requires a relatively large amount of small access cells (ABSs) to be deployed on streets with sufficient self-backhauling capacity as illustrated in Fig. 3, which in turn demands sophisticated RRM to maximize the spectrum efficiency. Due to the large number of base stations and directional antennas in the system, the conventional fixed frequency planning approach proposed in BuNGee is extremely complex. Moreover, system capacity can be largely constrained by the band size, because the user traffic has significant spatial variation in different directions of the street. The purpose of introducing self-organizing techniques, like SSQL, is to significantly reduce the complexity of the RRM design. With the capability of self-organizing RRM schemes, the requirement for a detailed frequency plan can be partially or even completely removed. It is assumed that all the entities in this work share the same pool of frequency resources. No 'hard' frequency plan is used and the interference control is done entirely by spectrum sensing and the SSQL algorithm. Thus, the requirement for frequency planning is completely removed, which in any case is extremely complicated because of the excessive number of small cells and the unpredictable small-scale propagation effects, and the 'in band' self-backhauling. In this section we will first only briefly introduce Q-learning, then a comprehensive introduction of the proposed SSQL algorithm will be given.

3.1. Single State Q-learning

The standard Q-learning model is represented in Fig. 3[27]. Here the learning process can be defined as a Markov Decision Process (MDP), where an agent perceives a set of distinct environment states $S = \{s_1, \cdots, s_n\}$, and a set of actions $A = \{a_1, \cdots, a_m\}$. At each discrete time *t*, the agent senses the environment $s_t \in S$. Based on s_t, an action $a_t \in A$ is selected and performed. By performing a_t, the environment changes to a new state s_{t+1} and a reward $r_t = r(s_t, a_t)$ is given to the agent.

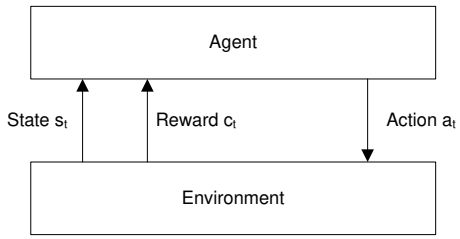

Fig. 3. Standard Q-learning Model

Therefore, a learning model normally consists of the following elements:

1. S: a finite set of environment states
2. A: a finite set of actions
3. r: $S \times A \rightarrow R$: a reward function
4. P: $S \times A \rightarrow \mathrm{P}(S)$: a state transition function

The objective is to find an optimized policy π^* and maximize its accumulated reward. The optimized value function $V^*(s)$ under the optimized policy $\pi^*$is defined as:

$$V^*(s) = \max_a (r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi^*}(s')) \quad (12)$$

where $P(s'|s,a)$ is the state transition probability from s to s' by taking an action a. γ is the learning rate $(0 \leq \gamma < 1)$. $r(s,a)$ is effectively the cumulative reward of state s and $\sum_{s' \in S} P(s'|s,a) V^{\pi^*}(s')$ is the expected feedback of the successor state s'. Then the optimized policy can be derived after computing the optimized value function $V^*(s)$:

$$\pi^*(s) = \operatorname{argmax}_a (r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi^*}(s')) \quad (13)$$

Q learning is used to compute $V^{\pi^*}(s')$ when the cognitive agent has lack of information regarding the state transition probability $P(s'|s,a)$ and the estimated reward $r(s,a)$. In other words, it allows the agent to choose among actions rather than system states based on the Q value. For a policy π, the Q value of an action in a specific system state can be obtained recursively as [27]

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r(s,a) + \gamma \max_a Q(s',a) - Q(s,a)) \quad (14)$$

It is clear from the above functions that in order to apply Q-learning techniques to any systems, the formulation of state-action pairs is important. However, the physical state from the RRM perspective refers to the number of resource blocks that are available to the user, which requires a centralized observation of the network. Furthermore, the state-action pair in Q learning cannot effectively model the physical states. In a queueing system, two state transitions occur on file arrival and departure, which requires only one action of resource allocation. On the other hand, a user may take multiple actions simultaneously when a number of files are in transmission. In order to make the learning system fully distributed and effectively model the network's physical behaviour, we introduce the Single-State Q-Learning where the formulation of state-action pairs is less of an issue.

