A Reinforcement Learning Framework for Autonomous Cell Activation and Customized Energy-Efficient Resource Allocation in C-RANs

Guolin Sun[1*], Gordon Owusu Boateng[1], Hu Huang[1] and Wei Jiang[2,3]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu-China
[2] German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany
[3] Department of Electrical and Information Technology (EIT), Technische University (TU) Kaiserslautern, Germany
[e-mail: guolin.sun@uestc.edu.cn]
*Corresponding author: Guolin Sun

Received December 12, 2018; revised February 18, 2019; accepted March 13, 2019; published August 31, 2019

Abstract

Cloud radio access networks (C-RANs) have been regarded in recent times as a promising concept in future 5G technologies where all DSP processors are moved into a central base band unit (BBU) pool in the cloud, and distributed remote radio heads (RRHs) compress and forward received radio signals from mobile users to the BBUs through radio links. In such dynamic environment, automatic decision-making approaches, such as artificial intelligence based deep reinforcement learning (DRL), become imperative in designing new solutions. In this paper, we propose a generic framework of autonomous cell activation and customized physical resource allocation schemes for energy consumption and QoS optimization in wireless networks. We formulate the problem as fractional power control with bandwidth adaptation and full power control and bandwidth allocation models and set up a Q-learning model to satisfy the QoS requirements of users and to achieve low energy consumption with the minimum number of active RRHs under varying traffic demand and network densities. Extensive simulations are conducted to show the effectiveness of our proposed solution compared to existing schemes.

Keywords: reinforcement learning; autonomous cell activation; resource allocation; cloud radio access network

http://doi.org/10.3837/tiis.2019.08.001 **ISSN : 1976-7277**

1. Introduction

Research on the fifth-generation (5G) mobile cellular communication technology indicates that the traffic density in crowded cities or hotspot areas will reach 20~Tbps/Km2 in the near future. It is expected that by 2020, mobile internet will need to be delivering 1GB of personalized data per user per day. Furthermore, traffic by 2030 is predicted to be up to 10,000 times greater than in 2010 and 100 Mbps end-user services will have to be supported [1]. To be able to support such demand, future mobile cellular networks are expected to be deployed in a very dense and multi-layered way. Ultra-dense small cell network (UDN) is considered as one of the most promising methods to meet the traffic volume requirement of 5G. The realization of this is simply done by the dense deployment of small cells in the hotspots, where immense traffic is generated [1]. However, this triggers a proportional consumption on energy. From the perspective of network operators, the increasing energy costs cannot sustain future network operations. From the environmental point of view, “greenness” can be more meaningful with a comprehensive evaluation that includes both energy savings and network performance, which is the basis for energy efficiency (EE) metrics.

Cloud radio access networks(C-RANs) have been proposed and regarded as a promising concept in the information and communications technology (ICT) area, where base-band units (BBUs) and radios are separated [2]. All DSP processors are moved into a central BBU pool in the cloud, and the distributed remote radio heads (RRHs) take the responsibility of compressing and forwarding the received radio signals from mobile users to BBUs through radio links. This will reduce the overall capital cost and operational cost, and make large-scale high-density network deployments possible. Especially, this centralized architecture makes it easy to collect and analyze statistics data of runtime system, as it motivates us to seek autonomous schemes for network energy management.

Recently, reinforcement learning (RL) has been advocated as a viable technology to enhance resource utilization. RL is a form of machine learning technique where a learning agent does not have a prior knowledge of the environment. To obtain low energy consumption on the RRHs and satisfy the QoS requirements of users under varying traffic demand and network densities, RL techniques are best to switch the RRHs on or off at defined time steps. While traditional solutions to optimizing networks such as greedy linear programming and greedy search satisfy instantaneous requirements of the system, RL agents survey the entire network taking into account every possible state [3]. For dynamic systems where conditions change periodically, the agent selects the most appropriate policy for allocating resource in real time. In the context of C-RAN architecture, the agent can be trained through each learning stage and then updates the trained data to determine the state of each RRH in each decision epoch to implement continuous control. This paper develops a framework for energy-efficiency where RL techniques are used to determine the power consumption states (sleep and active) for each RRH. The idea is to develop an autonomous cell activation scheme and a customized physical resource allocation scheme to achieve optimal network structure to reduce power consumption. The proposed framework can be realized in two steps: Firstly, we identify the active and inactive RRHs using a Q-learning based algorithm. Secondly, we set up a flexible resource allocation module based on the active RRH set by optimizing power and bandwidth allocation and control.

In other related works [4], [5], where power consumption is optimized over current timeslot or time frame, we present a cell activation RL-based framework which makes a

sequence of resource allocation decisions to minimize total energy consumption for the whole operational period. To efficiently solve this problem, we first use a Q-learning method to solve the cell activation problem and formulate the resource allocation problem for users as a convex optimization problem. Our motivation is to achieve the balance between EE and QoS to satisfy infrastructure providers (InPs) and mobile users via flexible power and bandwidth control or allocation, decoupled from cell activation techniques in a dense C-RAN system. In this paper, our main contributions can be summarized as:

- We propose an autonomous energy management framework using cell activation techniques for the customized network. We design a Q-learning model with a reduced size of state space set considering varying resource demand and user population.
- In this framework, we formulate the EE-QoS optimization as two models, fractional power control with bandwidth adaption and full power and bandwidth allocation.
- Considering physical resource allocation for a customized network, we optimize power and bandwidth jointly. We formulate the problem as a convex optimization with the aim of satisfying QoS requirements of user equipment (UEs) with the minimum number of active RRHs.

The remainder of this paper is organized as follows. In Section II, we present related works. Section III presents the system model in terms of network model, traffic model, energy model and utility model. Section IV provides the problem formulation and our proposed RL-based autonomous energy management framework. Simulation results and analysis are discussed in Section V. We conclude this work in section VI.

2. Related Work

The EE and QoS performance metric has become a design goal as the discussion on energy consumption continues to grow across every field. It has become a requirement for network engineers and scientists to develop systems that manage energy efficiently. Authors in [6] studied energy efficient wireless communications and identified energy-efficiency resource allocation as one of the key challenges of 5G. In C-RAN, baseband and processing functionality of a network are virtualized and shared among physical units. This architecture improves energy efficiency in the sense that the RRHs have less functions. In [7], authors considered RRH selection and power minimization jointly as the resource allocation problem in group sparse beamforming for green Cloud-RAN. Authors extended their work in [8] to reduce the computational complexity in selecting RRH using lagrangian dual methods. In [9], the effect of optimizing data-sharing and the compression on energy efficiency were studied in C-RAN. By minimizing the total power consumption in the network, they proved that a higher energy efficiency depends on the user target rate.

Intuitively, high density of active small cell base stations (sc-BSs) results in severe interference and also inefficient energy consumption. The inter-cell or inter-tier interference mitigation is the key to improve EE performances. Therefore, *Luo et al.* in [4] proposed a joint downlink and uplink mobile users access point (MU-AP) association and beamforming design for interference management and energy minimization in C-RAN. Authors in [10] also proposed an enhanced soft fractional frequency reuse scheme. In this scheme, they formulated a joint optimization problem with the resource block assignment and power allocation for interference mitigation in order to maximize EE performances in heterogeneous C-RAN. The joint rate allocation, routing, scheduling, power control and channel assignment problem was investigated in [5] with the aim of maximizing throughput

and achieving fairness of users. Joint optimization by cell activation or cell coverage adjustment, user association, and sub-carrier allocation has been investigated in [11] [12]. This was done under the constraints of maintaining an average sum rate and rate fairness. The authors argued that energy consumption was dependent on both the spatio-temporal variations of traffic demands and the internal hardware components of sc-BS.

