SCIENCE CHINA
Information Sciences

• **RESEARCH PAPER** •
Special Focus on Mobile Network Virtualization

April 2017, Vol. 60 040307:1–040307:16
doi: 10.1007/s11432-017-9044-x

Quantum entropy based tabu search algorithm for energy saving in SDWN

Chaowei WANG, Wuyang MEI, Xiangying QIN & Weidong WANG*

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876*, China*

Received January 13, 2017; accepted February 13, 2017; published online March 17, 2017

Abstract The energy consumption of the base station (BS) accounts for great proportion of the total wireless access network (WAN). Switching off the selected spare BSs with few network request would save a large amount of energy. It is difficult to deploy a BS energy saving strategy in existing network architecture due to the tightly coupled network devices. Therefore, we adopt the software defined wireless networks (SDWN) structure which is an sample of the wireless software defined networks (SDN). Then a novel quantum entropy based tabu search algorithm (QETS) is proposed to choose which BS to switch off, and it increases the search range and guarantee the convergence speed. The energy saving strategy can find the optimal solution with higher probabilities and can be deployed in centralized controller as a software. Theoretical analysis and simulation results show the QETS algorithm's gain over the greedy algorithm and quantum inspired tabu search algorithm (QTS) in terms of convergence.

Keywords energy saving, SDN, SDWN, quantum entropy, quantum inspired tabu search algorithm

Citation Wang C W, Mei W Y, Qin X Y, et al. Quantum entropy based tabu search algorithm for energy saving in SDWN. Sci China Inf Sci, 2017, 60(4): 040307, doi: 10.1007/s11432-017-9044-x

1 Introduction

With the development of the Internet era, the mobile Internet have experienced the explosive growth in recent years, which also brings in the great energy consumption [1]. Software defined network is a novel network structure in which the control and data planes are separated, and data stream can be feasibly scheduled. SDN can be used as a smart data pipe in practice. SDWN is an extension of SDN in the mobile networks. Different from the scenarios mentioned above, SDWN is more concerned with the application of SDN in mobile communication protocols and the attached users. Traditional switching protocol, OpenFlow in the backhaul network and core network are still employed. However, the switch need to support the related protocol in SDWN. In the RAN part, we need to simplify the existing AP, for example, the BS only keeps the signal transceiver and signal-processing unit. The Integrated Controller is responsible for the mobility management and power control. In order to guarantee the robustness of network and quality of service (QoS), operators always plan the network capacity which greatly exceeding the expected network request. During the idle time of network, there are plenty of network devices, including BS, is not fully utilized. Switching off a part of idle BSs has been adopted as an effective method to save energy.

* Corresponding author (email: wangweidong@bupt.edu.cn)

The dynamic BSs on/off problem has been proved as a NP-hard problem [2]. There is few effective algorithm that can solve NP-hard problem in polynomial time complexity.

The most fundamental algorithm is greedy algorithm. Son et al. [3] proposed a greedy algorithm without extra signaling cost and make simulation for the performance of the tradeoff between delay and energy consumption in. The convergence speed of greedy algorithm is fast, but it is also easily trapped into the local optimum. There are also some researchers designing the energy saving strategy based on the existing network architecture. Yaacoub et al. [4] designed an utility function revealing the energy consumption of BS and the terminal using D2D technology based on LTE-A structure in, the strategy made the energy saving of both terminals and BSs come true; In [5], the game theory was introduced into a distributed algorithm which could search the optimal solution with a certain probability, but its convergence speed is slow. It was also proved that the best Nash equilibrium is the global optimal solution. The network strategy couples with the network hardware in the form of software in current network architecture. In the condition of deploying any new strategy or configuring amount of network devices, it is essential to update the software that has been running in the network devices on a large scale. It not only costs lots of manpower and material, but also faces the problem caused by the unstable network status. Apparently, switching off BSs will unavoidably influence the QoS to some extent. So some of the energy saving strategies make contributions in guaranteeing QoS. For example, Niu et al. [6] used the M/G/1 queuing model to simulate the change of arrival rate in. The strategy took a compromise between total energy saving and average delay to guarantee QoS based on quantitative description. In [7], Han et al. designed a hybrid BS state model and analysed the change of the call blocking probability and the channel outage probability during the algorithm running and the algorithm ensured the loss of QoS in the acceptable range. These network strategy couples with the network hardware in current network architecture, which makes it difficult to update the strategy on a large scale. In [8, 9], the authors studied the traffic control and distributed controllers in SDN.

In this paper, we regard the BS energy saving problem as choosing set of active BSs with the system load constraint. Karp et al. [10] reveal the two main difficulties in finding an optimal solution to this problem: First, it requires high computational complex to directly find the optimal solutions among 2^n BS on/off state combinations, which n denotes the number of the BSs. Second, it needs a centralized controller to collect global information. In [11], Chiang et al. proposed QTS algorithm referring to the idea of the tabu search algorithm (TSA) and the quantum inspired evolutionary algorithm (QEA). QTS algorithm enhances local search ability and guarantee the convergence speed, but its global search ability is weak. Inspired by above works, we propose a centralized heuristic algorithm based on SDWN architecture which decouples the data plane and control plane. The main contributions are summarized as below.

- The SDWN which providing a centralized architecture to flexibly control network flow and effectively manage network resources is adopted. Our proposed BS energy saving strategy can be easily deployed in the controller and significantly reduce the overall energy consumption.
- The quantum mutation algorithm based on quantum entropy can dynamically adjust the probability of mutation according to the degree of qubit concentration. The proposed algorithm effectively enhances the global search ability and ensure the convergence speed in the meantime.

The rest of this paper is organized as follows. In Section 2, we introduce the architecture of SDWN and describe how the proposed algorithm can be implement in SDWN at the protocol-level. In Section 3, we present the system model followed by the problem formulation for the BS energy saving problem. In Section 4, we propose the QETS algorithm. Section 5 presents simulation results and discussions. Conclusions are drawn in Section 6.

2 Base station energy saving in SDWN

2.1 SDWN architecture

SDN is an emerging network paradigm dividing the network's control logic from the underlying routers and switches. The function of the control plane in SDN has been integrated to a controller entity

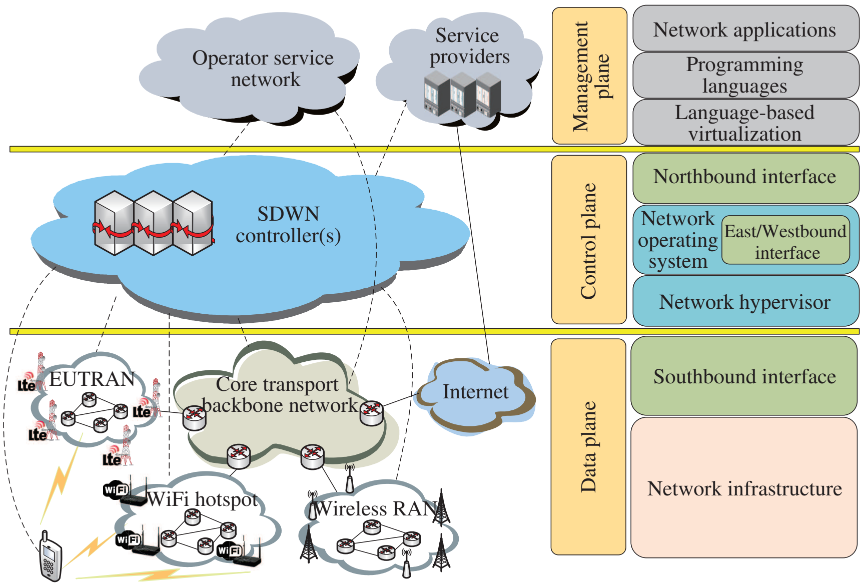

Figure 1 (Color online) Software defined wireless network architecture.