Single-State Q-Learning is originally proposed to solve stateless games in Computer Science [28]. A reformulation of the standard Q-learning algorithm is carried out so that the Q values of actions are effectively the estimation of the usefulness of the actions in the next step of the learning process. By maintaining a Q value for each action, the agent is able to select the action based entirely on its Q value and the Q value of the selected action will be updated by receiving a reward. The update function is defined as:

$$Q(a) \leftarrow Q(a) + \gamma(r(a) - Q(a)) \quad (15)$$

γ is the learning rate $(0 \leq \gamma < 1)$ and $r(a)$ is the immediate reward of choosing action a.

In our previous work [29], we have analysed the effect of learning rate on the changes of Q values. By considering the

function updated on action *a* with reward *r(a)* recursively by *t* iterations:

$$\begin{aligned} &Q_t(a) = (1-\gamma)^t Q_0(a) \\ &+\gamma r(a)((1-\gamma)^{t-1} + (1-\gamma)^{t-2} + \cdots + (1-\gamma)^0) \\ &= (1-\gamma)^t Q_0(a) + r(a) - r(a)(1-\gamma)^t \end{aligned} \tag{16}$$

$$\lim_{t\to\infty} Q_t(a) = r(a) \tag{17}$$

It can be concluded that the Q value of an action will ultimately stabilize if it receives the same reward value. By computing the time derivative of $Q_t(a)$, we have

$$\frac{d}{dt} Q_t(a) = (1-\gamma)^t (Q_0(a) + r(a)) \ln(1-\gamma) \tag{18}$$

This indicates that the slope of Q value decreases as more iterations t have been taken on action a. Unlike the reinforcement learning algorithm we have developed in [21], the historical information will not dominate the decision making even though the cognitive agent stabilizes on an action for a long time. The reward value can quickly change the converged Q value, which allows the cognitive agent effectively adapt to the dynamic environment.

Comparing with the following traditional update function, it can be seen that not only the state-action formation is not required, but also the information of the successor state s' is irrelevant. This reduces the complexity of the learning model and enhances the applicability of Q-learning to distributed wireless network.

In conventional Q learning, the Q values are usually arbitrarily initialized from the start, and updated iteratively by trial-and-error with a reward process. The Q table in the initial stage has a lack of information about the radio environment, thus an exploration approach is introduced to allow the cognitive agent make decisions from observation of the environment. The ε-Greedy method is commonly used to provide a random decision making with a probability of ε. In this paper, we use this method together with interference measurement to reduce the arbitrary effects of the initial Q values.

Instead of pursuing the optimized policy π^*, the objective of each agent i is to find the action with the highest estimated Q-value Q^*. In this case the reward $r(a)$ needs to be properly defined so that the feedback of taking an action reflects the successfulness of such actions correctly. Particularly in wireless communication systems, the reward is better to be associated with physical measurements of the system in order to facilitate the learning process. In this work, the link capacity is used as reward $r(a)$ to update the Q-values in equation (15). Therefore, if the HBS *m* takes the action *a*to use channel *u* subchannel *r* in order to transmit data to ABS *n*, then the update function of the self-backhaul link can be defined as:

$$Q_{n,u,r}^{m,l}(a) \leftarrow Q_{n,u,r}^{m,l}(a) + \gamma(C_{n,u,r}^{m,l} - Q_{n,q,r}^{m,l}(a)) \tag{19}$$

Similarly the update function of the access link can be defined as:

$$Q_{k,u,r}^{n,m}(a) \leftarrow Q_{k,u,r}^{n,m}(a) + \gamma(C_{k,u,r}^{n,m} - Q_{k,u,r}^{n,m}(a)) \tag{20}$$

where $C_{n,u,r}^{m,l}$, $C_{k,u,r}^{n,m}$ are defined by equation (9), (10).

It has been demonstrated in equation (16) and (17) that, the Q value varies according to the link capacity $C_{n,u,r}^{m,l}$ or $C_{k,u,r}^{n,m}$. In this context, the cognitive agent will stabilize on actions that frequently receive a high link capacity, while also reacting quickly to capacity changes.

3.2. SSQL-based RRM

As explained earlier, one of the most difficult problems for 5G high capacity density networks is the extremely complex RRM task generated by the utilization of a large number of small cells. This is especially true in our case where not only a large number of ABSs are deployed, but also the wireless self-backhauling directional links share the same pool of frequency resources with the access links. Thus, self-organising features are desirable for such networks. This section introduces our SSQL based RRM algorithms.

