Dynamic Base Station Sleep Control and RF Chain Activation for Energy-Efficient Millimeter-Wave Cellular Systems

Mingjie Feng, *Student Member, IEEE*, Shiwen Mao, *Senior Member, IEEE*, and Tao Jiang, *Senior Member, IEEE*

***Abstract*—In a millimeter-wave base station (BS) with hybrid precoding, the radio frequency (RF) chain is the main contributor to the energy consumption. As the traffic pattern varies over time, part of the RF chains in a BS or the entire BS can be turned OFF to prevent overheating and save energy. However, to guarantee the multiplexing gain of a hybrid precoding system, a certain number of RF chains needs to be activated. In addition, as BSs are turned OFF, the quality of service of users would be degraded. Thus, the number of active RF chains and the set of operating BSs should be carefully designed to achieve a tradeoff between data rate and energy consumption. We consider dynamic BS sleep control and RF chain activation to maximize the energy efficiency (EE) of a multicell millimeter-wave cellular systems. We formulate such a problem as an integer programming problem with two sets of variables. We first develop a centralized scheme, and then apply a primal decomposition to derive the optimal solution of the convex problem. Based on such solution, we derive the near optimal BS sleep control strategy with a greedy algorithm. With given BS ON–OFF states, the optimal user association and RF chain activation are obtained by solving a linear programming problem. We then propose a distributed scheme based on a matching between users and BSs, and show that the matching process converges. The proposed schemes are evaluated with simulations, showing that near optimal performance can be achieved. Compared with baseline schemes, the system EE can be significantly improved while the data rate loss due to BS sleep control and dynamic RF chain activation is relatively small.**

***Index Terms*—5G Wireless, millimeter wave (mmWave) communications, energy efficiency, BS sleep control, RF chain activation.**

Manuscript received June 4, 2018; accepted July 28, 2018. Date of publication August 1, 2018; date of current version October 15, 2018. This work was supported in part by the U.S. National Science Foundation under Grant CNS-1320664, in part by the Wireless Engineering Research and Engineering Center at Auburn University, in part by the National Science Foundation for Distinguished Young Scholars of China (NSFC) under Grant 61325004, and in part by the NSFC under Grant 61771216. The review of this paper was coordinated by Dr. D. Marabissi. *(Corresponding author: Shiwen Mao.)*

M. Feng and S. Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: mzf0022@auburn.edu; smao@ieee.org).

T. Jiang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: mzf0022@auburn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2018.2861899

I. Introduction

With the growing popularity of smart devices and massive Internet of Things (IoT) applications, the fifth generation (5G) wireless communication network is characterized by ubiquitous connection, extremely high traffic demand, and highly complicated network architecture. To support future data-intensive and delay-sensitive applications such as high-quality video streaming, unmanned vehicles, and online gaming, the 5G wireless network is expected to provide reliable wireless services with low delay and high data rate. In particular, the 5G wireless network is expected be provide 1000x data rate compared to current cellular systems [1]. Millimeter-wave (mmWave) communication is regarded as one of the enabling technologies of 5G wireless [2], [3], along with massive MIMO [4] and heterogeneous networks [5], [6]. Compared to current cellular systems with typical bandwidth less than 100 MHz, the mmWave communication operates on a much larger bandwidth. Such significantly increased bandwidth makes it possible to achieve a 1000x boost for data rate.

However, the large bandwidth comes at a price. Due to the high frequency and short wavelength of mmWave, the mmWave signals suffer from large attenuation and can be easily blocked by obstacles. To combat the large propagation loss, large antenna array and efficient spatial multiplexing must be supported to guarantee sufficient received power at each user. Due to the high cost and power consumption of mmWave devices, it is infeasible to implement full digital precoding in an mmWave BS, which requires a large number of radio frequency (RF) chains. To this end, analog-digital hybrid precoding [7] has been considered in many existing works. After baseband precoding, the user signals are processed by multiple RF chains, and the number of RF chains is much smaller than the number of antennas. Then, the RF signals generated by RF chains are directed by a set of phrase shifters to achieve highly directional transmission. This way, both analog domain directional transmission and digital domain spatial multiplexing are utilized to enhance the system performance with reduced cost.

Although the use of hybrid precoding avoids the deployment of a large number of RF chains, the energy consumption of supporting all the RF chains is still considerable. As mmWave systems are operating at much higher frequencies with extremely large bandwidth, according to Nyquist sampling theorem, the

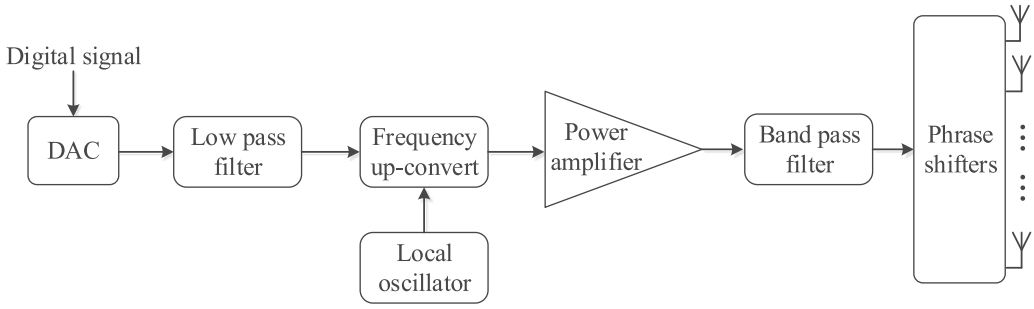

Fig. 1. Architecture of an RF chain at the transmitter of an mmWave BS.

analog to digital convertors (ADC)/digital to analog convertors (DAC) in an mmWave BS need to provide much higher sampling rate compared to sub-6 GHz systems. Since the power of an ADC/DAC is proportional to the sampling rate [8], an RF chain in an mmWave BS consumes much higher energy than one in existing cellular systems. In addition, due to the short wavelength and the need for highly directional transmission, an mmWave BS is expected to equip large number of antennas compacted in a limited space. Thus, it would be infeasible to equip mmWave BSs with dedicated cooling system. To prevent overheating, it is necessary to reduce the energy consumption of an mmWave BS. Due to the absence of a dedicated cooling system, the RF chain, which includes of ADC/DAC and power amplifier (PA) as shown in Fig. 1, is the dominant part of the energy consumption of an mmWave BS. Therefore, an efficient approach to prevent overheating and save energy is to deactivate part of the RF chains. However, to guarantee the multiplexing gain of a hybrid precoding system, a certain number of RF chains need to be activated. As in [7], the number of active RF chains should be no less than the number of users to be served. To achieve a good tradeoff between data rate performance and energy consumption, the RF chain activation should be adaptive to the traffic pattern. In particular, when the traffic load of a BS is low, only a few number of RF chains need to be activated.

On the other hand, the mmWave BSs can be dynamically turn on/off, i.e., with a sleep control, to further reduce the network energy consumption [11]. To provide ubiquitous high data rate services and maintain connectivity in case of blockage, mmWave BSs are expected to be densely deployed with certain level of redundancy [12]. As the traffic pattern changes, e.g., a large number of users move out of an office area during nighttime, a significant proportion of mmWave BSs would be under-utilized and can be turned off for energy saving. However, when BSs are turned off, some users cannot be served by the BS that provides the best link quality, resulting in degraded quality of service (QoS). For mmWave systems at 60 GHz, the QoS guarantee would be highly challenging due to extremely large propagation loss and vulnerability to blockage. Thus, the BS on/off operation mainly applies to mmWave systems at 73 GHz which suffers from less propagation loss and allows non line of sight (NLOS) transmission. In this paper, the BS sleep control is jointly considered with traffic-aware adaptive RF chain activation. As the RF chains are deactivated when the traffic load is low, an mmWave BS is turned off only when its traffic load is extremely low or even there is no users nearby. In addition, due to the dense deployment of mmWave BSs and highly directional transmission, a user can receive satisfactory QoS from another BS when its original serving BS is turned off. Thus, the QoS degradation brought by BS sleep control is expected to be small. However, to find a good tradeoff between energy consumption and user QoS, it is necessary investigate the actual relation between energy saving and QoS degradation and use it to determine the BS on/off pattern. Energy efficiency (EE), defined as the sum rate divided by the total power consumption, is considered as the optimization objective in previous works as an important metric to characterize the tradeoff between data rate performance and energy consumption.

