On Designing Energy-Efficient Heterogeneous Cloud Radio Access Networks

Qiang Liu, Tao Han, Nirwan Ansari, Gang Wu

Abstract—Heterogeneous cloud radio access network (H-CRAN) promises higher energy efficiency (EE) than the conventional cellular networks by centralizing the baseband signal processing into the baseband unit (BBU) pool hosted by cloud computing platforms. Because of the difference between H-CRAN and conventional cellular networks, existing energy-efficient networking mechanisms designed for conventional cellular networks cannot fully leverage H-CRAN in terms of reducing the network energy consumption. In this paper, we bridge this gap by proposing a radio resource management scheme to optimize the network energy efficiency (NEE) of H-CRAN. We develop a network energy consumption model that characterizes the energy consumption of radio access points (RAPs), fronthaul, and the BBU pool in H-CRAN. Based on the network energy consumption model, we formulate the NEE optimization problem with the consideration of the capacity constrained fronthaul. The NEE optimization problem is a mixed integer non-linear programming problem (MINLP). We propose the H-CRAN energy-efficient radio resource management (HERM) algorithm to solve the NEE optimization problem efficiently. Various properties of the proposed solution are derived and extensive simulations are conducted. The simulation results show that the HERM algorithm significantly improves the NEE of H-CRAN. As compared with a baseline algorithm in which the radio resource management is not optimized, HERM boosts the NEE by 59% under the dynamic network traffic. As compared to an energy-efficient radio resource allocation (ERA) algorithm which does not optimize the energy consumption of the BBU pool, the NEE of H-CRAN achieved by the HERM algorithm is up to 51% better than that by the ERA algorithm with network traffic dynamics.

Index Terms—Energy efficiency, Resource management, H-CRAN, Heterogeneous fronthaul.

I. INTRODUCTION

The fifth generation mobile communication system (5G) is expected to support massive connections with higher data rate, lower latency, ultra-higher reliability and higher energy efficiency [1]. Heterogeneous networks (HetNet) and cloud radio access networks (C-RAN) are promising technologies to reduce energy consumption and improve energy efficiency of cellular networks toward 5G [2], [3].

In HetNet, small cell base stations (ScBSs) are deployed to provide high-speed data transmissions in a small area while macro BSs are designed to provide seamless coverage to a large area. The dense deployment of ScBSs supports massive mobile connections and significantly increases the network throughput. However, such a network deployment may lead to extensive coverage overlaps among base stations (BSs) and severe inter-BS interference in HetNet. As a result, the advantages of HetNet will be neutralized by the interference and the complicated BS coordination required in ultra-dense mobile networks [4].

C-RAN decouples the RF and baseband signal processing to the remote radio head (RRH) and baseband unit (BBU), respectively. BBUs are centralized as a BBU pool which is implemented on cloud computing platforms. Owing to the centralized baseband signal processing, the channel state information (CSI) can be efficiently shared among BBUs. By leveraging the CSI information, C-RAN allows large-scale cooperative communications and thus mitigates the severe inter-BS interference. Moreover, since the BBU pool is implemented on cloud computing platforms, the computing resource for baseband signal processing can be dynamically provisioned based on the traffic demand [5]. However, the separation of the BBU and RRH imposes stringent capacity demand on fronthaul that carries the baseband signal between the BBU and RRH. Therefore, the limited fronthaul capacity may greatly impact the performance of C-RAN.

Heterogeneous cloud radio access network (H-CRAN), which leverages the centralized signal processing mechanism of C-RAN and the heterogeneous BS deployment scheme of HetNet, is a promising network architecture for next generation mobile communications [4], [6], [7]. As compared with C-RAN, H-CRAN introduces the enhanced RRHs (eRRHs) which support baseband signal processing functions. In H-CRAN, the eRRH is deployed to provide wide-area seamless coverage while the RRH is deployed to provision high-speed data transmission to a small area as shown in Figure 1. Since the eRRH is able to process baseband signal, the fronthaul capacity requirements are alleviated in H-CRAN [4], [8], [9].

H-CRAN combines the merits of both HetNet and C-RAN and thus has the potential of significantly enhance the energy
efficiency of mobile networks. However, designing energy-efficient H-CRAN is a challenging and yet undeveloped research topic. In H-CRAN, the baseband signal processing functions are partially implemented in the BBU pool. As a result, the energy consumption model of H-CRAN is fundamentally different from conventional cellular networks [10]. Owing to these differences, the existing energy-efficient networking solutions cannot be directly applied to H-CRAN. Moreover, the separation of the RF and baseband signal processing introduces a stringent requirement on the capacity of the fronthaul that connects the RRH/eRRH and BBU pool. Although optical fiber links are capable to provide high capacity, their availability is very limited due to the prohibitive deployment cost in large-scale H-CRAN [11]. Therefore, cost-effective wireless fronthaul will be indispensable in deploying H-CRAN [4], [12]. The wireless fronthaul may not provide as high capacity as optical fiber links and thus become the bottleneck of H-CRAN. Meanwhile, optical fiber fronthaul can also be the bottleneck when the network traffic loads are not well balanced. Hence, the capacity of fronthaul is a unique networking constraint which should be carefully considered in designing energy-efficient networking solutions in H-CRAN.