It is assumed that all the entities in our network, including HBSs, ABSs, and the MSs share the same pool of frequency resources. The cognitive engine is applied to RRM as demonstrated in Fig. 4. The cognitive agents observe the available channels based on interference measurement from distributed spectrum sensing [11], make decisions on channel selection, take actions on file transmission, and update the learning function based on the rewards from the action.

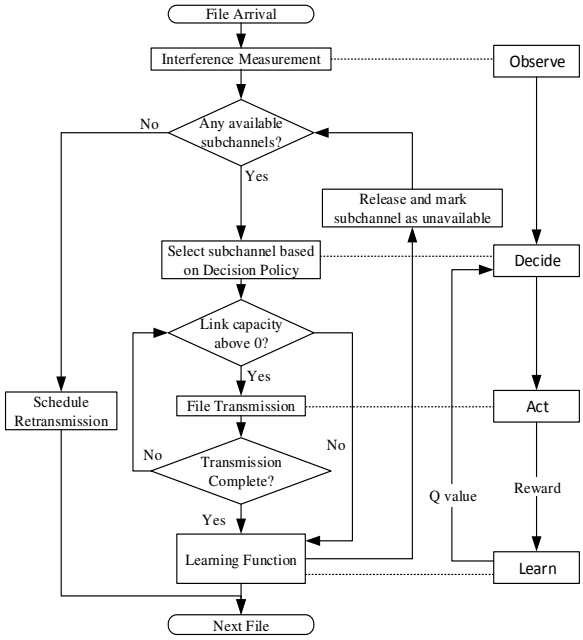

Fig. 4. Cognitive Engine

The SSQL-based RRM algorithm is illustrated in Fig. 4. The ε-greedy exploration algorithm is applied where the learning agent explores a random action with a probability of ε. An interference threshold I_{thre} is applied to obtain available channels from the spectrum pool, which will be specified in the simulation parameters.

The cognitive engine is proposed to be implemented at the ABS only, which is responsible for decision making on both the HBS-ABS backhaul and ABS-MS access links. This significantly simplifies the network architecture, making it widely applicable to different types of mobile devices, whilst also requiring fewer hardware amendments. In the basic cognitive radio approach, spectrum sensing (interference measurement) is operated on an ABS prior to data transmission, where the subchannels with an interference level higher than a predefined threshold are excluded from spectrum assignment. In the SSQL algorithm, the ABSs update the Q-values based on the Q learning function. The subchannels are assigned based on both spectrum sensing and Q-value vector. The SSQL algorithm uses a one-dimensional Q table because it only has a single state. This significantly reduces the computational complexity compared to multi-state Q learning, which should use a multi-dimensional matrix to model different system states and calculate state transition probability. In SSQL, the ABS should build two Q tables separately for access and backhaul subchannels. However, the implementation cost is reduced compared to the conventional approach, as the algorithm is largely simplified.

Algorithm 1.SSQL based RRM algorithm

1. let t = 0, assign random value to $Q(a)$
2. Interference measurement on the channels within allocated band
3. Obtain available action set a, where $\forall I(a) < I_{thre}$
4. Generate random number k between 0 and 1
5. **if** ($k<\varepsilon$)
6. select an random action a_t from a
7. **else**
8. select the action a_t with the maximum Q-value from a
9. **end if**
10. **if** on backhaul link
11. HBS assigns a_t for data transmission
12. **else if** on access link
13. MS verifies a_t and sends a response back to the ABS
14. ABS assigns a_t for data transmission
15. **end if**
16. receive reward $r(a_t)$ (link capacity)
17. update the Q-value vector: $Q(a_t) \leftarrow Q(a_t) + \lambda(r(a_t) - Q(a_t))$
18. t = t + 1

IV. Performance Evaluation and Discussion

This section presents the results obtained from an event driven simulation in MATLAB. The performance of our SSQL-based algorithm is compared with the fixed frequency planning approach described in Section 2.5, and a basic cognitive radio approach where the RRM entities rely only on spectrum sensing measurements at the ABSs with a -80 dBm interference threshold when making subchannel selection decisions [6]. Data traffic is modelled on the downlinks only, with other traffic types likely to have similar performance on uplinks. This is modelled using a file transfer based traffic model, where the file size and the inter-arrival time follow a Pareto distribution, which simulates a succession of packets delivered in the network [30]. The inter-arrival time and file size are modelled as long-tailed Pareto distribution. Any blocked or interrupted files will be back off for a random time and retransmitted until successfully delivered.