In this paper, we consider dynamic BS sleep control and RF chain activation to improve the EE of a multi-cell mmWave network.[1] We formulate such a problem as an integer programming problem, and propose centralized and distributed algorithms to obtain near-optimal solutions. For the centralized scheme, we first transform the original problem into a convex optimization problem by relaxing the integer constraint and applying a variable transformation. Then, a primal decomposition is applied to obtain the optimal solution to the convex problem. Based on the solution, we propose a greedy algorithm to obtain a near-optimal BS sleep control solution. Then, the optimal user association and RF chain activation under given BS ON-OFF states is derived. To reduce the overhead and complexity, we also propose a distributed scheme based on a matching between users and BSs. We show that the matching process converges and the outcome yields a stable matching. We evaluate the proposed schemes with simulations by comparing with several baseline schemes. The results show that near-optimal performance can be achieved by the proposed schemes. Compared to the schemes without any control, the system EE is significantly improved while the data rate loss due to BS sleep control and dynamic RF chain activation is small.

In the remainder of this paper, the system model and problem formulation are introduced in Section II. We present the cen-

[1]The dynamic BS sleep control and RF chain activation considered in this work can also be applied in hybrid precoding-based sub-6 GHz systems. However, compared to a BS operating at sub-6 GHz, the need for overheating prevention and energy saving in an mmWave BS is much more urgent due to the lack of cooling systems and the high energy consumption of ADCs/DACs.

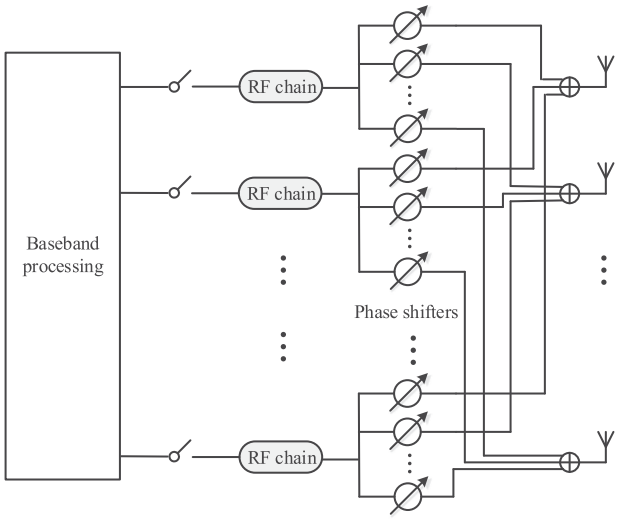

Fig. 2. System model of a hybrid precoding based mmWave BS with dynamic RF chain activation.

tralized and distributed solution algorithms in Section III and Section IV, respectively. The simulation results are presented and analyzed in Section V, followed by a discussion on related work in Section VI. We conclude this paper in Section VII.

II. Problem Formulation

We consider an mmWave cellular network consists of J BSs (indexed by $j = 1, 2, \ldots, J$), serving K UEs (indexed by $k = 1, 2, \ldots, K$). We define binary variable for user association as

$$a_{k,j} \doteq \begin{cases} 1, & \text{user } k \text{ is served by BS } j \\ 0, & \text{otherwise,} \end{cases}$$
$$k = 1, 2, \ldots, K, \ j = 0, 1, \ldots, J. \tag{1}$$

As shown in Fig. 2, analog-digital hybrid precoding is applied at each BS. The RF chains can be dynamically activated/deactivated to save energy. We assume that all RF chain in a BS are identical, then the total power consumption is determined by the number of active RF chains. To guarantee the multiplexing gain of a multi-user hybrid precoding mmWave system, the number of active RF chains, N_{RF}, should be no less than the number of users that can be simultaneously served [7], i.e., $N_{\mathrm{RF}} \geq \sum_{k=1}^{K} a_{k,j}$. Compared to a full-digital mmWave system where each antenna is allocated with a RF chain, the hybrid precoding architecture was shown to achieve near-optimal performance [7], [9], [10]. Thus, as long as $N_{\mathrm{RF}} \geq \sum_{k=1}^{K} a_{k,j}$ is satisfied, activating more RF chains brings marginal data rate gain while increases the system energy consumption. Due to this property, we assume the number of activated RF chains equals to the number of users served by a BS, i.e., $N_{\mathrm{RF}} = \sum_{k=1}^{K} a_{k,j}$. As a result, the number of activated RF chain in BS j is determined by the traffic load of BS j, the RF chain activation problem is *equivalent to* user association problem. Let M_j be the maximum number of RF chains that can be supported by BS j. Then, the number of user that can be served by BS j is upper bounded by M_j, we have $\sum_{k=1}^{K} a_{k,j} \leq M_j$.

We assume that each mmWave BS can be dynamically turned on/off to save energy, the BS ON-OFF indicator is defined by

$$b_j \doteq \begin{cases} 1, & \text{BS } j \text{ is turned on} \\ 0, & \text{BS } j \text{ is turned off,} \end{cases} \quad j = 1, 2, \ldots, J. \tag{2}$$

The power consumption of BS j consists of a static part and a dynamic part, given as

$$P_j = P_j^S + \sum_{k=1}^{K} a_{k,j} P_j^{\mathrm{RF}}, \tag{3}$$

where P_j^S is the constant power consumption whenever the BS is turned on, P_j^{RF} is power consumption of *one* RF chain. Since the power amplifier (PA) and ADC/DAC in a RF chain are the dominant source of power consumption in an mmWave BS, we assume that the dynamic part is determined by the number of active RF chains, which equals to the number of UEs served by BS j. Thus, the dynamic part is given by $\sum_{k=1}^{K} a_{k,j} P_j^{RF}$.

Let η be the efficiency of PA, the total transmitted power of BS j is given as

$$P_j^{\mathrm{T}} = \eta \sum_{k=1}^{K} a_{k,j} P_j^{\mathrm{RF}} \tag{4}$$

We assume equal power allocation among the baseband signals of different users.[2] Based on the SINR model in [7], the downlink data rate of user K when served by BS j is

$$R_{k,j} = \log_2 \left(1 + \frac{\frac{P_j^{\mathrm{T}}}{\sum_{k=}^{K} x_{k,j}} \left| \mathbf{u}_{k,j}^{\mathrm{H}} \mathbf{H}_{k,j} \mathbf{F}_j^{\mathrm{RF}} \mathbf{w}_{k,j}^{\mathrm{BB}} \right|^2}{\frac{P_j^{\mathrm{T}}}{\sum_{k=1}^{K} x_{k,j}} \sum_{k' \neq k} \left| \mathbf{u}_{k,j}^{\mathrm{H}} \mathbf{H}_{k,j} \mathbf{F}_j^{\mathrm{RF}} \mathbf{w}_{k',j}^{\mathrm{BB}} \right|^2 + \sigma^2} \right), \tag{5}$$

where $H_{k,j}$ is the channel gain between user k and BS j, $\mathbf{u}_{k,j}$ is the RF combiner of UE k when served by BS j. $\mathbf{F}_j^{\mathrm{RF}}$ is the RF precoder of BS j, which is implemented by analog phase shifters, $\mathbf{w}_{k,j}^{\mathrm{BB}}$ is the baseband precoding vector for UE k. We assume that each UE connects to a BS via LOS link or reflection-based NLOS link. A UE can measure and identify the signals from different BSs and reports its measured channel gain to nearby BSs. As the BS on-off switching and RF chain activation are operated in relatively large time scales, we use the *time-averaged* channel gain within an interval, rather than instantaneous channel gain, for scheduling purposes. Note that, the BS sleep control and RF chain activation are based on the measured average SINR, how to optimize precoding designs to enhance SINR is beyond the scope of this paper. In this paper, we use the algorithm in [7] to determine the values of $\mathbf{u}_{k,j}$, $\mathbf{F}_j^{\mathrm{RF}}$, and $\mathbf{w}_{k,j}^{\mathrm{BB}}$. Although the precoding design affects $R_{k,j}$, it has a limited impact on the BS energy consumption. This

[2]The power allocation among different users and different RF chains can be optimized. However, this is not the focus of this paper, we consider equal power allocation for simplicity.

is because the transmit power used in precoding is generated from RF chain, while the number of active RF chains is the dominant factor that determines the energy consumption. The EE gain provided by precoding is mostly achieved by increased data rate. As a result, the precoding design can be regarded as a problem that is independent of BS sleep control and RF chain activation. In this paper, we employ an efficient hybrid precoding scheme in existing work which effectively enhanced the data rate performance.

The time scales considered in this paper are three folds: the BS sleep control period, T_1; the user association and RF chain activation period, T_2; and the period of measuring average data rate, T_3. As turning on/off a BS takes considerable time and incurs additional energy consumption, it is infeasible to perform frequent BS sleep control. Thus, T_1 (e.g., 10 mins) is much large than T_2 (e.g., 1 min). The time averaged data rate of each user is measured and calculated within an interval of T_3 (e.g., 10 seconds) before the update user association.

