IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 12, 2017, accepted January 12, 2018, date of publication January 23, 2018, date of current version May 2, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2796126

Dynamic Femtocell gNB On/Off Strategies and Seamless Dual Connectivity in 5G Heterogeneous Cellular Networks

XIAOGE HUANG[1], SHE TANG[1], QIAN ZHENG[2], DONGYU ZHANG[1], AND QIANBIN CHEN[1]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]Vivo Communication Technology Co. Ltd., Dongguan 523000, China

Corresponding author: Xiaoge Huang (huangxg@cqupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61401053, in part by the 863 Project under Grant 2014AA01A701, in part by the Changjiang Scholars and Innovative Research Team in University under Grant IRT1299, in part by the Special Fund of Chongqing Key Laboratory, in part by the Advanced and Applied Basic Research Projects of Chongqing under Grant cstc2015jcyj40048, and in part by the Innovation Project of the Common Key Technology of Chongqing Science and Technology Industry under Grant cstc2015zdcyztzx40008.

ABSTRACT To meet the drastic growth of the mobile traffic, 5G network is designed to optimize the transmission efficiency and provide higher quality of service (QoS). Small cell is considered as a promising and feasible approach to meet the increasing traffic demand. For this purpose, in this paper, we study a dynamic Femtocell gNB (F-gNB) ON/OFF strategies in 5G heterogeneous cellular networks (HCNs), which aims at maximizing the network energy efficiency (NEE) by optimizing jointly the traffic load prediction, the cell association, and the dynamic F-gNB ON/OFF strategies with respected to the time-varying traffic load, while taking into account the load balancing and the outage probability of the network. However, the optimization problem is a nonconvex problem. In order to relax the computation complexity, the original optimization problem is divided into two steps: cell association with load balancing (CALB) scheme and energy efficiency-based dynamic F-gNBs ON/OFF (DFOO) strategies. Specifically, the proposed CALB scheme could guarantee the load balancing as well as the minimum signal-to-interference-plus-noise ratio requirement of user equipments (UEs). In addition, the proposed DFOO strategies consider the operation of base stations (BSs) according to the predicted time-varying traffic load from Markov procedure. Furthermore, dual connectivity-based seamless handover procedure is introduced to guarantee the transmission QoS of UEs. Simulation results illustrate that the proposed DFOO algorithm provide considerable improvement of NEE while ensuring the load balancing of the HCN.

INDEX TERMS Dual connectivity, dynamic F-gNBs ON/OFF strategies, heterogeneous cellular network, load balancing, network energy efficiency.

I. INTRODUCTION

Traffic will increase 1000× in the next decade. At the same time, users require faster data transmission experience and better QoS. 5G, a new generation mobile communication technology, applying in Enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLCC) and Massive Machine-Type Communications (mMTC), could ensure data-rate, latency, reliability, and energy efficiency [1]–[2]. To improve the communication performance, it needs to apply various technologies, such as ultra-dense network, massive MIMO, filter bank multicarrier (FBMC) and so on. Small cell networks which include microcells, picocells, and femtocells are considered as a promising solution to deal with the increasing traffic demands with low-cost, low power. Typically, when the traffic load of the macrocell is dramatically increasing, small cell networks could offload parts of the traffic load to prevent network from congesting and collapsing. However, a large number of small cells makes the cellular network extremely complicated and results in higher energy consumption. Consequently, seeking for high NEE is a trend for the next generation wireless communication. Designing a reasonable resource allocation scheme will be propitious to improve NEE [3]–[5].

A. RELATED WORK

The cell association scheme plays an important role in the performance of HCNs. Recently, there were various cell association schemes designed for HCNs [6]–[9]. Abbas *et al.* [6] allowed users belongs to the overload MBS to offload traffic load to small cell networks. The authors designed an analytical network model to relieve inter-cell interference (ICI) under the reverse frequency allocation (RFA) scheme. Muhammad *et al.* [7] developed a non-uniform heterogeneous cellular network (NuHCN) and proposed selection scheme that selectively mutes some small cells and covers end users by cell biasing to achieve the load balancing. Ao and Psounis [8] proposed an online algorithm to associate users with BSs in HCNs. Pervez *et al.* [9] designed a fuzzy Q-learning-based user-centric backhaul-aware cell association scheme. The scheme optimized user's association process in a context-aware and backhaul-aware manner.

In order to further improve NEE, the authors considered to maximize users' data rate by optimizing the detection operation and the power allocation, while taking into account the impact of the sensing accuracy and the interference limitation to primary users in [10]. As one of the most available ways for energy saving, BS on/off strategies were initially referred to in IEEE 802.11b. Recently, amount of existing literature discussed power saving in HCNs with BSs on/off strategies [11]–[15]. Inter-eNB energy saving and Inter-RAT energy saving scenarios were introduced in [11]. In [12], a scheme was proposed to maximize NEE by turning off base stations whose NEE is lower than a certain threshold. Su *et al.* [13] proposed BSs on/off strategies for multi-antenna multi-carrier small cell networks to reduce power consumption and maximize NEE. Kong *et al.* [14] considered the optimization problem by turning on only a fraction of BSs according to the activation ratio to minimize the network average power consumption per area in HCNs. In addition, an energy efficiency optimization problem was solved by alternating direction method, namely, rate maximization and power minimization respectively in [15]. Interference pricing mechanism was introduced to reduce the inter-cell interference and achieve a higher network performance. Cheng *et al.* [16] designed the statistical delay-bounded power allocation scheme to maximize the effective power efficiency (EPE). EPE is defined as the statistical-QoS-guaranteed throughput per unit power. All prior works focused on the tradeoff between energy consumption and transmission quality. However, there are very few efforts that trying to analyze the BS actions based on dynamic UEs behaviors jointly considering UEs behaviors prediction.

Due to the mobility of UEs as well as the associated dynamic action of BSs, the transmission quality during the handover procedure should be guaranteed. Realizing the seamless handover is one of the most important technologies which was discussed in [17]–[20]. In [18] and [19], a ''make-before-break'' scheme is proposed. In this scheme, UE should keep its connection with the source gNB (the name of the BS in 5G) until it is able to receive packets from the target gNB. i.e. the source gNB should continue to serve the UE parallelly with the target gNB until the target gNB is ready for the UE. In [20], a seamless handover procedure based on dual connectivity (DC) was discussed. The target gNB was considered as a secondary gNB (SgNB), and the source gNB was considered as the master gNB (MgNB) which initiates handover procedure. Finally, the SgNB replaced the MgNB, and released the connection between the UE and the original MgNB.