The key simulation parameters are listed in Table III. The base station antenna profile, gain and transmit power are defined in [23]. We consider a typical 20 MHz 4G spectrum in the 3.5 GHz licensed band.

TABLE I. SIMULATION PARAMETERS

Parameters		Values
Area		1350*1350m
Building size		75*75m
Street width		15m
Number of HBS/ABS/MS		5/96/2000
Transmit Power of HBS/ABS/MS		37/37/23 dBm
Antenna Gain HBS/ABS		19 dBi – 21 dBi / 17 dBi
Antenna Height HBS/ABS/MS		25/5/1.5 m
Antenna beams		HBS: 20; ABS: 2 (corner: 4)
Carrier Frequency		3.5 GHz
Spectrum Size		20 MHz
Number of Channels		4 (5 MHz each)
Number of Resource Blocks		120
Log-normal shadowing factor		6 dB
Noise floor		-114 dBm/MHz
Interference threshold		-80 dBm
Propagation Model	HBS-ABS	Ray-tracying [6]
	ABS-MS	WINNER II B1 [24]
Traffic Model	Inter-arrival time	Pareto distribution: α=4
	File size	100 kB

4.1. Fixed Frequency Planning

Radio Resource Management in BuNGee is achieved through a fixed frequency planning approach, which will be used as a baseline comparison. The details of the frequency plan are shown in Fig. 5. The entire downlink spectrum band is divided into four 5 MHz channels which are shared between the backhaul and access network. At the HBS side, 4 different channels are used for each group of 4 neighbouring beams in the order from channel 1 to channel 4. ABSs located at the top and bottom of the cell are designed to serve N-S streets, and ABSs on the left and right serve the E-W streets. The two ABS beams pointing in opposite directions should use two different channels. ABSs that serve N-S streets use two different channels from those that serve E-W streets.

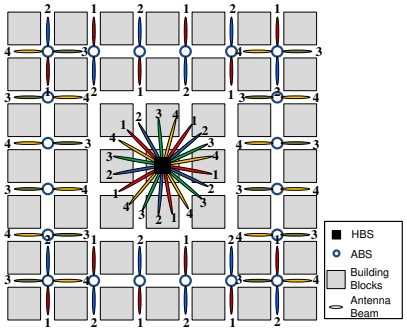

Fig. 5. Downlink Frequency Plan for Square Cells

It can be seen that the potential interference between the beams using the same channel have been minimized by the frequency plan where the immediate neighbouring beams utilize different frequency channels. Thus, by exploiting the natural isolation of building blocks, the novel hierarchical system is expected to deliver a high throughput density. Moreover, the channels allocated to a backhaul beam are different to the connected access beams on an ABS, which reduces the backhaul-access interference. The antenna beamwidth on both HBS and ABS is 30°. We can see from Fig. 3 that on a HBS the angle between neighbouring beams using the same channel is 72°, which is large enough to avoid interference between backhaul links.

4.2. Results and discussion

Fig. 6 shows the average link data rate per Hz in both the access and the backhaul network over up to 5000 simulation trials. The red curves represent the performance of the SSQL-based RRM algorithm and the blue curves represent the basic cognitive radio approach where only spectrum sensing takes place without frequency planning and learning. It can be seen that with SSQL, the data rates of the links in both the access network and the self-backhaul network are significantly higher than the basic cognitive radio approach. Not only does the SSQL scheme enable a faster convergence, but also it achieves a better performance after convergence.

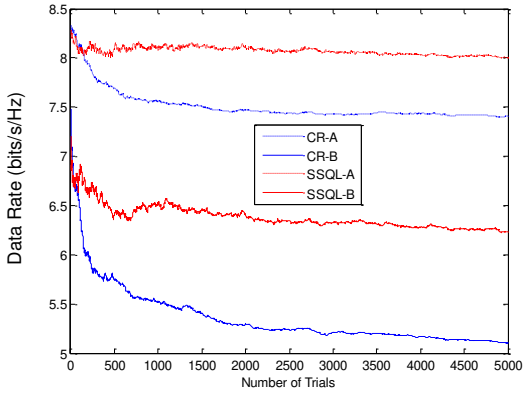

Fig. 6. Link capacity over 5,000 learning trials

The main end-to-end system performance measurements including both the access and backhaul network are system throughput density, delay and the probability of retry. The end-to-end downlink average throughput density can be defined as:

$$Thr_D = Thr_s / A_s \tag{21}$$

Where A_s is the service area. Thr_s is the system throughput that can be defined as:

$$Thr_s = \frac{\sum_{i=1}^{N_u} \sum_{k=1}^{n_i} C(t) \cdot t(k)}{t_s} \tag{22}$$