In this paper, we aim to achieve a good tradeoff between energy consumption and data rate with the objective of maximizing the EE of a multi-cell mmWave network through dynamic BS sleep control and RF chain activation. Since the number of activated RF chains in each BS equals to its traffic load and all RF chains consume the same amount of power, the RF chain activation control is directly determined by the user association strategy. Let $\mathbf{a}$ and $\mathbf{b}$ denote the $\{a_{k,j}\}$ matrix and the $\{b_j\}$ vector, respectively. We have

$$\mathbf{P1}: \max_{\{\mathbf{a},\mathbf{b}\}} \frac{\sum_{k=1}^{K}\sum_{j=1}^{J} a_{k,j} R_{k,j}}{\sum_{j=1}^{J} b_j \left(P_j^S + \sum_{k=1}^{K} a_{k,j} P_j^{RF}\right)} \tag{6}$$

$$\text{s. t.}: \sum_{j=0}^{J} a_{k,j} \le 1,\ k = 1, 2, \ldots, K \tag{7}$$

$$\sum_{k=1}^{K} a_{k,j} \le M_j,\ j = 1, \ldots, J \tag{8}$$

$$a_{k,j} \le b_j,\ k = 1, 2, \ldots, K,\ j = 1, 2, \ldots, J \tag{9}$$

$$a_{k,j} \in \{0, 1\},\ k = 1, 2, \ldots, K,\ j = 1, \ldots, J \tag{10}$$

$$b_j \in \{0, 1\},\ j = 1, 2, \ldots, J. \tag{11}$$

Constraint (7) indicates that each user can be served by at most one BS; constraint (8) defines the upper bound for the number of active RF chains in BS j as well as the number of users that can be served by BS j, it also serves to guarantee the QoS of users by limiting the interference from other users served by the same BS. Constraint (9) is due to the fact that a user can be served by BS j only when it is turned on.

III. Centralized Solution

As the BS sleep control and user association operate in different time scales, it is necessary to consider the two kinds of updates separately. We first derive the near-optimal BS sleep control strategy with a subgradient approach. Then, the optimal user association and RF chain activation strategy with given BS ON-OFF states are presented.

A. Near Optimal BS Sleep Control: A Subgradient Approach

Problem **P1** is an integer programming problem with two sets of variables, which is generally NP-hard. To make the problem tractable, we first relax the integer constraints by allowing $\{a_{k,j}\}$ and $\{b_j\}$ to take values in $[0, 1]$. However, the resulting problem is non-convex as the objective function is a difference of logarithmic functions. To this end, we define variable transformation $\tilde{b}_j = \log b_j$ and regard the traffic load $Q_j = \sum_{k=1}^{K} a_{k,j}$ as a constant. In particular, we use the value of Q_j at the end of the previous period as an estimation for the value of Q_j in the current period. With the transformation of $\mathbf{b}$ and the approximation of $Q_j = \sum_{k=1}^{K} a_{k,j}$, we have the following problem.

$$\mathbf{P2}: \max_{\{\mathbf{a},\tilde{\mathbf{b}}\}} \left\{ \log\left(\sum_{k=1}^{K}\sum_{j=1}^{J} a_{k,j} R_{k,j}\right) - \log\left(\sum_{j=1}^{J} e^{\tilde{b}_j}\left(P_j^S + Q_j P_j^{RF}\right)\right)\right\} \tag{12}$$

$$\text{s. t.}: \sum_{j=0}^{J} a_{k,j} \le 1,\ k = 1, 2, \ldots, K \tag{13}$$

$$\sum_{k=1}^{K} a_{k,j} \le M_j,\ j = 1, \ldots, J \tag{14}$$

$$\log a_{k,j} \le \tilde{b}_j,\ k = 1, 2, \ldots, K,\ j = 1, \ldots, J \tag{15}$$

$$0 \le a_{k,j} \le 1,\ k = 1, 2, \ldots, K,\ j = 1, \ldots, J \tag{16}$$

$$\tilde{b}_j \le 0,\ j = 1, \ldots, J. \tag{17}$$

Lemma 1. Problem **P2** is a convex optimization problem.

Proof: Consider the objective function of problem **P2**. The first part is a logarithmic function of a linear function of $\mathbf{x}$, which is obviously concave. The second part is a logarithmic function of sum of exponential functions, given as $-\log\left(\sum_{j=1}^{J} e^{\tilde{b}_j}\left(P_j^S + Q_j P_j^{RF}\right)\right)$, which is also a concave function [20]. As a sum of the two parts, the objective function of **P2** is concave.

In **P2**, constraint $\log a_{k,j} - \tilde{b}_j \le 0$ defines a convex set, all other constraints are linear. Therefore, Problem **P2** is a convex optimization problem. ■

The decision variables of **P2**, $\{a_{k,j}\}$ and $\{\tilde{b}_j\}$, are coupled due to the constraint $\log a_{k,j} \le \tilde{b}_j$. To this end, we decompose **P2** into two levels subproblems. For the lower level subproblem, the optimal solution of $\mathbf{a}$ with given $\tilde{\mathbf{b}}$ is derived. With the solution of the lower-level problem, we use a subgradient approach to obtain the optimal solution of $\tilde{\mathbf{b}}$ for the higher-level subproblem.

*1) Lower-Level Subproblem of **P2**:* With a given set of $\tilde{\mathbf{b}}$, the resulting lower-level subproblem is as follow

$$\mathbf{P3}: \max_{\{\mathbf{a},\tilde{\mathbf{b}}\}} \left\{ \log\left(\sum_{k=1}^{K}\sum_{j=1}^{J} a_{k,j}R_{k,j}\right) - \log\left(\sum_{j=1}^{J} e^{\tilde{b}_j}\left(P_j^S + Q_j P_j^{RF}\right)\right)\right\}$$
$$\text{s.t.}: \ (13)-(17). \tag{18}$$

Take a partial relaxation with respect to the constraint (15). The resulting dual problem is given by

$$\textbf{P3-Dual:} \ \min_{\{\boldsymbol{\lambda}\}} g(\boldsymbol{\lambda}), \tag{19}$$

where $\boldsymbol{\lambda}$ is the Lagrangian multiplier corresponding to constraint (15), and $g(\boldsymbol{\lambda})$ is given by

$$g(\boldsymbol{\lambda}) = \max_{\{\mathbf{a}\}} \left\{ \log\left(\sum_{k=1}^{k}\sum_{j=1}^{J} a_{k,j}R_{k,j}\right) - \log\left(\sum_{j=1}^{J} e^{\tilde{b}_j}\left(P_j^S + Q_j P_j^{RF}\right)\right) + \sum_{k=1}^{K}\sum_{j=1}^{J}\lambda_{k,j}\left(\tilde{b}_j - \log a_{k,j}\right)\right.$$

Since **P2** is a convex optimization problem, **P3-Dual** can be optimally solved by the following gradient approach.

$$\lambda_{k,j}^{[t+1]} = \left[\lambda_{k,j}^{[t]} + \frac{g(\boldsymbol{\lambda}^{[t]}) - g(\boldsymbol{\lambda}^*)}{\|\boldsymbol{\delta}_{\lambda}^{[t]}\|^2}\left(\log a_{k,j}^{[t]} - \tilde{b}_j^{[t]}\right)\right]^+, \tag{20}$$

where t is the index of iteration, $[\cdot]^+ \doteq \max\{0,\cdot\}$. $\boldsymbol{\delta}_{\lambda}^{[t]}$ denotes the vector of gradients of $\{\lambda_{k,j}\}$ given as $\left[\tilde{b}_1^{[t]} - \log a_{1,1}^{[t]}, \ldots, \tilde{b}_J^{[t]} - \log a_{K,J}^{[t]}\right]^T$. Note that, the value of $\boldsymbol{\lambda}^*$ is unknown before the solution is obtained, the mean of objective functions of the primal and dual problems as an estimate for $g(\boldsymbol{\lambda}^*)$.

In each iteration, with given $\{\lambda_{k,j}\}$, $g(\boldsymbol{\lambda})$ is obtained based on the solution of the Problem **P4**, given as

$$\mathbf{P4}: \max_{\{\mathbf{a}\}} \mathcal{L}(\mathbf{a},\boldsymbol{\lambda}) \quad \text{s.t.} \ (13),(14),\text{and}\ (16), \tag{21}$$

where $\mathcal{L}(\cdot)$ is the *Lagrangian function*. Due to the convexity of Problem **P4**, the KKT conditions can be used to derive its optimal solution. Then, the solution of problem **P4** and the optimal Lagrangian multipliers will be used in high-level problems.