Several studies in recent times also suggested a scheme, known as multiple base station scheduling (MBSS) [13]. Due to the computational complexity in MBSS, authors in [14] proposed a low complexity flexible flow scheduling algorithm to compensate for the energy consumption caused by increasing dimension of ultra-dense nodes. Trade-offs between QoS and EE for users with different traffics was presented in [15]. In [16], the authors studied the user association problem aiming at maximizing the network energy efficiency for the downlink of heterogeneous networks (HetNets). The goal of minimizing the system energy consumption and also maximizing the ratio of the peak-signal-to-noise-ratio was considered in [17] but only for QoE-aware energy efficiency (QEE) and QoE-aware spectral efficiency (QSE).

Reinforcement learning can be widely utilized in many applications with different optimization objectives, such as resource allocation in data centers, residential smart grid, embedded system power management and autonomous control [18]. The work in [18] developed a framework for solving the overall resource allocation and power management problem in cloud computing systems using deep reinforcement learning. *Shams et al* [19] proposed a Q-learning-based algorithm to achieve both energy efficiency and overall data rate. *Xu et al* proposed a framework which uses reinforcement learning to achieve optimal solution for power-efficient resource allocation for beamforming problem [20].

To the best of our knowledge, there are lack of solutions to maximize the EE performance in C-RANs where power and bandwidth are optimized jointly. The authors in [21] investigated energy-efficient power allocation and wireless backhaul bandwidth allocation in heterogeneous small cells. They formulated the problem as a non-convex non-linear programming problem and decomposed it into two sub-problems. Then, they proposed a sub-optimal low-complexity algorithm to solve the bandwidth allocation problem and a near optimal iterative resource allocation algorithm. However, these algorithms are still model-based method, which cannot autonomously produce optimal solutions in sequential time steps. In this paper, based on the traffic load prediction results and the current information, the power manager adopts the model-free RL technique to adaptively determine the suitable action for turning on/off of the RRHs and simultaneously reduce the power/energy consumption and improve QoS satisfaction.

3. System Model

The proposed autonomous cell activation and customized physical resource allocation schemes for energy consumption and QoS optimization framework is made up of RRHs, BBUs and UEs. The UEs are connected to the RRHs based on the execution of cell activation by the BBU pool. In this section, we present the network model, traffic model, energy model, and utility model.

3.1 Network model

The network model in the paper is based on the C-RAN architecture. In C-RAN, the BBUs are combined into a single resource pool, i.e BBU pool and shared among the RRHs [22]. All functions of the RAN are partially or completely integrated into the BBU pool in

the cloud. RRHs at different locations can access the functions from the virtual BBU pool. Let $\mathcal{J} = \{1,2, \ldots, J\}$ be a set of infrastructure nodes called RRHs. For each node $j \in \mathcal{J}$, a set of UEs are connected to them. We represent a set of UEs $\mathfrak{T} = \{1,2, \ldots, I\}$ as the mobile UEs connected to the RRHs. At each time interval, it is assumed that user $i \in \mathfrak{T}$ is connected to a RRH j. The spectrum bandwidth of RRH j is W_j Hz and the maximum transmit power of RRH j is P_j^{trans} watts. We denote the fraction of resource allocated to UE i from RRH j as $x_{ij} \in [0,1]$, where $x_{ij} = 0$ means that UE i is not associated with RRH j and $x_{ij} \neq 0$ means that UE i from RRH j is allocated a bandwidth proportion of x_{ij}.

In our system model, the path-loss is calculated as follows:

$$PathLoss = 20 * log(F) + 20 * log(D) + 32.4, \quad (1)$$

where F is the frequency band and D is the distance between a UE and a RRH. The shadowing small-scale fading is assumed as a Gaussian random variable with zero mean and standard deviation δ equal to 8dB [23].

For the resource allocation, the signal-to-interference-plus-noise-ratio (SINR) experienced by each UE i associated with a RRH j is modeled as;

$$\chi_{ij} = g_{ij} P_{ij}^{trans} / (\sum_{k,k \neq j} g_{ij} P_{ik}^{trans} + \sigma), \quad (2)$$

where g_{ij} is large-scale channel gain resulting from propagation loss and shadowing effects. σ is the power spectrum density of additive white Gaussian noise. P_{ij}^{trans} is the received signal power for UE i from RRH j. Next, we use the Shannon capacity formula to calculate the spectrum efficiency of UE i with RRH j as:

$$b_{ij} = log_2(1 + \chi_{ij}), \quad (3)$$

where χ_{ij} is the SINR of UE i from RRH j.

With the fraction of the bandwidth resource allocated to UE i from RRH j being x_{ij} and its transmission rate denoted by $\mathfrak{R}_{ij}$, we have $\mathfrak{R}_{ij} = W_j x_{ij} b_{ij}$. Based on equation (3), the transmission rate of UE i obtained from RRH j can be written as:

$$\mathfrak{R}_{ij} = W_j x_{ij} \log_2(1 + \chi_{ij}), \quad (4)$$

Since it is possible for a UE to associate with any RRH, the effective transmission rate $\mathfrak{R}_{ij}$ of UE i can be written as follow:

$$\mathfrak{R}_i = \sum_{j \epsilon J} |x_{ij}|_0 \, \mathfrak{R}_{ij}, \quad (5)$$

where $|x_{ij}|_0$ denotes an association indicator between UE i and RRH j. If $|x_{ij}|_0 = 0$, there is no association between the UE and RRH; otherwise, $|x_{ij}|_0 = 1$.

3.2. Traffic model

In our scenario, we monitor the spatial-time traffic distribution in the network over a 24-hour period. The number of active users and the traffic demands vary over this period. Using this model greatly increases the complexity of the traffic mode on the network. A traffic profile is based on the on-site measurements from the EU FP EARTH project [14]. An ideal traffic profile is configured based on the trapezoidal traffic pattern, which is a simple example of daily traffic pattern [14]. For a trapezoidal curve with a maximum value of one and different slopes, the traffic function is defined by the angular coefficient v.

$$f(t) = \begin{cases} 1 - vt; \left(0 \le t \le \frac{1}{v}\right) \\ 0; \\ 1 + v(t - T); \left(T - \frac{1}{v} \le t \le T\right) \end{cases} \quad (6)$$

where T represents a 24-hour scan period, v represents the slope and $f(t)$ is a normalized value between 0 and 1 as shown in Figure 2. If v is equal to 1/10, then we move the $f(t)$ to $f(t) + 12$, which is close to the real scenario. If the slope v is equal to 1/8, the traffic profile will be changed. Since the traffic changing trend is close to real situation when v is equal to 1/10, we will use the traffic function in that v as equal to 1/10.

3.3. Energy model

We define two power consumption states, sleep and active for each RRH. The active state combines the sum of power consumption of the transmit power and RRH power. Power consumption of the RRH at the sleep state is negligible. Therefore, we define the total power model for each RRH as follows:

$$P_j^{total} = \begin{cases} P_j^{active} + P_j^{trans} & j \epsilon j_A \\ P_j^{sleep} & j \epsilon j_S \end{cases}, \quad (7)$$

where P_j^{active} denotes the essential power consumption of the RRH j in the active state, which is necessary in order to maintain the basic operation of the RRH and P_j^{trans} is the used transmit power of RRH which is to ensure data transmission of user equipment (UE). If the RRH is not selected for transmission, it enters sleep mode. $\mathcal{J}_A \subseteq \mathcal{J}$ and $\mathcal{J}_S \subseteq \mathcal{J}$ denote the sets of active and sleep RRHs respectively. The total number of RRHs in the network is the sum of active set of RRHs and the sleep set of RRHs, i.e. $\mathcal{J}_A \cup \mathcal{J}_S = \mathcal{J}$.

Given a time $t = \{1,2,3 \dots, T\}$, a set of active RRHs $\mathcal{J}_A$ and a set of sleep RRHs $\mathcal{J}_S$, the total energy consumption of RRHs in the entire period can be expressed as:

$$E = \sum_{t=1}^{T} \left(\sum_{j \epsilon \mathcal{J}_A} P_j^{active} + \sum_{j \epsilon \mathcal{J}_A} P_j^{trans} + \sum_{j \epsilon \mathcal{J}_S} P_j^{sleep} \right) \quad (8)$$

In the C-RAN architecture, inactive RRHs are put to sleep in order to conserve energy. Our proposed Q-learning based cell activation scheme provides flexibility for managing energy consumption. The C-RAN control unit dynamically optimizes the total expected and cumulative energy during the entire operational period instead of the instantaneous energy consumption in a decision period.