$C(t)$ is the data rate of a link obtained at time t, and it is updated constantly in the simulation. $t(k)$ is the transmission time of the k^{th} file of an entity, and n_i is the total number of transmissions that have been finished by the i^{th} entity in the simulation. N_u is the total number of entities in the simulation. t_s is the simulation time.

The end-to-end delay can be obtained by:

$$Thr_s = \frac{\sum_{i=1}^{N_u} \sum_{k=1}^{n_i} t_T(k) \cdot t_B(k)}{N_u} \tag{23}$$

Where $t_T(k)$ is the transmission time of the k^{th} file and $t_B(k)$ is the backoff time of the k^{th} file. n_i is the total number of files that have been transmitted by the i^{th} entity in the simulation. N_u is the total number of entities in the simulation.
The end-to-end retry probability is also used in this simulation to describe the probability that the current file transmission request has been rejected by the system. The probability of retry at time t is obtained by:

$$P_{retry} = N_r / N_a(t) \tag{24}$$

Where $P_{retry}(t)$ is the probability of retry at time t. $N_r(t)$ is the total number of rejected file transmissions of the system by time t, and $N_a(t)$ is the total number of file transmission requests (including retries) of the system by time t.

Fig. 7 shows the performance of the aforementioned schemes. It can be seen that a downlink throughput density of 500Mbps/km^2 has been achieved by the SSQL-based scheme where the throughput density of the frequency planning and the basic cognitive radio approach is around 450Mbps/km^2. It is clear that not only is the SSQL-based RRM scheme is able to remove the requirements for frequency planning completely, while also achieving a better overall throughput density throughout because higher data rates have been reached at both the access links and the self-backhaul links. Furthermore, as the FDD system uses an additional frequency band of the same size for the uplinks, the entire system capacity density can reach 1 Gb/s/km^2, which achieves the ultra-dense network requirement in 5G.

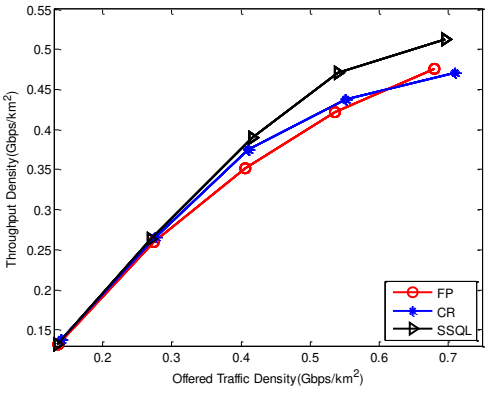

Fig. 7. System downlink throughput density versus downlink traffic density

Fig. 8 and Fig. 9 compare the end-to-end delay and retry probability of the system respectively. The system delay and retry probability of the SSQL-based scheme are significantly lower due to greatly improved link capacity, e.g. the retry probability of the SSQL-based approach is around 5% when the downlink offered traffic density is 0.4 Gbps/km^2. However, the retry probability of the frequency planning approach is about 27% at the same offered traffic level and the figure for the basic cognitive radio approach is around 12%. This result in particular highlights the problem of the frequency planning approach, with 27% of the subchannel selection decisions being initially wrong. This is a result of the reuse behaviour being difficult to predict even with a 'Manhattan Grid' type environment. In less regular high capacity density environments frequency planning will become even more difficult and less effective, or a more conservative reuse factor will be required. However, due to its inherent flexibility, the SSQL scheme should be applicable in various network scenarios for the delivery of self-organized RRM for dynamic spectrum management, even when there are irregular building heights, street layouts ABS placement and user distributions.