B. CONTRIBUTION

In this paper, we investigate the tradeoff between energy consumption and data rate with dynamic F-gNBs on/off strategies, while taking into account the load balancing and the QoS requirement of the HCN. Firstly, the traffic load of UEs in the hot region is formulated as the tidal effect which could be predicted by the proposed enhanced Markov-based scheme. In addition, femtocell gNBs (F-gNBs) on/off strategies with respect to UEs behaviors as well as the load balancing-based cell association is proposed. The main contributions of the paper are the following:

1) Firstly, traditionally, the traffic load with respected to UEs behaviors is considered as a Poisson distribution. However, this model is too idealized for the real scenario, especially in the hot region in which the traffic load of UEs should be considered the tidal effect. In this paper, an enhanced Markov-based prediction (EMP) scheme is proposed to predict the timely UEs traffic load in the hot region.
2) Secondly, the load-balancing is taken into consideration according to the dynamic traffic load. A load balancing-based cell association algorithm is implemented which optimizes the UEs-BSs association while ensuring the transmission QoS of the HCN.
3) Thirdly, in order to further maximize NEE under the consideration of the F-gNB model, a dynamic F-gNBs on/off strategies is discussed which could optimize the tradeoff between energy consumption and data rate under the QoS requirement.
4) Finally, a seamless handover scheme based on dual connectivity is proposed to ensure the handover between the source gNB and the target gNB.

The rest of the paper is organized as follows. Section 2 describes the system model. Section 3 presents the enhanced Markov-based prediction scheme, the load balancing-based cell association algorithm, the dynamic F-gNBs on/off strategies as well as the seamless handover procedure respectively. In Section 4, simulation results are presented and discussed. Finally, Section 5 concludes the paper.

II. SYSTEM MODEL

In this section, we introduce the network scenario, the tidal effect, the femtocell gNBs on/off model, and the path loss model.

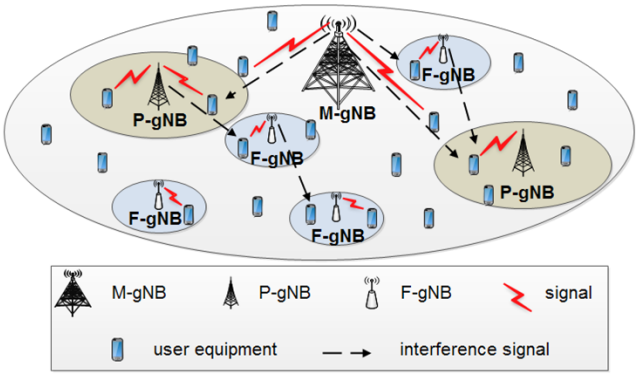

FIGURE 1. Network scenario model.

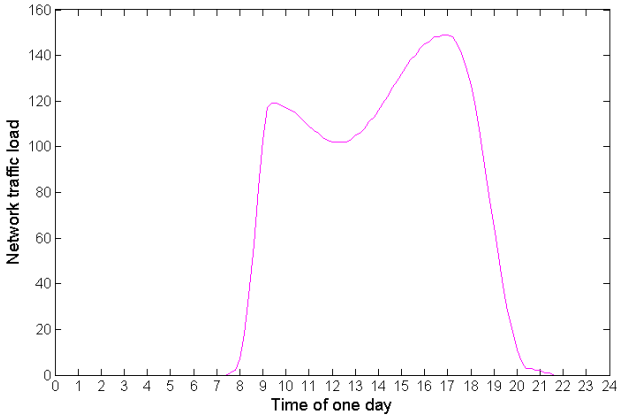

FIGURE 2. The network traffic load is changed by the time during a day.

A. NETWORK SCENARIO

We consider a 5G heterogeneous cellular network with 3 tiers, macrocell gNB, picocell gNB and femtocell gNB. As shown in Figure 1, each tier models is a particular type of gNBs: tier 1 is consisted of macrocell gNBs (M-gNB), tier 2 is comprised of picocell gNBs (P-gNB) and tier 3 is composed of femtocell gNBs (F-gNB). Different types of gNBs have different radius of coverages. P-gNBs, F-gNBs and UEs are randomly distributed in the coverage of the M-gNB. The total number of gNBs and UEs is M and K respectively. The transmission power of gNB m is denoted as P_m, $1 \leq m \leq M$. In the scenario, M-gNB and P-gNB hold "on" mode all the time. F-gNBs are designed to have two working modes, i.e. "on" and "off". The state of a gNB is γ_m. $\gamma_m = 1$ means gNB m is in "on" mode, otherwise $\gamma_m = 0$. The state of a F-gNBs are related with the data traffic, the number of UEs as well as parameters which could effect the transmission QoS.

B. TIDAL EFFECT

In the communication system, the tidal effect refers to the fluctuation of the traffic load in the hot region, such as the traffic load of the Central Business District (CBD) area between working hours and closing hours due to the dynamic migration with respect to the time. Specially, the traffic load of UEs is modeled as the tidal effect model in this paper, the peaks and troughs of the traffic load is changing during the time. For instance, in the morning, from 8:00 a.m. to 12:00 a.m., the traffic load is an increasing function with the time which comes from the indoor activity of UEs, such as working, shopping and so on. In order to meet the large traffic load requirement during the working hours, the F-gNBs and P-gNBs are considered as a better option which could ensure the transmission QoS in the ultra dense network. Particularly, in the rest time, the traffic load requirement decreases sharply, and turn to 0 during the night time. The traffic load in the hot region is formulated to the tidal effect, as shown in Figure 2.

C. F-GNB ON/OFF POWER CONSUMPTION MODEL

The power consumption mode of the hardware model for the BS is presented in [21], which consists of microprocessor block, power amplifier (PA) block, radio frequency (RF) block and field-programmable gate array (FPGA) block. The total hardware power consumption can be expressed as:

$$P_{total} = P_{mp} + P_p + P_t + P_f \tag{1}$$

where P_{mp}, P_p, P_t and P_f are the power consumption of the microprocessor, the PA, the RF and the FPGA, respectively.

We consider two modes for F-gNBs operation: "on" and "off" mode.

1) "On" mode: a F-gNB is in full operation, the power consumption is P_{total}.
2) "Off" mode: a F-gNB is turned down, but it still consumes a certain amount of power to maintain the function to be activated.

Therefore, the power consumption model is expressed as:

$$E_m(\gamma_m) = \alpha P_m \gamma_m + (1-\beta)P^0\gamma_m + \beta P^0 \tag{2}$$

where m denotes the index of F-gNBs and $\gamma_m = 0, 1$ represents the state of F-gNB m. Here, α is the portion of the power consumption due to feeder losses and power amplifier. β indicates the inactive level, where $0 < \beta < 1$. P_m is the transmission power consumption of F-gNB m. P^0 represents the baseline power consumption of F-gNB m in "off" mode. In this paper, the total consumption power of F-gNB m is $E_m = \beta P^0$, which is approximated to zero when F-gNB m is turned off (i.e. $\gamma_m = 0, \beta \to 0$). Otherwise, $E_m = P^0 + \alpha P_m$ when F-gNB m is working (i.e. $\gamma_m = 1$).