*2) Higher-Level Subproblem of **P2**:* In the following, we derive the optimal $\tilde{\mathbf{b}}$ with a subgradient approach. To begin with, we first demonstrate that the duality gap between primal and dual problems of **P3** is zero.

Lemma 2. Strong duality holds for problem **P3**.

Proof: For the constraints of **P3**, the inequalities (15) and other linear functions define a feasible region. Then, there is a feasible $\mathbf{a}$ that satisfies all constraints. Therefore, **P3** is strictly feasible. Then, the Slater's condition holds, resulting in strong duality. ■

Denote $f(\mathbf{a})$ as the objective function of **P3**. The higher-level subproblem of problem **P2** aims to find the optimal $\tilde{\mathbf{b}}$ by solving the following problem.

$$\mathbf{P5}: \max_{\{\tilde{\mathbf{b}}\}} f(\mathbf{a}(\tilde{\mathbf{b}})). \tag{22}$$

Lemma 3. The optimal solution of Problem **P5** can be obtained with the following update.

$$\tilde{\mathbf{b}}^{[t+1]} = \tilde{\mathbf{b}}^{[t]} + \frac{f(\tilde{\mathbf{b}}^{[t]}) - f(\tilde{\mathbf{b}}^*)}{\|\boldsymbol{\theta}^{[t]}\|^2}\boldsymbol{\theta}^{[t]}, \tag{23}$$

where $\boldsymbol{\theta}^{[t]} = \left[\sum_{k=1}^{K}\lambda_{k,1}^{*}{}^{[t]} - \frac{P_1 e^{\tilde{b}_1{}^{[t]}}}{\sum_{j=1}^{J} P_j e^{\tilde{b}_j{}^{[t]}}}, \ldots, \sum_{k=1}^{K}\lambda_{k,J}^{*}{}^{[t]} - \frac{P_J e^{\tilde{b}_J{}^{[t]}}}{\sum_{j=1}^{J} P_j e^{\tilde{b}_j{}^{[t]}}}\right]^T$, $P_j = P_j^S + \sum_{k=1}^{K} a_{k,j}P_j^{RF}$.

Proof: The objective function of problem **P2** has two parts, given as $\log(\sum_{k=1}^{K}\sum_{j=0}^{J} a_{k,j}R_{k,j})$ and $-\log(\sum_{j=1}^{J} e^{\tilde{b}_j}P_j)$, respectively. Based on the principle of primal decomposition, we can maximize the two parts separately and combine the results to derive the subgradient for obtaining the optimal solution.

Let $E^*(\tilde{\mathbf{b}})$ be the optimal value of the *first* part as a function of $\tilde{\mathbf{b}}$. Suppose $\mathbf{a}^*(\tilde{\mathbf{b}}')$ is the optimal solution to problem **P2** with a given $\tilde{\mathbf{b}}'$ and let $\mathbf{a}$ be another feasible solution with a given $\tilde{\mathbf{b}}$. Then, the following inequalities and equalities hold.

$$\begin{aligned} E^*(\tilde{\mathbf{b}}') &\overset{(1)}{=} E(\mathbf{a}^*(\tilde{\mathbf{b}}')) = \mathcal{L}(\mathbf{a}^*,\boldsymbol{\lambda}^*(\tilde{\mathbf{b}}')) \overset{(2)}{\geq} \mathcal{L}(\mathbf{a},\boldsymbol{\lambda}^*(\tilde{\mathbf{b}}')) \\ &= E(\mathbf{a}) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}')(\tilde{\mathbf{b}}' - \boldsymbol{\varphi}_k) \\ &= E(\mathbf{a}) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}')(\tilde{\mathbf{b}} - \boldsymbol{\varphi}_k) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}')(\tilde{\mathbf{b}}' - \tilde{\mathbf{b}}) \\ &\overset{(3)}{\geq} E(\mathbf{a}) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}')(\tilde{\mathbf{b}}' - \tilde{\mathbf{b}}), \end{aligned} \tag{24}$$

where $\boldsymbol{\varphi}_k = [\log a_{k,1}, \log a_{k,2}, \ldots, \log a_{k,J}]^T$ and $\boldsymbol{\lambda}_k^*(\mathbf{b}')$ is the kth row of $\boldsymbol{\lambda}^*(\mathbf{b}')$.

In (24), equality (1) is due to strong duality, inequality (2) is due to the optimality of $\mathbf{a}^*$, inequality (3) is due to the constraints of problem **P4** and the nonnegativity of $\boldsymbol{\lambda}$.

In particular, we have

$$\begin{aligned} E^*(\tilde{\mathbf{b}}') &\geq \max_{\{\mathbf{a}|\boldsymbol{\varphi}\leq\tilde{\mathbf{b}}\}}\left\{E(\mathbf{a}) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}' - \mathbf{b})\right\} \\ &= E^*(\tilde{\mathbf{b}}) + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^*(\tilde{\mathbf{b}}' - \tilde{\mathbf{b}}). \end{aligned} \tag{25}$$

We conclude that the vector $[\sum_{k=1}^{K}\lambda_{k,1}^{*}{}^{[t]}, \ldots, \sum_{k=1}^{K}\lambda_{k,J}^{*}{}^{[t]}]^T$ is a subgradient of $\tilde{\mathbf{b}}$ as a function of $E^*(\tilde{\mathbf{b}})$.

Algorithm 1: Centralized BS Sleep Control Algorithm.

```
Initialize b̃ and λ ;
do
    do
        Solve problem P4 with the standard Lagrangian dual method ;
        Update λ as in (20) ;
    while (λ does not converge);
    Update b̃ as in (23) ;
while (b̃ does not converge);
for τ = 1 : J do
    Search the τ BSs with largest values of b̃_j ;
    Set b_j = 1 for the τ BSs ;
    Calculate the EE under the given τ, EE(τ) ;
end
τ = arg max_{τ} EE(τ) ;
Set b_j = 1 for the corresponding τ BSs.
```

Consider the *second* part of the objective function of problem **P2**, $-\log(P_0+\sum_{j=1}^{J} e^{\tilde{b}_j} P_j)$. It is a differentiable concave function. Based on the property of gradient, we have

$$-\log\left(\sum_{j=1}^{J} e^{\tilde{b}_j} P_j\right) \leq -\log\left(\sum_{j=1}^{J} e^{\tilde{b}'_j} P_j\right) + \sum_{j=1}^{J} \frac{e^{\tilde{b}_j} P_j (\tilde{b}'_j - \tilde{b}_j)}{\sum_{j=1}^{J} e^{\tilde{b}_j} P_j}. \quad (26)$$

As a global variable across the two parts, the optimal value of $\tilde{\mathbf{b}}$ can be obtained by combining (25) and (26). Then, $\boldsymbol{\theta}$ is a subgradient for the objective function of problem **P2**. Thus, the update given in (23) can optimally solve problem **P5**. ■

3) Near Optimal Solution of $\mathbf{b}$*:* The optimal solution of $\tilde{\mathbf{b}}$ for problem **P2** is in the range of $[0, 1]$. To obtain the integer solution and determine the actual BS ON-OFF states, we develop a heuristic scheme to obtain a near optimal integer solutions of $\mathbf{b}$.

In the update of $\tilde{b}_j^{[t]}$ given by (23), the first part of the subgradient, $\sum_{k=1}^{K} \lambda_{k,j}^{*\,[t]}$, indicates the data rate gain of all users served by BS j with the current value of $\tilde{b}_j$. The second part, $-\frac{P_j e^{\tilde{b}_j\,[t]}}{P_0+\sum_{j=1}^{J} P_j e^{\tilde{b}_j\,[t]}}$, indicates the power consumption of BS j. After convergence, a large value of $\tilde{b}_j$ indicates that a BS can provide high sum rate to users at a relatively low power, i.e., high EE can be achieved by turning the BS on.

Due to this property, we propose a greedy scheme to find the set of BSs to be turned on. Let τ be an integer between 0 and J. For any value of τ, the first τ BSs with the largest values of $\tilde{b}_j$ are turned on. We calculate the values of EE with different values of τ, and select the τ that generates the largest value of EE. When calculating the EE with given $\mathbf{b}$, the user association is determined by solving a linear programming problem, which will be presented in the following part. The procedure of the greedy scheme is presented in Algorithm 1.