In this paper, we study the network energy efficiency (NEE) of the downlink data transmission in H-CRAN with constrained fronthaul capacity. In order to analyze the NEE of H-CRAN, we build a comprehensive energy consumption model that characterizes the energy consumption of RRHs, eRRHs, fronthaul, and the BBU pool in H-CRAN. Based on the energy consumption model, the resource block (RB) assignments and corresponding power allocations determines the network energy consumption. Therefore, we formulate the NEE optimization problem to investigate the optimal RB assignment and power allocation strategy that minimizes the network energy consumption in H-CRAN. On formulating the problem, we consider multiple practical constraints such as users’ minimum data rates, the maximum transmission power of RRHs and eRRHs, and the maximum fronthaul capacity. The NEE optimization problem turns out to be a mixed integer non-linear programming (MINLP) problem which is generally NP-hard [13].

We propose the H-CRAN energy-efficient radio resource management (HERM) algorithm to solve the NEE problem efficiently. The HERM algorithm first transforms the objective function of the NEE optimization problem from a divisional-form to a subtractive-form by using the Dinkelbach algorithm [14]. Then, the ℓ₀-norm approximation method is adopted to approximate the non-convex ℓ₀-norm as ℓ₁-norm [15]. Next, the HERM algorithm optimizes the RB assignments and corresponding power allocation based on the Lagrangian-dual method [16], [17]. The optimality and convergence of the HERM algorithm is analyzed and proved. We also discuss the practicality of the HERM algorithm in terms of implementing the algorithm in a practical H-CRAN. The performance of the HERM algorithm is evaluated with extensive simulations, which show that the HERM algorithm significantly improves the NEE of H-CRAN. As compared with a baseline algorithm in which the radio resource management is not optimized, the HERM algorithm boosts the NEE by 59%. As compared to an energy-efficient radio resource allocation (ERA) algorithm which does not optimize the energy consumption of the BBU pool, the NEE of H-CRAN achieved by the HERM algorithm is up to 51% better than that by the ERA algorithm. Moreover, the simulation results also reveal several insights on optimizing energy efficiency of cloud-based radio access networks.

The main contributions of this paper can be summarized as follows.

- We develop a comprehensive network energy consumption model that characterizes the energy consumption of RRHs, eRRHs, fronthaul, and the BBU pool in H-CRAN. The network energy consumption model is essential for studying the energy efficiency of H-CRAN.
- We mathematically formulate and analyze the NEE maximization problem in H-CRAN with the consideration of multiple practical constraints such as users’ minimum data rates, the maximum transmission power of RRHs and eRRHs, and the maximum fronthaul capacity. The problem formulation and analysis lay a foundation for investigating the performance and tradeoff in optimizing the energy efficiency of H-CRAN.
- We design the HERM algorithm to solve the NEE maximization problem efficiently. The optimality and convergence of the HERM algorithm is analyzed and proved. The implementation of the HERM algorithm in practical H-CRAN is discussed.
- We validate the HERM algorithm via extensive network simulations. The simulation results are well analyzed and reveal several insights on optimizing energy efficiency of cloud-based radio access networks.

The remainder of this paper is organized as follows. In Section II, we briefly review the related works. In Section III, we describe the system model of H-CRAN and formulate the NEE optimization problem. In Section IV, we design the HERM algorithm to solve the problem. In Section V, the practical implementation of the HERM algorithm in H-CRAN is discussed. In Section VI, we evaluate the algorithm via extensive simulations, and analyze the results. Finally, we conclude our works in Section VII.

II. RELATED WORK

Energy-efficient mobile networks have been extensively studied over recent years [18], [19]. In this section, we briefly overview on energy-efficient networking solutions in conventional mobile networks, e.g., LTE, and the cloud based radio access network.

A. Energy Efficiency in Conventional Mobile Networks

In a mobile network, the radio access network, which is mainly composed of BSs, is the major energy consumer [20]. Therefore, most of energy-efficient networking solutions for conventional mobile networks focus on reducing the energy consumption of BSs. These solutions can be classified into three categories [18], [21], [22]. The first one includes radio source management algorithms that optimize the energy efficiency of mobile networks [23]–[25]. The second category
designs the BS sleep mode algorithms, which dynamically switch BSs to the low-power mode according to their traffic loads [26]–[28]. The third category aims to power mobile BSs with renewable energy [29]–[32]. Although these solutions provide some insights in designing energy-efficient mobile networks, they are not directly applicable to cloud-based radio access networks because of the architectural differences between conventional mobile networks and C-RAN.