D. PATH LOSS MODEL

The HCN is consisted of k tiers, such as the M-gNB, P-gNBs, and F-gNBs. BSs belong to different tiers may have different transmission power, spatial density as well as transmission range. Therefore, the received SINR from gNB m is given by:

$$\lambda_k^m = \frac{P_m L(d_{k,m})\gamma_m}{\sum\limits_{m'=1,m'\neq m}^{M} P_{m'} L(d_{k,m'})\gamma_{m'} + \delta_n^2} \tag{3}$$

where δ_n^2 is the noise power, $d_{k,m}$ is the distance between UE k ($1 \leq k \leq K$) and gNB m, $L(\cdot)$ is the path loss function. The path loss model of M-gNBs, P-gNBs and F-gNBs are formulated respectively as follow:

$$\begin{aligned} L_M(d) &= 34 + 40\log_{10}(d)dB \\ L_P(d) &= 34 + 40\log_{10}(d)dB \\ L_F(d) &= 37 + 30\log_{10}(d)dB \end{aligned} \tag{4}$$

According to Shannon's formula, the achievable rate of gNB m can be calculated as:

$$R_m = \sum_{k=1}^{K} B \cdot \log_2(\lambda_k^m) \tag{5}$$

III. LOAD BALANCING-BASED DYNAMIC F-GNB ON/OFF STRATEGIES

In this paper, we aim to maximize NEE while taking into account load balancing, cell association, dynamic F-gNBs on/off strategies and transmission QoS by optimizing the tradeoff between energy consumption and data rate. The optimization problem could be divided into three sub-problems, namely, UEs behavior prediction based on Markov model, load balancing-based optimal cell association and dynamic F-gNB on/off strategies.

A. UEs BEHAVIOR PREDICTION BASED ON MARKOV PROCEDURE

Various prediction algorithms have been proposed in [22] and [23]. Fazio *et al.* [22] proposed a pattern prediction algorithm based on Hidden Finite State Markov Chains (HFSMC) which considers UEs handover direction. The probabilistic link prediction based on Markov model which lies in its integration of multiple timescales with local and semi-global correlated structural evolution, in tandem with the interactions occurring in real-world social networks was introduce in [23]. In this paper, an enhance Markov-based prediction (EMP) scheme is proposed to predict the timely UEs traffic load in the hot region. In the hot region, the mobile traffic load can be formulated as the tidal effect model in which the peaks and troughs of the traffic load is time-varying. In order to predict the dynamic behavior of UEs, EMP algorithm is introduced which could estimate the traffic load of HCNs according to the mobility of UEs, and calculate the instantaneous transmission QoS based on current network state. The prediction procedure can be concluded as follow:

1) Step 1: Generate the transition matrix $\prod_{K\times K}(t)$ at time t according to learning time and the historical state of UEs traffic load.
2) Step 2: Calculate the total traffic load of the network at time $t+1$ based on the traffic load state and transition matrix at time t.

We assume that the traffic load state of the network has K levels. Apparently, the set of the state is denoted as $\{1, 2, \ldots, K-1, K\}$. The state transition process is shown in Figure 3.

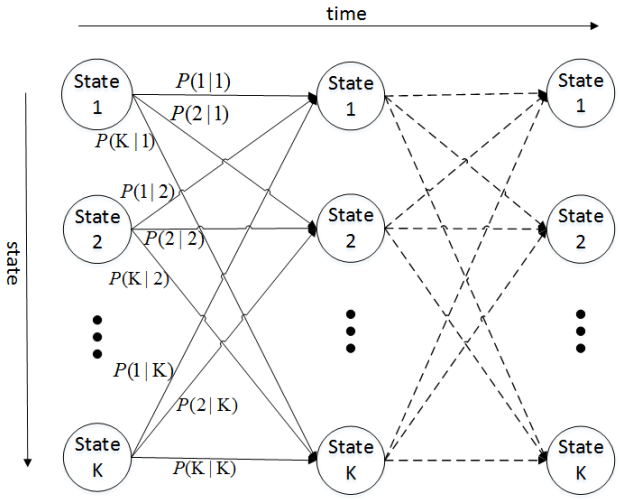

FIGURE 3. State transition diagram.

The learning time T could be divided into T_s learning intervals. Each learning interval contents Q individual slots. s_t^i denotes the traffic load state at leaning interval t ($0 < t \leq T_s$), slot i ($0 < i \leq Q$). The network traffic load can be described by matrix as follow:

$$\begin{array}{c} \\ 1 \\ \vdots \\ t \\ \vdots \\ T_s \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & \cdots & i & \cdots & Q \end{array} \\ \begin{pmatrix} s_1^1 & \cdots & s_1^i & \cdots & s_1^Q \\ \vdots & \ddots & & & \vdots \\ s_t^1 & & s_t^i & & s_t^Q \\ \vdots & & & \ddots & \vdots \\ s_{T_s}^1 & \cdots & s_{T_s}^i & \cdots & s_{T_s}^Q \end{pmatrix} \end{array}$$

The network traffic load state transition can be expressed as:

$$P(s_{t+1}^i = k' \mid s_t^i = k) = \frac{F(s_{t+1}^i = k')}{F(s_t^i = k)} \tag{6}$$

where $F(s_t^i = k)$ calculates the number of the traffic load state $s_t^i = k$ in network traffic load matrix. Therefore, the transition matrix could be written as:

$$\prod_{K\times K}(t) = \begin{pmatrix} P(1 \mid 1) & \cdots & P(K \mid 1) \\ \vdots & \ddots & \vdots \\ P(1 \mid K) & \cdots & P(K \mid K) \end{pmatrix}$$

Assume that the probability of the state at learning interval t is $P(t) = \{P_t(1), P_t(2), \ldots, P_t(K)\}$. Then, the probability of the state at learning interval $t+1$ can be calculate by $P(t+1) = P(t) \times \prod_{K\times K}(t)$.

Based on the EMP algorithm, we could obtain the network traffic load at $t+1$ learning interval and dynamically adjust F-gNBs on/off strategies with respect to UEs behavior to maximize NEE while satisfying QoS constraints.

