IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE EDGE COMPUTING AND MOBILE CLOUD COMPUTING: ADDRESSING HETEROGENEITY AND ENERGY ISSUES OF COMPUTE AND NETWORK RESOURCES

Received July 31, 2019, accepted August 30, 2019, date of publication September 4, 2019, date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939348

Arrival Rate-Based Average Energy-Efficient Resource Allocation for 5G Heterogeneous Cloud RAN

YIZHONG ZHANG, (Student Member, IEEE), GANG WU, (Member, IEEE), LIJUN DENG, (Student Member, IEEE), AND JINGWEI FU, (Student Member, IEEE)

National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Gang Wu (wugang99@uestc.edu.cn)

This work was supported in part by Innovation Fund of National Key Lab of S&T on Communications, the National Science Foundation of China (Grant No. 61771106), the Sichuan Science and Technology Program (Grant No. 2018HH0138), and the National Key R&D Program (2018YFB1800800).

ABSTRACT Heterogeneous cloud radio access network (H-CRAN) needs more elegant design to achieve higher energy efficiency and spectral efficiency than traditional cloud radio access networks. In this paper, we propose an energy-efficient resource allocation algorithm by taking into account the impact of arrival rates of various user traffic. Firstly, based on the power consumption model of the H-CRAN, the average energy-efficiency of the whole network is adopted as the optimization objective with multiple constraints of maximum transmit power, average power, and minimum data rate of each users, etc. In order to solve the non-convex and non-deterministic polynomial time- hardness (NP-hard) problem, we transform the objective function into analyzable multiple sub-problems by using fractional programming and norm approximation. Secondly, by the Lyapunov optimization method, we turn the original problem into a problem of system stability. Thirdly, we derive the closed expression of the optimal power allocation matrix and the optimal user association matrix with the Lagrangian dual decomposition. We propose a two-layer iterative algorithm to balance the power consumption and energy efficiency with a designed control factor. Both theoretical bound of average energy efficiency and length of data queuing are derived. Finally, the comprehensive numerical results demonstrate of convergence of the proposed algorithm and verify the performance gain by proposed energy-efficient resource allocation scheme.

INDEX TERMS Heterogeneous cloud radio access network (H-CRAN), resource allocation, green communication, average energy efficiency, Lyapunov optimization.

I. INTRODUCTION

It is a trend that millions more base stations (BSs) and billions more smart devices will be connected in the fifth generation (5G) wireless network [1]. The higher demand of functionality and data rate is challenging the power consumption of 5G. According to statistics in [2], the rise of power consumption in mobile communications rise up to 20% per year. Thus, the energy efficiency of 5G networks is expected to be increased 100× times to reduce power consumption [3]. Along with the explosive amount of data traffic, the machine-type communication (MTC) is one of the key areas in 5G. The massive MTC (mMTC) scenario specifies at least one million devices should be supported per square kilometer [4]. Traditional access technologies are no longer able to meet the require with the numerous deployment of sensors, accessories, and tools, which gives a rise to the Internet of things (IoT). With exponential growth of the IoT intelligent devices, the demand of massive calculation and storage is challenging 5G wireless communication [5]–[7]. Heterogeneous cloud radio access network (H-CRAN) is one of the most promising access technologies in IoT due to its scalability, flexibility and compatibility. The H-CRAN is composed of a series of remote radio heads (RRHs) and central baseband units (BBUs). Macro BSs are connected to the BBU pool through the backhaul of the X2/S1 interface, and the RRHs are connected to the BBU pool through the wireless fronthaul link. The interferences between RRHs can

The associate editor coordinating the review of this manuscript and approving it for publication was Guanding Yu.

be handled and even eliminated by cooperation in BBU pool. Control and data planes are separated [8]. Control functionalities are shifted to high power nodes (HPNs). RRHs are only utilized for signal transmission and reception, and the remaining important functions and procedures of the upper layers are implemented in the cloud BBU pool [9]–[11]. In addition, the centralized BBU pool facilitates the interaction of different cells and regions, enabling efficient cooperation between HPNs (e.g., macro or micro base stations) and low-power RRH nodes.

Access technology is one of the significant issues in H-CRAN. Due to the deployment of ultra-dense RRHs in the hotspot area of HetNets, each user can access numerous RRHs and HPNs, which introduces the complexity of scheduling [12]. The association rules between HPNs/RRHs and users have a great impact on performance of H-CRANs. First, the cooperation between high-power HPN nodes and low-power RRH nodes need to be scheduled. Second, although the centralized cloud computing-based cooperative processing techniques are considered in BBU pool, the cross-tier interference between RRHs and HPNs are critical and urgently need to be mitigated. To commercialize H-CRAN, an optimization allocation strategy for user access is urgently required. Furthermore, a typical feature of the forthcoming 5G is to provide diverse services with multiple class of quality of service (QoS) demand. Diverse requirements in three main scenarios require different system performance [13]–[15]. Essentially, the arrival rate and their requirements for network performance (e.g., maximum tolerable delay, minimum rate requirement, probability of burst, etc.) vary with different services. Therefore, the arrival rate is a significant factor to be considered.

There are some of related work on user access and resource allocation in H-CRAN [16]–[22]. In [16], the average energy efficiency optimization problem for delay-aware traffic in downlink C-RAN is studied. The original optimization problem is transformed by Lyapunov optimization method. The optimal beamforming vector and cloud-based resource allocation are obtained through WMMSE method. However, this work is limited to the traditional C-RAN architecture and does not take user traffic into account. An average energy efficiency optimization problem under the traditional network architecture is studied in [17]. The model of maximizing the average energy efficiency under the conditions of average transmit power and peak power is established, and the expressions of the average energy efficiency of the system and the delay with respect to the control factor are derived. However, this work is limited to the traditional network structure and can not be extended for the H-CRAN. The authors in [18] explored the system energy efficiency optimization problem under the C-RANs architecture. The authors assume that the fronthaul links are heterogeneous and try to minimize the system energy consumption. In [19], a sparse algorithm is proposed based on beamforming with channel matrix. The authors in [20] and [21] maximize overall network throughput under a series of restrictions. However, [18]–[21] overlook the impact of arrival rate on resource allocation and the sleep mode of fronthaul link. The authors in [22] propose an energy efficient radio resource management algorithm in H-CRAN. The consumption model of H-CRAN is established and the problem is solved by Lagrangian dual method. However, in our work the problem is simplified by Lyapunov optimization, which adapts a more concise factor to control the performance of system.

In this paper, we propose an efficient resource allocation algorithm called Arrival Rate based Average Energy-efficient (ARAE) dynamic resource allocation which introduces the arrival rate of users under H-CRAN architecture. The optimization parameters are the association matrix and power allocation matrix, and the optimization objective is the system average energy efficiency. First, the optimization problem is modeled under the constraints of average power, maximum power, minimum data rate, etc. For this non-convex and non-deterministic polynomial-time hardness (NP hard) problem, we first transform the objective function by fractional programming and norm approximation. Then, by using the Lyapunov optimization method, the original optimization is turned into a problem of system stability. Finally, based on the Lagrangian dual decomposition, the closed expressions of the optimal power allocation matrix and the optimal user association matrix are derived theoretically. With the help of gradient method, a two-layer iterative algorithm is proposed. A control factor V is introduced to balance the power consumption and energy efficient of the system. More importantly, this paper theoretically analyzes the performance of the algorithm. The average energy efficiency and the average data queue is proportional to V with the rate $O(1/V)$ and $O(V)$, respectively.

The major contributions can be summarized as follows.