3.4. Utility model

Based on the objective of the proposed scheme, we can know the precondition of saving energy consumption of network is to ensure that we satisfy the QoS requirement of UEs. In order to offer better QoS to UEs, the required transmission rate should be guaranteed. We consider measuring the satisfaction of a UE i with a sigmoid function, which can be expressed in [24] as:

$$\xi(\Re_i) = \frac{1}{1 + e^{-\tau(\Re_i - \Re_i^{min})}}, \quad (9)$$

where $\Re_i^{min}$ is the minimum rate demand required by the UE i and τ is a constant deciding the steepness of the satisfactory curve. In addition, $\Re_i$ is the real transmission rate for UE i, which is determined by the network infrastructure, transmission power, noise, interference and many other related factors. It is easy to verify that: 1) $\xi\left(\Re_i^{min}\right)$ is a monotonic increasing

function with respect to $\Re_i$, because individual UEs will feel more satisfied if they receive higher throughput above their minimum demand and vice versa; 2) $\xi\left(\Re_i^{min}\right)$ of each UE i is scaled between 0 and 1, i.e. $\xi\left(\Re_i^{min}\right) \in (0, 1)$.

The analysis shows it results in a trivial solution using linear utility function (5) for throughput maximization as in [25], in which each RRH serves only its strongest user. While throughput is optimal, this is not a satisfactory solution for many reasons. Instead, we seek a utility function which achieves load balancing and user fairness naturally. A logarithmic function, in particular, is a very common choice of utility function. The resulting objective function with logarithmic utility is defined as;

$$U_i(\Re_i) = \log(\Re_i), \tag{10}$$

4. Problem Formulation

The framework of our proposed energy consumption optimization system has three hierarchies as shown in **Fig. 1**. Firstly, user association between UEs and RRHs are established through user admission control. Then, the RL agent executes cell activation using the Q-learning technique to select the active RRH set. Resource allocation module uses the active RRH set to execute radio resource allocation based on the needed active RRH set for satisfying the QoS requirement of UEs. The result of radio resource allocation serves as the reward that is fed back into the Q-Learning-based cell activation module. The RL agent dynamically monitors the change in user population, distribution, QoS demand and resource utilization of UEs caused by the dynamics in the UEs' number and their location. Once learning is completed, the agent executes cell activation autonomously as the action of the Q-learning algorithm for minimizing energy consumption. The resource allocation module also performs energy management and QoS satisfaction based on the set of active RRHs obtained from cell activation. The admission control and association result in association between UEs and RRHs. We demonstrate autonomous cell activation and customized resource allocation modules of the system framework one by one.

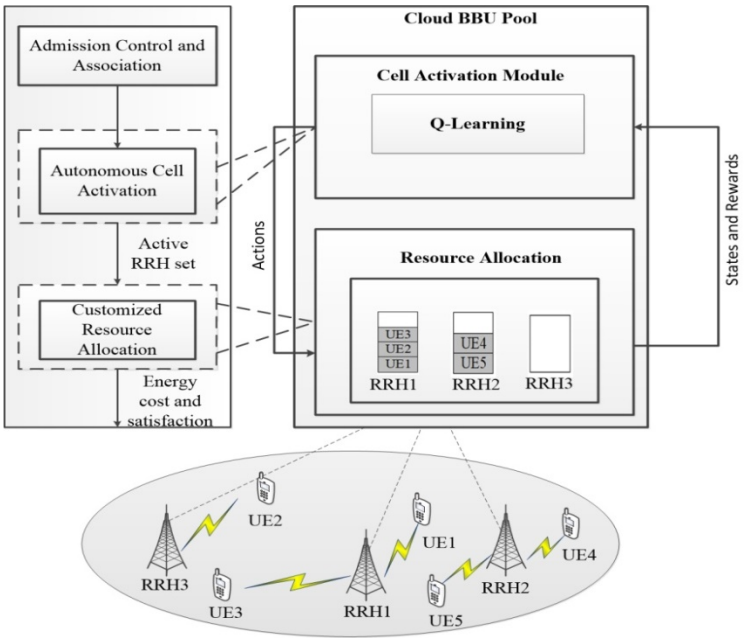

Fig. 1. System framework

4.1. Autonomous cell activation

In this section, we present the Q-learning-based autonomous cell activation framework, which minimizes the number of active RRHs to achieve low energy consumption while ensuring that the demand of each UE can be satisfied by the set of active RRHs. Unlike most of the previous works that presented algorithms optimizing a certain objective (such as power consumption) for the current timeslot (or time frame), our proposed Q-learning-based framework makes a sequence of cell activation decisions to minimize total energy consumption while satisfying QoS demand of UEs for the whole operational period. In our framework, the RL agent can turn off some RRHs in order to minimize energy consumption if the available UEs can be satisfied by few number of RRHs. It can also turn on some RRHs if the current active RRHs cannot satisfy the requirements of some UEs. These on-off switching decisions are made by the RL learning agent deployed in the BBU cloud.

Q-learning: Reinforcement learning technique is a form of machine learning which does not need much labeled data to make decisions. There are a number of reinforcement learning technique variations such as Q-learning, deep Q-learning and double Q-learning. One of the most well-known and generally applicable implementations of reinforcement learning is Q-learning [26]. Q-learning is a model-free reinforcement learning algorithm whereby an agent interacts solely with an environment, without requiring additional information about the environment except for awareness of the environment states, possible (enabled) actions from its current state, and the obtained rewards after performing an action.

In Q-learning, we define a matrix-like Q-table, which has the form $Q : S \times A \rightarrow R$ where S is the set of possible states in the environment, A is the set of actions that are possible for those states and R is the reward obtained after performing the action. The Q-table $Q(s, a)$ with $s \in S$ and $a \in A$ maps state-action pairs to the maximum discounted future reward R' when performing action a from state s. The Q-value which can be looked up in the Q-table can be expressed as follows:

$$Q(s', a') = \max R' \tag{11}$$

where a' is the action of next state s' and R' is the discounted future reward.

The letter Q is derived from the word "quality," as the Q-function represents the quality score for performing an action in a certain state. The ideal policy π_{ideal} for an agent to follow to maximize the future (discounted) reward from state s is to always choose the action with the highest Q-value as follows:

$$\pi_{ideal}(s) = arg\, max_{a \epsilon A_s} Q(s, a) \tag{12}$$

where A_s is the set of actions that are enabled in state s.

The idea of Q-learning is that, we iteratively approximate the Q-table. Consider a single transition performed by a reinforcement learning agent: (s, a, r, s') where s denotes the previous state of the agent, a is the chosen action by the agent when being in state s, r is the obtained reward for performing action a in state s, and s' represents the resulting state the environment is in after the agent performed action a. We can express the Q-value of state-action pair (s, a) in terms of the next state s' using the Bellman equation as:

$$Q(s, a) = r + \gamma \max_{a' \epsilon A_{s'}} Q(s', a') \tag{13}$$

We begin to formulate the Q-learning-based cell activation problem of our wireless network scenario by defining network states, actions and reward in the context of the generic semi-markov decision process (SMDP) framework [27].

State(*s*): As mentioned above, the purpose of Q-learning is to minimize the number of active RRHs while satisfying the QoS requirements of the UEs. Based on this, the state space needs to reflect the on-off states of RRHs and their bandwidth occupancy. Therefore, we define the state space of the agent to include the on-off state of RRHs and the proportion of bandwidth resources occupied by the current RRHs. Each RRH has two states, the active state and the sleep state. The two states for any RRH j can be expressed as $M_j \in \{0, 1\}$. $M_j = 0$ indicates that RRH j is in the sleep state while $M_j = 1$ indicates that RRH j is in the active state. We use θ to represent the proportion of total system bandwidth resources occupied by all of the UEs on all of the RRHs. Since θ is a continuous value, it leads to infinite states. Therefore, in order to reduce the size of state spaces in Q-table, the range of θ from zero to one is partitioned into eight non-overlapped subzones averagely uniformly. In summary, the state space for wireless network scenario can be expressed as s= $[M_1, M_2..M_J, \theta_1, \theta_2..\theta_J]$ with $o(2^J)$ discrete combinational states.