B. LOAD BALANCING-BASED OPTIMAL CELL ASSOCIATION SCHEME

In order to maximize NEE of HCNs while considering the target SINR, the load-balancing and the outage probability, we propose a optimal cell association load balancing (CALB) algorithm. The utility function is defined as:

$$\omega_k^m = \theta_k^m \cdot NEE_k^m \tag{7}$$

where θ_k^m is the access factor which is related to the available resource and the scheduling method, showing the probability of UE k successfully access to gNB m. NEE_k^m is the energy efficiency of UE k belongs to gNB m. Let L_m define the current load of gNB m, L_m^{max} define the maximum acceptable load of gNB m, and L_m^0 define the maximum number of UEs of gNB m could be served in each Transmission Time Interval (TTI). If $L_m > L_m^0$, gNB m is not considered as an over-loaded gNB in the following TTI, some UEs could be selected due to the dynamic LB scheduling scheme. A transmission time in the LTE include two slots, according for 1ms. By Rounding Robin Scheduling, we have $L_m^{max} = 2L_m^0$. Consequently, θ_k^m is updated by the follow equations:

$$\theta_k^m = \begin{cases} \dfrac{L_m^{\max} - L_m}{L_m^{\max}} & if \quad L_m < L_m^0 \\ \dfrac{L_m^{\max} - L_m}{L_m^{\max}} \dfrac{L_m^0}{L_m} & if \quad L_m \geq L_m^0 \end{cases} \tag{8}$$

The gNB with smaller traffic load will get larger access factor (more likely to be selected it). When a gNB is full-loaded, θ_k^m reduces to zero, and users cannot connect with the overload gNB. In this paper, NEE_k^m is the energy efficiency of UE k from gNB m.

$$NEE_k^m = \frac{R_k^m}{E_m} = \frac{B \cdot \log_2(1 + SINR_k^m)}{E_m} \tag{9}$$

To maximize NEE, we propose an optimal cell association load-balancing (CALB) algorithm. For UE k, the main steps of the CALB algorithm can be concluded as:

- Step 1: Calculate the candidate gNBs which could meet the SINR constraint for UE k, and establish the candidate cell list C_k^m. The freedom degree $\phi(k) = |C_k^m|$ is denoted as the size of the candidate cell.
- Step 2: Repeat step 1, until every UE obtains its candidate cell list. Broadcast the candidate cell list information to the nearby gNB.
- Step 3: The UE with smaller freedom degree has higher priority to select a gNB. During cell selection process, gNBs will constantly send feedback to UEs with the updated gNBs information which includes the current load, the maximum load and downlink transmission power.
- Step 4: Calculate θ_k^m and ω_k^m based on the feedback information.
- Step 5: Associate to the gNB with largest ω_k^m and send access requirment to this gNB.

Let $\Phi(k)$ denote the freedom degree of UE k, which is the number of F-gNB m could be selected by UE k under the SINR constraint. The proposed CALB algorithm is described in Algorithm 1.

Algorithm 1 Optimal Cell Association Load-Balancing (CALB) Algorithm

Initialization:
$\Phi = 0_{K\times 1}, X = 0_{K\times M}$;
Cell Selection
for $1 \leq k \leq K$ **do**
  Generate the candidate list for UE k, $C_k^m = \{1 \leq m \leq M \mid SINR_k^m \geq SINR_k^{thr}\}$;
  $\Phi(k) = |C_k^m|$;
  for $k \in C_k^m$ **do**
    Broadcast C_k^m to the nearby gNBs;
  end for
end for
Cell Association
repeat
Find UE k which has minimal freedom degree, $k^* = arg_k \min(\Phi)$;
Calculate $\omega_k^m = \theta_k^m \cdot NEE_k^m$;
$m^* = arg_m \max(\omega_{k^*}^m)$;
Associates UE k^* with gNB m^*, $X(k^*, m^*) = 1$;
$L_{m^*} = L_{m^*} + 1$;
if $L_{m^*} = L_m^{max}$
for $m^* \in C_k^m$ **do**
  UE k $(1 \leq k \leq K)$ deletes the gNB m^* in its list, $C_k^m = 0$;
end for
end if
Update $\Phi(k)$, $|C_k^m|$.
until all UEs associate with gNB

C. DYNAMIC F-GNB ON/OFF STRATEGIES

Depend on the solution from CALB algorithm, we could obtain the optimal cell association scheme. A dynamic F-gNBs on/off (DFOO) strategies is proposed in this section. The optimization of maximizing NEE with DFOO strategies under the load-balancing and the target SINR constraints can be formulated as:

$$\begin{aligned} \max_{\gamma, X} \; & NEE = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K} R_k^m \theta_k^m}{\sum_{m=1}^{M} E(\gamma_m)} \\ s.t. \; & C_1: \; X_k^m \in (0,1), \quad \forall k, \; m \\ & C_2: \; \sum_{m=1}^{M} X_k^m \leq 1, \quad \forall k, \; m \\ & C_3: \; \sum_{m=1}^{M} X_k^m \leq L_m^{max}, \quad \forall k, \; m \\ & C_4: \; P_{out} < \varrho \\ & C_5: \; SINR_k^m \geq SINR_k^{th}, \quad \forall k, \; m \end{aligned} \tag{10}$$

where $R_k^m \theta_k^m$ represents the effective transmission data of the UE k. The constraint C_1 indicates whether UE k connects with gNB m. The constraint C_2 ensures that a UE can only connect with one gNB simultaneously. The constraint C_3 means gNB m cannot be overloaded, the maximum number of UEs could be associated with gNB m is L_m^{max}. The constraint C_4 denotes that the outage probability of the total network could not exceed a given threshold. The constraint C_5 ensures the UE k could be served by gNB m, if $SINR_k^m$ is larger than the SINR threshold of UE k.

To decrease the power consumption and ensure the transmission requirement with respect to the dynamic traffic load of UEs, the DFOO strategies consist of two steps: turning on and turning off. Generally, the optimal F-gNB on/off strategies can be found by exhaustive search which results in high complexity and time-consuming. To deal with this problem the proposed DFOO makes the reversed direction decision based on the previous decisions. The DFOO strategies is summarized by the following steps.

''Turning off'' process:

- Step 1: Based on the CALB algorithm, find F-gNB m with the minimum NEE. Turn off F-gNB m.
- Step 2: Repeat CALB algorithm, update NEE of F-gNBs and sort in descending order.
- Step 3: Repeat Step1 and Step 2 until obtain the maximum NEE which satisfies the outage probability constraint.

Notice that in the ''hot region'' and the ''hot time'' the down link traffic load is changing during the time which could be model as the tidal effect model. Once the transmission QoS can not be guaranteed according to the increasing traffic load requirement, we decide to turn on parts of sleeping F-gNBs.

''Turning on'' process:

- Step 1: Estimate the traffic load in the coming interval according to the EMP.
- Step 2: If the traffic load has a drastic increasing, turn on parts of sleeping F-gNBs to ensure the load-balancing and the outage probability constraint of the network.
- Step 3:Turn on F-gNBs with respect to the sleeping order of F-gNBs, thus the lastest sleeping F-gNB has the highest priority to be turned on. Repeat CALB algorithm and update the association between UEs and available gNBs.
- Step 4: Repeat Step 2 and Step 3 until obtain the maximum NEE.