B. Optimal User Association and RF Activation With Given BS ON-OFF States

Due to the fact that user association is updated more frequently than that of BS sleep control, the user association is mostly determined when the BS ON-OFF states are given. Then, the user association problem is given by

$$\mathbf{P6}: \max_{\{\mathbf{a}\}} \sum_{k=1}^{K}\sum_{j=1}^{J} a_{k,j} R_{k,j} \quad \text{s.t.}: (7)-(10). \quad (27)$$

When the BS ON-OFF states are determined, each user can only by connect to BSs that are turned on. Thus, only the active BSs need to be considered. Let Ω be the set of BSs that are turned on, $\Omega = \{j \,|b_j = 1\}$. We re-assign the indices for active BSs as $\{j = 1, \ldots, |\Omega|\}$. Relax the integer constraints on $\{a_{k,j}\}$, the resulting problem is given as

$$\mathbf{P7}: \max_{\{\mathbf{a}\}} \sum_{k=1}^{K}\sum_{j=1}^{|\Omega|} a_{k,j} R_{k,j} \quad (28)$$

$$\text{s.t.}: \sum_{j=0}^{|\Omega|} a_{k,j} \leq 1,\ k = 1, 2, \ldots, K \quad (29)$$

$$\sum_{k=1}^{K} a_{k,j} \leq M_j,\ j = 0, 1, \ldots, |\Omega|, \quad (30)$$

$$0 \leq a_{k,j} \leq 1,\ k = 1, \ldots, K,\ j = 0, \ldots, |\Omega|. \quad (31)$$

By vectorizing $\{a_{k,j}\}$, problem **P7** can be transformed to a linear programming problem with unimodular constraint matrix. Similar to the case in [22], we can proof that the solution of **P7** are all integers. Thus, the optimal solution of **P7** is also optimal to **P6**.

C. Mobility Aware Adaptive Selection of Update Periods

To deal with the non-convexity of the original problem **P1**, we use the value of $\sum_{k=1}^{K} a_{k,j}$ in the previous period to approximate its value in the current period. However, the effectiveness of such approximation largely depends on the user mobility. In high mobility scenarios, the traffic load of each BS may drastically change, leading to inaccurate estimation of $\sum_{k=1}^{K} a_{k,j}$. To deal with this challenge, each active BS monitors its traffic load based on the user association outcome. If the change of the traffic load of a BS is above a threshold, the BS sends a request to the network controller for BS on-off switching. If the a certain amount of BSs request to update the current BS on-off states, the network controller reduces the value of T_1 to enable more frequent BS on-off switching, so as to deal with the challenge of high user mobility.

IV. Distributed Solution

To reduce the overhead and computation complexity brought by the centralized scheme in case of large scale networks, we propose a distributed solution algorithm in this section. We consider a matching between users and BSs and use the outcome of the matching to determine the user association and BS ON-OFF states. We assume that the each user has a preference list of users for the BSs. The order of BSs in the list is determined by the achievable data rates of connecting to different BSs, i.e., the values of $R_{k,j}$. The strategy of a user is always propose to the BS that is in the top of its preference list. On the other hand,

each BS has a preference list determined by the values of $R_{k,j}$ of different users. Each BS also has a waiting list that indicates its current selection. To maximize the sum rate under constraint $\sum_{k=1}^{K} a_{k,j} \leq M_j$, a BS always keeps the top M_j users in its waiting list if the number of received proposals is larger than M_j, and keeps all users in the waiting list otherwise.

The matching has three stages. In the first stage, all users propose to the BSs according to their preference list. The BSs respond to the proposals based on the evaluated data rates of the users. In the second stage, a user proposes to another BS if one of the following two cases happens.

Case 1: The user is rejected by a BS.

Case 2: The user can obtain a higher utility when served by another BS j' and one of the following condition holds: (i) BS j' still has room in its waiting list, i.e., $Q_{j'} < M_{j'}$, (ii) $Q_{j'} = M_{j'}$, there is a user k' in the waiting list of BS j' such that $R_{k,j'} > R_{k',j'}$.

When a user k is rejected by BS j, it proposes to the top BS among other BSs other than BS j. The BSs receive the new proposals and make comparison between the new ones with those already in the waiting list. Then, the BSs respond to the users based on the outcome of comparison. Based on the decisions of BSs, users then make another period of proposals if Case 1 or Case 2 holds. The propose, hold, and reject actions are continued until the user association result is converged.

In the last stage, we determine the BS ON-OFF states. Based on the outcome of the matching, the BS with smallest value of EE is turned off. Here, the EE of a BS is defined by the sum rate of its serving users divided by its power consumption. Suppose BS j is selected to be turned off, the users originally served by BS j are regarded as being rejected and will propose to other BSs. Then, another series of propose, hold, reject actions will be carried out until a new convergence is reached. After the new convergence is achieved, the system EE under the new BS ON-OFF states and user association is evaluated. If the system EE is improved by turning off BS j, we continue the process by picking up the BS with smallest EE and evaluate the effect of turning off that BS. If the system EE is decreased by turning off BS j, the process is terminated and BS j is not set to be turned off.

A. Convergence Analysis

Definition 1. In a stable matching, there is no such a user-BS pair (user k, BS j') that is not matched, while (i) user k prefers BS j' compared to its current serving BS j, (ii) BS j' prefers user k over user k', while user k' is in the waiting list of BS j'; or BS j' still has room for more users ($\sum_{k=1}^{K} a_{k,j'} < M_{j'}$). In other words, there is no such a user-BS pair that both of them have a better and feasible option than their current one(s) [27].

Lemma 4. The users entering the waiting list of a BS is non-decreasing in the preference list of the BS.

Proof: Based on the strategy of a BS, if the waiting list of the BS is not full, all the received proposals will be put into the waiting list. If the waiting list is full, the BS compares the new incoming proposals with those in the waiting list, and keeps the proposal(s) that bring higher data rate. Thus, the proposals accepted by a BS is non-decreasing in its preference list. ■

Theorem 1. The matching process converges and the outcome yields a stable matching.

Proof: Suppose for contradiction. Based on Definition 1, there must be a user k who prefers BS j' compared to its current serving BS j, and BS j' will accept the proposal of user k. Since user k is currently served by BS j, user k has not been rejected by BS j, it must be the case that a higher performance can be achieved by switching to BS j', i.e., $R_{k,j'} > R_{k,j}$. Since BS j' will accept the proposal of user k, one of the following cases must be satisfied (i) $Q_{j'} < M_{j'}$, (ii) $Q_{j'} = M_{j'}$ and there is another user k' that is served by BS j' such that $R_{k,j'} > R_{k',j'}$. For case (i), since user k prefers BS j' over BS j, it should have proposed to BS j' before BS j. Since $Q_{j'} < M_{j'}$, BS j' would have accepted the proposal of user k, and user k would be served by BS j', which contradicts to the fact that user k is currently served by BS j. For case (ii), user k is not in the waiting list of BS j' and user k' is in, while user k ranks higher than user k', according to Lemma 4, the only explanation is that user k has never proposed to BS j' before. However, since user k prefers BS j' over BS j, it must have proposed to BS j' before BS j, which is also a contradiction. Thus, the hypothesis for contradiction does not hold, a stable matching can be achieved. ■

V. Simulation Study

The performance of the proposed schemes is evaluated with MATLAB simulations. We consider a 500 m $\times$ 500 m area, the BSs are randomly located in the area. We consider two different user distributions, uniform and non-uniform. For the uniform distribution, users are randomly distributed in the area; for the non-uniform case, we divide the area into multiple subareas and the number of users in each area is a random variable. The system bandwidth is 1 GHz. Each user is subject to random blockage. The path loss components when users are served by LOS and NLOS links are 2 and 4, respectively. The channel model is based on the baseline model presented in [24]. We make comparisons with several baseline schemes. The first one is the scheme without BS sleep control, termed BS always ON, and we assume that all RF chains are activated. In the other two schemes, the proposed dynamic BS sleep control is applied, but the numbers of active RF chains are fixed. We also derive an upper bound by solving Problem **P2** under different combinations of Q_j and select one with the largest value of objective function.