B. Energy Efficiency in Cloud-based Radio Access Networks

As cloud-based radio access networks emerge as a promising networking architecture for next generation mobile networks, energy-efficient C-RAN becomes an important research topic. Many have addressed energy efficiency problems in C-RAN from different aspects. Liu et al. [33] proposed a network energy consumption model and designed a joint user association and resource allocation algorithm to optimize the energy efficiency in H-CRAN. Peng et al. [34] investigated the enhanced soft fractional frequency reuse (S-FFR) scheme for energy-efficient H-CRAN. This scheme considers the RRH and fronthaul energy consumption. Tang et al. [35] studied the cross-layer resource allocation problem in C-RAN to minimize the network energy consumption. Pompili et al. [36] proposed an elastic resource utilization framework for C-RAN. This framework improves C-RAN energy efficiency by dynamically grouping virtual base stations into clusters according to traffic condition.

The limited fronthaul capacity has been considered in optimizing the energy efficiency of C-RAN. In order to minimize the transmission power of C-RAN, Vu et al. [37] proposed the jointly coordinated beamforming and admission control scheme in downlink C-RAN with fronthaul capacity constraints. Dhifallah et al. [38] studied energy efficient beamforming and backhaul power control schemes in C-RAN with heterogeneous backhaul. Oliva et al. [39] proposed a novel architecture named Xhaul which integrates the fronthaul and backhaul and provides an adaptive, shareable and cost-efficient transport network solution.

These existing works usually focus on the energy efficiency of certain networking devices, e.g., RRH and fronthaul. Our paper studies the network energy efficiency in H-CRAN based on a comprehensive network energy consumption model that characterizes the energy consumption of all networking devices in the downlink H-CRAN. Moreover, in our paper, multiple practical networking constraints have been considered in designing the energy efficient networking solutions.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first describe the system model including the network architecture, wireless communication model, fronthaul model, and network energy consumption model. Then, we mathematically formulate the network energy efficiency (NEE) maximization problem and analyze its properties.

A. System Model

1) Network Architecture: We consider a H-CRAN that consists of multiple RRHs and eRRHs. The RRHs only have the RF signal processing functions while the eRRHs are able to process both RF and baseband signals. The baseband signal processing functions of RRHs are carried by the BBU pool implemented on a cloud computing platform. The RRHs and eRRHs are connected to the BBU pool via heterogeneous fronthaul, e.g., optical fiber cables and wireless communication links, which facilitates the joint resource management. Denote $\mathcal{M}_s$, $\mathcal{M}_c$, and $\mathcal{K}$ as the sets of RRHs, eRRHs, and UEs, respectively. Then, $\mathcal{M} = \mathcal{M}_s \cup \mathcal{M}_c$ is the set of radio access points (RAPs) including both RRHs and eRRHs in the H-CRAN. The number of RAPs and UEs are $M = |\mathcal{M}|$ and $K = |\mathcal{K}|$, respectively.

| NOTATION |
|-----------------|-----------------|-----------------|-----------------|
| $\alpha_{k,m,n}$ | The RB assignment indicator |
| $A$             | The RB assignment matrix |
| $P$             | The power allocation matrix |
| $\mathcal{K}$   | The set of UEs |
| $\mathcal{M}$   | The set of RAPs |
| $\mathcal{N}$   | The set of RBs |
| $C_m$           | The effective fronthaul capacity of the $m$th RAP |
| $\eta_{NEE}$    | The network energy efficiency (NEE) |
| $g_{k,m,n}$     | The channel gain for $m$th RAP to $k$th UE on $n$th RB |
| $P_T$           | The network energy consumption |
| $P_B$           | The energy consumption of the BBU pool |
| $P_{m}^f$       | The energy consumption of the $m$th fronthaul |
| $P_{m}^r$       | The energy consumption of the $m$th RRH |
| $P_{st}$        | The static energy consumption of the entire network |
| $p_{k,m,n}$     | The allocated power for $m$th RAP to $k$th UE on $n$th RB |
| $R_T$           | The sum-rate of the network |
| $R_{k,m,n}$     | The maximum achievable data rate of $m$th RAP to $k$th UE on $n$th RB |

2) Wireless Communication Model: We adopt the orthogonal frequency division multiple access (OFDMA) in H-CRAN downlink transmission. We consider the large-scale channel fading such as distance-dependent path loss and shadowing, and assume the CSI can be obtained by the BBU pool for optimizing the radio resource allocation [3]. Denote $B_s$ and $B_e$ as the frequency bandwidth allocated to the RRH and eRRH, respectively. For the sake of simplicity of the presentation, we assume $B_s = B_e$ and use $B_e$ to represent the total bandwidth in either RRH or eRRH. The spectrum reuse factors for both RRHs and eRRHs are 1. That is, the total bandwidth in an individual RAP, e.g., RRH and eRRH, is $B_e$. The bandwidth in a RAP is divided into multiple resource blocks (RBs), and each RB is exclusively assigned to one UE in a resource allocation interval. Denote $\mathcal{N}$ as the set of RBs in a RAP. Then, the number of RBs in a RAP is $N = |\mathcal{N}|$.