According to the DFOO strategies, the HCN can reasonably and dynamically adjust F-gNBs ''on'' mode or ''off'' mode, while reducing the computation complexity and time-consuming. More precisely, the dynamic F-gNB on/off strategies is described in Algorithm 2.

Algorithm 2 Dynamic F-gNBs On/Off (DFOO) Strategies

Initialization:
$t = 0, x = 1$, gNB $\Gamma = 0_{1\times M}$, The state of gNB $\gamma = I_{1\times M}$;
Turning off
repeat
Caculate the $NEE(\gamma)$ by CALB algorithm;
Caculate the NEE^m for all active gNBs based on (10).
Find the gNB with minimal NEE, $m^* = arg_m \min NEE^m$;
Sort gNB in decreasing order, and switch off in turns, $\gamma(m^*) = 0, \Gamma(x) = m^*$;
$x = x + 1, t = t + 1$;
until $NEE(\gamma_{t-1}) > NEE(\gamma_t)$
return $\Gamma = \Gamma_{t-1}, \gamma = \gamma_{t-1}, x = x - 1$;
Turning on
Execute EMP algorithm;
if $P_{out} \geq \varrho$
repeat
Turn on gNB with reversed direction $m = \Gamma(x)$, $\gamma(m) = 1$;
Execute CALB algorithm;
Caculate $NEE(\gamma)$;
$x = x - 1$;
until maximum NEE
end if

D. DUAL CONNECTIVITY BASED SEAMLESS HANDOVER

Offering urgent data transmission with low latency and massive data transmission are the main challenges for 5G HCNs. In the traditional handover process, when a source gNB (S-gNB) sends handover request signal to a target gNB (T-gNB), UE will be interrupted from the source gNB until Random Access Channel (RACH) process is completed. During this process, UE needs to obtain mobility control information, resource allocation in downlink and uplink, and synchronisation signal about the target gNB from the source gNB. In this paper, in order to decrease the transmission latency and realize seamless handover, we propose a dual connectivity based seamless handover scheme (DCHO) to deal with UEs behavior. The DCHO procedure is divided into handover preparation step, handover execution step and handover completion step as shown in Figure 4. When F-gNB m is selected to be turned off, the user k belong to F-gNB m will start the following handover process.

Handover preparation:

- Step 1: According to DFOO algorithm, UE k should connect to T-gNB m'. UE k sends the measure information to S-gNB m which includes RSRP and RSRQ of cells, tracking area code, and PLMN identity list.
- Step 2-4: S-gNB m decides to add the T-gNB m' as a secondary gNB and sends the response request to T-gNB m', T-gNB m' returns the request ACK to S-gNB m.

Handover execution:

- Step 5: S-gNB informs UE k the RRC reconfiguration.(i.e. Additional function from UE RRC layer to deal with signals from T-gNB m').
- Step 6-7: Inform S-gNB m and T-gNB m' to be prepared.
- Step 8: UE k connects to T-gNB m' through RACH. UE k can receive data from T-gNB m'.

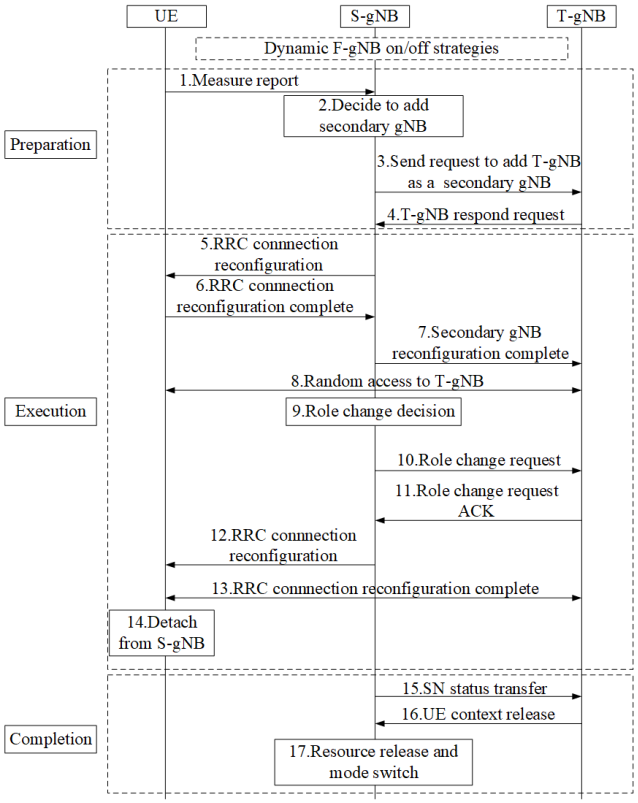

FIGURE 4. **Handover signal process based on the DC architecture.**

- Step 9-11: S-gNB m exchanges the role with T-gNB m'.
- Step 12: S-gNB m informs UE k the RRC reconfiguration (i.e. UE RRC layer only deal with signals from T-gNB m').
- Step 13-14: UE k informs T-gNB m' that the RRC reconfiguration completes and detaches from S-gNB m.

Handover complete:

- Step 15: S-gNB m transfers the rest of the data to T-gNB m'.
- Step 16-17: T-gNB m' informs S-gNB m to release the context of UE k stored in S-gNB m. Then, S-gNB m will be turned off.

In the handover preparation step (i.e. step 1-4), UE k needs to establish RLC and MAC layers connect to T-gNB m' before adding T-gNB m' as the secondary gNB, and establish one PDCP layer link to receive and handle SDUs (Service Date Unit) from two RLC layers. For the handover execution step (i.e. step 5-14), the role will exchange between S-gNB m and T-gNB m'. Before the executing handover step, the PDCP layer bear in S-gNB m is split, and PDCP PDUs (Protocol Data Unit) can be transmitted from S-gNB m to T-gNB m'. Therefore, UE k can receive data from both S-gNB m and T-gNB m'. After step 13, the PDCP layer bear in T-gNB m' is split, and PDCP PDUs can be transmitted from T-gNB m' to S-gNB m. Finally, in the handover completion step (i.e. step 15-17), UE k has detached from the S-gNB m which means data bear is not spilt from T-gNB m' to S-gNB m. Based on the above discussions, the data transmission during handover process is not interrupted in dual connectivity based seamless handover procedure.

IV. SIMULATION RESULT

In this section, we analyse the performance of various proposed algorithms in previous section. In the simulation, we assume the M-gNB is located in the center of a $200m$ radius area, and other gNBs and UEs are randomly distributed in this area. A comparison between the proposed DFOO algorithm and the baseline algorithms in the literature is shown to evaluate the system performance in various aspects. The simulation parameters are described in Table.1.