Fig. 3 shows the EE performance under different numbers of BSs with uniform user distribution. Compared with the always ON scheme, the EE can be effectively improved with dynamic sleep control, especially when the number of BSs is large. When the number of BSs is small, the performance of scheme with half RF chains activated is the lowest, since the total number of users that can be served is limited by the total number of active RF chains. The performance of the scheme with all RF chains activated grows at a slower rate compared to the half activation scheme and the proposed scheme. This indicates that dynamic RF chain activation can effectively improve the system EE when the BSs are deployed with a certain level of redundancy. The proposed schemes achieve the higher EE than other schemes since a combination of dynamic BS sleep control and RF chain activation is employed. Compared to the upper bound, the

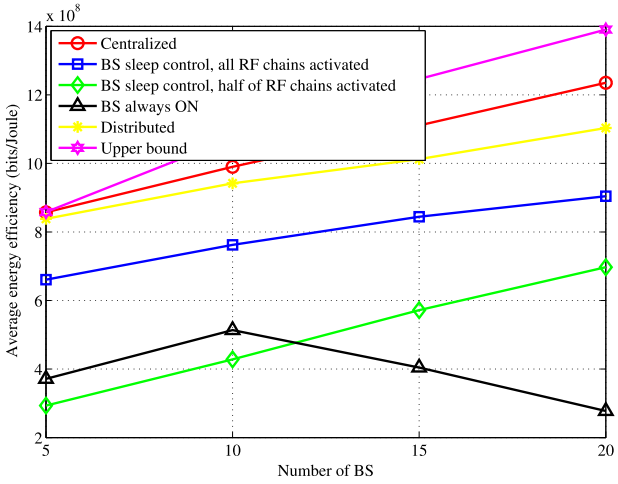

Fig. 3. Average EE under different numbers of BSs: Uniform user distribution and the number of users is 60.

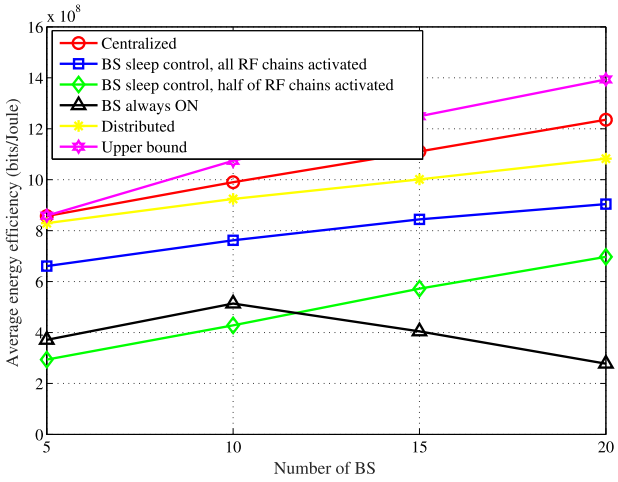

Fig. 4. Average EE under different numbers of BSs: Non-uniform user distribution and the number of users is 60.

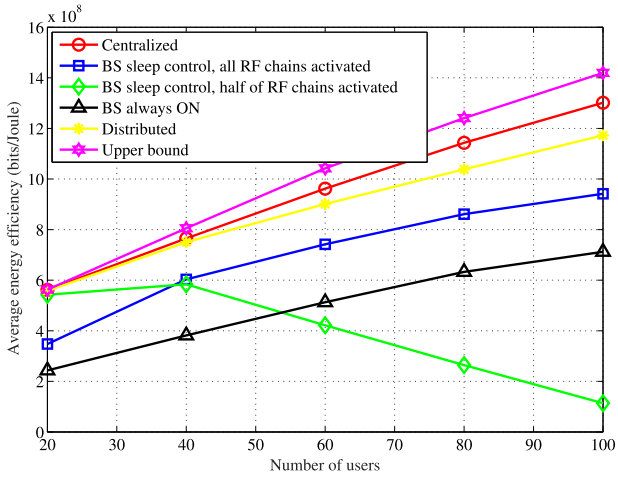

Fig. 5. Average EE under different numbers of users: Uniform user distribution and the number of BSs is 10.

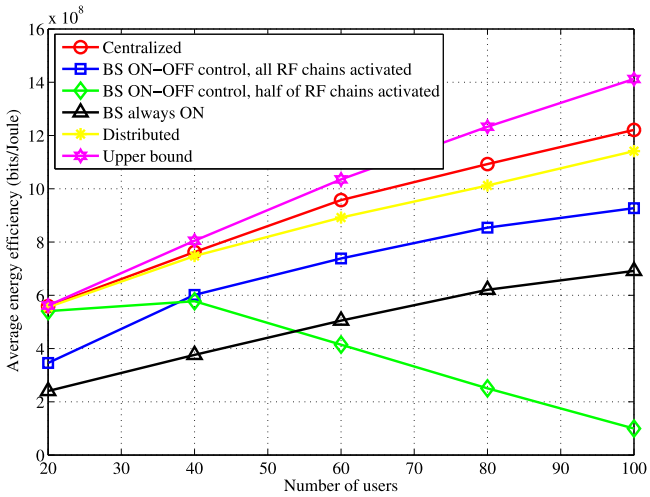

Fig. 6. Average EE under different numbers of users: Non-uniform user distribution and the number of BSs is 10.

performance gap is small, showing that near-optimal solution can be obtained with Algorithm 1. In particular, such gap approaches zero when the number of mmWave BSs is small. This is because in a non-ultra-dense network, the coverage overlap of neighboring BSs is small. For most users, there is a dominant BS that provides much higher data rate compared to other BSs. Thus, the ON-OFF states of neighboring BSs has limited impact on each other. As a result, the proportion of non-integer solution of **P2**, $0 < a_{k,j} < 1$, is small. With the constraint $a_{k,j} \leq b_j$, the proportion of non-integer b_j's would be limited. As the ON-OFF control of different BSs are mostly independent to each other, the set BSs with the largest values of $\tilde{b}_j$ is expected to provide highest EE when they are turned on.

The EE performance under different numbers of BSs with non-uniform user distribution is shown in Fig. 4, where similar trend among different schemes can be observed. The performance gaps between different schemes are enlarged, since the traffic load of each BS is unevenly distributed, the case of under-utilized BSs and RF chains is increased.

Fig. 5 and Fig. 6 show the EE performance under different numbers of users. When the number of users is small, the scheme with half RF chains activated achieves higher EE than that of the scheme with all RF chains activated, since only a small amount of RF chains are required to serve these users. However, as the number of users grow, the performance of the scheme with all RF chains activated becomes better than that of the scheme with half RF chains activated. The proposed scheme achieves the best performance since it is adaptive to the specific traffic pattern.

Figs. 7 and 8 show the data rates of different schemes with varying BS densities. It can be seen that the sum rate increases as more BSs are deployed, due to the higher average SINR of users. As expected, the BS always ON scheme provides highest sum rate since each user can be served by the BS that provides largest rate. The performance of the centralized scheme is close to that of BS always ON, since only the BSs that are not energy efficient are turned off. As a result, the sum rates of users served by these BSs is relatively small, the data rate loss due to turning off these BSs is limited. The distributed scheme also achieves

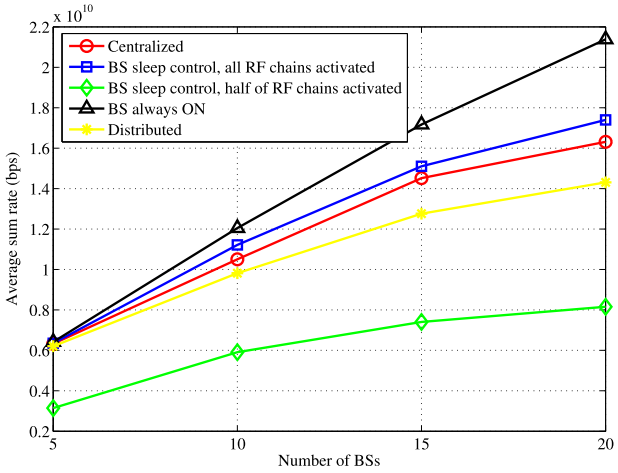

Fig. 7. Average sum rate under different numbers of BSs: Uniform user distribution and the number of users is 60.

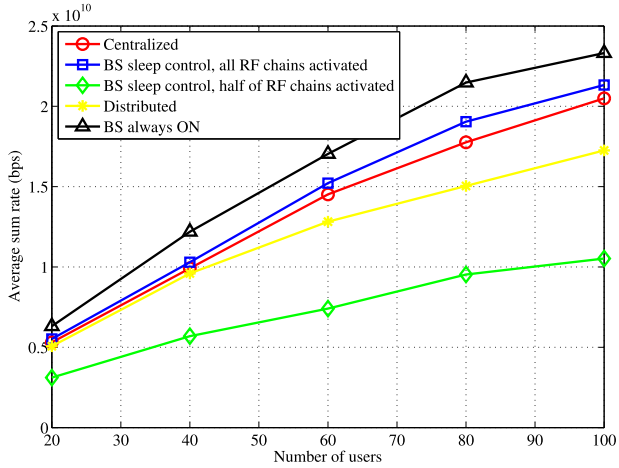

Fig. 9. Average sum rate under different numbers of users: Uniform user distribution and the number of BSs is 15.