Action(*a*): In this paper, the goal of Q-learning-based cell activation is to minimize the number of active RRHs by switching off some RRHs when a few number of RRHs can satisfy the demand of the UEs. The action to be performed is the switching decision that is made by the agent on the RRHs. That is, the agent makes a corresponding switching action of RRHs according to the current state. Each RRH corresponds to two actions, switching on or off. For any RRH j, the two actions can be represented as $N_j \in \{0, 1\}$, $N_j = 0$ indicates switching off RRH j to turn it to sleep, and $N_j = 1$ indicates switching on RRH j to turn it to active. The action space can be expressed as a = $[N_1, N_2, .. N_J]$.

Reward(*r*): Reward is the feedback received from the environment after performing an action in a certain state. Therefore, the reward needs to reflect the purpose of the Q-learning algorithm; in our case, the satisfaction of the user's service quality and the energy consumption minimization of the wireless network. Since the optimal strategy of Q-learning is to find the action with the largest value in the Q-table for each state, satisfaction of user QoS and energy consumption minimization of the wireless network after the action is performed gives the agent the largest Q-value. We define the reward as follows:

$$R(s, a) = \frac{1}{E} + \omega * \xi(\cdot) \tag{14}$$

where $\xi(\cdot)\epsilon[0,1]$ is an indicator to show the satisfaction of UEs, with the utility function defined in (9) and $\omega > 0$. After obtaining the set of active RRHs through Q-learning-based cell activation, we focus on how to allocate power and bandwidth to the UEs through customized resource allocation module in the next sub-section.

4.2. Customized resource allocation

In the resource allocation module, we set up different objective functions for different resource allocation schemes. In this paper, we propose a scheme where the QoS requirements of the UEs are satisfied with a limited amount of resource available. It is assumed that, the proportion of bandwidth occupied by UEs is equal to the transmit power consumption. On the other hand, a scheme to maximize the throughput of UEs by fully utilizing the available resource is considered.

4.2.1. Fractional power control with bandwidth allocation

In this resource allocation scheme, we assume that the value of transmit power of unit bandwidth is the equal for all RRHs. The objective of this scheme is to minimize the usage of power and bandwidth while ensuring QoS satisfaction of UEs. Based on this, we define the objective function as the sum of the occupied resource in the whole system B as follows:

$$min\ B = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij} \tag{15}$$

such that;

$$x_{ij} \epsilon [0,1], \quad \forall\, i\epsilon\{1,2,\dots,I\}, \forall\, j\epsilon\{1,2,\dots,J\} \tag{16}$$

$$\sum_{j=1}^{J} |x_{ij}|_0 = 1, \quad \forall\, i\epsilon\{1,2,\dots,I\} \tag{17}$$

$$\sum_{i=1}^{I} x_{ij} \leq 1\ , \ \forall\, j\epsilon\{1,2,\dots,J\} \tag{18}$$

$$\sum_{j=1}^{J} x_{ij} * w * \log_2 (1 + \chi_{ij}) \geq d_i, \quad \forall i\epsilon\{1,2,\dots,I\} \tag{19}$$

where x_{ij} is the proportion of bandwidth occupied by UE i from RRH j and d_i is the demand of UE i . We assume all UEs have equal demand. Constraint (16) states that, the fraction of resource allocated to UEs ranges from 0 to 1.Constraint (17) states that a UE can only associate with one RRH simultaneously. This is because; we assume each UE has only one interface at a time. $|x_{ij}|_0$ denotes an association indicator between UE i and RRH j. If $|x_{ij}|_0 = 0$, there is no association between the UE and RRH; otherwise, $|x_{ij}|_0 = 1$. Constraint (18) means the proportion of bandwidth occupied by the UEs should not exceed one. This is because all occupied bandwidth of UEs on each RRH should not be more than the total bandwidth of the associated RRH. Constraint (19) indicates that bandwidth resource occupied by UE should be greater than its QoS requirement. This optimization problem is a mixed non-linear integer programming problem and can be solved efficiently using existing MATLAB solver YALMIP [28]. In this solver, the mixed integer linear programming (MILP) is used and appropriate for solving this resource allocation optimization problem.

4.2.2. Full power and bandwidth allocation

In this resource allocation scheme, our objective is to maximize the throughput of the network system. Instead of fractional power control with bandwidth adaptation, the network would allocate as much power and bandwidth as possible to UEs to maximize throughput while observing user fairness. We define the objective function of this scheme as follows:

$$max \quad T = \sum_{i=1}^{I} U(\Re_i) \tag{20}$$

$$\sum_{i=1}^{I} P_{ij}^{trans} = P_j^{trans}, \quad \forall\, j\epsilon\{1,2,\dots,J\} \tag{21}$$

$$\sum_{i=1}^{I} x_{ij} = 1, \quad \forall\, j\epsilon\{1,2,\dots,J\} \tag{22}$$

$$\sum_{j=1}^{J} \Re_{ij} \geq d_i, \quad \forall\, i\epsilon\{1,2,\dots,J\} \tag{23}$$

$$\sum_{j=1}^{J} |x_{ij}|_0 = 1, \quad \forall\, i\epsilon\{1,2,\dots,I\} \tag{24}$$

$$x_{ij} \epsilon [0,1], \forall i\epsilon\{1,2,\dots,I\}, \forall\, j\epsilon\{1,2,\dots,J\} \tag{25}$$

where P_{ij}^{trans} is the transmit power allocated to UE i from RRH j, P_j^{trans} is the maximum RRH power, x_{ij} indicates the bandwidth resource proportion allocated to UE i from RRH j, and d_i is the rate demand of UE i. Constraint (21) indicates that RRH j associated with UE i should run out of resource to UEs for maximizing throughput. Constraint (22) means the sum of bandwidth resource allocated to all UEs associated with RRH j should be equal to one. This is because in maximizing throughput, a lot of bandwidth must be used. From constraint (23), we can state that the achieved throughput of UEs should be greater than their minimum QoS demands. Constraint (24) means that one UE can only associate with one RRH

simultaneously. Constraint (25) states that, the fraction of resource allocated to users ranges from 0 to 1. This is a convex optimization problem and can be efficiently solved using existing MATLAB solver CVX [29].

The proposed algorithm framework is summarized in detail as follows; In step 1, in line 1-2, the association between UEs and RRHs take place before UEs request resource. In step 2 from line 3-9, the Q-table of the Q-learning algorithm is initialized and iterated for each decision epoch as the demand of UEs changes. Actions are selected randomly initially as learning is ongoing. After some time, actions are selected based on the maximum Q-value to obtain the set of active RRHs. In step 3, in line 10-12, we obtain an optimal energy consumption based on the set of active RRHs customized by solving the resource allocation models using MILP for fractional power control with bandwidth adaptation or CVX for full power and bandwidth allocation. Lastly, we observe the reward and update the Q-table in step 4.