such as SDN controller or network operating system (NOS). SDWN, as a wireless version of SDN in the field of mobile communication, inherits the SDN's key features and rebuilds the original mobile network in SDN-style. The current distributed control plane of wireless networks is suboptimal for managing the limited spectrum, allocating radio resources, completing handover mechanisms, managing interference, and performing efficient load balancing between cells, etc. The centralized architecture in SDWN improves network management in controlling the increasingly bloated network devices. In addition, SDWN provides an opportunity for making it easier to deploy strategies and manage different types of wireless networks, such as WLANs, WiFi hotspot, other RAN and the future ones [12–15]. We refer to the network architecture of 3GPP Evolved Packet System to introduce a well established and understand system architecture of SDWN. The SDWN architecture can be divided into the data plane, control plane, management plane following a bottom-up approach and the corresponding relationship among network elements and each layer's functions are also shown in Figure 1. In which, a dashed line in the figure is used for control plane, and the solid one denotes a user plane connection.

SDWN infrastructure, similarly to a traditional network, is composed of a set of networking equipment, i.e., switches, routers, middlebox appliances and wireless access points in wireless radio access network (RAN). Different from traditional equipment, which is embedded control logic to take autonomous decisions, the networking devices in SDWN are simply forwarding elements, and controlled by the centralized controller through southbound interface. In SDWN, complete and on-limits API defines the Functions of Radio Resource Management Module (RRM) coupled in access devices and the RRM is integrated in the controller. In cellular network scenarios, User Equipment (UE) can be redirected to the low-traffic BS nearby by RRM. If the BS is equipped with multi-antenna, the controller can also decide whether to use the antenna booster signal or spatial reuse. On the other hand, FlowVisor proposes the idea of BS recourses slice from time slot dimension, subcarrier dimension, and power consumption dimension through BS virtualization. For example, part of time slot can be allocated for the virtual BS, and at the same time, for the controlling the base station transmission power.

SDWN controller is a logically centralized controller entity in flow level. It should have such capabilities as data transmission, mobility management and coordinated multiple points, which is equivalent to the functional part of Mobility Management Entity (MME) and Policy and Charging Rules Function (PCRF) in LTE. Network hypervisor realizes the hardware virtualization which makes it possible for different virtual machines to share the limited hardware resources and help controller flexibly distribute the network resources. NOS is to provide abstractions, essential services and common APIs to the developers. The generic functionality is provided by NOS included network state and network topology, device discovery and distribution of network configuration. Network application developers do not have to understand the technology details of underlying hardware, but make use of APIs provided by NOS, which is in favor of rapid strategy deployment. Due to the limited operating ability of one controller, the east/westbound interface is proposed to combined a lot of distributed controllers as a controller entity to complete larger scale and more complex network flow controlling. The northbound interface, which is different from the southbound interface and tightly tied to the forwarding elements, is mostly a software ecosystem. An abstraction brought by northbound interface would allow network applications not to depend on specific implementations. Though the details of northbound is still in discussion, a common consensus is that northbound APIs are indeed important but it is indeed too early to define a single standard right now [12].

In management plane, developers use programing language to abstract the inner details of the controller functions and data plane behavior, at the same time, language-based virtualization provide different levels of abstraction. Based on programmable network fundamental, developers can use the programming language to develop SDWN applications to realize functions. For example, in the Radio Access Network (RAN), applications involve the functions such as interface management, wireless resource management, and data unloading. In the mobile backhaul network, it involves backhaul network resource management, traffic engineering, monitoring, and other functions. While in the mobile core network, the function like MME, GW-C, PCRF, HSS and AAA in the existing network can be a web application. These web applications run on NOS and distribute the control command through controller to the underlying network equipment, so as to realize the program logic written by the network manager.

2.2 Key techniques of BS energy saving in SDWN

The basic target of BS energy saving is that control plane switches off part of BSs depending on the network state. There are three key techniques in BS energy saving: BS energy saving strategy, handover and BS switch off. First, the control plane of wireless network collects network information and initiate BS energy saving strategy according to certain judgment conditions. Then the strategy will get the set of switching off BS and the corresponding plan of handover, which is the basis of the future network actions. Finally, the control plane will lead related UEs, target BS, source BS and other network elements to finish the handover. After confirming the user resource in source BS has been fully released, the control plane will switch off the source BS and finish the BS energy saving procedure.

2.2.1 *BS energy saving strategy*

The base station energy saving strategy is based on the network status parameters, and regards the most amount of the saving energy as the main target. In order to study the impact of user migration and BS-off behavior,the energy consumption before the saving strategy should be formulated. And the user migration and BS-off behavior is on the basis of the result of the strategy directly.

Limited by existing network architecture, the majority of BS energy saving strategy is distributed algorithm which has complex signaling interaction and low efficiency. These algorithms are hard to get the optimal solution because of the lack of a complete network vision. The centralized algorithms in SDWN architecture are expected to get better performance in energy saving and execution efficiency. For example, the greedy algorithm and SWES algorithm in [3] select the energy saving and QoS as the judgment condition to find the best solution in current iteration. But these algorithms have narrow searching space and are easy to trap into local optimum. Only when the solution space is small, these algorithms would find out the optimal solution. Nevertheless, these algorithms still are widely used because of the low complexity and high execution efficiency.

Quantum tabu algorithm [11], as one kind of intelligent group algorithm, is combined with the characteristics of quantum evolutionary algorithm and tabu algorithm. It uses the changes of the probability of two basic states to characterize the probability distribution of the BS's on/off state. Based on the gap between the current optimal solution and the optimal historical solution, Quantum tabu algorithm can update the qubit by quantum revolving door, and get the final result after several iterations. However,