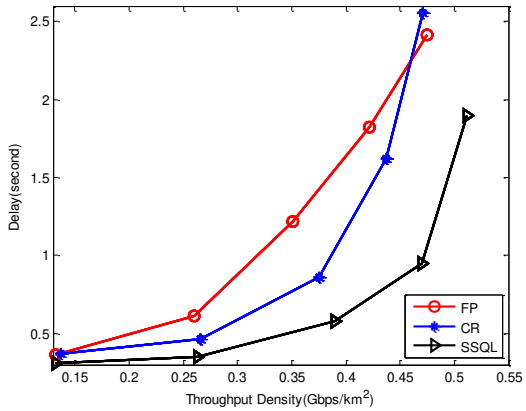

Fig. 8. System delay versus system downlink throughput density

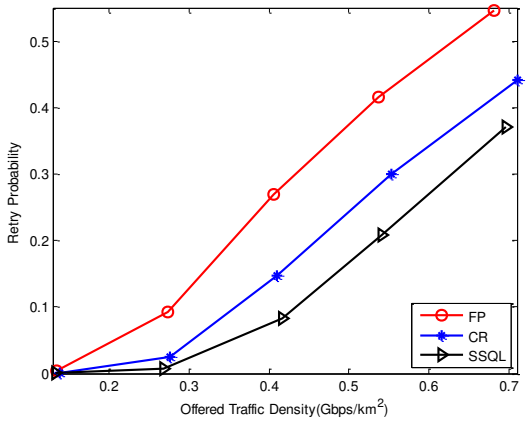

Fig. 9. Probability of retry versus system downlink traffic density

V. CONCLUSIONS

In this paper, a novel SSQL-based RRM algorithm has been developed for dual-hop 5G ultra-dense mobile networks, where the access and self-backhaul network share the same radio spectrum. The SSQL-based RRM algorithm is able to completely remove the need and complexity of frequency planning while achieving a significantly better performance. It is also proven to effectively control the inter-cell and access-backhaul interference in a fully distributed manner, providing a highly efficient way of supporting the self-organization requirements in 5G networks. Compared with traditional Q-learning algorithms, it can be seen that not only is a state-action formation not required, but also the information of the successor state s' is irrelevant. This reduces the complexity and convergence time of the learning model, and enhances the applicability of Q-learning to distributed wireless networks. By introducing the SSQL-based algorithm, the data rates of the links in both the access and self-backhaul networks are significantly higher than with the basic cognitive radio approach. Not only does the SSQL-based scheme enable a faster convergence, it achieves a better data rate performance after convergence. The retry probability of SSQL scheme is 15% lower than the frequency planning scheme, and 5% lower than the basic cognitive radio scheme. This also contributes to significant lower system delay at medium traffic levels. Furthermore, we demonstrate the downlink throughput density of SSQL scheme reaches above 0.5 Gbps/ km^2. Thus, in a FDD system where uplinks use an equal spectrum band, a 1 Gbps/ km^2 capacity density can be achieved for the 5G ultra-dense network deployments.

ACKNOWLEDGEMENT

This research work has been funded by the European Commission's Seventh Framework Programme (FP7) projects under grant agreement No. 248267, Beyond Next Generation Mobile Broadband (BuNGee), and No. 318632, Aerial Base Stations with Opportunistic Links for Unexpected and Temporary Events (ABSOLUTE).

REFERENCES

[1] H. Droste, G. Zimmermann, M. Stamatelatos, N. Lindqvist, O. Bulakci, J. Eichinger, *et al.*, "The METIS 5G Architecture: A Summary of METIS Work on 5G Architectures," in *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*, 2015, pp. 1-5.

[2] D. Guoru, W. Jinlong, W. Qihui, Y. Yu-Dong, L. Rongpeng, Z. Honggang, *et al.*, "On the limits of predictability in real-world radio spectrum state dynamics: from entropy theory to 5G spectrum sharing," *Communications Magazine, IEEE,* vol. 53, pp. 178-183, 2015.

[3] M. Song, C. Xin, Y. Zhao, and X. Cheng, "Dynamic spectrum access: from cognitive radio to network radio," *Wireless Communications, IEEE,* vol. 19, pp. 23-29.

[4] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *Network, IEEE,* vol. 28, pp. 27-33, 2014.

[5] C. Clancy, J. Hecker, E. Stuntebeck, and T. O'Shea, "Applications of Machine Learning to Cognitive Radio Networks," *Wireless Communications, IEEE,* vol. 14, pp. 47-52, 2007.