- We analyse and formulate the optimization problem of maximising average energy efficient in H-CRAN under the constrains of average power, maximum power, average arrival rate, minimum data rate, and limits of RBs. The impact of arrival rate is considered in this problem.
- We propose an efficient dynamic resource allocation algorithm under H-CRAN. The non-convex optimization problem is converted by fractional programming and norm approximation, and solved by introducing Lyapunov optimization and Lagrangian dual decomposition. We mathematically derive the closed expressions of the optimal power allocation matrix and the optimal user association matrix. Moreover, the theoretical boundaries of average energy efficiency and average arrival data queue length of the proposed ARAE algorithm are given for analysis.
- An extensive system-level simulation is established to evaluate the performance of ARAE algorithm. The simulation results are well analysed and show good convergence. Different tradeoff can be achieved by control the value of factor V.

The remainder of this paper will be organized as follows. In Section II, we formally describe the system model.

TABLE 1. Notations of parameters.

Notations	Meanings
$\mathcal{M}$	The set of HPN/RRHs
$\mathcal{K}$	The set of users
$\mathcal{N}$	The set of RBs
W	The bandwidth of the system
$\alpha_{k,m,n}(t)$	Indicate if kth user accesses to mth RRH in nth RB
$p_{k,m,n}(t)$	The power of mth RRH allocated to kth user in nth RB
$\boldsymbol{\alpha}(t)$	Association matrix between users and RBs at slot t
$\boldsymbol{P}(t)$	Power allocation matrix at slot t
$g_{k,m,n}(t)$	Channel gain between kth user and mth RRH in nth RB
$R_{k,m,n}(t)$	Achievable rate of kth user accessed to mth RRH in nth RB
$R^T(t)$	Sum of achievable rate at slot t
$P^T(t)$	Sum of power consumption at slot t
V	Control factor

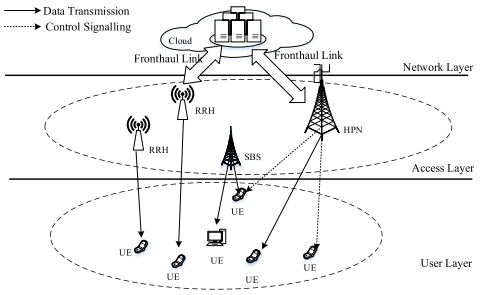

FIGURE 1. Network model.

In Section III, the proposed ARAE is devised. The performance of the ARAE is discussed in Section IV. Simulation results are presented in Section V. Finally, Section VI contains the conclusion.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Different from the traditional network architecture, the power consumption of the H-CRANs is mainly composed of three parts: HPNs/RRHs consumption, fronthaul link consumption, and BBU pool consumption. The notations of resource allocation model are shown in Table 1.

Fig.1 shows a double-layer H-CRAN system, containing one HPN and M RRHs. There are K users in the system and each user can arbitrarily access the HPN/RRHs to transmit data. The system has N resource blocks (RBs) to support the communication. Assume that the scheduler in the cloud acquires all information including channel state information (CSI). All the users utilize orthogonal frequency division multiplexed (OFDM) to access the HPN/RRHs and thus there is no interference between users. Traffic data for each user arrives randomly and independently, waiting for downlink scheduling. Here we assume that the arrival rate obeys the Poisson distribution with mean λ.

A detailed consumption model including network model, transmission model, and consumption model is proposed in [22]. The total system power consumption can be expressed as

$$P^T(t) = \sum_{m\in\mathcal{M}} P_m(t), \tag{1}$$

where

$$P_m(t) = P_m^s(t) + (\Delta_m + \rho_m) p_m(t) + \left(P_m^{F,A} - P_m^{F,S}\right) \|p_m(t)\|_0. \tag{2}$$

Here $P_m^S(t) = P_m^{B,S}(t) + P_m^{R,S}(t) + P_m^{F,S}(t)$ is the total static power consumption of RRHs, fronthaul link and cloud BBU pools. ρ_m denotes the power consumption factor and Δ_m represents power conversion factor of HPN/RRHs in active mode. $p_m(t) = \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} \alpha_{k,m,n}(t) p_{k,m,n}(t)$ is the total power consumption of mth RRH at slot t. The overall achievable rate of the system can be formulated as

$$R^T(t) = \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} \alpha_{k,m,n}(t) R_{k,m,n}(t), \tag{3}$$

where

$$R_{k,m,n}(t) = \frac{W}{N}\log_2\left(1 + \frac{p_{k,m,n}(t) g_{k,m,n}(t)}{W\cdot\sigma^2/N}\right), \tag{4}$$

is the maximum achievable rate for user k and $\alpha_{k,m,n}(t) \in \{0,1\}$ is an association indicator between kth user and mth RRH in nth RB. Indicator $\alpha_{k,m,n}(t) = 1$ if and only if the nth RB of mth RRH is allocated to kth user.

To allocate resources efficiently, different from [24]–[26], we define the objective function as the average power consumption, which is the ratio of the total long-term sum of transmission data $\overline{R^T}$ to the total long-term system energy consumption $\overline{P^T}$, i.e.,

$$\eta_{EE} = \frac{\overline{R^T}(\boldsymbol{\alpha},\boldsymbol{P})}{\overline{P^T}(\boldsymbol{\alpha},\boldsymbol{P})} = \lim_{t\to\infty}\frac{\frac{1}{t}\sum_{\tau=0}^{t-1} R^T(\tau)}{\frac{1}{t}\sum_{\tau=0}^{t-1} P^T(\tau)}, \tag{5}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}(0), \boldsymbol{\alpha}(1), \cdots, \boldsymbol{\alpha}(t-1)]$ and $\boldsymbol{P} = [\boldsymbol{P}(0), \boldsymbol{P}(1), \cdots, \boldsymbol{P}(t-1)]$ are the user association matrix and power allocation matrix, respectively. Our target is to find the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{P}$ to maximize η_{EE}, while satisfying the constraints of average transmit power of HPN/RRHs, instantaneous power peak, user arrival rate, minimum transmit rate of users, and limits of RBs. The whole problem can be formulated as

$$\begin{aligned}
\mathcal{P}_1:\ &\max_{\boldsymbol{\alpha},\boldsymbol{P}}\ \eta_{EE} = \frac{\overline{R^T}(\boldsymbol{\alpha},\boldsymbol{P})}{\overline{P^T}(\boldsymbol{\alpha},\boldsymbol{P})}\\
&s.t.\ \mathcal{C}_1: \overline{p_m} = \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{\tau=t-1}\mathbb{E}\{p_m(\tau)\} \le p_m^{avg},\\
&\quad \forall m\in\mathcal{M}\\
&\quad \mathcal{C}_2: p_m(t) \le p_m^{\max},\quad \forall m\in\mathcal{M}\\
&\quad \mathcal{C}_3: \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{A_k(\tau)\}\\
&\qquad \le \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{R_k(\tau)\},\quad \forall k\in\mathcal{K}\\
&\quad \mathcal{C}_4: \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\alpha_{k,m,n}(t) R_{k,m,n}(t) \geqslant R_k^{min},\\
&\quad \forall k\in\mathcal{K}\\
&\quad \mathcal{C}_5: \sum_{k\in\mathcal{K}}\alpha_{k,m,n}(t) \le 1,\quad \forall m\in\mathcal{M},\ \forall n\in\mathcal{N}\\
&\quad \mathcal{C}_6: \alpha_{k,m,n}(t) = \{0,1\},\quad \forall k\in\mathcal{K},\ \forall m\in\mathcal{M},\\
&\quad \forall n\in\mathcal{N},
\end{aligned} \tag{6}$$

where constraint $\mathcal{C}_1$ means the average transmit power of RRHs should be less than target p_m^{avg}. Constraint $\mathcal{C}_2$ restricts the maximum transmit power of each RRH. To ensure all users can be scheduled, constraint $\mathcal{C}_3$ means the average data rate of users should lower than the achievable rate. $\mathcal{C}_4$ is the constraint of minimum transmit rate of users. Constraints $\mathcal{C}_5$ and $\mathcal{C}_6$ mean that each RB is allocated to at most one user.