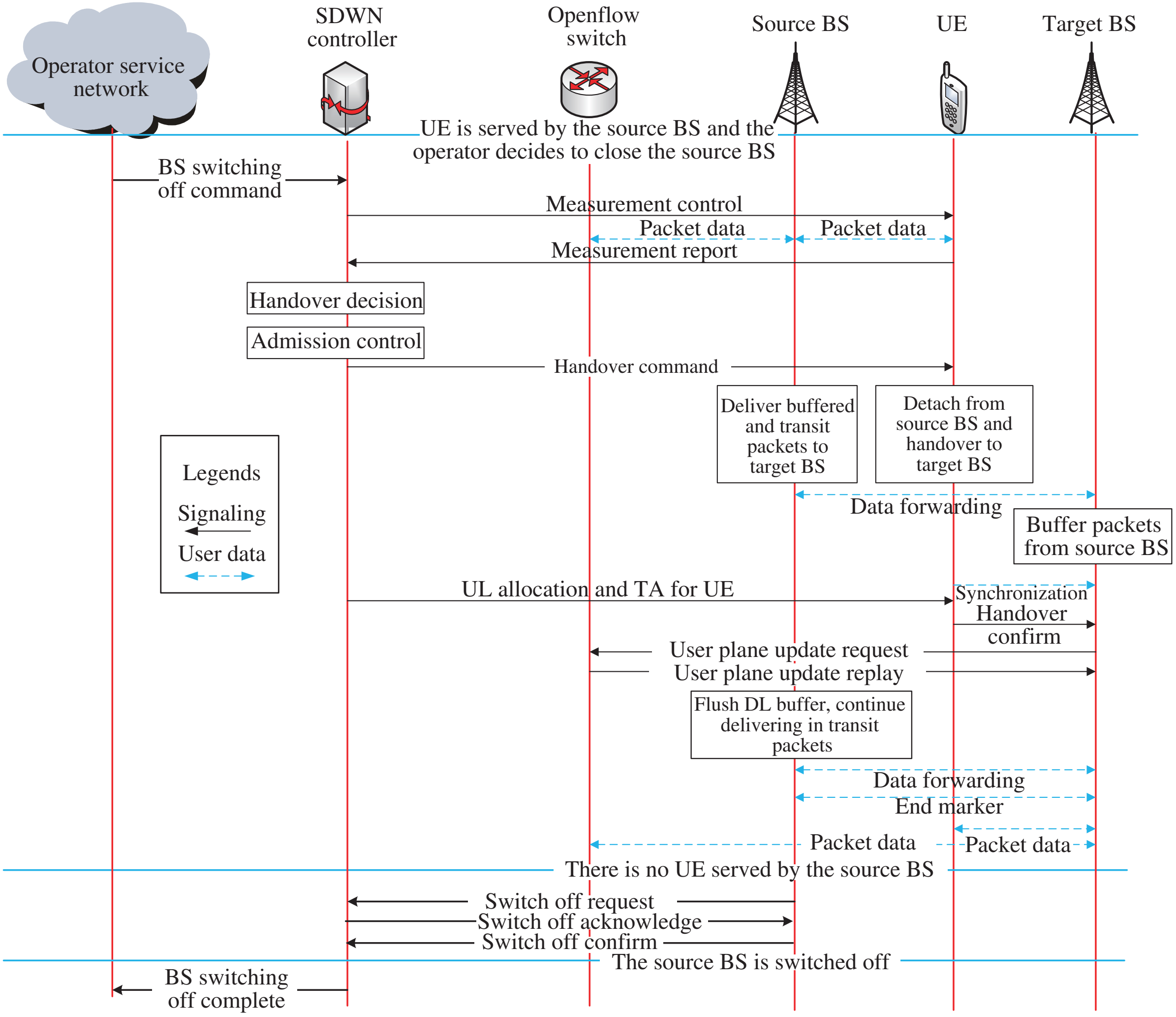

Figure 2 (Color online) Massage sequence chart of BS switching off in SDWN.

the taboo mechanism of the algorithm is complicated and it reduces the search efficiency. Lacking of genetic variation mechanism, quantum tabu leads the algorithm into local optimum. But the combination of quantum gate and tabu table in the algorithm can improve the qubit updating efficiency significantly. Compared with the evolution algorithm, quantum tabu can improve the calculation speed.

2.2.2 *Handover*

It is important to note that the handover in this paper is an active behavior of control plane, which is the active mobility management. Mobility management mainly include two functions: location management and handover control. The location management help the system rapidly locate the user location and accurately establish the data path. The handover control manages the trigger condition of handover, the establishment of new route and network resource release. The handover control aims to maintain continuous session at the cost of the short break of physical connection.

We realize the improvements based on general switching process for the LTE X2 interface. The whole BS energy saving strategy is shown in Figure 2. EPC is simplified in SDWN and some network elements like MME, SGW, PGW are cancelled. Taking the aforementioned network elements and the control logic coupled on eNodeB away, a unified logic controller is produced. Compared to SDN, SDWN is equipped not only with the forward function, but also the standard protocol of SDN. And SDWN also has some peculiar management functions in the mobile network, such as PCRF, CoMP, etc. Taking SDWN with one single controller for example, the controller can not only be a single controller on hardware, but also a single logic unified controller achieved by cloud technology. This kind of structure requires high quality control link to ensure the reliability and low latency of the communication between controller and all wireless nodes and data forwarding nodes. The controller in this kind of SDWN structure also needs to have strong ability of processing or extension ability to manage a large number of network nodes. The advantage is that when switching among several control nodes, control information exchange is not needed. Controller only needs to interact with RAN nodes.

The controller periodically analyse the information report of network state and user requirements to make switching decisions. By means of distributing flow table, the controller changes the specific flow or all of the flow related to UE in order to change the routing rules. In Figure 2, when the operator network application monitor the judgment condition has been met, it will send the strategy active signal to controller to launch the BS energy saving strategy. In the phase of handover, the handshake signal, sequence number signal and resource release signal have been simplified because the controller integrate the control plane functions of MME and eNodeB. In SDWN, signal interaction in the OpenFlow switch and software defined BS are of OpenFlow command and flow table. So the handover in SDWN could eliminate the 2xRTT between different BSs. On the other hand, the signaling delay brought by new link establishment also ceases to exist because of the centralized control. So the signal interaction of handover in SDWN has been simplified comparing with LTE. And switching off BSs on the premise of guarantee QoS also means giving priority to switch off the BS with fewer UEs, which is the same with the BS energy saving target.

2.2.3 *BS switching off*

For mobile networks, the base station energy consumption has accounted for 60%. Taking LTE network as an example, eNodeB is the basic component unit and the important network node. It mainly includes BBU and RRU. BBU is mainly in charge of coordination management of each system in the station, including the function of allocation of resources, environmental monitoring, baseband data demodulation, channel coding and decoding. This part has taken FPGA products with low energy consumption and mature technology, so the improvement of energy consumption is limited. Key components in RRU are power amplifier (PA), filter, power supply module, etc. Functions in RRU include the implementation of BBU resource allocation instruction, downlink signal power amplification, filtering, uplink signal noise removing, power amplifier, etc. Power amplifier module takes the largest proportion of energy consumption for macro base station, usually from 40% to 70%. The floating of energy consumption is often associated with base station working state, and the power consumption of the rest modules is lower and stable. It is easy to see that the energy saving proportion is the largest when shutting off the radio power amplifier module. In SDWN, parts of the BS control function in BBU are separated from the hardware system, however, the radio frequency system in charge of signal transceiver and processing function can be retained. Because the hardware related to the control plane will be cancelled, the energy consumption ratio of radio frequency system in the whole BS will increase greatly. In the second place, in the distributed system, the radio frequency system is generally located in remote place, so a lot of air conditioning energy can be saved during the spare time.

We compare the BS switching off procedure in LTE and SDWN in Figure 3. In LTE, BS to be switched off needs to exchange signaling with its neighboring BSs with network parameters and confirm information. This tedious self-organization signal interaction has lower execution efficiency. And the signal compensation in distributed network is easy to generate the repetitive signal coverage problem, which would increase the power consumption of neighbor BSs. In SDWN, the controller maintains the whole network view and monitors the network state of network equipment though southbound interface. The controller makes the signal compensation decision according to the BS location, BS load, the types of equipment and other network parameters. The neighbor BSs with less UE would be selected to serve the UE belonging to the switching off BS which could solve the repetitive signal coverage problem and get the optimal network configuration. Besides, the signal interaction about BS switching off in SDWN is simple, which means the high execution efficiency.