Algorithm 1: The RL-based framework

1 Initialization Request of UE;
2 Build initial association among UE and RRH;
3 Initialize the Q-table;
4 **For** each decision epoch *t* do (UE demand changed by traffic model)
/* Operation */
5 Randomly select an activation action,
6 Otherwise $a' = \text{argmax}Q(s, a)$;
7 where $Q(\cdot)$ is estimate by energy consumption and UE Satisfaction;
8 Execute activation action a;
9 Obtain the set of active RRHs L;
/* Resource allocation */
10 Obtain the optimal energy consumption based on the given L
11 by solving power allocation Using MILP for fractional power with bandwidth adaptation or CVX for full power and bandwidth allocation;
12 Reconfigure association of the network;
/* Update */
13 Observe the reward r_k and the new state s';
14 Store the state transition(s, a, r, s');
15 Update Q-table by $Q(s, a) = r + \gamma \max Q(s', a')$
16 **End**

5. Performance Evaluation

5.1 Scenario configuration

To evaluate the performance of our proposed algorithm, we perform the numerical simulations using MATLAB. Two solvers are used as shown in the algorithm specification to solve the physical resource allocation problem namely: MILP solver [28] and CVX solver

[29]. The simulation parameters are provided based on LTE standards and listed in **Table 1** below.

With the specified actors in the defined system model, we consider three RRHs as a cluster which is connected to one BBU pool in the network. In each cluster, the BBU takes over the on-off action using the Q-table generated from the Q-learning agent. Inter-cluster resource allocation is controlled by a multiple-agent controller. The number of RRHs is assumed based on the use case scenario in the experiment. The two use cases defined in this paper are assuming fixed number of RRHs from the performance evaluation with changing traffic demand perspective and changing number of RRHs from the performance evaluation with changing network density perspective.

Table 1. Simulation Parameters

Parameter and units	Values
Number of RRH, J	~18
RRH coverage, c	200 m
System Bandwidth, W_j	20 MHz
Maximum transmit power per RRH, P_j^{trans}	1Watt
Channel gain, g_{ij}	9dB
Noise power spectrum density, σ	-174 dBm/Hz
Carrier Frequency band, F	2.4 GHz
Path loss model	32.4+20lg(F)+20lg(D)
Shadowing effects, δ	0-8 dB, random
Angular coefficient in traffic model, v	0.1
Number of UEs in traffic model, I	4-192
UE sensitivity	-120dBm
Power consumption(active state), P_j^{active}	6.8Watt
Power consumption (sleep state), P_j^{sleep}	4.3Watt
UE demand, d_i	1Mbps
Steepness coefficient in satisfaction model, τ	1
weighted factor in reward function, ω	100

We also assume the maximum number of RRHs for each BBU cloud to be 3, and at most 18 in total in the whole system. To be more realistic in our work, we set the system bandwidth of RRHs at 20MHz. The threshold of UE sensitivity is set at -120dBm for edge UEs. As specified in the network model, the RRH coverage in the network is 200m. The number of UEs ranges from 4 to 192 according to the traffic model [14]. The user demand of 1Mbps is equal for each of the UEs. Each UE is considered to have only one interface. The energy consumption largely depends on the transceiver power settings, traffic load and the active duration of RRH. As specified in the energy model, the power consumption of RRH is set at 6.8W, and 4.3W for the active, and idle states respectively [20]. The transmitter power per RRH in active state is 1.0W, while in sleep state it is negligible and therefore, eliminated from our model. In addition, two utility functions are adopted in terms of throughput/data rate in (10) and QoS satisfaction in (9).

In order to evaluate our proposed model and algorithms, we define four different schemes. We define scheme I as simple on-off cell activation with the simple nearest-RRH association, which is also called the simple on-off scheme. We define scheme II as the cell activation with the load ordering-based heuristic scheduling algorithm. In scheme III, Q-learning based cell activation algorithm is used, and the fractional power with bandwidth adaptation is solved by the MILP solver, which is identified as Q-learning with MILP (Q-learning-MILP). Lastly in algorithm scheme IV, we use Q-learning to make a cell activation decision, but the full power and bandwidth allocation is solved by the CVX solver, which is identified as Q-learning with CVX (Q-learning-CVX). We compare our proposed algorithm with the simple on-off scheme and the heuristic scheduling algorithm because of the following reasons; all the three algorithms are model-free and take the dynamics of traffic distribution into consideration. However, the simple on-off scheme is a baseline algorithm that prefers nearest-association of UEs and RRHs. If there are no UEs near to an RRH, the RRH is switched off and vice versa. The difference between the heuristic scheduling based algorithm and the proposed Q-learning algorithm is that, the heuristic algorithm is based on static policy, i.e. there is no feedback to the former after scheduling. The learning agent in the Q-learning based algorithm receives feedback in the form of a reward. As the traffic distribution changes, the learning agent selects an optimal solution to the problem.

The objective of this paper is to optimize the wireless network energy consumption and radio resource occupancy while satisfying the QoS requirement of UEs. The simulation results can be classified into the following metrics; the number of active RRHs, transmit power cost, accumulated total energy consumption and average user QoS satisfaction. The normalized number of active RRHs can be used to evaluate the effect of cell activation. For Q-learning-MILP scheme, we assume that, the used bandwidth proportion is equal to transmit power cost proportion. The transmit power cost can be used to evaluate the effect of our radio resource allocation model. Accumulated total energy consumption can be used to evaluate optimized effect of the entire wireless network's energy consumption. Since UEs only care about their QoS satisfaction, we can use the QoS satisfaction metric to evaluate the effect of user satisfaction. For these four performance criteria, we develop two aspects of evaluation in the experiment. One is that we observe the performance of 24 hours-in-a-day-based traffic model to evaluate the performance of our proposed algorithm with changing traffic demand. Another is that we observe the performance with changing network density to evaluate the extension of our proposed algorithm. We define the density as the number of UEs over the number of RRHs.

5.2. Performance evaluation with changing load

In this simulation, we configure 18 RRHs which can be considered as 6 clusters with 3RRHs each in a coverage area of 400m-by-600m. The user demand of the individual UEs do not change but the total user demand changes based on the traffic model of 24-hours-in-a-day. Considering one hour as a decision cycle, we observe the performance in 24 cycles/hours on the above-mentioned evaluation metrics. The result presented in **Fig. 2** is the number of active RRHs against 24 hours-in-a-day-based traffic model in terms of UE population. The traffic of each hour changes leading to the states of RRHs changing between active and sleep. The results of the number of active RRHs, as illustrated in **Fig. 2**, show that the number of active RRHs correlates positively with the traffic volume.

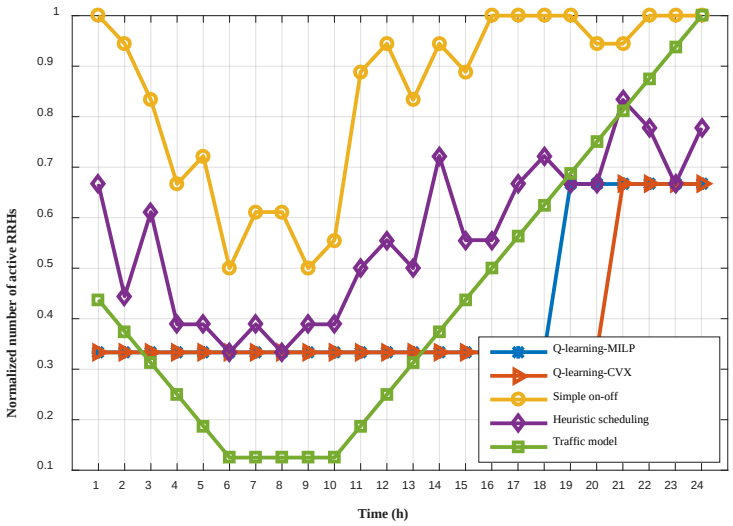

Fig. 2. Normalized number of active RRHs

In all 4 schemes, a change in trend of traffic results in relatively same proportional change in trend in the number of active RRHs. Cell activation based Q-learning schemes outperform scheme I and scheme II. Performance of Q-learning-MILP is nearly the same as that of Q-learning-CVX in cell activation. The gain is higher in heuristic scheduling than Simple on-off. While the simple on-off increases the number of active RRHs up to 100%, the cell activation-based Q-learning schemes increase to less than 70% with the same traffic profile. It can be concluded that, cell activation-based Q-learning schemes can support fewer number of active RRHs compared with the simple on-off and heuristic scheduling schemes, making the cell activation-based Q-learning schemes better than the simple on-off and heuristic scheduling schemes in this scenario.