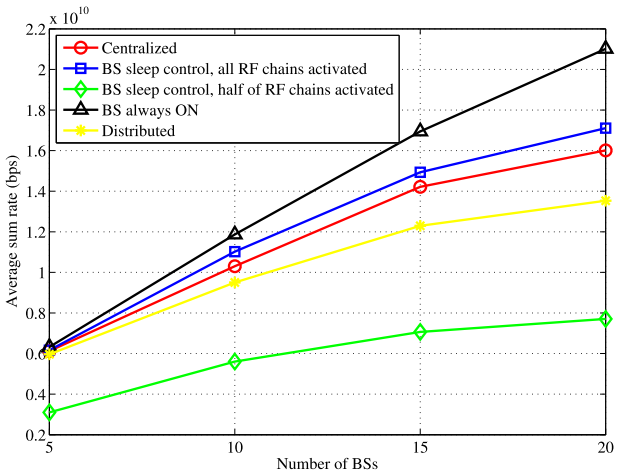

Fig. 8. Average sum rate under different numbers of BSs: Non-uniform user distribution and the number of users is 60.

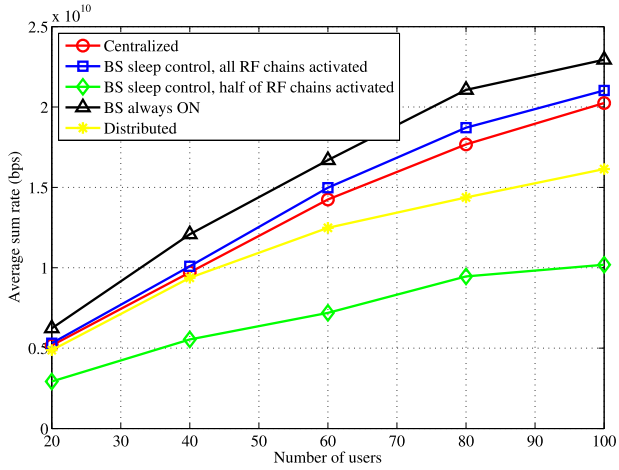

Fig. 10. Average sum rate under different numbers of users: Non-uniform user distribution and the number of BSs is 15.

a good performance, since the matching process optimizes user association aiming to improve data rate, and the BSs with small values of EE are turned off. Comparing Fig. 7 to Fig. 8, we find that data rate under non-uniform user distribution is lower than that with uniform distribution, and the gaps between different schemes are larger. This is because the traffic loads of BSs are significantly different from each other, the BSs and RF chains are not efficiently utilized to serve the users.

The data rate performance under different numbers of users is shown in Fig. 9 and Fig. 10. It can be seen that the schemes with more RF chains activated and more BSs turned on achieve better performance. For the scheme with half of the RF chains activated, part of the users are not served when the number of users is large, showing that the RF chain activation should be adaptive to the traffic pattern. The performance gaps between different schemes first increase and then decrease. This is because when the number of users is large, most BSs would be turned on and most RF chains would be activated to serve these users. The performance of the proposed schemes is close to the one with all RF chains activated and the BS always ON scheme, showing that the data rate loss brought by dynamic RF chain activation and BS sleep control is small and a good tradeoff between energy and data rate is achieved.

An example of the convergence of the distributed scheme is shown in Fig. 11. It can be seen that the matching process converges after the propose, reject, and hold actions in the first and second stage. After the first convergence, some BSs are turned off, the matching continues and finally converges after a few periods.

VI. Related Work

MmWave communication is a key technology of 5G. The modeling of mmWave communication was introduced in [24]. An overview of signal processing issues can be found in [28]. To analyze the impact of blockage, a stochastic geometry based analytical framework was presented in [29]. The solutions to deal with link blockage was analyzed in [30] and effective link scheduling algorithms were developed in [13]–[17]. Due to hardware limit, hybrid precoding is expected to be applied in mmWave systems. The technical aspects and design issues of hybrid precoding based system were discussed in [9], [31],

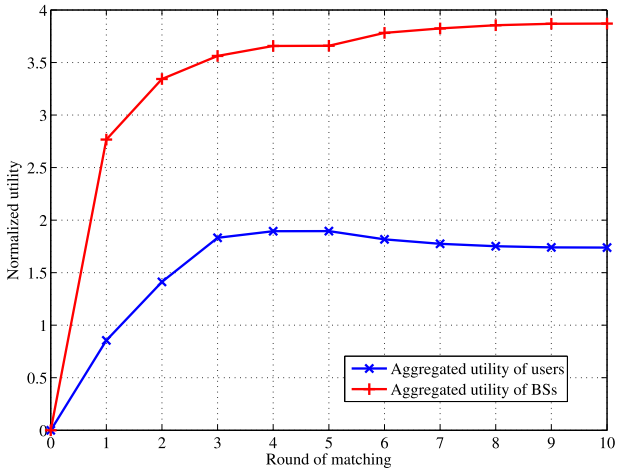

Fig. 11. Convergence of matching process in the distributed scheme.

and the applications in multi-user systems were investigated in [7], [32]. In [7], the analog precoding is performed based on a codebook that indicates different transmission directions, while the analog precoding in [32] is based on the angle of arrival estimation. The models in [7], [32] assumed that the number of RF chains equals to the number of served users to simplify analysis and notation. In this paper, we consider traffic load aware dynamic RF chain activation for energy saving.

As a major approach to improve EE, BS sleep control has been investigated in different wireless networks[33]–[35], [37], [38]. In [34], the BS sleep control is performed in a cooperative pattern. When a BS is turned off, the neighboring BSs monitor its traffic load and informs the BS to wake up when the traffic load exceeds a threshold. In [35], the sleep control for large scale cellular network with hybrid energy supplies was considered with the objective of minimizing on-grid energy cost. With a stochastic geometry analytical framework, a low complexity joint BS on/off operation and on-grid energy purchase policy was proposed. In [38], the BS sleep control in a two-tier HetNet was studied by analyzing the distribution and density of users. Using the stochastic geometry model, the optimal BS operation thresholds are derived under different user distributions. Although these solutions are efficient in sub-6 GHz systems, they cannot be directly applied in mmWave systems. Unlike a traditional cellular BS in which the static part (e.g., cooling system, hardware circuits) is the main source of energy consumption, the RF chains are the major contributor to the energy consumption of an mmWave BS, due to the high power of ADCs/DACs and lack of dedicated on-site cooling system. Thus, turning off under-loaded BSs alone cannot effectively save the energy of an mmWave system. The RF chain activation is key factor that need to be considered to enable energy efficient mmWave network. The RF chain activation has been considered in a few works in the context of sub-6 GHz systems. In [36], a traffic aware RF chain and user selection scheme was proposed for energy saving in a multiuser MIMO system. Compared to existing works, we jointly optimize BS sleep control and RF chain activation to enhance the EE of an mmWave network.

User association in mmWave networks has been considered in recent works [39]–[42]. Due to the vulnerability to link blockage, the efficient operation of an mmWave network can be easily interrupted, posting new challenges to the user association schedule. In [39], user association and power allocation were jointly considered to improve the EE of an mmWave network, and a dual decomposition approach was applied to obtain the solution. In [40], the challenge of frequent handover caused by high mobility was addressed by tracking the mobility patterns of users. In [41], a Markov chain model was employed to characterize the dynamic nature of mmWave channel, and a fairness aware user association scheme was proposed. Besides SINR, the traffic load of each BS is another major factor that needs to be considered in user association as it determines the amount of resource that can be allocated to each user. To prevent overloading of some BSs, a load balancing aware scheme for mmWave network was presented in [42]. Compared to these works, we consider joint BS sleep control and user association with the objective of maximizing the EE of an mmWave network.

VII. Conclusion

We considered dynamic BS sleep control and RF chain activation to enhance the EE of a multi-cell mmWave network. Based on a special property of hybrid precoding based mmWave system, we proposed a traffic aware dynamic scheduling architecture. We proposed both centralized and distributed schemes to derive the near-optimal solutions to the BS sleep control and RF chain activation problem. Compared to the benchmark schemes without any control, the proposed schemes achieve significant EE gain with small data rate loss, showing that a good tradeoff between energy consumption and data rate is achieved.