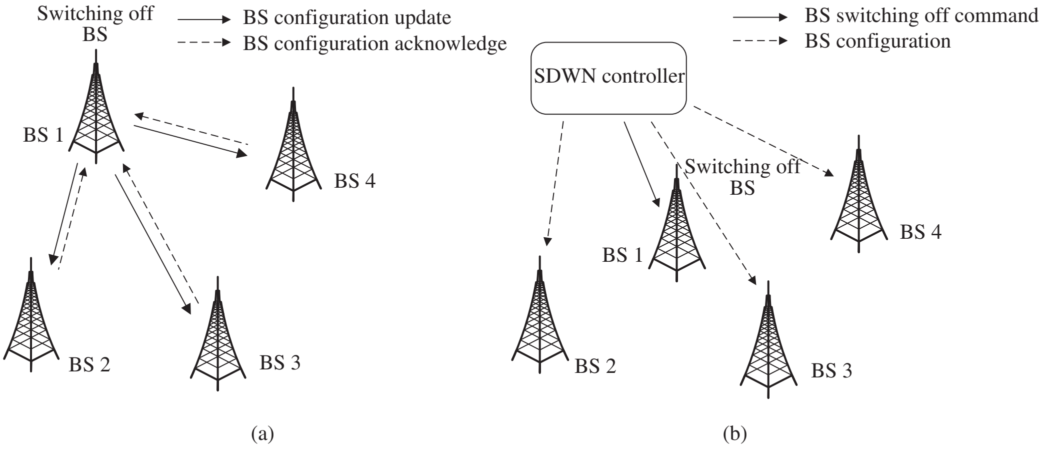

Figure 3 The comparison of the BS switching off procedure in LTE and SDWN. (a) The BS switching off in LTE; (b) the BS switching off in SDWN.

3 System model and problem fomulation

3.1 System model

Our BS switching off strategy is just related to the system parameters of UE, BS and the wireless channel, so the current communication system model also can be applied to SDWN. We focus on downlink communication as the most of the data flow in the mobile network. Assume all of the BSs, denoted by $B = \{b_1, b_2, \ldots b_n\}$ lies in the two-dimensional area Ω. Reference to [16], we assume the traffic arrival rate of UE located at location a follows an independent Poisson distribution with mean arrival rate $\lambda(a)$, and its average request file size is assumed to be an exponentially distributed random variable with mean $1/\mu(a)$. The traffic load of UE is defined as $\gamma(a) = \lambda(a)/\mu(a)$. Since the optimal user association problem has been discussed by many researchers, i.e. [17]. Considering the space limitation, our paper simplifies the user association problem and mainly introduces the switching off strategy. Reference to [16], a UE chooses to be served by the BS whose signal is the strongest,

$$b = \arg \max_{i \in B_{\text{on}}} g(i, a) \cdot P_i^{\text{tx}}, \tag{1}$$

where b is the BS that the UE at location a chooses to associate with, B_{on} denotes the set of active BSs, P_i^{tx} denotes the transmission power of BS i, $g(i, a)$ is the average channel gain from BS i to UE at location a including the path loss, multipath fading, log-normally distributed shadowing. The channel capacity from BS $b \in B_{\text{on}}$ to UE at location a can be described by Shannon's formula:

$$C(a, B_{\text{on}}) = W \cdot \log_2(1 + \text{SINR}_b(a, B_{\text{on}})), \tag{2}$$

where W is the channel bandwidth. $\text{SINR}_b(a, t)$ denotes the received signal to interference plus noise ratio at location a from BS b that is given by

$$\text{SINR}_b(a, B_{\text{on}}) = \frac{g(b, a) \cdot P_b^{\text{tx}}}{\sum_{i \in B_{\text{on}} \setminus \{b\}} g(i, a) \cdot P_i^{\text{tx}} + \sigma^2}, \tag{3}$$

where σ^2 is the noise power, every UE served by the BS b would gain the total traffic load of it, so we set the traffic load density namely $\zeta_b(a) = \gamma(a)/C(a, B_{\text{on}})$, which means that the fraction of time required to deliver traffic loads from BS b to location a. In the serving area of BS b, which is denoted by Ω_b, the system load of BS b can be defined as the fraction of time to serve the total traffic load can be defined as follows:

$$\rho_b(a, B_{\text{on}}) = \int_{\Omega_b} \frac{\gamma(a)}{C(a, B_{\text{on}})} \mathrm{d}a. \tag{4}$$

3.2 Problem formulation

3.2.1 *BS energy consumption model*

We adopt a BS energy consumption model that has been applied in [3]. This model divided the total BS energy consumption into two components: the fixed power consumption and adaptive power consumption,

$$E_b = \begin{cases} (1-h_b)\rho_b P_b + h_b P_b, & b \in B_{\text{on}}, \\ 0, & b \notin B_{\text{on}}, \end{cases} \tag{5}$$

where E_b denotes the total energy consumption of BS b, P_b is the maximum operational power of BS b including every aspects of the energy consumption, e.g., transmit antennas, power amplifiers, cooling equipment, baseband unit and so on. h_b denotes the proportion of the fixed power consumption in the total energy consumption of BS b. When $h_b = 1$, the BS b will consume the fixed power no matter how high the traffic load is. If $h_b \in (0,1)$, the adaptive power consumption will be a part of the total energy consumption besides the constant power consumption. This type is much closer to the real condition. When h_b goes to another extreme, e.g., $h_b = 0$, the BS model is referred to as energy-proportional BS. The energy consumption of BS totally scales with the BS load. When there is no user served by the BS, its energy consumption will reduce to 0 under ideal condition. Apparently, this solution is not practical, e.g., the BS load mainly influence the BS total energy consumption by giving rise to the power amplifiers consumption which just amounts to 55%–60% of the total energy consumption [18].

3.2.2 *Energy saving strategy*

In this paper, our objective is to minimize the total energy expenditure in cellular networks. Our algorithm mainly turns off some BSs with relatively lower traffic load level. The energy saving problem considering the BS switching operations can be formulated as

$$\begin{aligned} &\min_{B_{\text{on}}} E(B_{\text{on}}) = \sum_{b \in B_{\text{on}}} E_b, \\ &\text{s.t.} \quad 0 \leqslant \rho_b \leqslant \rho_b^{\text{th}},\ \forall b \in B_{\text{on}},\ B_{\text{on}} \neq \emptyset. \end{aligned} \tag{6}$$

The BS energy saving algorithm will have negative impacts on the system stability and QoS. Reference to [16], we set a system load threshold $\rho_b^{\text{th}} \leqslant 1$ to make sure the influence brought by algorithm within the acceptable range. When we set a low threshold value, the system will get better QoS and have the larger capacity to cope with the bursty traffic arrivals. But on the other hand, the energy-saving effect will be greatly reduced by the demand of the QoS. In contrast, if we choose a loose threshold, we can obtain more energy saving from turning off more BSs. So taking a proper threshold on the basis of the real network condition can save energy as much as we can in the premise of guaranteeing the system QoS.