In **Fig. 3**, we consider the gains of transmit power cost. Based on the above assumption in Q-Learning-MILP that the used bandwidth proportion is equal to transmit power cost proportion, we can know that the change in trend of radio resource occupancy proportion is the same as transmit power cost. From **Fig. 3**, it is observed that the transmit power cost of Q-learning-CVX is greater than the other three schemes while in schemes I, II and III, the transmit power cost is almost the same. This is so because the objective of Q-learning-CVX is maximizing throughput. Therefore, this scheme will need more radio resource, while the other three schemes allocate radio resource by their QoS satisfaction requirements to save radio resource. It can be deduced that, when network is in the limited resource situation Q-learning-MILP is more suitable. On the other hand, when there is abundance of resource, Q-learning-CVX is more suitable.

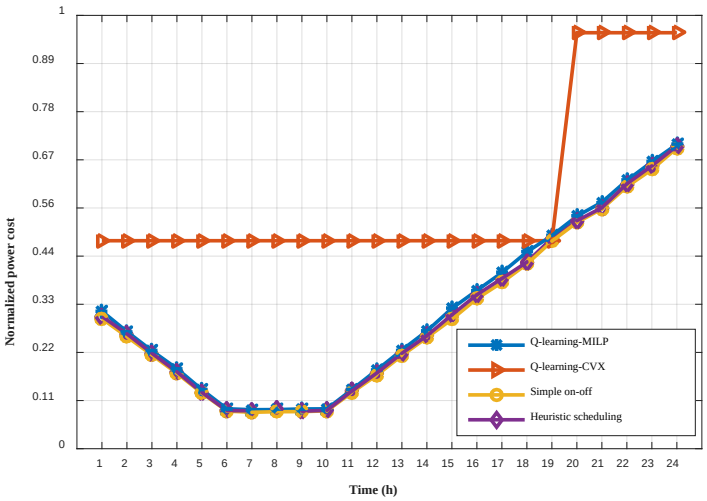

Fig. 3. Normalized transmit power cost

To illustrate the achieved total energy consumption of the proposed algorithm, a simulation is done with two schemes, Q-learning-MILP and Q-learning-CVX compared with the existing heuristic scheduling scheme and simple on-off schemes. The normalized total energy consumption is compared, the result being illustrated in **Fig. 4**. In this paper, we consider two aspects of energy cost including energy cost of active RRHs and transmit power cost. The total energy consumption is the sum of energy cost of active RRHs and transmit power cost. From **Fig. 4**, the normalized total energy consumption of the simple on-off scheme is just above 0.83, which corresponds to 1800KJ in actual value, whiles it is about 0.67, which corresponds to 1600KJ in the Q-Learning-based schemes. The proposed Q-learning-MILP algorithm outperforms the others having the least total energy consumption and is closely near to the Q-learning-CVX. This is because, the transmit power cost of Q-learning-CVX is more than Q-learning-MILP. The Q-learning based algorithms outperform scheme I and scheme II, while heuristic scheduling scheme outperforms simple on-off scheme.

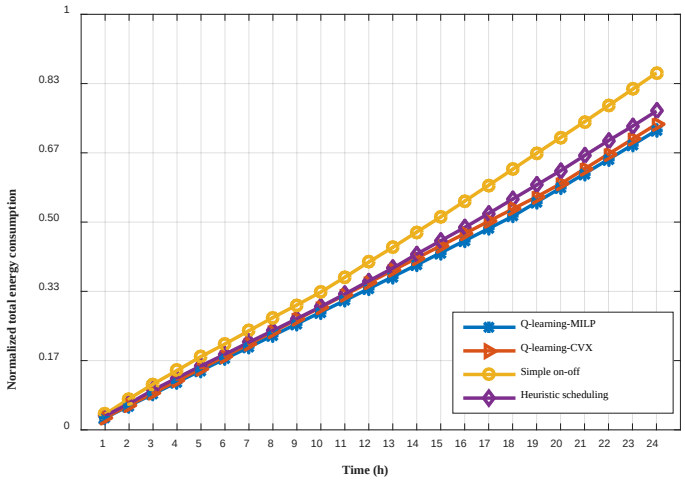

Fig. 4. Normalized total energy consumption

In order to check if our proposed algorithm can satisfy QoS requirement of all UEs, a simulation is done to observe QoS satisfaction with the four schemes. The result is illustrated in **Fig. 5**. Obviously, Q-learning-CVX scheme outperforms the other schemes in terms of satisfaction. This is because; the Q-learning-CVX scheme uses as much radio resource available to satisfy UEs. In other words, Q-learning-CVX sacrifices its transmit power and bandwidth resource to achieve higher satisfaction and throughput.

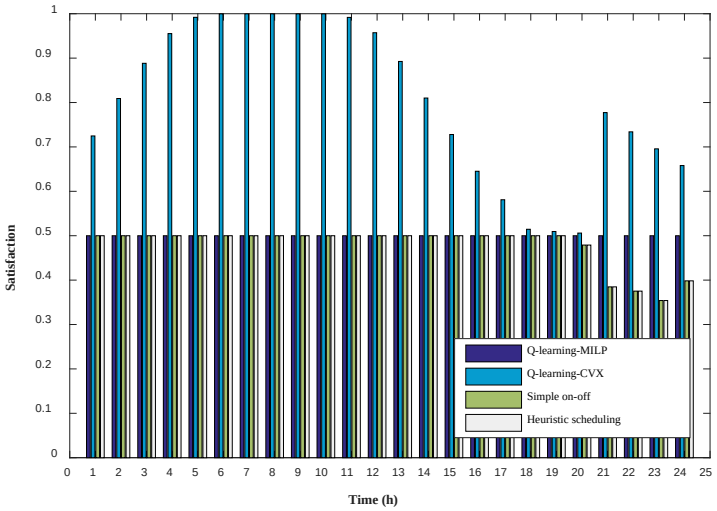

Fig. 5. User satisfaction

The satisfaction rate of Q-learning-MILP hinges at 50% for each hour, meaning the Q-learning-MILP scheme only cares about satisfying the minimum QoS requirement of the UEs. For simple on-off and heuristic scheduling schemes, it is observed that, the QoS requirements of all UEs are satisfied under light network load. At hour 1-19, the satisfaction rate is 50%. As the network load increases further, say at hour 20-24, satisfaction of the UEs begin to drop to as low as 35% at hour 23. In summary, we observed the performance of the four schemes in terms of number of active RRHs, transmit power cost, accumulated total energy consumption and QoS satisfaction. Based on the above discussion, we conclude that our proposed algorithm works better than the other schemes with changing traffic demand. To check the scalability of our proposed algorithm for other scenarios, we extend the evaluation with changing network density.

5.3. Performance evaluation with clusters

In this evaluation, we configure 6 density values based on the ratio of the number of UEs to the number of RRHs. We set the number of RRHs range from 3 to 18 by adding 3 RRHs for each density value change. In order to show the change of network load from light to heavy, we set the number of UEs as 4, 16, 36, 64, 100 and 144. Then the value of density is 4/3, 8/3, 16/3, 20/3, 24/3. Let each density be divided by the biggest value of density as normalization, so the value of density is normalized as 1/6, 2/6, 3/6, 4/6, 5/6, 1. The higher the network density, the heavier the network load.

The result presented in **Fig. 6** is the number of active RRHs against the value of normalized network density from 1/6 to 1. A change in density leads to the status of RRHs changing between active and sleep. The results of the number of active RRHs, as illustrated in **Fig. 6** show that the number of active RRHs correlates positively with the density. Q-learning based cell activation schemes outperform schemes I and II just like the evaluation

with changing traffic demand. The performance of Q-learning-MILP is same with Q-learning-CVX in cell activation. The gain is higher in heuristic scheduling than simple on-off scheme. However, when the network load is very heavy, for instance at a density of 1, the number of active RRHs is the same for all four schemes. In conclusion, Q-learning-based schemes use less number of RRHs under light network load but increase to the same level as the other schemes under heavy load.