It is worthy to point out that we focus more on a long-term performance in $\mathcal{C}_3$. Hence the instantaneous arrival queue may larger than achievable rate at some slot. This happens when one's arrival data queue is too large to manage at that slot. However, the remaining data queue will be removed to next slot. Finally, the average energy-efficiency performance of the system achieves optimal since the average rate constrain of $\mathcal{C}_3$ is satisfied.

III. PROPOSED SOLUTION

Optimization objective in $\mathcal{P}_1$ is non-convex and NP hard [27]. In this section, $\mathcal{P}_1$ is transformed into a convex optimization problem through fractional programming, norm approximation and Lyapunov optimization. With the help of Lagrangian dual decomposition, we propose a resource allocation algorithm based on arrival rate.

A. FRACTIONAL PROGRAMMING

Note that the optimization objective η_{EE} is a ratio of two nonlinear functions, which leads to a dilemma of analysis. We utilize fractional programming to turn the objective into an equivalent linear form. Suppose the objective achieves the maximum value η_{EE}^{opt} when $\boldsymbol{\alpha}, \boldsymbol{P}$ are equal to optimal $\boldsymbol{\alpha}^*, \boldsymbol{P}^*$, respectively, i.e.,

$$\eta_{EE} = \frac{\overline{R^T}\left(\boldsymbol{\alpha}^*, \boldsymbol{P}^*\right)}{\overline{P^T}\left(\boldsymbol{\alpha}^*, \boldsymbol{P}^*\right)} = \max_{\boldsymbol{\alpha}, \boldsymbol{P}} \frac{\overline{R^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right)}{\overline{P^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right)}. \tag{7}$$

According to [31], the optimization objective η_{EE} achieves its maximum value if and only if

$$\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{P}} \overline{R^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) - \eta_{EE}^{opt} \overline{P^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) \\ &\quad = \overline{R^T}\left(\boldsymbol{\alpha}^*, \boldsymbol{P}^*\right) - \eta_{EE}^{opt} \overline{P^T}\left(\boldsymbol{\alpha}^*, \boldsymbol{P}^*\right) = 0, \end{aligned} \tag{8}$$

where $\boldsymbol{\alpha}, \boldsymbol{P}$ are the any feasible solutions that satisfy the constraints of $\mathcal{C}_1 \sim \mathcal{C}_6$. Therefore, the optimization problem is equivalent to

$$\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{P}} \overline{R^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) - \eta_{EE}^{opt} \overline{P^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) \\ &s.t.\ \mathcal{C}_1 \sim \mathcal{C}_6. \end{aligned} \tag{9}$$

It is hard to predict the range of η_{EE}^{opt} in advance. Similar to [16] and [18], for slot $t \in \{1, 2, \cdots\}$, we define the average energy efficiency $\eta_{EE}(t)$ as

$$\eta_{EE}(t) = \frac{\sum_{\tau=0}^{t-1} R^T\left(\alpha(\tau), P(\tau)\right)}{\sum_{\tau=0}^{t-1} P^T\left(\alpha(\tau), P(\tau)\right)}. \tag{10}$$

Especially, $\eta_{EE}(0) = 0$. The discrete variable $\alpha_{k,m,n}(t)$ is related in continuous domain. Now the optimization problem $\mathcal{P}_1$ can be formulated as

$$\begin{aligned} \mathcal{P}_2:\ &\max_{\boldsymbol{\alpha}, \boldsymbol{P}} \overline{R^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) - \eta_{EE}(t) \overline{P^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) \\ &s.t.:\ \mathcal{C}_1 \sim \mathcal{C}_5, \alpha_{k,m,n}(t) \in [0, 1]. \end{aligned} \tag{11}$$

B. NORM APPROXIMATION

Note that in (2), $P^T(t)$ has a $\mathcal{L}_0$-norm. Hence $\mathcal{P}_2$ is non-convex and NP-hard [27]. To approximately represent $\mathcal{L}_0-$norm by a continuous function [28]–[30], we use a popular approximation:

$$\|p_m(t)\|_0 \approx \sum_{\tau} \left(1 - e^{-\beta |p_m(\tau)|}\right), \tag{12}$$

where $p_m(\tau)$ is the τth row of $\boldsymbol{p_m}$, and the approximation are strictly equal when β approaches zero. Thus the value of β actually establishes a trade-off between accuracy and smoothness. The larger β leads to a better approximation, and the smaller β results in a smoother approximation.

Note that (12) still has an exponential form. By the first order Taylor expansion series of (12), which is approximately equal to

$$e^{-\beta |p_m(\tau)|} \approx 1 - \beta p_m(\tau). \tag{13}$$

However, (13) holds only when $1 - \beta p_m(\tau)$ is very close to zero. Thus, β should be a dynamic value to avoid too many local optimum (when β is too large) and no local optimum (when β is too small). In this paper, β is updated by

$$\beta(\tau + 1) = \frac{1}{p_m(\tau) + \varepsilon}, \tag{14}$$

where ε is a very small positive regularization factor for controlling the accuracy and robustness of the approximate solution.

It is worthy to point out that $p_m(\tau)$ means the power of the nth RB in mth RRH at slot τ. The larger value of $p_m(\tau)$ means the larger load the fronthaul link has. According to (21), when $p_m(\tau)$ is decreasing, the corresponding $\beta(\tau)$ increases, which will lead to a smaller value of $p_m(\tau + 1)$ at slot $\tau + 1$. At last the fronthaul link is forced to the sleep mode as $p_m(\tau)$ goes to zero.

Combing (2), (12), (13), and (14), it yields that

$$\overline{P^T}(t) = \sum_{m \in \mathcal{M}} \left[P_m^S(t) + \phi_m p_m(t)\right], \tag{15}$$

where $\phi_m = \Delta_m + \rho_m + \beta_m \left(P_m^{F,A} - P_m^{F,S}\right)$ stands for the effective power conversion factor of HPN/RRHs. Then the optimal problem $\mathcal{P}_2$ becomes

$$\begin{aligned} \mathcal{P}_3:\ &\max_{\boldsymbol{\alpha}, \boldsymbol{P}} \overline{R^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) - \eta_{EE}(t) \overline{P^T}\left(\boldsymbol{\alpha}, \boldsymbol{P}\right) \\ &s.t.:\ \mathcal{C}_1 \sim \mathcal{C}_5, \quad \alpha_{k,m,n}(t) \in [0, 1]. \end{aligned} \tag{16}$$

C. LYAPUNOV OPTIMIZATION

Note that $\mathcal{C}_3$ is a constraint about the queue of arrival rate and the average power of BS, we can transform $\mathcal{P}_3$ with the help of Lyapunov optimization. Lyapunov optimization, as an effective method, has been widely used to solve the problem between queue and system stability [24], [25]. The key point of Lyapunov optimization is to introduce control factors and establish quadratic Lyapunov functions. The optimal objective and the system stability are generally two contradictory factors and need a compromise. The system stability can be adjusted conveniently through jointly optimizing the valusers of control factors. Here are some basic assumptions:

1) STABILITY ASSUMPTION

A queue $U(t)$ is stable if and only if

$$\lim_{t\to\infty}\frac{\mathbb{E}\{|U(t)|\}}{t}=0. \tag{17}$$

2) INDEPENDENT AND IDENTICALLY DISTRIBUTED (I.I.D) ASSUMPTION

At slot $t\in\{1,2,\cdots\}$, power allocation matrix $\boldsymbol{P}(t)$ and association matrix $\boldsymbol{\alpha}(t)$ are independent and randomly choose a value from its space respectively.