4 The proposed QETS algorithm

The quantum computation is the algorithm that run on the quantum computer. Similar to classic bits in traditional computer, the qubit in quantum computer may be the basis state “0” or “1”, but may also be in any superposition of both state [19]. Additionally, the act of measure or observing a qubit will make the quantum system collapse to one of its basis status. As people have not developed the ideal quantum computer yet, some researchers try to improve the traditional algorithm with introducing the characteristic of quantum computation.

The quantum inspired tabu search algorithm, which combines the tabu search algorithm (TSA) and the quantum inspired evolutionary algorithm (QEA), is the one of the most effective quantum intelligent algorithm. TSA uses tabu list to record the historical optimal solution and selectively search the area around the historical optimum, which can effectively avoid premature convergence [20]. In QEA, the population in evolutionary algorithm has been replaced by quantum chromosomes which is the qubit vector coding. The evolution and mutation of the population is realized by the quantum gate [21]. The quantum entropy based tabu search algorithm combining the virtues of the above two algorithms not only avoid trapping into premature convergence, but also can fast get the optimal solution with high probability.

4.1 Quantum coding

The qubit $|\varphi_i\rangle$ in QETS can be formulated by Dirac as

$$|\varphi_i\rangle = \alpha_i|0\rangle + \beta_i|1\rangle = \begin{bmatrix}\alpha_i\\ \beta_i\end{bmatrix}, \quad i = 1, 2, \ldots, n, \tag{7}$$

where the complex numbers α_i and β_i denote the probability amplitude of the basis qubit status $|0\rangle$ and $|1\rangle$, respectively. After being measured, the qubit will collapse to $|0\rangle$ and $|1\rangle$ with certain probabilities of $|\alpha_i|^2$ and $|\beta_i|^2$. In the BS off problem, $|\beta_i|^2$ denotes the probability of the ith BS switching on in the current iteration. Similarly, $|\alpha_i|^2$ is the probability of ith BS switching off in the current iteration. Additionally, both of them meet the normalization condition as (8)

$$|\alpha_i|^2 + |\beta i|^2 = 1, \tag{8}$$

where n is the number of BSs. The quantum chromosome in QETS is represented by a qubit. Each population is produced by measuring a string of qubits as

$$q = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_n\\ \beta_1 & \beta_2 & \cdots & \beta_n\end{bmatrix}. \tag{9}$$

A quantum register with n qubits stores 2^n quantum states at the same time. The set of the qubits denote all of the BS switch on/off combinations which coexist with probability. Apparently, the quantum parallelism can significantly improve the searching efficiency.

4.2 Quantum rotation gate and tabu list

An important step in QETS is to update the qubits, which uses the quantum rotation gate to change the value of qubit [21]. We represent the mathematical expression of the operation as follows.

$$\begin{bmatrix}\alpha_i^t\\ \beta_i^t\end{bmatrix} = U\begin{bmatrix}\alpha_i^{t-1}\\ \beta_i^{t-1}\end{bmatrix} = \begin{bmatrix}\cos(\Delta\theta) & -\sin(\Delta\theta)\\ \sin(\Delta\theta) & \cos(\Delta\theta)\end{bmatrix}\begin{bmatrix}\alpha_i^{t-1}\\ \beta_i^{t-1}\end{bmatrix}, \tag{10}$$

where t denotes the number of iteration and $\Delta\theta$ is the quantum rotation angle. Moreover, $\Delta\theta$ has a great influence upon the convergence and optimality of the proposed algorithm.

In QETS, the tabu list is used to tabu the qubits prohibited from updating their probability distribution, which differs from general tabu list, which is used to record recent moves. Therefore tabu size is dynamic. s^b and s^w, which are binary strings, denote the best solution and the worst solution in current iteration. s_i^b and s_i^w are the on/off status of the ith BS in s^b or s^w. Compare the two binary strings s^b and s^w, if s_i^b and s_i^w are the same, the ith qubit q_k would be placed in tabu list. In QETS, only n-qubit need to be updated, which differs from QEA that needs n by m-qubit. This is why QETS is more effective than QEA [11]. The tabu length is set to 1. The tabu conditions and the quantum rotation angle are shown in Table 1.

4.3 Population produce

The intensification searching and diversification searching coexists in QEA, but they are in conflict with each other. In QETS, we proposes a mutation mechanism based on quantum entropy for dynamically

Table 1 Quantum rotation gate lookup table

s_i^b	s_i^w	q_i locates in first or third quadrant	$q_i \in T$	$\Delta\theta$
0	0	True	True	0
0	1	True	False	$-\theta$
1	0	True	False	$+\theta$
1	1	True	True	0
0	0	False	True	0
0	1	False	False	$+\theta$
1	0	False	False	$-\theta$
1	1	False	True	0

adjusting the proportion of the diversification searching. The quantum entropy denotes the gap between the probability of the two basis status of qubit, which is defined as follows:

$$H\left(q_i^t\right) = -\left|\alpha_i^t\right|^2 \log_2\left(\left|\alpha_i^t\right|^2\right) - \left|\beta_i^t\right|^2 \log_2\left(\left|\beta_i^t\right|^2\right), \tag{11}$$

where q_i^t is the ith qubit in the tth iteration.

Apparently, the quantum entropy values between 0 and 1. When the gap between the $|\alpha_i|^2$ and $|\beta_i|^2$ becomes larger, the quantum entropy is reduced. Therefore, the quantum entropy reflects the aggregation extent of qubit. When the qubits tend to converge to a certain state, which means the algorithm tend to be trapped into the local optimum, QETS would expand the searching scope. One solution is denoted by $s = \{x_1, x_2, \ldots, x_n\}$, where x_i is the ith BS on/off switch status, which values 0 or 1. When x_i is equal to 0, it means that the ith BS is in off mode and vise versa. The pseudocode of the qubit state decision algorithm, which is an important part of QETS, is as follows:

Algorithm 1 Qubit state decision algorithm

```
begin
  for $i = 2$ to $n$ do
    $c, f \in$ random number in $[0, 1]$
    if $c \geqslant H(q_i^t)$ then
      if $f \leqslant |\alpha_i^t|^2$ then
        $x_i = 1$
      else
        $x_i = 0$
      end if
    else
      If $f \leqslant |\alpha_i^t|^2$ then
        $x_i = 0$
      else
        $x_i = 1$
      end if
    end if
  end for
end
```

With the running of QETS, the probability distribution of the qubits tends to converge to the historical optimal solution. The gap between $|\alpha_i|^2$ and $|\beta_i|^2$ gets larger while the quantum entropy of qubit gets smaller. In QETS, the algorithm generates a random number c between 0 and 1. If $c < H(q_i^t)$, the qubit state decision algorithm will decide the BS state with the probability of $|\alpha_i|^2$ to be compared with another random number $f \in [0, 1]$, when $f \leqslant |\alpha_i^t|^2$, set $x_i = 0$, otherwise $x_i = 1$. However, when the qubit was in condensed state, which means $c \geqslant H(q_i^t)$, QETS will use the opposite way to decide the BS state. The mutation mechanism can adaptively adjust the search range depending on the quantum cluster degree, which can help QETS effectively escape from the local optimum.