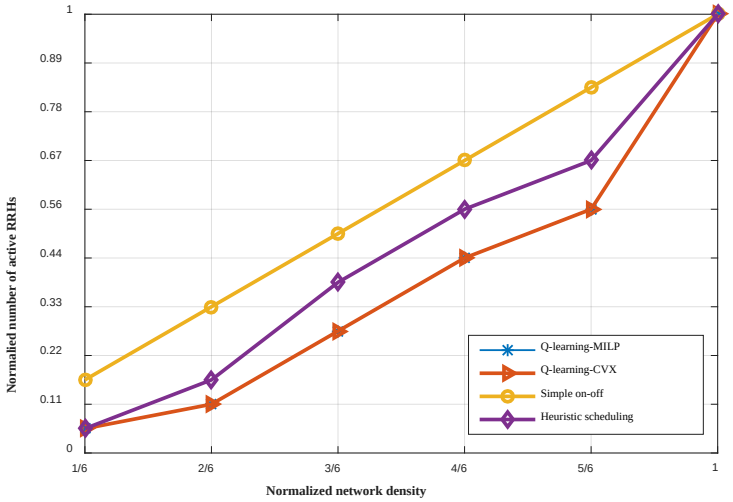

Fig. 6. Normalized number of active RRHs with density changing

In this simulation, an evaluation on total energy consumption is done, with the result illustrated in **Fig. 7**. From **Fig. 7** the proposed Q-learning-MILP algorithm outperforms the others and is closely near to the Q-learning-CVX. This is because, the transmit power cost of Q-learning-CVX is more than Q-learning-MILP. Q-learning based algorithm outperforms schemes I and II, while heuristic scheduling scheme outperforms simple on-off scheme. It is deduced that, the Q-learning-based algorithms attain lower energy costs than the simple on-off and heuristic scheduling schemes even with increasing density.

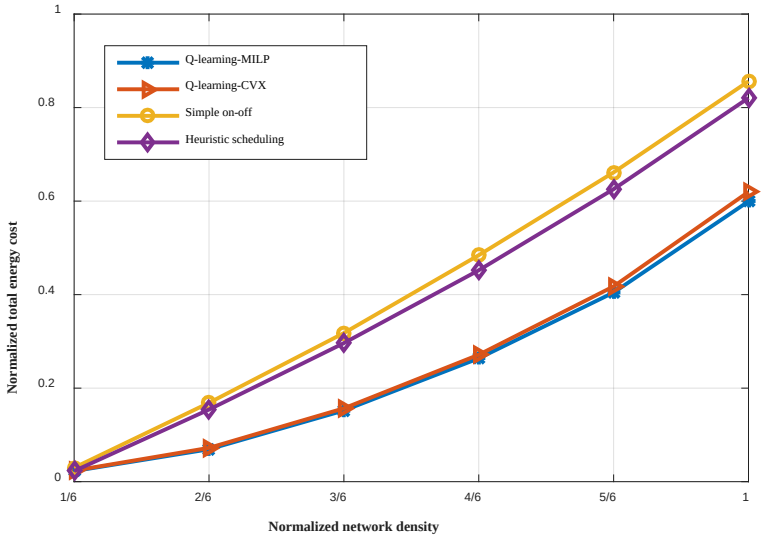

Fig. 7. Normalized total energy cost with density changing

In this simulation, we check the ability of our proposed algorithm to satisfy the QoS requirement of all UEs. A simulation is done to observe QoS satisfaction with the four schemes. The result illustrated in **Fig. 8** show that Q-learning-CVX scheme outperforms the other schemes in terms of satisfaction.

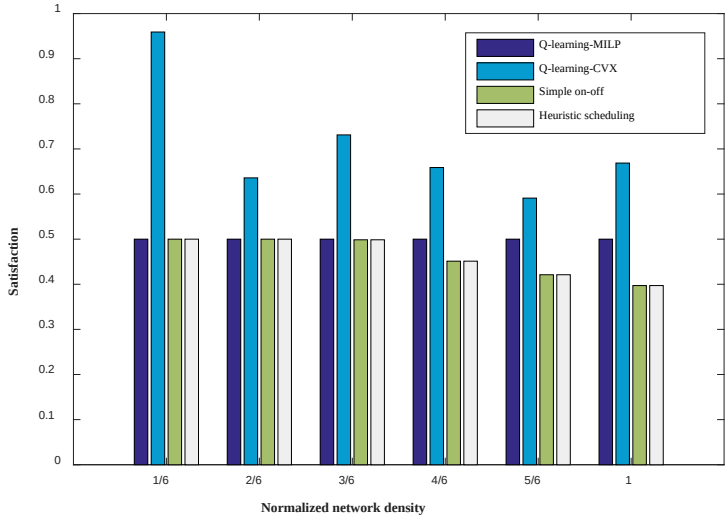

Fig. 8. Satisfaction with density changing

The satisfaction of Q-learning-MILP rests at 50% for each density change, implying that Q-learning-MILP can satisfy the minimum QoS requirement of all UEs without focusing on maximizing throughput.For the simple on-off and heuristic scheduling schemes, we observe that with light network load or at lower densities i.e. for example normalized network density 1/6, 2/6 and 3/6, the QoS requirement of all UEs are satisfied, the satisfaction level being 50%. As the network load increases with increasing density, these two schemes' satisfaction levels drop to as low as 40%. It can be concluded that, the Q-Learning based schemes attain higher satisfaction gains under heavy network loads than the simple on-off and heuristic scheduling schemes. In a summary, we observed the performance of the four schemes in terms of four metrics namely; number of active RRHs, transmit power cost, accumulative total energy consumption and QoS satisfaction with density changing. Based on the above discussion, we conclude that our proposed algorithm performs better than the others even under the changing density scenario.

6. Conclusion

In this paper, we proposed a generic framework of autonomous cell activation and customized physical resource allocation schemes for energy consumption and QoS optimization in wireless networks. In the cell activation scheme, we set up a Q-learning model to satisfy the QoS requirements of users and to achieve low energy consumption with the minimum number of active RRHs under varying traffic demand. In the customized physical resource allocation scheme, we formulated the EE-QoS optimization problem as fractional power control with bandwidth adaptation and full power and bandwidth allocation models. Under the fractional power control with bandwidth adaptation model, we minimized bandwidth resource usage while satisfying user QoS with limited resource. In the full power and bandwidth allocation model, we maximized the system throughput while kept fairness

among users by utilizing all bandwidth resource available. The proposed schemes, Q-learning-CVX and Q-learning-MILP were compared with the existing simple on-off and heuristic scheduling schemes. Simulation results showed that, the proposed Q-learning based schemes outperform the other existing schemes in terms of energy consumption and user satisfaction.

Acknowledgment

This work is supported by National Natural Science Research Foundation of China, Grant, no. 61771098, by the Science and Technology Planning project of Sichuan Province, China, under grant, no. 2016GZ0075, by the Fundamental Research Funds for the Central Universities under grant, no. ZYGX2014J060, and the ZTE Innovation Research Fund for Universities Program 2016.