3) BOUNDEDNESS ASSUMPTION

The expectation of the overall power $P^T(t)$ and the achievable rate $R^T(t)$ are bounded, i.e.,

$$\begin{aligned} P^T_{min}&\le\mathbb{E}\left\{P^T(t)\right\}\le P^T_{max},\\ R^T_{min}&\le\mathbb{E}\left\{R^T(t)\right\}\le R^T_{max}. \end{aligned} \tag{18}$$

Since the practical limitation, we also assume that there exists a positive integer θ, such that

$$\begin{aligned} \mathbb{E}\left\{y^2_m(t)\right\}&\le\theta,\\ \mathbb{E}\left\{R^2_k(t)\right\}&\le\theta,\\ \mathbb{E}\left\{A^2_k(t)\right\}&\le\theta, \end{aligned} \tag{19}$$

where $y_m(t)$ is the difference of actual power at slot t. The detailed definition will be discussed next.

A virtual power queue $Y_m(t)$ is defined to convert constrain $\mathcal{C}_1$ to the stability of the queue $Y_m(t)$. Considering the meaning of actual power, $Y_m(t)$ should satisfy

$$Y_m(t+1)=\max\{Y_m(t)+y_m(t),0\},\quad \forall m\in\mathcal{M}, \tag{20}$$

where $y_m(t)=p_m(t)-P^{avg}_m$. Under the help of $Y_m(t)$, we have the following Lemma 1.

Lemma 1: If $Y_m(t)$ is stable under an certain allocation algorithm, then the algorithm satisfies constrain $\mathcal{C}_1$.

Proof: The proof is presented in Appendix A.

Lemma 1 turns constrain $\mathcal{C}_1$ to the stability of queue $Y_m(t)$. Denote $\boldsymbol{Q}(t)=[Q_1(t),Q_2(t),\cdots,Q_k(t)]$ as the actual data queue at slot t. Due to the actual meaning, $Q_k(t)$ should satisfy

$$Q_k(t+1)=\max\left\{Q_k(t)-\frac{R_k(t)}{W},0\right\}+\frac{A_k(t)}{W}. \tag{21}$$

Note that both $Q_k(t)$ and $Y_m(t)$ have the same form, then we define a joint matrix of actual data queue $Q(t)$ and virtual power queue $Y_m(t)$ as $\Theta(t)=[Q(t),Y_m(t)]$. The quadratic Lyapunov function is defined as

$$L(\Theta(t))=\frac{1}{2}\left\{\sum_{k\in\mathcal{K}}Q^2_k(t)+\sum_{m\in\mathcal{M}}Y^2_m(t)\right\}, \tag{22}$$

and the Lyapunov penalty function is defined as

$$\begin{aligned} &\Delta L(\Theta(t))-V\left[\overline{R^T}(t)-\eta_{EE}(t)\,\overline{P^T}(t)\right]\\ &=L(\Theta(t+1))-L(\Theta(t))-V\left[\overline{R^T}(t)-\eta_{EE}(t)\,\overline{P^T}(t)\right]. \end{aligned} \tag{23}$$

Based on Lyapunov penalty function, we have the following Lemma 2.

Lemma 2: Lyapunov penalty function has an upper bound

$$\begin{aligned} &\mathbb{E}\{\Delta L(\Theta(t))\mid\Theta(t)\}\\ &\quad+V\mathbb{E}\left\{\eta_{EE}(t)\,\overline{P^T}(t)-\overline{R^T}(t)\mid\Theta(t)\right\}\\ &\le L_m+\sum_{m\in\mathcal{M}}Y_m(t)\,\mathbb{E}\left\{p_m(t)-P^{avg}_m\mid\Theta(t)\right\}\\ &\quad-\sum_{k\in\mathcal{K}}Q_k(t)\,\mathbb{E}\{A_k(t)-R_k(t)\mid\Theta(t)\}\\ &\quad+V\mathbb{E}\left\{\eta_{EE}(t)\,\overline{P^T}(t)-\overline{R^T}(t)\mid\Theta(t)\right\}. \end{aligned} \tag{24}$$

Proof: The proof is presented in Appendix B.

According to Lemma 2, minimize Lyapunov penalty function is equivalent to

$$\begin{aligned} \mathcal{P}_4:\ &\max_{\boldsymbol{\alpha},\boldsymbol{P}}\sum_{k\in\mathcal{K}}X_k(t)R_k(t)-\sum_{m\in\mathcal{M}}Z_m(t)p_m(t)\\ &s.t.:\mathcal{C}_2,\quad\mathcal{C}_4, \end{aligned} \tag{25}$$

where

$$Z_k(t)=Y_m(t)+V\phi_m\eta_{EE}(t), \tag{26}$$

$$X_k(t)=Q_k(t)/W+V. \tag{27}$$

D. LAGRANGIAN DUAL DECOMPOSITION

According to [32], the duality gap of the optimization problem is approaching to zero when the number of RBs is large. We use Lagrangian dual decomposition method to solve $\mathcal{P}_4$. The Lagrangian dual function can be written as

$$\begin{aligned} g(\boldsymbol{\mu},\boldsymbol{\gamma})=&\max_{\boldsymbol{\alpha},\boldsymbol{P}}\mathcal{L}(\alpha(t),P(t),\boldsymbol{\mu},\boldsymbol{\gamma})\\ &s.t.:\ \alpha_{k,m,n}\in[0,1], \end{aligned} \tag{28}$$

where

$$\begin{aligned}\mathcal{L}(\alpha(t), P(t), \mu, \gamma) &= \sum_{k\in\mathcal{K}} X_k(t) R_k(t) - \sum_{m\in\mathcal{M}} Z_m(t) p_m(t) \\ &+ \sum_{k\in\mathcal{K}} \mu_k \left(\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \alpha_{k,m,n}(t) R_{k,m,n}(t) - R_k^{min} \right) \\ &- \sum_{m\in\mathcal{M}} \gamma_m \left(\sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} \alpha_{k,m,n}(t) p_{k,m,n}(t) - P_m^{max} \right). \end{aligned} \tag{29}$$

Parameter μ_k and γ_m are Lagrangian multipliers. The dual form of $\mathcal{P}_4$ is

$$\begin{aligned}\mathcal{P}_5: \ &\max_{\boldsymbol{\mu},\boldsymbol{\gamma}} g(\boldsymbol{\mu}, \boldsymbol{\gamma}) \\ &s.t.: \boldsymbol{\mu} \geq 0, \quad \boldsymbol{\gamma} \geq 0. \end{aligned} \tag{30}$$

Note that the optimization object in $\mathcal{P}_5$ is convex and can be decomposed to a series independent subproblems

$$g(\mu, \gamma) = \max_{\{\alpha, P\}} [\sum_{n\in\mathcal{N}} g_n(\mu, \gamma) - \sum_{k\in\mathcal{K}} \mu_k R_k^{min} + \sum_{k\in\mathcal{K}} \gamma_m P_m^{max}], \tag{31}$$

where

$$\begin{aligned} g_n(\mu, \gamma) = \max_{\{\alpha, P\}} \sum_{k\in\mathcal{K}} \sum_{m\in\mathcal{M}} &[(\mu_k + X_k(t)) \alpha_{k,m,n}(t) R_{k,m,n}(t) \\ &- (\gamma_m + Z_m(t)) \alpha_{k,m,n}(t) p_{k,m,n}(t)]. \end{aligned} \tag{32}$$

Thus the optimal solution of power allocation is given by

$$p_{k,m,n}^*(t) = \max\left\{ \frac{B_0}{\ln 2} \cdot \frac{\mu_k + X_k(t)}{\gamma_m + Z_m(t)} - \frac{B_0 \sigma^2}{g_{k,m,n}(t)}, 0 \right\}, \tag{33}$$

and the optimal indication matrix $\boldsymbol{\alpha}(t)$ is given by

$$\alpha_{k,m,n}^* = \begin{cases} 1, & k = \arg\max_{k\in\mathcal{K}} T_{k,m,n}(t), \ T_{k,m,n}(t) > 0 \\ 0, & others, \end{cases} \tag{34}$$

where

$$\begin{aligned} T_{k,m,n}(t) = &(\mu_k + X_k(t)) R_{k,m,n}\left(p_{k,m,n}^*(t)\right) \\ &- (\gamma_m + Z_m(t)) p_{k,m,n}^*(t). \end{aligned} \tag{35}$$

Both μ_k and γ_m can be computed through the gradient descent method. In nth iteration, μ_k and γ_m are updated by

$$\mu_k(n) = \mu_k(n-1) - \delta_\mu(n) \cdot \nabla\mu_k(n), \quad \forall k \in \mathcal{K} \tag{36}$$

$$\gamma_m(n) = \gamma_n(n-1) - \delta_\gamma(n) \cdot \nabla\gamma_m(n), \quad \forall m \in \mathcal{M} \tag{37}$$

where $\delta_\mu(n)$ and $\delta_\gamma(n)$ is the learning rate of μ_k and γ_m, respectively.