4.4 QETS algorithm description

The proposed algorithm is described below:

Algorithm 2 Quantum entropy based tabu search algorithm

```
1 begin
2     $t = 0$
3     $T = \emptyset$
4     Initialize $q(t)$
5     Initialize the local optimal solution $s^b$ and the local worst solution $s^w$
6     Initialize the historical optimal solution $M$ and $E(M)$
7     while (not termination-condition) do
8         $t = t + 1$
9         Produce neighborhood mapping set $Q_t$ by multiple measure $q(t-1)$
10        Delete $s \in \{s \in Q_t | \rho(s) > \rho_b^{\mathrm{th}}\}$
11        Evaluate $E(s)$
12        Update the local optimal solution $s^b$ and the local worst solution $s^w$
13        Update the global optimal solution $M$ and $E(M)$
14        Update tabu list $T$
15        Update $q(t)$
16    end while
17 end
```

First, the parameters used in QETS should be initialized. The tabu list is empty at the beginning of the algorithm.

$$q(t) = \begin{bmatrix} \alpha_1^t & \alpha_2^t & \cdots & \alpha_n^t \\ \beta_1^t & \beta_2^t & \cdots & \beta_n^t \end{bmatrix}$$

denotes the set of the qubits in the tth iteration, all α and β are initialized with $1/\sqrt{2}$ [17], which represents the probability of BS switching either on or off are the same. The initial value of s^b and s^w are the binary string of which each place is 1. M is the historical optimal solution and its initial value is the same as that of s^b.

The main program involving lines 7–16 of QETS is executed cyclically until the termination condition is satisfied. We take the tth iteration as the example. The set of populations Q_t in the tth iteration, is produced by multiply measuring $q(t-1)$ with m times:

$$Q_t = \{s_t^1, s_t^2, \ldots, s_t^j, \ldots, s_t^m\}, \tag{12}$$

$$s_t^j = \{x_1^j, x_2^j, \ldots, x_i^j, \ldots, x_n^j\}, \tag{13}$$

where s_t^j denotes the solution produced by the jth measurement in the tth iteration. The each place of the binary string s_t^j is x_i^j, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$, which represents the ith BS on/off switch status in the jth measurement. Therefore, Q_t is the set of the combinations of BS on/off switch. The measurement applies the qubit state decision algorithm. Delete the solution unsatisfying the system load constraint condition from Q_t. Then evaluate the total energy consumption of the rest of solution with the objective function $E(s)$.

From all of the neighborhood solutions, choose the solution with the minimum energy consumption stored into s^b and the solution with the maximum energy consumption stored into s^w. Comparing the energy consumption of s^b and M, if s^b saves more energy than M does, M will be replaced by s^b. Finally update $q(t)$ with the quantum rotation gate. The angle of the quantum rotation gate can be obtained by searching the table. $\Delta\theta$ should be designed in compliance with the BS switching off problem and it is usually set to a small value to prevent premature convergence.

5 Simulation results

In this section, we analyze the proposed BS switching off strategy from several perspectives. We adopt Matlab as our simulation software. We verify the algorithms which could run in the SDWN controller. The network topology is composed of 19 LTE hexagonal macro BSs with one BS in the center and two laps seamless extended out around the central BS. We also adopt wrap-around technique to avoid edge effects [22]. Users are allocated randomly in BS serving area. Considering the base station switching off strategy is often used in low-level network traffic, the user number of each BS is less than 5. Transmission power of base station is 43 dBm, based on linear relationship between transmission and operational power consumption [3]. We can get the maximal operational powers for macro cell of 865 W according to [23]. The macro propagation model in urban is using the carrier frequency of 2000 MHz, 5 MHz bandwidth and a base station antenna height of 15 m above average rooftop level. As we have mentioned in Section 3, we adopt a Poisson point access, whose arrival rate at location a is denoted by $\lambda(a)$ and average file size $1/\mu(a) = 100$ kB. In order to guarantee the QoS and robustness, we set the system load threshold for all BSs as $\rho^{\mathrm{th}} = 0.6$.

5.1 Convergence of the proposed algorithm

In this paper, we compare QETS with the greedy algorithm proposed in [3], the QTS algorithms proposed in [11] and the exhausted search algorithm. As shown in Figure 4, we use the energy saving ratio to show the energy efficiency of the algorithm, it is defined as follows:

$$E_{\mathrm{ratio}} = 1 - \frac{E_{\mathrm{algorithm}}}{E_{\max}}, \tag{14}$$

where $E_{\mathrm{algorithm}}$ represents the energy consumption after the algorithm has complete and $E_{\max}$ denotes the maximum energy consumption when all BSs are turned on.

From Figure 4 we can see that QETS algorithm has the best energy saving effectiveness, which has found out the optimal solution. The suboptimal solution is reached by the greedy algorithm, QTS algorithm is the worst. In terms of convergence speed, QTS algorithm is the fastest, and next is the greedy algorithm, QETS is the slowest. Therefore, QETS can get the solution with better energy saving effectiveness at the cost of searching time. After analysing a large number of simulation data, we find that QETS can converge to the optimal solution with high probability around 20 iterations. From a practical point of view, we just consider the results comparison within 80 iterations.

As we have mentioned in Section 4, the angle of quantum rotation gate $\Delta\theta$ is a key parameter that affects QETS optimality and convergence speed, its positive and negative variations are determined by the direction of rotation. The issues of the direction of rotation have been illustrated in [21]. A large rotation angle means that, at each iteration, the quantum rotation gate makes the probability distribution of existing BSs switch combination rapidly get close to the historical optimal solution. However, it increases the velocity of convergence at the cost of the extent of searching. Conversely, start with a small rotation angle means slower convergence speed. We were motivated to choose some representative value of θ to compare. As shown in Figure 5, when $\theta = 0.001$, convergence speed is slow, the optimal solution is achieved around the 75th iteration. When $\theta = 0.006$, QETS converges faster, but the algorithm is trapped into local optimum around the 50th iteration, which is very close to the optimal solution. When $\theta = 0.01$, convergence speed is fast, but it is trapped into local optimum at the 5th iteration. Since the searching space is limited, it's hard to escape from local optimum to search a better solution. When $\theta = 0.003$, around 25th iteration, QETS gets the optimal solution.

As we can search the optimal solution with probability, a series of experiments with different values of θ were simulated 500 times, and the results are listed in the Table 2. Optimal solution denotes the number of optimal solution that QETS got in 500 experiments. 95% energy saving indicates the number of solutions of which energy saving up to 95% of the optimal solution in the experiments. 97.5% energy saving is defined in similar way. When $\theta = 0.003$, the probability of algorithm achieving the optimal solution is greater than 50%, and the probability of getting a near-optimal solution is more than 90%.

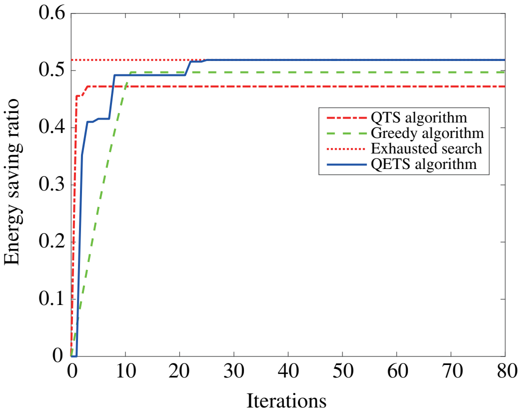

Figure 4 (Color online) The performance of the energy saving algorithm ($h_b = 0.5$, $\rho^{\rm th} = 0.6$, $\lambda(x) = 0.1$, $\theta = 0.003$, $m = 200$).