TABLE 1. **Simulation parameter.**

Parameter name	Value
Macro cell radius	$200m$
M-gNB transmission power	$46dBm$
P-gNB transmission power	$35dBm$
F-gNB transmission power	$20dBm$
M-gNB load (L_m^{max}, L_m^0)	$80, 40$
P-gNB load (L_m^{max}, L_m^0)	$16, 8$
F-gNB load (L_m^{max}, L_m^0)	$8, 4$
probability of outage threshold	0.01
noise power of AWGN	$-174dBm/HZ$

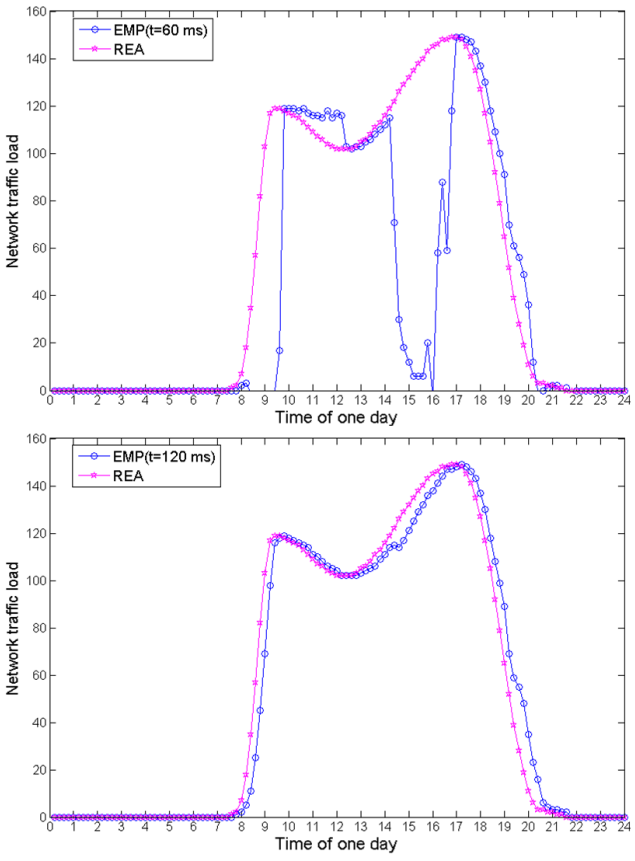

FIGURE 5. **The prediction accuracy with different learning time 60ms and 120ms.**

Figure 5 shows that the traffic load gap between the EMP algorithm and the realistic scenario. As shown in the figure,

compared with the realistic scenario, the accuracy of the prediction algorithm is related with the length of the learning time. Notice that if the learning time is shorter than a certain value, the prediction result is not reliable. Consequently, the prediction result has a significant influence on the CALB algorithm and DFOO strategies, which could cause a serious decreasing in system performance. However, longer learning time causes the increasing in complexity of the EMP algorithm. According to the simulation result the learning time $T = 120ms$ is preferable than $T = 60ms$.

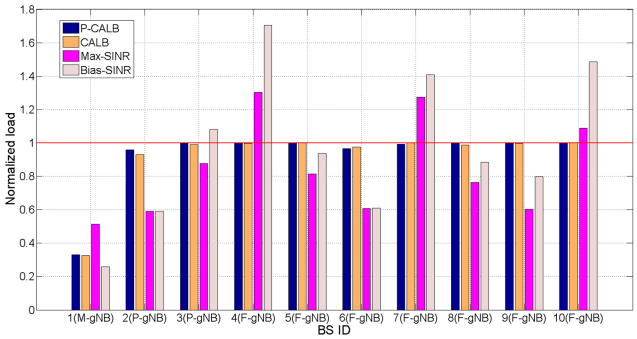

FIGURE 6. Normalize load of different algorithms.

In Figure 6, we compare the normalized traffic load of gNBs with the proposed P-CALB/CALB algorithm, the Max-SINR algorithm and the Bias-SINR algorithm. In the Max-SINR algorithm UE k selects gNB m on the basis of receiving SINR from the gNB. The gNB m should not only meet the SINR threshold, but also provide the maximum SINR to UE k among all gNBs. In the Bias-SINR algorithm, based on the bias factor, UEs would associate to the gNB with lower transmission power. Both the Max-SINR and Bias-SINR algorithms are similar to the greedy algorithm which only considers UEs benefits instead of the load-balancing of the whole network. The result shows that the normalized traffic load of the gNB 4, 7, 10 are overloaded by the Max-SINR scheme. Similarly, the normalized traffic load of the gNB 3, 4, 7, 10 are overloaded by the Bias-SINR scheme. Compared with the traditional cell association algorithm, the proposed CALB which could achieve the load balancing (the normalized load of each gNB is less or equal to 1). In addition, in the P-CALB algorithm, we firstly obtain the traffic load by the EMP algorithm, and then follow the same steps as the CALB algorithm based on the predicted traffic load. It is observed that the P-CALB also could achieve the load balancing.

Figure 7 illustrates the performance of network throughput, energy consumption (EC), and network energy efficiency (NEE) in different algorithms. In the no association off algorithm (NAO) cell-association is based on the Max-SINR algorithm, and F-gNBs without traffic load will be turned off. In the lowest association off algorithm (LAO), cell-association is based on the Max-SINR algorithm, and half of

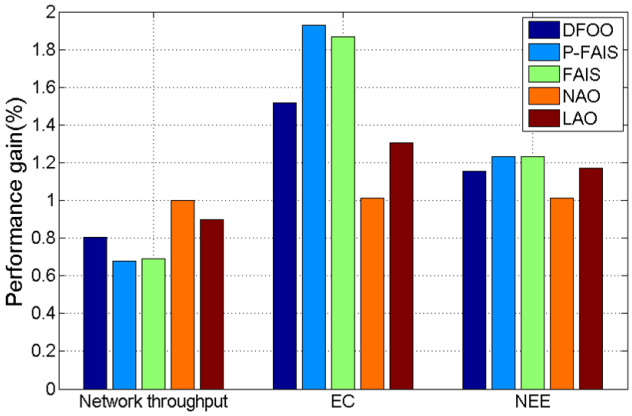

FIGURE 7. Performance gain of different algorithms.