The whole proposed ARAE dynamic resource allocation algorithm is shown in Algorithm 1.

Algorithm 1 Proposed Algorithm ARAE

For each slot t, compute $Q_k(t)$, $Y_m(t)$, $\eta_{EE}(t)$, $\boldsymbol{H}(t)$
Initialize ε_1, ε_2, $\eta_{EE} = 0$, β_m and other const
Initialize iteration $i = 1$
while $i \leq ITER\_MAX_1$ **do**
  if $\left|\beta_m^{(i)} - \beta_m^{(i-1)}\right| / \beta_m^{(i-1)} \leq \varepsilon_1$ **then**
    return $\beta^* = \beta^{(i)}$, $\{\boldsymbol{\alpha}^*, \boldsymbol{P}^*\} = \{\boldsymbol{\alpha}^{(i)}, \boldsymbol{P}^{(i)}\}$;
  else
    while $j \leq ITER\_MAX_2$ **do**
      $j = 1$;
      Compute $\boldsymbol{P}^j$, $\boldsymbol{\alpha}^j$ according to (33), (34);
      if $\boldsymbol{P}^j$ and $\boldsymbol{\alpha}^j$ converge **then**
        $\{\boldsymbol{\alpha}^{(i)}, \boldsymbol{P}^{(i)}\} = \{\boldsymbol{\alpha}^{(j)}, \boldsymbol{P}^{(j)}\}$;
        break;
      else
        $j = j + 1$;
        Compute gradient, update μ_k and γ_m;
      end if
      Update β_m according to (14);
      Compute $X_k(t)$ and $Z_m(t)$;
    end while
  end if
end while
return $\{\boldsymbol{\alpha}^*, \boldsymbol{P}^*\}$, $\eta_{EE}(t+1)$;

IV. PERFORMANCE ANALYSIS

To analyze the theoretical performance of proposed ARAE, we need the following lemma 3. The detailed proof can be found on [33] and [34].

Lemma 3: Assume that the expectation of arrival rate is λ. *For any* $\varepsilon > 0$, *if* $\mathcal{P}_1$ *has any solution satisfying* $\mathcal{C}_1 \sim \mathcal{C}_6$, *and the boundedness assumptions* (18) *and* (19) *hold, then for any* $\delta > 0$, *there exists an optimal allocation strategy, such that*

$$\mathbb{E}\left\{R^{T*}(t)\right\} \leq \mathbb{E}\left\{P^{T*}((t))\right\}\left(\eta_{EE}^{opt} + \delta\right), \tag{38}$$

$$\mathbb{E}\left\{y_m^*(t) \mid \Theta(t)\right\} = \mathbb{E}\left\{y_m^*(t)\right\} \leq \delta, \tag{39}$$

$$\mathbb{E}\left\{R_k^*(t) \mid \Theta(t)\right\} = \mathbb{E}\left\{R_k^*(t)\right\} \geq \lambda + \varepsilon, \tag{40}$$

where $y_m^*(t)$, $R_k^*(t)$, $R^{T*}(t)$, $P^{T*}(t)$ *are the optimal value for the optimal allocation strategy.*

Under Lemma 2 and Lemma 3, the theoretical performance of ARAE is analyzed, as specified in Theorem 1.

Theorem 1: If an allocation strategy generated by ARAE satisfies Lemma 2 and Lemma 3, then
(1) The virtual power queue $Y_m(t)$ *satisfies Lemma 1.*
(2) The average energy efficiency η_{EE} *satisfies*

$$\eta_{EE} \geq \eta_{EE}^{opt} - \frac{L_m}{VP_{\min}^T}. \tag{41}$$

(3) The length of average data queue $\overline{Q(t)}$ *satisfies*

$$\overline{Q(t)} \leq \frac{L_m + V\left(R_{\max}^T - \eta_{EE}^{opt} P_{\min}^T\right)}{\varepsilon / W}. \tag{42}$$

Proof: The proof is presented in Appendix C.

TABLE 2. Default simulation parameters.

Simulation Parameters		Value
Carrier frequserncy(GHz)		2(HPN),3.5(RRHs)
Bandwidth(MHz)		10
Number of RBs		20
ISD(m)		500
Minimum distance (m)	RRHs-RRHs	40
	RRHs-HPN	75
Power of RRHs (W)	Maximum transimt power	0.8
	Average transmit power	0.6
	Static power of ARAE	3.5
	Static power of JRS	9.2
Power of HPN (W)	Maximum transimt power	20
	Average transmit power	10
	Static power of ARAE	84
	Static power of JRS	84
Power of Fronthaul link (W)	Power of Active mode	37
	Power of Sleep mode	22.5
Power of BBU pool (W)	Static power	12
Path Loss (d in km)	HPN - user	PL=128.1+37.6log(d)
	RRHs - user	PL=128.1+37.6log(d)
Variance of shadow fading (dB)	HPN	8
	RRHs	10
N_0 (dBm/Hz)		-174
Δ_m		4.75
ρ_m		5.676

Equation (41) indicates that the average energy efficiency η_{EE} can asymptotically achieve the optimal value η_{EE}^{opt} by increasing the value of V. Meanwhile, (41) and (42) show that the average power efficiency and the average actual data queue is proportional to V with the rate $O(1/V)$ and $O(V)$, respectively. This is mainly because there is a nonlinear relationship between the transmission power and the achievable rate, which means that the system has more impact on η_{EE} with the increasing of V. Moreover, the increase of η_{EE} will lead to a decrease of the transmission power and the achievable data rate, resulting in the increasing length of data queue.

V. NUMERICAL RESULTS

In this section, a system-level simulation is established. We verify the correctness of the ARAE algorithm and theoretical derivations in section IV. Then the impact of control factor V and two important energy factors are evaluated. The default simulation parameters are shown in Table 2 [24], [35], [36]. The simulation is compared with a baseline algorithm [16]:

- **Joint Resource Scheduling (JRS) for Delay-Aware Traffic Algorithm** [16]: The JRS algorithm proposes an EE joint resource scheduling scheme for delay-aware traffic. Since JRS does not contain the factors Δ_m and ρ_m, they are evaluated in different arrival rates.

Assume that the HPN/RRHs and the cloud BBU pool are connected by fiber, and the capacity of the fronthaul link is not limited. The high-power HPN is located in the center of hexagon cell and is responsible for controlling signaling and data transmission. The low-power RRHs and users are evenly distributed in the simulation area.

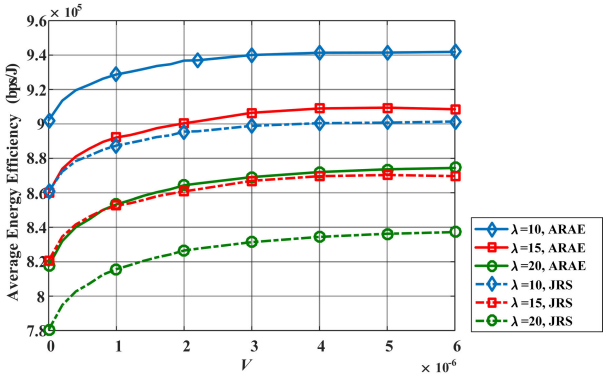

FIGURE 2. System energy efficiency against V.