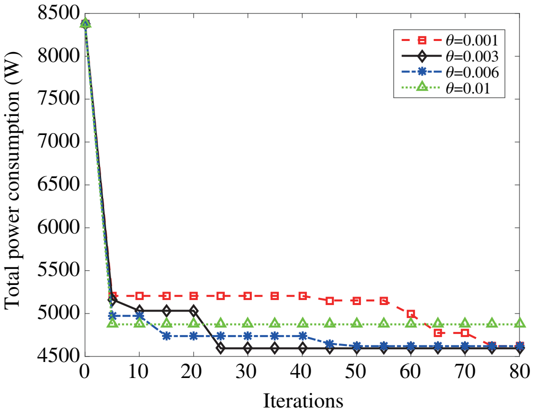

Figure 5 (Color online) Comparison of convergence speed of QETS with θ ($h_b = 0.5$, $\rho^{\rm th} = 0.6$, $\lambda(x) = 0.1$, $\theta = 0.003$, $m = 200$).

Table 2 The result of 500 times QETS experiments with different θ ($h_b = 0.5$, $\rho^{\rm th} = 0.6$, $\lambda(x) = 0.1$, $m = 200$)

θ	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
Optimal solution	84	178	281	130	194	162	152	92	138	109
97.5% energy saving	296	385	412	350	381	346	324	294	338	305
95% energy saving	406	432	451	384	416	399	387	379	363	356

Besides, the number of its optimal solution or near-optimal solution is the largest in different θ. Therefore, when $\theta = 0.003$, QETS is more easily to get the optimal solution or the near-optimal one. It is worth mentioning that, no matter how much θ value we set in QETS, the probability of getting the 95% energy saving solution is more than 62%. It shows that even QETS fails to get the optimal solution, it is still a good energy-saving algorithm very meaningful in practice.

Quantum population can be achieved through m times of the measurement. The objective is to find the best energy saving solutions from the m combinations of the BSs on/off state. The larger the population is, the larger the solution searching space is. And the probability of finding a better solution increases accordingly. However, it will also prolong the calculation time. Intuitively, it can be seen from Figure 6. with the decrease of the population, the most energy saving can be achieved by QETS within 80 iterations on a declining curve. Furthermore, convergence to the optimal solution can be realized when the population is up to 200. When the population continues to increase, calculation time of the algorithm increases significantly. In practice, in order to achieve the tradeoff between execution time and effectiveness of QETS, we set $m = 200$.

Switching off a BS needs to migrate its UE to other BSs, it needs UE handovers and more signaling overhead. In practical network, if the BSs switching off strategy adapts to the dynamic flow of the network, we have to consider whether the signaling overhead threats the network stability. As the BS switching off strategy mainly works when the network traffic is fairly low, the network traffic fluctuation is quite small. Then we only need to monitor the real-time change of network traffic through SDWN controller and restart the strategy when the traffic changes a lot. Furthermore, the collected data in [24] shows that network traffic is in constant change during a certain period of time. According to [5], in a period of time, e.g., an hour, we can assume that the network traffic model is stable, and the implementation of our strategy in each period is of a quasi-static manner based on the average channel gain and expected traffic arrival.

5.2 Energy saving performance

In this part, we discuss the performance of QETS under various parameter configurations in quasi-static environment.

Figure 7 shows the energy saving ratio by different algorithms versus the traffic arrival rate. It can

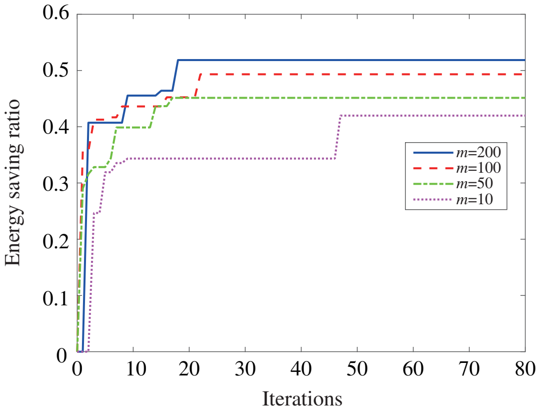

Figure 6 (Color online) The convergence and optimality of QETS with m ($h_b = 0.5$, $\rho^{\text{th}} = 0.6$, $\lambda(x) = 0.1$, $\theta = 0.003$).

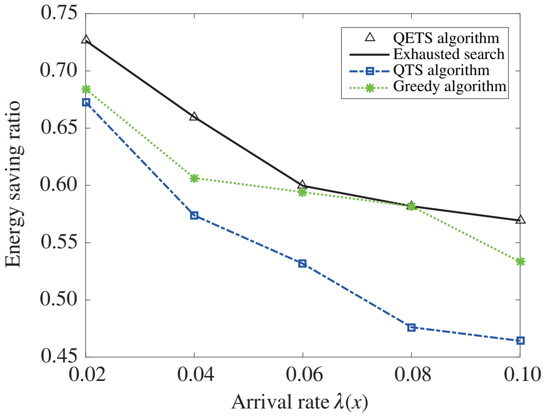

Figure 7 (Color online) The energy saving ratio by different algorithms versus the arrival rate ($h_b = 0.5$, $\rho^{\text{th}} = 0.6$, $\theta = 0.003$, $m = 200$).

be observed that the energy saving ratio decreases with the increasing of the arrival rate.It can be noted that the effect of the greedy algorithm is related to the actual network topology and network load, it is also affected by the starting point of the greedy algorithm. Therefore, different level of network load will lead to certain fluctuation of results. More importantly, with the increase of arrival rate, system load increases, the load threshold will lead to the reduce of potential solution space. Then, expected results can also be attained through greedy searching algorithm. Thanks to the load threshold, less BSs can be switched off. The energy saving ratio decreases with the arrival rate as expected. Besides, we notice that QETS algorithm achieves the optimal performance with the five arrival rate models. Besides, there is increasing gap between QTS and QETS, it is because the quantum mutation mechanism in QTS algorithm has no QETS, it tends to fall into local optimum. When the load increases, the search solution space becomes smaller, it is very easy for QTS algorithm to find a worse solution at the beginning. Even after several iterations, the convergence is achieved. Then the QTS is unable to find a better solution, so the gap between QTS and QETS algorithm increases the network load.

In Figure 8, we illustrate the energy saving ratio achieved by different algorithms versus the proportion of the fixed power consumption. When a BS is switched off, its system load is transferred to neighboring BSs, which gives a rise to the adaptive power consumption of the target BSs, that means a part of the adaptive power consumption of source BS is transferred to the target BSs. Furthermore, the reduction of the fixed power consumption makes great contribution to the energy-saving. As shown in Figure 8, the larger energy saving ratio can be expected from the larger proportion of the fixed power consumption. QETS algorithm achieves the best energy saving performance, the greedy algorithm inferior and QTS worst. Furthermore, when the proportion of the fixed power consumption increases, the gap between the greedy solution and QTS becomes smaller.

Figure 9 shows the energy saving ratio by different algorithms versus the threshold of the system load. Intuitively, more BS on/off switch combination under the load constraints can be achieved when the system load threshold gets larger. In the terms of the ability to accept users, each BS is allowed to serve more users which leads to larger solution space. As shown in Figure 9, the energy saving ratio increases with the system load threshold. Additionally, there is turning down for QTS. That is because the quantum state of QTS easily aggregates to the historical optimal solution and the search space is limited.