References

[1] Qualcomm, "The 1000x data challenge." [Online] Available: https://www.qualcomm.com/1000x, Accessed on: Aug. 1, 2018.
[2] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[3] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–449, 2013.
[4] M. Feng and S. Mao, "Harvest the potential of massive MIMO with multi-layer techniques," *IEEE Netw.,* vol. 30, no. 5, pp. 40–45, Sep./Oct. 2016.
[5] M. Feng, T. Jiang, D. Chen, and S. Mao, "Cooperative small cell networks: High capacity for hotspots with interference mitigation," *IEEE Wireless Commun.,* vol. 21, no. 6, pp. 108–116, Dec. 2014.
[6] W. Wang and Q. Zhang, "Local cooperation architecture for self-healing femtocell networks," *IEEE Wireless Commun.,* vol. 21, no. 2, pp. 42–49, Apr. 2014.
[7] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.,* vol. 14, no. 11, pp. 6481–6494, Nov. 2015.
[8] D. Zhang, C. Svensson, and A. Alvandpour, "Power consumption bounds for SAR ADCs," in *Proc. IEEE Eur. Conf. Circuit Theory Des.*, Linkoping, Sweden, Aug. 2011, pp. 556–559.
[9] S. Han, C-l. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun.,* vol. 53, no. 1, pp. 186–194, Jan. 2015.
[10] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.
[11] M. Feng, S. Mao, and T. Jiang, "Base station ON-OFF switching in 5G wireless networks: Approaches and challenges," *IEEE Wireless Commun.,* vol. 24, no. 4, pp. 46–54, Aug. 2017.

[12] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE,* vol. 102, no. 3, pp. 366–385, Mar. 2014.
[13] Z. He, S. Mao, and T. S. Rappaport, "On link scheduling under blockage and interference in 60 GHz ad hoc networks," *IEEE Access J.*, vol. 3, pp. 1437–1449, 2015.
[14] Z. He and S. Mao, "A decomposition principle for link and relay selection in dual-hop 60 GHz networks," in *Proc. Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Apr. 2016, pp. 1683–1691.
[15] Z. He, S. Mao, S. Kompella, and A. Swami, "On link scheduling in dual-hop 60 GHz mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11180–11192, Dec. 2017.
[16] I.-K. Son, S. Mao, Y. Li, M. Chen, M. X. Gong, and T. S. Rappaport, "Frame-based medium access control for 5G wireless networks," *Springer Mobile Netw. Appl. J.*, vol. 20, no. 6, pp. 763–772, Dec. 2015.
[17] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. Annu. IEEE Int. Conf. Comput. Commun.*, Orlando, FL, USA, Mar. 2012, pp. 2149–2157.
[18] J. Qiao, X. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun.,* vol. 53, no. 1, pp. 209–215, Jan. 2015.
[19] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A.V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.,* vol. 33, no. 10, pp. 2052–2069, Oct. 2015.
[20] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, U.K.: Cambridge Univ. Press, 2004.
[21] W. Wang, Y. Chen, Q. Zhang, and T. Jiang, "A software-defined wireless networking enabled spectrum management architecture," *IEEE Commun. Mag.,* vol. 54, no. 1, pp. 33–39, Jan. 2016.
[22] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for energy efficient massive MIMO HetNets," in *Proc. IEEE INFOCOM'16,* San Francisco, CA, USA, Apr. 2016, pp. 1395–1403.
[23] Q. Xue, X. Fang, and C.-X. Wang, "Beamspace SU-MIMO for future millimeter wave wireless communications," *IEEE J. Sel. Areas Commun.,* vol. 35, no. 7, pp. 1564–1575, Jul. 2017.
[24] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.,* vol. 65, no. 1, pp. 6481–6494, Jan. 2017.
[25] Y. Wang, S. Mao, and T. S. Rappaport, "On directional neighbor discovery in mmWave networks," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst.,* Atlanta, GA, USA, Jun. 2017, pp. 1704–1713.
[26] J. Qiao, L. X. Cai, X. S. Shen, and J. W. Mark, "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Trans. Wireless Commun.,* vol. 10, no. 11, pp. 3824–3833, Nov. 2011.
[27] R. W. Irving, "An efficient algorithm for the "Stable Roommates" problem," *J. Algorithms*, vol. 6, no. 6, pp. 577–595, Dec. 1985.
[28] R. W. Heath *et al.*, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.,* vol. 10, no. 3, pp. 436–453, Apr. 2016.
[29] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.,* vol. 14, no. 2, pp. 1100–1114, Feb. 2015.
[30] M. Feng, S. Mao, and T. Jiang, "Dealing with link blockage in mmWave networks: D2D relaying or multi-beam reflection?" in *Proc. IEEE Annu. Int. Symp. Personal, Indoor, Mobile Radio Commun,* Montreal, QC, Canada, Oct. 2017, pp. 1–5.
[31] A. Alkhateeb, J. Mo, N. Gonzlez-Prelcic, and R.W. Heath, "MIMO precoding and combining solutions for millimeter-mave systems," *IEEE Commun.,* vol. 52, no. 12, pp. 122–131, Dec. 2014.
[32] L. Zhao, D. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Sel. Areas Commun.,* vol. 35, no. 7, pp. 1576–1690, Jul. 2017.
[33] Z. Niu, "TANGO: Traffic-aware network planning and green operation," *IEEE Wireless Commun.,* vol. 18, no. 5, pp. 25–29, Oct. 2011.
[34] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.,* vol. 12, no. 5, pp. 2126–36, May 2013.
[35] Y. Che, L. Duan, and R. Zhang, "Dynamic base station operation in large-scale green cellular networks," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 12, pp. 3127–3141, Dec. 2016.
[36] X. Zhang, S. Zhou, Z. Niu, and X. Lin, "RF chain and user selection for multiuser MIMO systems under random data arrival," in *Proc. IEEE Wireless Commun. Netw. Conf.,* New Orleans, LA, USA, Mar. 2015, pp. 872–877.
[37] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for green massive MIMO HetNets," *IEEE Trans. Wireless Commun.,* vol. 16, no. 11, pp. 7319–7332, Nov. 2017.
[38] S. Cai, Y. Che, L. Duan, J. Wang, S. Zhou, and R. Zhang, "Green 5G heterogeneous networks through dynamic small-cell operation," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 5, pp. 1103–1115, May 2016.
[39] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.,* vol. 35, no. 9, pp. 1936–1947, Sep. 2017.
[40] A. S. Cacciapuoti, "Mobility-aware user association for 5G mmWave networks," *IEEE Access J.*, vol. 5, pp. 21497–21507, 2017.
[41] S. Goyal, M. Mezzavilla, S. Rangan, S. Panwar, and M. Zorzi, "User association in 5G mmWave networks," in *Proc. IEEE Wireless Commun. Netw. Conf.,* San Francisco, CA, USA, Mar. 2017, pp. 1–6.
[42] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks," *IEEE/ACM Trans. Netw.,* vol. 23, no. 3, pp. 836–850, Jun. 2015.

Mingjie Feng (S'15) received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2010 and 2013, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. He was a Visiting Student in the Department of Computer Science, Hong Kong University of Science and Technology, in 2013. His research interests include cognitive radio networks, heterogeneous networks, massive MIMO, mmWave network, and full-duplex communication. He is a recipient of a Woltosz Fellowship at Auburn University.

Shiwen Mao (S'99–M'04–SM'09) received the Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. He is a Samuel Ginn Distinguished Professor and the Director of the Wireless Engineering Research and Education Center, Auburn University, Auburn, AL, USA. His research interests include wireless networks and multimedia communications. He is a Distinguished Speaker for the IEEE Vehicular Technology Society. He was the recipient of the 2017 IEEE ComSoc ITC Outstanding Service Award, the 2015 IEEE ComSoc TC-CSR Distinguished Service Award, the 2013 IEEE ComSoc MMTC Outstanding Leadership Award, and the NSF CAREER Award in 2010. He is a co-recipient of the 2017 Best Conference Paper Award of IEEE ComSoc MMTC, the Best Demo Award from IEEE SECON 2017, the Best Paper Awards from IEEE GLOBECOM 2016 and 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.

Tao Jiang (M'06–SM'10) received the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in April 2004. He is currently a Distinguished Professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan, China. From August 2004 to December 2007, he was with some universities, such as Brunel University and University of Michigan-Dearborn. He has authored or co-authored more than 300 technical papers in major journals and conferences and nine books/chapters in the areas of communications and networks. He was or is a symposium technical program committee member of some major IEEE conferences, including INFOCOM, GLOBECOM, and ICC, etc. He was invited to serve as TPC Symposium Chair for the IEEE GLOBECOM 2013, IEEEE WCNC 2013, and ICCC 2013. He was or is an Associate Editor for some technical journals in communications, including in the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE INTERNET OF THINGS JOURNAL, and he is currently the Associate Editor-in-Chief of China Communications, etc.