[6] ETSI Technical Report 101 534 v1.1.1: Broadband Radio Access Networks (BRAN); Very high capacity density BWA

networks; System architecture, economic model and derivation of technical requirements.
[7] N. Saquib, E. Hossain, and K. Dong In, "Fractional frequency reuse for interference management in LTE-advanced hetnets," *Wireless Communications, IEEE,* vol. 20, pp. 113-122.
[8] F. Guidolin, A. Orsino, L. Badia, and M. Zorzi, "Statistical analysis of non orthogonal spectrum sharing and scheduling strategies in next generation mobile networks," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, 2013, pp. 680-685.
[9] A. Orsino, G. Araniti, A. Molinaro, and A. Iera, "Effective RAT Selection Approach for 5G Dense Wireless Networks," in *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*, 2015, pp. 1-5.
[10] L. Militano, A. Orsino, G. Araniti, A. Molinaro, and A. Iera, "Overlapping coalitions for D2D-supported data uploading in LTE-A systems," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2015, pp. 1716-1720.
[11] S. Maharjan, Z. Yan, Y. Chau, and S. Gjessing, "Distributed Spectrum Sensing in Cognitive Radio Networks with Fairness Consideration: Efficiency of Correlated Equilibrium," in *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*, 2011, pp. 540-549.
[12] N. ul Hassan, S. Hussain, Y. Chau, and D. Lingjie, "Tradeoff between spectrum cost and quality of service in a cognitive radio network," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, 2013, pp. 1179-1184.
[13] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *Vehicular Technology, IEEE Transactions on,* vol. 48, pp. 1676-1687, 1999.
[14] S.-M. Senouci and G. Pujolle, "Dynamic channel assignment in cellular networks: a reinforcement learning solution," presented at the International Conference on Telecommunications, Feb. 2003.
[15] A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *Vehicular Technology, IEEE Transactions on,* vol. 59, pp. 1823-1834, 2010.
[16] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, 2009, pp. 1893-1898.
[17] C. Wu, K. Chowdhury, M. D. Felice, and W. Meleis, "Spectrum management of cognitive radio using multi-agent reinforcement learning," presented at the Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Industry track, Toronto, Canada, 2010.
[18] K. L. A. Yau, P. Komisarczuk, and P. D. Teal, "Applications of Reinforcement Learning to Cognitive Radio Networks," in *Communications Workshops (ICC), 2010 IEEE International Conference on*, 2010, pp. 1-6.
[19] K. L. A. Yau, P. Komisarczuk, and P. D. Teal, "Enhancing network performance in Distributed Cognitive Radio Networks using single-agent and multi-agent Reinforcement Learning," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, 2010, pp. 152-159.
[20] L. Kleinrock, "Queueing Systems - Volume I: Theory," *John Wiley & Sons,* 1975.
[21] T. Jiang, D. Grace, and Y. Liu, "Two-stage reinforcement-learning-based cognitive radio with exploration control," *Communications, IET,* vol. 5, pp. 644-651.
[22] Q. Zhao and D. Grace, "Application of Cognition based Resource Allocation Strategies on a Multi-hop Backhaul Network," in *IEEE International Conference on Communication Systems*, 2012.
[23] "BuNGee Deliverable: D1.2, Baseline BuNGee Architecture," January 2010.
[24] P. Kyösti, J. Meinila, L. Hentila, X. Zhao, and T. Jamsa, "WINNER II Channel Models v1.2," IST-WINNER D1.1.22007.
[25] C. A. Balanis, *Antenna Theory, Analysis and Design*, 2nd ed. New York: Wiley, 1997.
[26] A. Papadogiannis and A. G. Burr, "Multi-beam assisted MIMO - A novel approach to fixed beamforming," in *Future Network & Mobile Summit (FutureNetw), 2011*, pp. 1-8.
[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Mass., London: MIT Press, 1998.
[28] S. Kapetanakis and D. Kudenko, "Reinforcement learning of coordination in cooperative multi-agent systems," presented at the Eighteenth national conference on Artificial intelligence, Edmonton, Alberta, Canada, 2002.
[29] Q. Zhao, D. Grace, and T. Clarke, "Transfer learning and cooperation management: balancing the quality of service and information exchange overhead in cognitive radio networks," *Transactions on Emerging Telecommunications Technologies,* vol. 26, pp. 290-301, 2015.
[30] J. D. Chimeh, M. Hakkak, and S. A. Alavian, "Internet Traffic and Capacity Evaluation in UMTS Downlink," in *Future Generation Communication and Networking*, 2007, pp. 547-552.