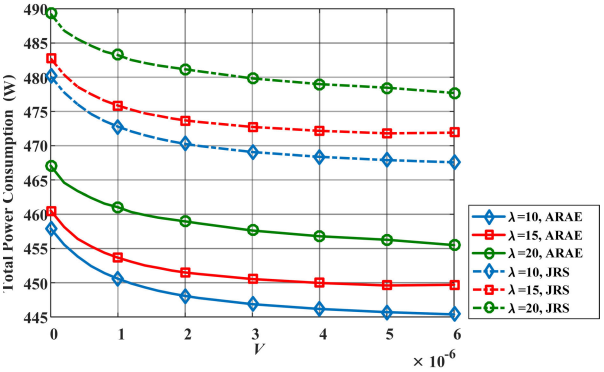

FIGURE 3. System power consumption against V.

Fig.2 shows the relationship between the average energy efficiency η_{EE} and the control factor V. As V increases, η_{EE} increases with the speed $O(1/V)$ roughly, which verifies the correctness of the theoretical derivation (41). In addition, η_{EE} decreases with the increasing of arrival rate λ for a fixed V. The reason is that the system needs more energy to support the data traffic for a larger λ. Hence the total energy consumption increases, and leads to a reduction of η_{EE}.

In Fig.3, as the control factor V increases, the total energy consumption decreases. This is mainly because the power and the maximum achievable rate are in a nonlinear relationship. From (22) the definition of Lyapunov's penalty function, it can be referred that the increase of the control factor V indicates that the system wants a higher energy efficiency, and the total energy consumption needs to be reduced. Hence the system can achieve asymptotically optimal performance through the adjustment of V. In addition, as the arrival rate λ increases, the total energy consumption increases as well in order to support the larger traffic.

As shown in Fig.4, the length of average data queue $\overline{Q(t)}$ increases with the speed of $O(V)$ roughly as V increases. Fig.2 and Fig.4 indicate that there is a tradeoff between η_{EE} and $\overline{Q(t)}$, which can be expressed by $(O(1/V), O(V))$ quantitatively. Therefore, if the network should work in an ideal or predefined stable state for a long time, a suitable control factor V must be chosen. Specifically, if the system

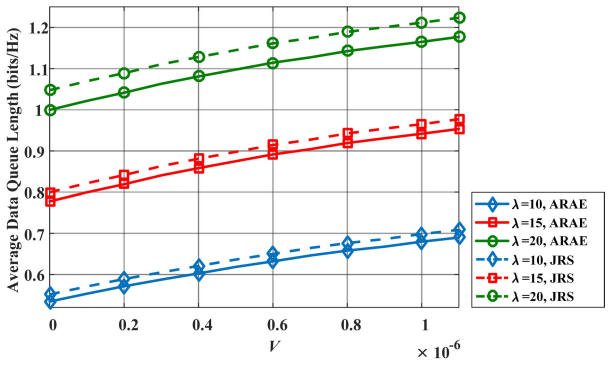

FIGURE 4. Average arrival data queue length against V.

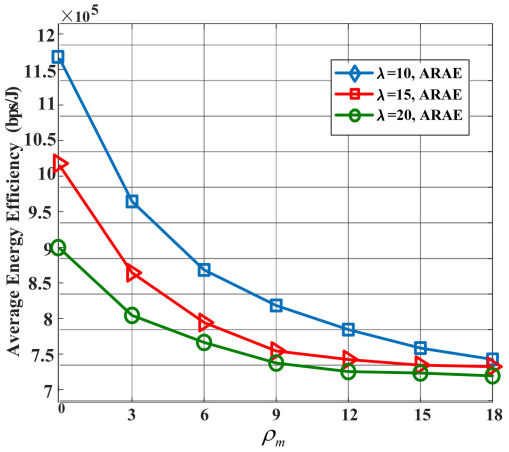

FIGURE 5. Energy factor ρ_m against average EE.

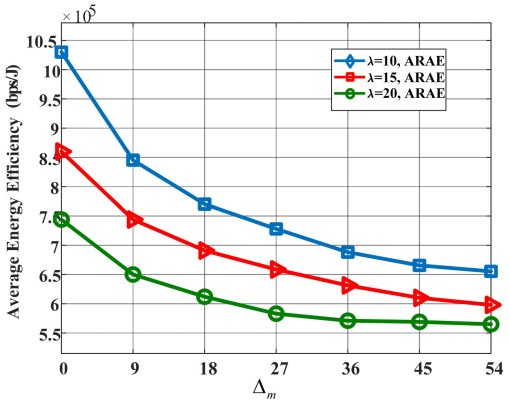

FIGURE 6. Energy factor Δ_m against average EE.

requires higher energy efficiency, V should be larger; conversely, if the system is delay-sensitive, the corresponding V should be smaller.

Fig.5 shows that η_{EE} decreases with the increase of ρ_m. That is because ρ_m stands for the ability of processing signals of BBU pool. The BBU pool will need more energy to processing the same signal as the increasing of ρ_m, leading to a lower performance of η_{EE}. The performance of a lower λ can achieve much higher η_{EE} than higher λ, e.g., 41% better in η_{EE} between $\lambda = 10$ and $\lambda = 20$ when $\rho_m = 3$. However, this gap shrinks with the weaker ability of processing signals in BBU pool. Hence the ability of BBU pools is one of key issues in H-CRAN.

Fig.6 illustrates the relationship between factor Δ_m and η_{EE}. It is observed that a larger Δ_m leads to a worse performance of η_{EE}. Since Δ_m represents energy conversion ability of HPN/RRHs in active mode, and a small Δ_m means the HPN/RRHs are more efficient in converting energy for transmission. That is, less power will be needed to transmit the same data traffic in HPN/RRHs. In addition, for a fixed Δ_m, the increase of λ leads to a poorer performance of η_{EE}. That is because for the same level of conversion ability, the overload will lower the energy efficient if a larger traffic rate is required.

VI. CONCLUSION

In this paper, we discuss the resource allocation problem in H-CRAN. A resource allocation algorithm has been proposed based on arrival rate. The power consumption model is established, and the original optimization problem is transformed by fractional programming, norm approximation, and Lyapunov optimization. The solution is based on the Lagrangian dual decomposition and gradient descent method. The simulation results show the validity of theoretical derivation and the influence of control factor V. The total power consumption cuts down by controlling the value of V, which has a significant impact of green communication.

APPENDIX A
PROOF OF LEMMA 1

According to (20), $Y_m(t+1) \geq Y_m(t) + y_m(t)$. Calculating expectation for $t \in \{0, 1, \cdots, K\}$ on both side, we have

$$\mathbb{E}\{Y_m(K)\} \geq \sum_{t=0}^{K} \mathbb{E}\{p_m(t)\} - KP^{avg}. \tag{43}$$

One constructs the inequality

$$\lim_{K\to\infty} \frac{\mathbb{E}\{Y_m(K)\}}{K} \geq \lim_{K\to\infty} \frac{1}{K} \sum_{t=0}^{K} \mathbb{E}\{p_m(t)\} - P^{avg} \tag{44}$$

via dividing both sides of (43) by K and letting $K \to \infty$. Combine Jensen's inequality

$$0 \leq |\mathbb{E}\{Y_m(K)\}| \leq \mathbb{E}\{|Y_m(K)|\} \tag{45}$$

and

$$\lim_{K\to\infty} \mathbb{E}\{|Y_m(K)|\}/K = 0, \tag{46}$$

the expectation

$$\lim_{K\to\infty} |\mathbb{E}\{Y_m(K)\}|/K = 0 \tag{47}$$

is obtained according to squeezed theorem. Hence,

$$\lim_{K\to\infty} \sum_{t=0}^{K-1} \mathbb{E}\{p_m(t)\} \leq P^{avg}. \tag{48}$$