Denote $p_{k,m,n}$ and $g_{k,m,n}$ as the transmission power and channel gain from the $m$th RAP to the $k$th UE on the $n$th RB, respectively. Then, the maximum achievable data rate [40] of the $k$th UE using the $n$th RB in the $m$th RRH be calculated as

$$R_{k,m,n} = \frac{B_e}{N} \log_2 \left( 1 + \frac{p_{k,m,n}g_{k,m,n}}{\sigma B_e / N} \right),$$

(1)
where $\sigma$ is the power spectrum density of the aggregated interference, including the co-channel interference among RAPs and the additive white Gaussian noise.

Denote $\alpha_{k,m,n}$ as the RB assignment indicator. If the $n$th RB of the $m$th RAP is assigned to the $k$th UE, $\alpha_{k,m,n} = 1$; otherwise, $\alpha_{k,m,n} = 0$. Then, we construct two $K \times M \times N$ variable matrices, $\mathbf{A} = [\alpha_{k,m,n}]_{K \times M \times N}$ and $\mathbf{P} = [p_{k,m,n}]_{K \times M \times N}$, to represent the RB assignment and power allocation, respectively. Note that user association is indicated in the expression of $\alpha_{k,m,n}$, i.e., $\sum_{n \in N} \alpha_{k,m,n} > 0$ indicates that the $k$th UE is associated with the $m$th RAP. Therefore, the sum-rate of the network $R_T$ can be expressed as

$$R_T (\mathbf{A}, \mathbf{P}) = \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \alpha_{k,m,n} R_{k,m,n}. \quad (2)$$

3) Fronthaul Model: The digitized in-phase and quadrature samples of the baseband signals are transmitted from RRHs and eRRHs to the BBU pool over the common public radio interface (CPRI) interface in the mobile networks [9], [41]. The transmission of the digitized in-phase and quadrature samples via fronthaul requires more than 10 times of the original wireless data bandwidth [42]. As a result, fronthaul may be the bottleneck of cloud-based radio access networks. The functional split can potentially alleviate the transmission bottleneck by shifting some baseband signal processing functions from the BBU pool to RRHs [42]-[44]. Therefore, eRRHs may demand less fronthaul capacity than RRHs because eRRHs have the baseband signal processing functions. Meanwhile, since heterogeneous fronthaul is considered in the H-CRAN, the capacity of fronthaul varies depending on the transmission technologies, e.g., optical fiber and wireless communications. Therefore, we model the fronthaul capacity constraint as

$$\sum_{k \in K} \sum_{n \in N} \alpha_{k,m,n} R_{k,m,n} \leq C_m, \forall m \in M, \quad (3)$$

where $C_m$ is defined as the effective fronthaul capacity of the $m$th RAP. Denote $\lambda$ as the ratio of the bandwidth required from transmitting the digitized in-phase and quadrature samples over the original wireless data bandwidth. Let $C_{m,\text{max}}$ be the fronthaul capacity of the $m$th RAP. Then, $C_m = C_{m,\text{max}}/\lambda$. The value of $C_m$ depends on the fronthaul transmission technologies. For example, optical fiber fronthaul has a larger effective capacity than wireless fronthaul.

4) Network Energy Consumption Model: Because of the architectural difference between the conventional mobile networks and H-CRAN, the energy consumption model designed for the conventional mobile networks is not appropriate for H-CRAN. Hence, we build the a comprehensive network energy consumption model that characterizes the energy consumption of RRHs, eRRHs, the BBU pool, and fronthaul in H-CRAN. The network energy consumption model is expressed as

$$P^T = \sum_{m \in M} (P^R_m + P^F_m) + P^B, \quad (4)$$

where $P^R_m$, $P^F_m$, and $P^B$ denote the energy consumption of the $m$th RAP, the fronthaul of the $m$th RAP, and the BBU pool, respectively.