F-gNBs with lower traffic load will be turned off. NAO algorithm is considered as the baseline algorithm. In the optimal FBSs active/ideal switch strategies control algorithm (FAIS), we sort the EE of gNBs by a decreasing order, and turn off the gNB in turns from the bottom. Similarly, the EMP algorithm is used to predict the traffic load of the network in P-FAIS algorithm. Specially, the proposed DFOO algorithm consists of two steps: turn on and turn off. When network traffic load is low, We could turn off some F-gNBs to improve NEE. When the traffic load of the network is increasing according to the time, we should turn on parts of sleeping gNBs to ensure the transmission QoS. The gNBs which are turned off in the latest time will be woken up firstly. Among those algorithms, NAO algorithm could achieve the highest NEE. The proposed DFOO is a sub-optimal algorithm which could achieve good performance in both network throughput, EC and NEE with low computation complexity.

Figure 8 shows the dynamic cell association during a day at time 8:00 a.m, 12:00 a.m, 16:00 p.m, and 20:00 p.m, where the star represents the M-gNB, diamonds represent P-gNBs, triangles represent F-gNBs, and points represent UEs. In the scenario, the M-gNB and P-gNBs are turned on all the time, whereas F-gNBs could be turned off according to the dynamic traffic load. Obviously, there are more active F-gNBs at time 12:00 a.m and 16:00 p.m due to higher traffic load in the "hot time". Notice that F-gNBs without UEs will be turned off.

In Figure 9, it is observed that the proposed DFOO algorithm could obtain the similar NEE to the FAIS and the P-FAIS algorithm which is higher than the baseline algorithm NAO. Particularly, with the increasing number of UEs, NEE is increasing in all algorithms.

In Figure 10, we compare NEE of the proposed DFOO algorithm with different M-gNB transmission power. Obviously, NEE will be improved with increasing P_m. Since more UEs prefer to associating to the M-gNB with increasing P_m. Therefore, more F-gNBs could be turned off, which result in dramatic increasing of NEE.

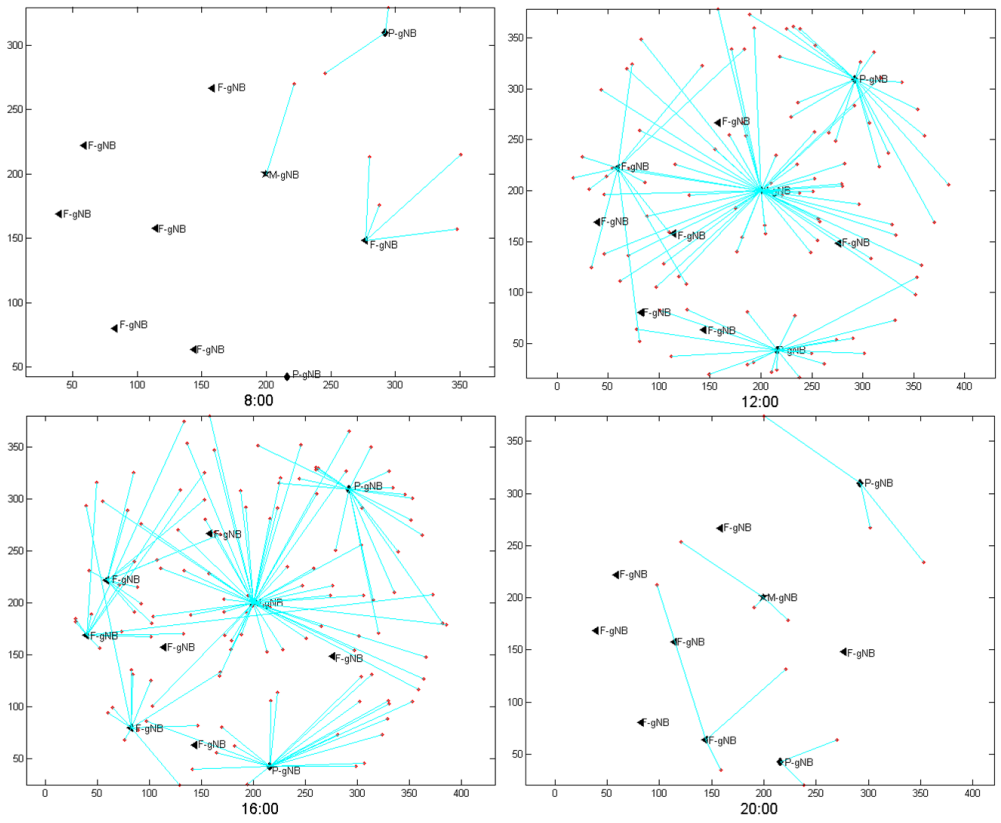

FIGURE 8. UEs-gNBs association at time 8:00 a.m, 12:00 a.m, 16:00 p.m, 20:00 p.m.

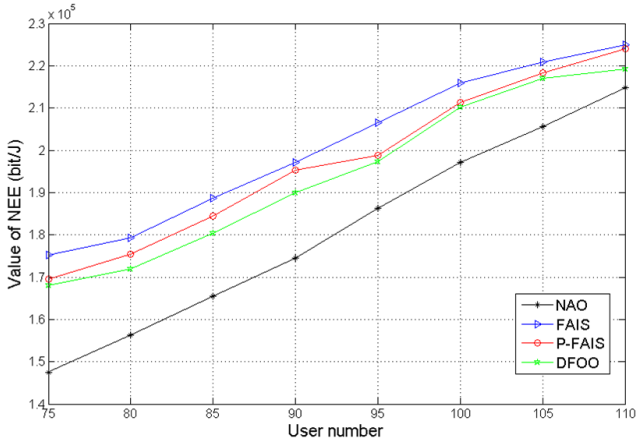

FIGURE 9. NEE V.S the number of UEs of different algorithms.

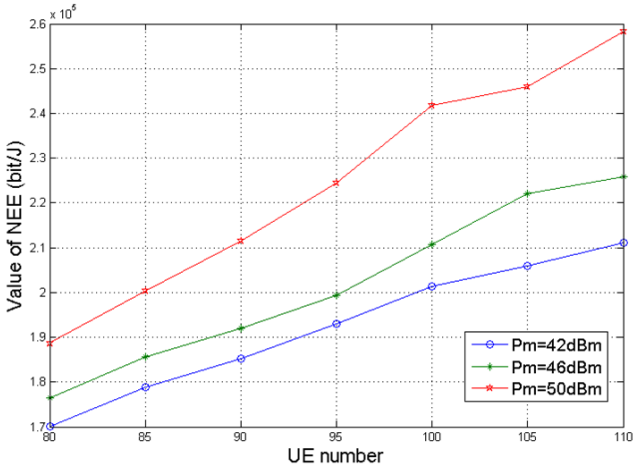

FIGURE 10. NEE V.S the number of UEs with different P_m.