Lemma 1 has been proved. □

APPENDIX B
PROOF OF LEMMA 2

Obviously the inequality

$$\left\{\max\left[Q-R,0\right]^2+A\right\}^2 \le Q^2+R^2+A^2-2Q\left(R-A\right) \tag{49}$$

always holds. Applying (49) to (21) by taking square on both sides and calculating the sum of all users, it yields that

$$\sum_{k\in\mathcal{K}}\left(\frac{Q_k^2\left(t+1\right)-Q_k^2\left(t\right)}{2}\right) \le \sum_{k\in\mathcal{K}}\frac{R_k^2\left(t\right)+A_k^2\left(t\right)}{2W^2}-\sum_{k\in\mathcal{K}}\frac{Q_k\left(t\right)}{W}\left(R_k\left(t\right)-A_k\left(t\right)\right). \tag{50}$$

Similarly, the virtual power queue $Y_m(t)$ satisfies

$$\sum_{m\in\mathcal{M}}\left(\frac{Y_k^2\left(t+1\right)-Y_k^2\left(t\right)}{2}\right) \le \sum_{m\in\mathcal{M}}\frac{\left(p_m\left(t\right)-P_m^{avg}\right)^2}{2}-\sum_{m\in\mathcal{M}}Y_m\left(t\right)\left(p_m\left(t\right)-P_m^{avg}\right). \tag{51}$$

Hence, the difference of quadratic Lyapunov function holds

$$\begin{aligned} &L\left(\Theta\left(t+1\right)\right)-L\left(\Theta\left(t\right)\right) \\ &\quad\le \sum_{m\in\mathcal{M}}\frac{\left(p_m\left(t\right)-P_m^{avg}\right)^2}{2}+\sum_{k\in\mathcal{K}}\frac{R_k^2\left(t\right)+A_k^2\left(t\right)}{2W^2} \\ &\qquad+\sum_{m\in\mathcal{M}}Y_m\left(t\right)\left(p_m\left(t\right)-P_m^{avg}\right) \\ &\qquad-\sum_{k\in\mathcal{K}}\frac{Q_k\left(t\right)}{W}\left(R_k\left(t\right)-A_k\left(t\right)\right) \end{aligned} \tag{52}$$

by combining (50) and (51). Applying (52) to the expectation of (23) yields

$$\begin{aligned} &\mathbb{E}\left\{\Delta L\left(\Theta\left(t\right)\right)\mid\Theta\left(t\right)\right\} \\ &\qquad+V\mathbb{E}\left\{\eta_{EE}\left(t\right)\ \overline{P^T}\left(t\right)-\overline{R^T}\left(t\right)\mid\Theta\left(t\right)\right\} \\ &\quad\le L_m+\sum_{m\in\mathcal{M}}Y_m\left(t\right)\mathbb{E}\left\{p_m\left(t\right)-P_m^{avg}\mid\Theta\left(t\right)\right\} \\ &\qquad-\sum_{k\in\mathcal{K}}Q_k\left(t\right)\mathbb{E}\left\{A_k\left(t\right)-R_k\left(t\right)\mid\Theta\left(t\right)\right\} \\ &\qquad+V\mathbb{E}\left\{\eta_{EE}\left(t\right)\ \overline{P^T}\left(t\right)-\overline{R^T}\left(t\right)\mid\Theta\left(t\right)\right\}, \end{aligned} \tag{53}$$

where

$$\begin{aligned} L_m \ge \frac{1}{2}\sum_{m\in\mathcal{M}}&\mathbb{E}\{p_m\left(t\right)-P_m^{avg}\mid\Theta\left(t\right)\}^2 \\ &-\frac{1}{2W^2}\sum_{k\in\mathcal{K}}\mathbb{E}\left\{R_k^2\left(t\right)+A_k^2\left(t\right)\mid\Theta\left(t\right)\right\}. \end{aligned} \tag{54}$$

Then Lemma 2 has been proved. □

APPENDIX C
PROOF OF THEOREM 1

Applying (38) and (40) to (53) yields

$$\begin{aligned} &\mathbb{E}\left\{\Delta L\left(\Theta\left(t\right)\right)\mid\Theta\left(t\right)\right\} \\ &\qquad+V\mathbb{E}\left\{\eta_{EE}\left(t\right)\ \overline{P^T}\left(t\right)-\overline{R^T}\left(t\right)\mid\Theta\left(t\right)\right\} \\ &\quad\le L_m-\varepsilon\sum_{k\in\mathcal{K}}\frac{Q_k\left(t\right)}{W}-V\eta_{EE}^{opt}\mathbb{E}\left\{P^{T*}\left(t\right)\right\} \\ &\qquad+V\eta_{EE}\left(t\right)\mathbb{E}\left\{P^{T*}\left(t\right)\right\} \\ &\quad\le L_m-V\eta_{EE}^{opt}\mathbb{E}\left\{P^{T*}\left(t\right)\right\}+V\eta_{EE}\left(t\right)\mathbb{E}\left\{P^{T*}\left(t\right)\right\} \end{aligned} \tag{55}$$

by letting $\delta\to 0$. Since $\Delta L\left(\Theta\left(t\right)\right)$ has an upper bound θ, i.e., $\Delta L\left(\Theta\left(t\right)\right)\le\theta$, the difference of $E\left\{L\left(\Theta\left(t\right)\right)\right\}$ is also bounded by

$$\mathbb{E}\left\{L\left(\Theta\left(t+1\right)\right)\right\}-\mathbb{E}\left\{L\left(\Theta\left(t\right)\right)\right\}\le\theta \tag{56}$$

via calculating difference in (55). Calculating the sum for $t\in\{0,1,\cdots,K\}$ to obtain that

$$\mathbb{E}\left\{L\left(\Theta\left(K\right)\right)\right\}-\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}\le K\theta. \tag{57}$$

Note that the inequality $\mathbb{E}\left\{\left|Y_m\left(K\right)\right|^2\right\}\ge\mathbb{E}^2\left\{\left|Y_m\left(K\right)\right|\right\}$ always holds, and then applying (57) to (22) yields

$$\mathbb{E}\left\{\left|Y_m\left(K\right)\right|\right\}\le\sqrt{2K\theta+2\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}}. \tag{58}$$

The stability constrain

$$\lim_{K\to\infty}\frac{\mathbb{E}\left\{\left|Y_m\left(K\right)\right|\right\}}{K}=0 \tag{59}$$

is achieved via dividing both sides on (58) by K and let $K\to\infty$. So far, it has been proved that the $Y_m\left(t\right)$ is stable and satisfies Lemma 1.

On the other hand, a similar inequality

$$\begin{aligned} &\mathbb{E}\left\{L\left(\Theta\left(K\right)\right)\right\}-\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\} \\ &\qquad+V\mathbb{E}\left\{\sum_{t=0}^{K-1}\mathbb{E}\left\{\eta_{EE}\left(t\right)P^T\left(t\right)\right\}-\sum_{t=0}^{K-1}\mathbb{E}\left\{R^T\left(t\right)\right\}\right\} \\ &\quad\le K\left[L_m-V\eta_{EE}^{opt}\mathbb{E}\left\{P^{T*}\left(t\right)\right\}\right] \\ &\qquad+V\mathbb{E}\left\{P^{T*}\left(t\right)\right\}\sum_{t=0}^{K-1}\eta_{EE}\left(t\right) \end{aligned} \tag{60}$$

can be obtained by calculating expectations for different slot t as (55). Note the fact that

$$\begin{aligned} \lim_{k\to\infty}\left[\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}\left\{\eta_{EE}P^T\left(t\right)\right\}-\lim_{k\to\infty}\sum_{t=0}^{K-1}\mathbb{E}\left\{R^T\left(t\right)\right\}\right] \\ =\eta_{EE}\overline{P^T}-\overline{R^T}=0, \end{aligned} \tag{61}$$