The power consumption of a RAP is composed of the static and dynamic power consumption. The static power consumption is required to run the RAP while the dynamic power consumption is usually proportional to the traffic load [45]. Therefore, the energy consumption of the $m$th RAP can be written as

$$P^R_m = P^F_m + \Delta_m \sum_{k \in K} \sum_{n \in N} \alpha_{k,m,n} R_{k,m,n}, \quad (5)$$

where $P^F_m$ denotes the static energy consumption of the $m$th RAP and $\Delta_m$ is the energy factor of the $m$th RAP characterizing the relationship between the dynamic power consumption and the traffic load. The value of $\Delta_m$ is determined by the type of the RAP. With the baseband signal processing function, an eRRH consumes more dynamic power than a RRH. Therefore, an eRRH has a larger energy factor than a RRH.

On modelling the energy consumption of fronthaul, we consider the time-division multiplexing (TDM) fronthaul which can be switched to the sleep mode for energy savings [46], [47]. The power consumption of fronthaul is dominated by its static power consumption [10]. Therefore, we do not model the dynamic power consumption which depends on the traffic load carried by fronthaul. Then, the fronthaul power consumption of the $m$th RAP can be expressed as

$$P^F_m = \{ P^{F,A}_m, I_m \neq 0; P^{F,S}_m, I_m = 0 \}, \quad (6)$$

where $P^{F,A}_m$ and $P^{F,S}_m$ denote the fronthaul power consumption of the $m$th RAP in the active and sleep mode, respectively.

The power consumption of the BBU pool is closely related to the computing workloads for the baseband signal processing [10]. Since the BBU pool is implemented on cloud computing platforms, we assume that, given a RAP in H-CRAN, there is a corresponding virtual machine (VM) initialized to process baseband signals for the RAP [36], [48]. Hence, the energy consumption of the BBU pool is $P^B = \sum_{m \in M} P^{B}_m$, in which $P^{B}_m$ is the power consumption of the VM that processes baseband signals for the $m$th RAP. The power consumption of a VM in the BBU pool depends on the traffic load in its corresponding RAP. Therefore, the power consumption of the VM that serves the $m$th RAP is modeled as

$$P^{B}_m = P^{B,S}_m + \rho_m \sum_{k \in K} \sum_{n \in N} \alpha_{k,m,n}, \quad (7)$$

where $P^{B,S}_m$ is the static power of the $m$th RAP’s corresponding VM. $\rho_m$ is an energy factor characterizes the relationship between the VM power consumption and the radio resource utilization [10]. The value of $\rho_m$ is determined by the computing architecture and hardware equipment of the BBU pool [10].
Based on the above analysis, the network energy consumption model of H-CRAN expressed in Eq. (4) can be rewritten as

\[
P^T(A, P) = \sum_{m \in M} (P_{m}^{R,S} + P_{m}^{F,S} + P_{m}^{B,S}) + \sum_{m \in M} \rho_{m} \sum_{k \in K} \sum_{n \in N} \alpha_{k,m,n} + \sum_{m \in M} (\Delta_{m} I_{m} + (P_{m}^{F,A} - P_{m}^{F,S})) ||I_{m}||_0.
\]

(9)

In the following analysis, we define \( P_{\text{static}} = \sum_{m \in M} (P_{m}^{R,S} + P_{m}^{F,S} + P_{m}^{B,S}) \) for the sake of simplicity.

B. Optimization Problem Formulation

Our objective is to obtain the optimal RB assignment \( A \) and power allocation \( P \) that maximize the NEE of H-CRAN without violating the practical constraints such as the minimum data rate requirement, the fronthaul capacity, and the maximum transmission power. The NEE of H-CRAN is defined as

\[
\eta_{\text{NEE}}(A, P) = \frac{R^T(A, P)}{P^T(A, P)},
\]

(10)

Then, the NEE optimization problem can be mathematically formulated as

\[
\max_{(A,P)} \eta_{\text{NEE}}(A, P)
\]

s.t.

\[
C_1 : \sum_{m \in M} \sum_{n \in N} \alpha_{k,m,n} R_{k,m,n} \geq r_{k}^{\text{min}}, \forall k \in K \nonumber \]

\[
C_2 : \sum_{k \in K} \sum_{n \in N} \alpha_{k-m,n} p_{k,m,n} \leq p_{k}^{\text{max}}, \forall m \in M \nonumber \]

\[
C_3 : \sum_{k \in K} \sum_{n \in N} \alpha_{k-m,n} R_{k,m,n} \leq C_{m}, \forall m \in M \nonumber \]

\[
C_4 : \sum_{k \in K} \sum_{n \in N} \alpha_{k-m,n} \leq 1, \forall m \in M, n \in N \nonumber \]

\[
C_5 : \alpha_{k-m,n} \in \{0, 1\}, \forall k \in K, m \in M, n \in N.
\]

(11)

where \( r_{k}^{\text{min}} \) and \( p_{k}^{\text{max}} \) are the minimum data rate requirement of the \( k \)-th UE and the maximum transmission power of the \( m \)-th RAP, respectively. In the optimization problem, \( C_1 \) are the minimum data rate constraints of UEs; \( C_2 \) are the maximum transmission power constraints of RAPs; \( C_3 \) are the fronthaul capacity constraint. \( C_4 \) and \( C_5 \) regulate the RB assignment in H-CRAN. \( C_4 \) restricts that a RB can be assigned to at most one UE. \( C_5 \) imposes the binary RB assignment in H-CRAN.