V. CONCLUSION

In this paper, we investigated the NEE problem in 5G HCNs. The DFOO algorithm was proposed to optimize jointly the traffic load prediction, the cell association, and the dynamic F-gNBs on/off strategies. Specially, the EMP algorithm was introduced to prediction network traffic load and the CALB algorithm considered load-balancing of the network and minimum SINR requirement of UEs. In addition, we introduced dual connectivity based seamless handover procedure to transfer UEs from the S-gNB to the T-gNB. Our work provided a new insight into improve NEE and guarantee the QoS in the 5G HCN.

REFERENCES

[1] C. Sexton, N. J. Kaminski, J. M. Marquez-Barja, N. Marchetti, and L. A. DaSilva, "5G: Adaptable networks enabled by versatile radio access technologies," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 688–720, 2nd Quart., 2017.

[2] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
[3] X. G. Huang, S. Z. Tang, F. Zhang, and Q. B. Chen, "Base stations multi-level sleep strategies and load balancing in heterogeneous cellular networks," *EURASIP J. Wireless Commun. Netw.*, to be published.
[4] B. Cao, Y. Li, C. Wang, G. Feng, S. Qin, and Y. Zhou, "Resource allocation in software defined wireless networks," *IEEE Netw.*, vol. 31, no. 1, pp. 44–51, Feb. 2017.
[5] X. G. Huang, L. Shi, C. L. Zhang, D. Y. Zhang, and Q. B. Chen, "Distributed resource allocation with imperfect spectrum sensing information and channel uncertainty in cognitive femtocell networks," *EURASIP J. Wireless Commun. Netw.*, vol. 201, Dec. 2017.
[6] Z. H. Abbas, F. Muhammad, and L. Jiao, "Analysis of load balancing and interference management in heterogeneous cellular networks," *IEEE Access*, vol. 5, pp. 14690–14705, 2017.
[7] F. Muhammad, Z. H. Abbas, and F. Y. Li, "Cell association with load balancing in nonuniform heterogeneous cellular networks: Coverage probability and rate analysis," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5241–5255, Jun. 2017.
[8] W. C. Ao and K. Psounis, "Approximation algorithms for online user association in multi-tier multi-cell mobile networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2361–2374, Aug. 2017.
[9] F. Pervez, M. Jaber, J. Qadir, S. Younis, and M. A. Imran, "Fuzzy Q-learning-based user-centric backhaul-aware user cell association scheme," in *Proc. 13th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Valencia, Spain, 2017, pp. 1840–1845.
[10] X. Huang, B. Beferull-Lozano, and C. Botella, "Quasi-Nash equilibria for non-convex distributed power allocation games in cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3326–3337, Jul. 2013.
[11] *Evolved Universal Terrestrial Radio Access (E-UTRA); Potential Solutions for Energy Saving for E-UTRAN*, document 3GPP TR 36.927, V14.0.0, Mar. 2017.
[12] X. G. Huang, Z. F. Zhang, W. P. Dai, Q. Huang, and Q. B. Chen, "Energy-efficient femtocells active/idle control and load balancing in heterogeneous networks," in *Proc. 11th EAI Int. Conf. Commun. Netw. China*, Sep. 2016, pp. 237–247.
[13] L. Su, C. Yang, Z. Xu, and A. F. Molisch, "Energy-efficient downlink transmission with base station closing in small cell networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 4784–4788.
[14] F. Kong, X. Sun, V. C. M. Leung, Y. J. Guo, Q. Zhu, and H. Zhu, "Queue-aware small cell activation for energy efficiency in two-tier heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.
[15] K. Davaslioglu, C. C. Coskun, and E. Ayanoglu, "New algorithms for maximizing cellular wireless network energy efficiency," in *Proc. Inf. Theory Appl. Workshop (ITA)*, La Jolla, CA, USA, 2016, pp. 1–10.
[16] W. Cheng, X. Zhang, and H. Zhang, "Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3092–3107, Dec. 2016.
[17] *Study on Small Cell Enhancements for E-UTRA and E-UTRAN—Higher Layer Aspects*, document 3GPP TR 36.842, V12.0.0, Dec. 2013.
[18] *Comparison of 0ms Interruption Solutions*, document 3GPP TSG-RAN WG2 #98, TDoc R2-1704854, Huawei, HiSilicon, May 2017.
[19] *0 ms Interruption Support During Handover Procedure in NR*, document 3GPP TSG-RAN WG2 NR AH#2, TDoc R2-1706625, Ericsson, Jun. 2017.
[20] *SgNB to MgNB Reconfiguration for 0ms Interruption Handover*, document 3GPP TSG-RAN WG2 #98, TDoc R2-1704853, Huawei, HiSilicon, May 2017.
[21] I. Ashraf, F. Boccardi, and L. Ho, "Power savings in small cell deployments via sleep mode techniques," in *Proc. IEEE 21st Int. Symp. Pers., Indoor Mobile Radio Commun. Workshops*, Instanbul, Turkey, Sep. 2010, pp. 307–311.
[22] P. Fazio, M. Tropea, and S. Marano, "A distributed hand-over management and pattern prediction algorithm for wireless networks with mobile hosts," in *Proc. 9th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Sardinia, Italy, 2013, pp. 294–298.
[23] S. Das and S. K. Das, "A probabilistic link prediction model in time-varying social networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

XIAOGE HUANG received the Ph.D. degree (Hons.) from the Group of Information and Communication Systems, Institute of Robotics and Information and Communication Technologies, University of Valencia, Spain. In 2013, she joined the Group of Wireless Communication Technology, Chongqing University of Posts and Telecommunications, as an Associate Professor. Her research interests include convex optimization, centralized and decentralized power allocation strategies, game theory, and cognitive radio networks.

SHE TANG received the B.S. degree from Wuhan Polytechnic University, Hubei, China, in 2016. He is currently pursuing the M.S. degree in information and communication engineering with the Wireless Transmission Laboratory, Chongqing University of Posts and Telecommunications, Chongqing, China. His main research interest is energy saving in small cell network.

QIAN ZHENG received the B.S. and M.S. degrees from the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, in 2011 and 2014, respectively. Her current research interests include device-to-device, vehicle-to-everything, and new-radio technologies.

DONGYU ZHANG received the B.S. degree in electronic and information engineering from the Zhengzhou University of Light Industry, Henan, China, in 2016. He is currently pursuing the M.S. degree in electronics and communication engineering with the Wireless Transmission Laboratory, Chongqing University of Posts and Telecommunications, Chongqing, China. His main research interests are cognitive radio network, wireless communication, and small cell network.

QIANBIN CHEN received the Ph.D. degree in communication and information systems from the University of Electronic Science and Technology of China, Chengdu, China, in 2002. He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, and also the Director of the Chongqing Key Laboratory of Mobile Communication Technology. He has authored or co-authored over 100 papers in journals and peer-reviewed conference proceedings, and has co-authored seven books. He holds 47 granted national patents.