References

[1] N. Bhushan, D. Malladi, J. Li and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82-89, Feb. 2014. Article (CrossRef Link)

[2] China Mobile, "C-RAN: the road towards green RAN," *White Paper*, 2011. Article (CrossRef Link)

[3] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," *MIT Press, Cambridge, MA*, Feb. 1998. Article (CrossRef Link)

[4] S. Luo, R. Zhang and T. J. Lim, "Downlink and uplink energy minimization through user association and beam forming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 494–508, Feb. 2015. Article (CrossRef Link)

[5] J. Tang, G. Xue and W. Zhang, "Cross-layer optimization for end-to-end rate allocation in multi-radio wireless mesh networks," *Wireless Networks*, vol. 15, no. 1, pp. 53–64, Feb. 2009. Article (CrossRef Link)

[6] S. Buzzi, C. L. I, T. E. Klein, H. V. Poor, C. Yang and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697-709, Apr. 2016. Article (CrossRef Link)

[7] Y. Shi, J. Zhang and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Transactions on Wireless Communication*, vol. 13, no. 5, pp. 2809–2823, May 2014. Article (CrossRef Link)

[8] Y. Shi, J. Zhang, W. Chen and K. B. Letaief, "Enhanced group sparse beamforming for green cloud-RAN: a random matrix approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2511-2524, Nov. 2017. Article (CrossRef Link)

[9] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016. Article (CrossRef Link)

[10] M. Peng, K. Zhang, J. Jiang, J. Wang and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015. Article (CrossRef Link)

[11] Lin, Yicheng, Bao, Wei, Yu, Wei, Liang and Ben, "Optimizing user association and spectrum allocation in HetNets: a utility perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1025 – 1039, Jun. 2015. Article (CrossRef Link)

[12] Wei-Sheng Lai , Tsung-Hui Chang and Ta-Sung Lee, "Joint power and admission control for spectral and energy efficiency maximization in heterogeneous OFDMA networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 3531 – 3547, May 2016. Article (CrossRef Link)

[13] Koudouridis, G.P., H. Gao and Legg, P., "A centralised approach to power on-off optimization for heterogeneous networks," in *Proc. of IEEE Vehicular Technology Conference (VTC Fall)*, pp. 3-6, Sept. 2012. Article (CrossRef Link)
[14] Sun, G., Addo, P. C., Wang, G., and Liu, G., "Energy efficient cell management by flow scheduling in ultra-dense networks," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 9, pp. 4108-4122, Sept. 2016. Article (CrossRef Link)
[15] X. Zhang, J. Zhang, Y. Huang and W. Wang, "On the study of fundamental trade-offs between QoE and energy efficiency in wireless networks," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 3, pp. 259-265, Apr. 2013. Article (CrossRef Link)
[16] A. Mesodiakaki, F. Adelantado, L. Alonso and C. Verikoukis, "Energy-efficient context-aware user association for outdoor small cell heterogeneous networks," in *Proc. of IEEE Int. Conf. on Commun. (ICC)*, pp. 1614–1619, Jun. 2014. Article (CrossRef Link)
[17] Y. Xu, R. Hu, L. Wei and G. Wu, "QoE-aware mobile association and resource allocation over wireless heterogeneous networks," in *Proc. of IEEE Global Commun. Conf. (GLOBECOM)*, pp. 4695–4701, Dec. 2014. Article (CrossRef Link)
[18] H. Li, T. Wei, A. Ren, Q. Zhu and Y. Wang, "Deep reinforcement learning: framework, applications and embedded implementations," in *Proc. of IEEE/ACM International Conference on Computer-aided Design (ICCAD), Irvine, CA*, pp. 847-854, Oct. 2017. Article (CrossRef Link)
[19] F. Shams, G. Bacci and M. Luise, "Energy-efficient power control for multiple-relay cooperative networks using Q-learning," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1567-1580, Mar. 2015. Article (CrossRef Link)
[20] X. Zhiyuan, Y. Wang and J. Tang, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. of IEEE International Conference on Communications (ICC)*, pp. 1 – 6, May 2017. Article (CrossRef Link)
[21] H. Zhang, H. Liu, J. Cheng and V.C.M. Leung, "Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1705-1716, 2018. Article (CrossRef Link)
[22] Checko A., Christiansen H. L., Yan Y., et al. "Cloud RAN for mobile networks—A technology overview," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405-426, Firstquarter 2015. Article (CrossRef Link)
[23] A. Moubayed, A. Shami and H. Lutfiyya, "Wireless resource virtualization with device-to-device communication underlaying LTE Network," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 734-740, Dec. 2015. Article (CrossRef Link)
[24] C. Xu, T.Li, M. Sheng, et al, "Self-organized dynamic caching space sharing in virtualized wireless networks," in *Proc. of IEEE Globecom Workshops (GC Wkshps)*, pp.1-6, Dec. 2016. Article (CrossRef Link)
[25] Q. Ye, B. Rong, Y. Chen and M. AI-Shalash, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706-2716, Jun. 2013.
[26] D. A. Duwaer, "On the deep reinforcement learning for data-driven traffic control," *LD Software, Eindhoven*, 2016. Article (CrossRef Link)
[27] Baxter A. Laurence, "Markov decision processes: discrete stochastic dynamic programming," *Technometrics*, vol. 37, no. 3, pp.353, 1995. Article (CrossRef Link)
[28] J. Lofberg, "YALMIP: a toolbox for modeling and optimization in MATLAB," in *Proc. of the IEEE International Symposium on Computer-Aided Control System Design (CACSD04), Taipei, Taiwan*, Oct. 2004. Article (CrossRef Link)
[29] Michael Grant and Stephen Boyd. "CVX: Matlab software for disciplined convex programming, version 2.0 beta," 2013. Article (CrossRef Link)

Guolin Sun received his B.S., M.S. and Ph.D. degrees all in Communication and Information Systems from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2000, 2003 and 2005 respectively. After Ph.D. graduation in 2005, Dr. Guolin has got eight years industrial work experience on wireless research and development for LTE, Wi-Fi, Internet of Things (ZIGBEE and RFID, etc.), Cognitive radio, Localization and navigation. Before he joined the School of Computer Science and Engineering, University of Electronic Science and Technology of China as an Associate Professor in Aug. 2012, he worked in Huawei Technologies Sweden. Dr. Guolin Sun has filed over 40 patents, and published over 40 scientific conference and journal papers, acts as TPC member of conferences. Currently, he serves as a vice-chair of the 5G oriented cognitive radio SIG of the IEEE (Technical Committee on Cognitive Networks (TCCN)) of the IEEE Communication Society. His general research interest is software defined networks, network function virtualization, radio resource management.

Gordon Owusu Boateng received his Bachelor degree in Telecommunications Engineering from the Kwame Nkrumah University of Science and Technology, Kumasi-Ghana, West Africa, in 2014. He is currently studying MSc. Computer Science and Technology in University of Electronic Science and Technology of China (UESTC). From 2014 to 2016, he worked under sub-contracts for Ericsson (Ghana) and TIGO (Ghana). He is also a member of the Mobile Cloud-Net Research Team – UESTC. His interests include Mobile/Cloud Computing, 5G Wireless Networks, D2D Communications, Data Mining and SDN.

Hu Huang received his Bachelor of Information and Computing Science from Qingdao Agricultural University, China in 2013 and Master of Computer Science and Technology from the University of Electronic Science and Technology of China (UESTC) in 2018. Now he is working as a Software Developer for Hisense, China. His research interests include Internet of Things, SDN and 5G.

Wei Jiang received his Ph.D degree from Beijing University of Posts and Telecommunications (BUPT) in 2008. In March 2008, he worked for 4 years in Central Research Institute of Huawei Technologies, in the field of wireless communications and 3GPP standardization. In September 2012, he joined the Institute of Digital Signal Processing, University of Duisburg-Essen, Germany, where he was a Postdoctoral researcher and worked for EU FP7 ABSOLUTE project and H2020 5G-PPP COHERENT project. Since October 2015, he joined the Intelligent Networking Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, as a senior researcher and works for H2020 5G-PPP SELFNET project. Meanwhile, he also works for the Department of Electrical and Information Technology (EIT), Technische University (TU) Kaiserslautern, Germany, as a senior lecturer. He served as a vice Chair of IEEE TCCN special interest group (SIG) “Cognitive Radio in 5G”. He is the author of more than 30 papers in top international journals and conference proceedings, and has 27 patent applications in wireless communications, most of which have already been authorized in China, Europe, United States or Japan. He wrote a chapter “From OFDM to FBMC: Principles and Comparisons” for the book “Signal Processing for 5G: Algorithms and Implementations” (Wiley, 2016). His present research interests are in digital signal processing, multi-antenna technology, cooperative communications, 5G, and machine learning.