In the NEE optimization problem, the fractional-form objective function is non-convex. The \( l_0 \)-norm in the network energy consumption model imposes additional difficulties on solving the problem. With the binary RB assignment, the NEE optimization problem is a mixed integer non-linear programming (MINLP) problem which is NP-hard and difficult to be solved [13], [23], [34]. In the following section, we will solve the problem and analyze the performance of the solution.

IV. THE HERM ALGORITHM

In this section, we present the H-CRAN energy-efficient radio resource management (HERM) algorithm to solve the NEE problem efficiently. The flow chart of the HERM algorithm is shown in Fig. 2. The HERM algorithm first transforms the objective function of the NEE optimization problem to a subtractive-form using the Dinkelbach algorithm [14]. Then, the \( l_0 \)-norm approximation method is adopted to approximate the non-convex \( l_0 \)-norm as \( l_1 \)-norm [15]. Next, the HERM algorithm optimizes the RB assignments and corresponding power allocation based on the Lagrangian-dual method [16], [17].

A. Problem Transformation

The objective function of the NEE optimization problem is in the fractional-form, which makes the problem highly non-convex. In order to solve this problem efficiently, we use the Dinkelbach algorithm to transform the objective function from the fractional-form into the subtractive-form [14]. We denote \((A^*, P^*)\) as the optimal solution to the following problem:

\[
\max_{(A,P)} \frac{R^T(A, P)}{P^T(A, P)} - q^* \cdot P^T(A, P)
\]

s.t. \( C_1 \sim C_5 \).

(12)

Lemma 1. When \( R^T(A^*, P^*) - q^* P^T(A^*, P^*) = 0, (A^*, P^*) \) is the optimal solution to the NEE optimization problem.

Proof: This lemma is proved in Appendix A.

According to Lemma 1, if the optimal \( q^* \) is obtained, the NEE optimization problem can be transformed to the problem in Eq. (12). That is, the solution to the problem in Eq. (12) solves the NEE optimization problem. Therefore, we apply the Dinkelbach algorithm to obtain the optimal \( q^* \) iteratively. At the beginning, the \( q \) is initialized. In each iteration, the problem in Eq. (12) with fixed \( q \) will be solved. By solving the problem in Eq. (12), \( q \) will be updated and converged to the optimal NEE \( q^* \).

B. \( l_0 \)-norm Approximation

After the above problem transformation, the NEE optimization problem is still non-convex because of the \( l_0 \)-norm in the network energy consumption model \( P^T(A, P) \). The \( l_0 \)-norm of a vector calculates the number of non-zero entries in the vector, and can be approximated as

\[
||X||_0 = \sum_{i=1}^{m} \beta_i |X_i|,
\]

(13)

where \( X_i \) is the \( m \)-th component of vector \( X \) and \( \beta_i \) is the weight of \( X_i \) [15]. Since \( I_m \geq 0 \), based on this \( l_0 \)-norm approximation, the optimal solution to the transformed NEE optimization problem is still non-convex.
approximation method, we remove the $l_0$-norm in the energy consumption model by letting
\[ ||I_m||_0 \approx \beta_m I_m. \] (14)
Here, $\beta_m$ is the weight of the fronthaul link of the $m$th RAP.

As a result of the approximation, the network energy consumption can be expressed as
\[ \bar{\beta} \]

\[ \bar{P}^T (A, P) = P_{\text{static}}^T + \sum_{A \in M} \frac{\rho_m}{\beta_m} \sum_{k \in K} \sum_{n \in N} \alpha_{k,m,n} \]
\[ + \sum_{m \in M} (\Delta_m + \beta_m (P_{FA}^m - P_{FS}^m))I_m. \] (15)

Hence, the NEE maximization problem is approximately transformed to
\[ \mathcal{P}_0 : \max_{\{A, P\}} R^T (A, P) - q^* \cdot \bar{P}^T (A, P) \]
\[ \text{s.t. } C_1 \sim C_5. \] (16)
$\beta_m$ in Eq. (15) is calculated as
\[ \beta_m = h(I_m, \tau) = \frac{\xi}{I_m + \tau}, \] (17)
where $I_m$ is from the previous iteration. $\tau$ and $\xi$ are constant regularization factors which regulate the stability and control precision of the $l_0$-norm approximation. The value of $\beta_m$ is iteratively updated based on Eq. (17) until it converges in the HERM algorithm.