6 Conclusion

This paper presents a BS switching off strategy for SDWN, where some of the lightly-loaded BSs is switched off to save energy in cellular networks. It can be easily deployed at the SDWN controller in the form of software. The problem is analyzed and the network model is established. We proposed the QETS

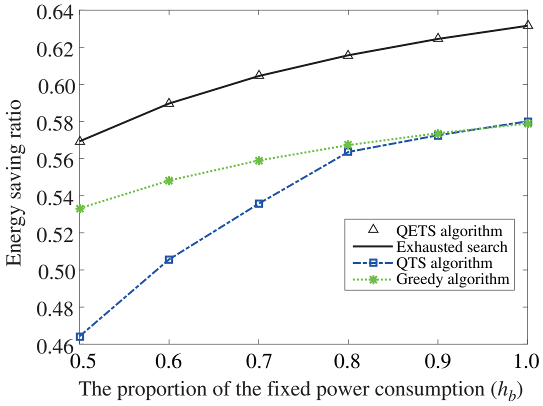

Figure 8 (Color online) The energy saving ratio achieved by different algorithms versus the proportion of the fixed power consumption ($\rho^{\text{th}} = 0.6$, $\lambda(x) = 0.1$, $\theta = 0.003$, $m = 200$).

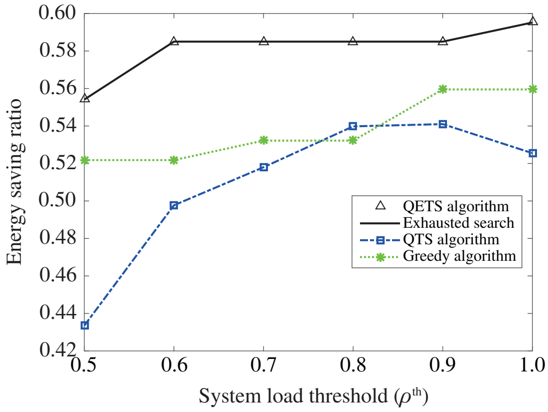

Figure 9 (Color online) The energy saving ratio by different algorithms versus the threshold of the system load ($h_b = 0.5$, $\lambda(x) = 0.1$, $\theta = 0.003$, $m = 200$).

algorithm to search the switching off BS candidates as joint optimization problem. Quantum mutation mechanism based on the quantum entropy is introduced into QTS, which makes it easier to get the optimal solution with high probability of fast convergence. Meanwhile, quantum mutation mechanism greatly expand the searching space of the algorithm. We also adopted the exhaustive search algorithm as the benchmark with the optimal performance, and the proposed QETS, QTS and greedy algorithm are compared by simulation. Simulation results show that the QETS algorithm approaches the optimal performance with higher probability and less iterations.

Acknowledgements This work was jointly supported by National High-Tech R&D Program of China (863) (Grant No. 2015AA01A705) and State Grid (Grant of "Research and Application of Key Technologies in Smart Grid Park Energy Management and Optimization for Smart City").

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Marsan M A, Meo M. Network sharing and its energy benefits: a study of European mobile network operators. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Atlanta, 2013. 2561–2567

2 Wong W T, Yu Y J, Pang A C. Decentralized energy-efficient base station operation for green cellular networks. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Anaheim, 2012. 5194–5200

3 Son K, Kim H, Yi Y, et al. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. IEEE J Sel Areas Commun, 2011, 29: 1525–1536

4 Yaacoub E. Achieving green LTE-A HetNets with D2D traffic offload and renewable energy powered small cell BSs. In: Proceedings of IEEE Online Conference on Green Communications (OnlineGreencomm), Tucson, 2014. 1–6

5 Zheng J C, Cai Y M, Chen X F, et al. Optimal base station sleeping in green cellular networks: a distributed cooperative framework based on game theory. IEEE Trans Wirel Commun, 2015, 14: 4391–4406

6 Niu Z S, Guo X Y, Zhou S, et al. Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control. IEEE J Sel Areas Commun, 2015, 33: 641–650

7 Han F, Safar Z, Liu K J R. Energy-efficient base-station cooperative operation with guaranteed QoS. IEEE Trans Commun, 2013, 61: 3505–3517

8 Wu X C, Wu C M, Lin C T, et al. A multipath resource updating approach for distributed controllers in software-defined network. Sci China Inf Sci, 2016, 59: 092301

9 Hu Y N, Wang W D, Gong X Y, et al. On the feasibility and efficacy of control traffic protection in software-defined networks. Sci China Inf Sci, 2015, 58: 120104

10 Karp R M. Reducibility among combinatorial problems. In: Proceedings of Symposium on the Complexity of Computer Computations, New York, 1972. 85–103

11 Chiang H P, Chou Y H, Chiu C H, et al. A quantum-inspired tabu search algorithm for solving combinatorial optimization problems. Soft Comput, 2013, 18: 1–11

12 Bernardos C J, De L O A, Serrano P, et al. An architecture for software defined wireless networking. IEEE Wirel Commun, 2014, 21: 52–61

13 Jiang X X, Du D H C. PTMAC: a prediction-based TDMA MAC protocol for reducing packet collisions in VANET. IEEE Trans Veh Technol, 2016, 65: 9209–9223

14 Zhou Z Y, Ota K, Dong M X, et al. Energy-efficient matching for resource allocation in D2D enabled cellular networks. IEEE Trans Veh Technol, 2016, doi: 10.1109/TVT.2016.2615718

15 Yao Y, Cheng X, Yu J, et al. Analysis and Design of a Novel Circularly Polarized Antipodal Linearly Tapered Slot Antenna. IEEE Trans Antenn Propag, 2016, 64: 4178–4187

16 Oh E, Son K, Krishnamachari B. Dynamic base station switching-on/off strategies for green cellular networks. IEEE Trans Wirel Commun, 2013, 12: 2126–2136

17 Hossain M F, Munasinghe K S, Jamalipour A. Distributed inter-BS cooperation aided energy efficient load balancing for cellular networks. IEEE Trans Wirel Commun, 2013, 12: 5929–5939

18 Auer G, Giannini V, Desset C, et al. How much energy is needed to run a wireless network? IEEE Wirel Commun, 2011, 18: 40–49

19 Loss D, Divincenzo D P. Quantum computation with quantum dots. Phys Rev A, 1997, 57: 120–126

20 Glover F, Marti R. Tabu search. Gen Inform, 1998, 106: 221–225

21 Han K H, Kim J H. Quantum-inspired evolutionary algorithms with a new termination criterion, H_ε gate, and two-phase scheme. IEEE Trans Evol Computat, 2004, 8: 156–169

22 IEEE 802.16m evaluation methodology document (EMD). IEEE: Technical Report. IEEE 802.16m-08/004r5, 2009

23 Son K, Oh E, Krishnamachari B. Energy-aware hierarchical cell configuration: from deployment to operation. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, 2011. 289–294

24 Marsan M A, Chiaraviglio L, Ciullo D, et al. Optimal energy savings in cellular access networks. In: Proceedings of IEEE International Conference on Communications Workshops, Dresden, 2009. 1-5