it is sufficient to derive

$$\frac{L_m}{V}-\eta_{EE}^{opt}\mathbb{E}\left\{P^{T*}\left(t\right)\right\}+\mathbb{E}\left\{P^{T*}\left(t\right)\right\}\eta_{EE}\ge 0 \tag{62}$$

via dividing both sides of (60) by VK and letting $K \to \infty$. Indeed, (62) is equivalent to

$$\eta_{EE} \geq \eta_{EE}^{opt} - \frac{L_m}{VP_{min}^T}. \tag{63}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left\{L\left(\Theta\left(K\right)\right)\right\} - \mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\} \\
&\quad + V\mathbb{E}\left\{\sum_{t=0}^{K-1}\mathbb{E}\left\{\eta_{EE}\left(t\right)P^T\left(t\right)\right\} - \sum_{t=0}^{K-1}\mathbb{E}\left\{R^T\left(t\right)\right\}\right\} \\
&\leq K\left[L_m - V\eta_{EE}^{opt}\mathbb{E}\left\{P^{T*}\left(t\right)\right\}\right] - \frac{\varepsilon}{W}\sum_{t=0}^{K-1}\sum_{k\in\mathcal{K}}\mathbb{E}\left\{Q_k\left(t\right)\right\} \\
&\quad + V\mathbb{E}\left\{P^{T*}\left(t\right)\right\}\sum_{t=0}^{K-1}\eta_{EE}\left(t\right).
\end{aligned} \tag{64}$$

If both sides on (64) are divided by $\varepsilon K/W$ and let $K \to \infty$, then

$$\overline{Q\left(t\right)} \leq \frac{L_m + V\left(R_{max}^T - \eta_{EE}^{opt}P_{min}^T\right)}{\varepsilon/W}. \tag{65}$$

The Theorem 1 has been proved. □

REFERENCES

[1] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[2] M. Webb, "SMART 2020: Enabling the low carbon economy in the information age," *Climate Group. London*, vol. 1, no. 1, p. 1, Jan. 2008.

[3] *IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, ITU Standard M.2083-0, Sep. 2015.

[4] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.

[5] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Bus. Horizons*, vol. 58, no. 4, pp. 431–440, 2015.

[6] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[7] J. M. Khurpade, D. Rao, and P. D. Sanghavi, "A survey on IOT and 5G network," in *Proc. Int. Conf. Smart City Emerg. Technol. (ICSCET)*, Mumbai, India, 2018, pp. 1–3.

[8] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar./Apr. 2015.

[9] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: Key challenges for the radio-access network," *IEEE Veh. Technol. Mag.*, vol. 8, no. 3, pp. 47–53, Sep. 2013.

[10] M. A. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Z. Granville, J. Rochol, L. DaSilva, and C. B. Both, "Resource sharing in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 74–82, Jun. 2015.

[11] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[12] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: Advances and challenges," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 66–73, Jun. 2015.

[13] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[14] Alliance NGMN. (Feb. 2015). *5G White Paper. Next Generation Mobile Networks*. [Online]. Available: https://ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf

[15] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[16] K. Wang, W. Zhou, and S. Mao, "Energy efficient joint resource scheduling for delay-aware traffic in cloud-RAN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[17] Y. Li, M. Sheng, Y. Shi, X. Ma, and W. Jiao, "Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 5921–5931, Nov. 2014.

[18] O. Dhifallah, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Joint hybrid backhaul and access links design in cloud-radio access networks," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Boston, MA, USA, Jan. 2015, pp. 1–5.

[19] X. Zhang, M. Jia, X. Gu, and Q. Guo, "An energy efficient resource allocation scheme based on cloud-computing in H-CRAN," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4968–4976, Jun. 2019.

[20] Y. L. Lee, L. Wang, T. C. Chuah, and J. Loo, "Joint resource allocation and user association for heterogeneous cloud radio access networks," in *Proc. 28th Int. Teletraffic Congr. (ITC)*, Würzburg, Germany, 2016, pp. 87–93.

[21] M. Ali, Q. Rabbani, M. Naeem, S. Qaisar, and F. Qamar, "Joint user association, power allocation, and throughput maximization in 5G H-CRAN networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9254–9262, Oct. 2017.

[22] Q. Liu, T. Han, N. Ansari, and G. Wu, "On designing energy-efficient heterogeneous cloud radio access networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 721–734, Sep. 2018.

[23] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. Di Girolamo, and G. Giuliani, "RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network," in *Proc. Future Netw. Mobile Summit*, Lisboa, Portugal, 2013, pp. 1–8.

[24] D. Sabella, A. de Domenico, E. Katranaras, M. A. Imran, M. di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access*, vol. 2, pp. 1586–1597, 2014.

[25] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[26] Y. Chen, X. Wen, Z. Lu, and H. Shao, "Energy efficient clustering and beamforming for cloud radio access networks," *Mobile Netw. Appl.*, vol. 22, no. 3, pp. 589–601, Jun. 2017.

[27] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[28] Y. Gu, J. Jin, and S. Mei, "ℓ_0 norm constraint LMS algorithm for sparse system identification," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 774–777, Sep. 2009.

[29] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. ICML*. vol. 98, Jul. 1998, pp. 82–90.

[30] P. Sudhakar and R. Gribonval, "A sparsity-based method to solve permutation indeterminacy in frequency-domain convolutive blind source separation," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat. (ICASS)*, Berlin, Germany, 2009, pp. 338–345.

[31] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.

[32] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[33] R. Berry, E. Modiano, and M. Zafer, *Energy-Efficient Scheduling under Delay Constraints for Wireless Networks* (Synthesis Lectures on Communication Networks), vol. 5, no. 2. San Rafael, CA, USA: Morgan & Claypool, 2012, pp. 1–96.

[34] M. J. Neely, "Dynamic optimization and learning for renewal systems," *IEEE Trans. Autom. Control*, vol. 58, no. 1, pp. 32–46, Jan. 2013.

[35] X. Chu, D. Lopez-Perez, Y. Yang, and F. Gunnarsson, *Heterogeneous Cellular Networks: Theory, Simulation and Deployment*. Cambridge, U.K.: Cambridge Univ. Press, 2013, pp. 10–280.
[36] M. Gruber, O. Blume, D. Ferling, D. Zeller, M. A. Imran, and E. C. Strinati, "EARTH—Energy aware radio and network technologies," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Tokyo, Japan, Sep. 2009, pp. 1–5.

YIZHONG ZHANG received the B.S. degree from the University of Electronic Science and Technology of China, in 2017, where he is currently pursuing the M.S. degree with the National Key Laboratory of Science and Technology on Communication. His research interests include noncoherent communication, NOMA, and resource allocation.

GANG WU (M'05) received the B.Eng. and M.Eng. degrees from the Chongqing University of Post and Telecommunications, Chongqing, China, in 1996 and 1999, respectively, and the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2004. In June 2004, he joined the UESTC. He was a Research Fellow with the Positioning and Wireless Technology Centre, Nanyang Technological University, Singapore, from November 2005 to February 2007. He was a Visiting Professor with the Georgia Institute of Technology, Atlanta, GA, USA, from October 2009 to September 2010. He is currently a Professor with the National Key Laboratory of Science and Technology on Communications, UESTC. His research interest includes PHY/MAC techniques for 5G. He was a co-recipient of the IEEE Globecom 2012 Best Paper Award. He is also an Associate Editor of *Science China Information Sciences*.

LIJUN DENG received the B.S. degree from Huaqiao University, China, in 2017. She is currently pursuing the M.S. degree with the National Key Laboratory of Science and Technology on Communication, University of Electronic Science and Technology of China. Her research interests include wireless communication and machine learning.

JINGWEI FU received the B.S. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2017. She is currently pursuing the M.S. degree with the National Key Laboratory of Science and Technology on Communication, University of Electronic Science and Technology of China. Her research interests include the 5G new multiple access technology, cooperative communications, and machine learning.