Note that $I_m$ represents the traffic load in the $m$th fronthaul. In the above iterative updates of $\beta_m$, the fronthaul with lighter traffic loads (smaller $I_m$) will be assigned larger weights. This weight assignment will further reduce the traffic load in the fronthaul in the next iteration, and eventually force the fronthaul into the sleep mode.

### C. Optimal RB and Power Allocation

In this part, we design the RB and power allocation algorithm to solve Problem $\mathcal{P}_0$ by using the Lagrangian-dual method [16]. We first replace the binary variable, $\alpha_{k,m,n}$, in Problem $\mathcal{P}_0$ with a continuous variable $\bar{\alpha}_{k,m,n} \in [0, 1]$. Therefore, Problem $\mathcal{P}_0$ is relaxed to
\[ \mathcal{P}_1 : \max_{\{A, P\}} R^T (A, P) - q^* \cdot \bar{P}^T (A, P) \]
\[ \text{s.t. } C_1 \sim C_4, \]
\[ \bar{\alpha}_{k,m,n} \in [0, 1], \forall k \in K, m \in M, n \in N. \] (18)

Then, we prove that the solution to the relaxed problem obtained through the RB and power allocation algorithm solves Problem $\mathcal{P}_0$.

When the number of RBs is large, Problem $\mathcal{P}_1$ satisfies the time-sharing conditions. As a result, the duality gap between the primal and dual problems approaches zero [49]. Hence, the RB and power allocation algorithm is designed based on the Lagrangian-dual method [16].

The Lagrangian of Problem $\mathcal{P}_1$ can be expressed as
\[ \mathcal{L} (A, P, \mu, \gamma, v) = \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \bar{\alpha}_{k,m,n} R_{k,m,n} \]
\[ -q \left [ P_{\text{static}}^T + \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \bar{\alpha}_{k,m,n} (\Delta_{k,m,n}^E p_{k,m,n} + \rho_m) \right ] \]
\[ + \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \bar{\alpha}_{k,m,n} R_{k,m,n} - \gamma_m \min \]
\[ - \sum_{m \in M} \gamma_m \sum_{k \in K} \sum_{n \in N} \bar{\alpha}_{k,m,n} p_{k,m,n} - \rho_m \max \]
\[ - \sum_{m \in M} \sum_{k \in K} \sum_{n \in N} \bar{\alpha}_{k,m,n} R_{k,m,n} - C_m, \] (19)
where $\mu = \{\mu_k | k \in K\}$, $\gamma = \{\gamma_m | m \in M\}$, and $v = \{v_m | m \in M\}$ are the Lagrange multipliers corresponding to the minimum data rates of UEs $\rho_m \min$, the maximum transmission power of RAPs $P_{\max}$, and the fronthaul capacity $C_m$, respectively. We define $\Delta_{k,m,n}^E = \Delta_k^E + \beta_m (P_{FA}^m - P_{FS}^m)$ to simplify the mathematical presentation.

The Lagrangian dual function is
\[ g(\mu, \gamma, v) = \max_{\{A, P\}} \mathcal{L} (A, P, \mu, \gamma, v) \] (20)
and the Lagrangian dual problem can be expressed as
\[ \min_{\{\mu, \gamma, v\}} g(\mu, \gamma, v) \]
\[ \text{s.t. } \mu \geq 0, \gamma \geq 0, v \geq 0. \] (21)

Based on the Lagrangian-dual method, we decompose the dual problem into $N$ independent sub-problems with the same structure expressed as
\[ g_n(\mu, \gamma, v) = \sum_{m \in M} g_n(\mu, \gamma, v) - q^* P_{\text{static}}^T - \sum_{k \in K} \mu_k \rho_m \min \]
\[ + \sum_{m \in M} \gamma_m P_{\max} + \sum_{m \in M} v_m C_m, \] (22)
where
\[ g_n(\mu, \gamma, v) = \max_{\{A, P\}} \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \bar{\alpha}_{k,m,n} \left [ (1 + \mu_k - v_m) R_{k,m,n} \right ] \]
\[ - (\gamma_m + q^* \Delta_{k,m,n}^E) p_{k,m,n} - \rho_m \]. (23)

Let
\[ f(A, P) = \sum_{k \in K} \sum_{m \in M} \sum_{n \in N} \bar{\alpha}_{k,m,n} \left [ (1 + \mu_k - v_m) R_{k,m,n} \right ] \]
\[ - (\gamma_m + q^* \Delta_{k,m,n}^E) p_{k,m,n} - \rho_m \]. (24)

Then, we can derive
\[ \frac{\partial f(A, P)}{\partial P} = \bar{\alpha}_{k,m,n} \left [ (1 + \mu_k - v_m) \frac{\partial R_{k,m,n}}{\partial p_{k,m,n}} - \gamma_m + q^* \Delta_{k,m,n}^E \right ]. \] (25)

When $\bar{\alpha}_{k,m,n} = 1$, the optimal power allocation can be calculated by setting $\frac{\partial f(A, P)}{\partial P} = 0$ according to the Karush-Kuhn-Tucker (KKT) condition [16]. So the optimal power allocation is
\[ p_{k,m,n}^* = \left [ \omega_{k,m,n} - \frac{\sigma B_e}{N g_{k,m,n}^*} \right ]^+, \] (26)
where $[x]^+$ is defined as $\max\{x, 0\}$. The optimal power allocation $p_{k,m,n}^*$ is in the form of the multi-level water filling.
The water filling level is expressed as
\[ \omega_{k,m,n} = \frac{B_e}{\ln 2} \cdot \frac{1 + \mu_k - u_m}{\gamma_m + q^* \Delta E_m^m}. \] (27)

The water filling level \( \omega_{k,m,n} \) depends on the Lagrange multipliers and weighted factors, \( q^* \) and \( \beta_m \). For example, \( u_m \) controls the power allocation in the \( m \)th RAP so that the traffic load in fronthaul does not exceed its capacity. That is, \( u_m \) may constrain the maximum transmission power of the \( m \)th RAP.

After deriving \( p^*_k,m,n \), Eq. (24) can be expressed as
\[ f(A, P^*) = \sum_{k \in K} \sum_{m \in M} \tilde{\alpha}_{k,m,n} H_{k,m,n}, \] (28)

where
\[ H_{k,m,n} = (1 + \mu_k - u_m) R^*_k,m,n - (\gamma_m + q^* \Delta E_m^m) p^*_k,m,n - q^* \rho_m. \] (29)

Given \( m \) and \( n \), we define \( K = \{ k | H_{k,m,n} > 0, k \in K \} \). Based on Eqs. (26) and (24), the optimal RB assignment can be determined as
\[ \tilde{\alpha}^*_{k,m,n} = \begin{cases} 1, & k = \arg \max_{k \in K} H_{k,m,n} \\ 0, & \text{otherwise}. \end{cases} \] (30)

After we obtain the optimal power allocation and RB assignment, we use the sub-gradient method [16] to solve the Lagrange dual problem shown in Eq. (21). The sub-gradient of dual function in the \( l \)th iteration is
\[ \nabla \mu_k(i) = \sum_{m \in M} \sum_{n \in N} \tilde{\alpha}_{k,m,n} R_{k,m,n} - r_{k,n}^{\min}, \forall k \in K, \]
\[ \nabla \gamma_m(i) = P_{\max} - \sum_{k \in K} \sum_{n \in N} \tilde{\alpha}_{k,m,n} p_{k,m,n}, \forall m \in M, \]
\[ \nabla u_m(i) = C_m - \sum_{k \in K} \sum_{n \in N} \tilde{\alpha}_{k,m,n} R_{k,m,n}, \forall m \in M. \] (31)

The dual variables in the \( l \)th iteration are updated as
\[ \mu_k(i) = [\mu_k(i - l) - \delta_\mu(i) \times \nabla \mu_k(i)]^+, \forall k \in K, \]
\[ \gamma_m(i) = [\gamma_m(i - l) - \delta_\gamma(i) \times \nabla \gamma_m(i)]^+, \forall m \in M, \]
\[ u_m(i) = [u_m(i - l) - \delta_v(i) \times \nabla u_m(i)]^+, \forall m \in M. \] (32)

where \( \delta_\mu(i), \delta_\gamma(i), \) and \( \delta_v(i) \) are the positive step sizes. If the step sizes satisfy the infinite travel condition, it is guaranteed that dual variables will converge to the optimal solution [16].

The pseudo code of the RB and power allocation algorithm is shown in Alg. 1. At the beginning of the algorithm, we initialize Lagrange multipliers and the iteration step sizes. Then, we calculate the power allocation and RB assignment based on Eqs. (26) and (30). Next, Lagrange multipliers are updated according to Eqs. (31) and (32). When Lagrange multipliers converge, the optimal power allocation and RB assignment are derived.

**Lemma 2.** \( \{A^*, P^*\} \) maximizes \( \eta_{NEE}(A, P) \) of Problem \( \mathcal{P}_0 \).

**Proof:** The proof is presented in Appendix B. ■

The proof of the Lemma 2 means that, using the RB and power allocation algorithm, we can obtain the optimal solution to Problem \( \mathcal{P}_0 \) by solving the relaxed problem in which the binary RB assignment is replaced by the fractional RB assignment.

**D. The Pseudo Code of the HERM Algorithm**

The HERM algorithm integrates the problem transformation, \( l_0 \)-norm approximation, and the RB and power allocation algorithm to solve the NEE optimization problem. The pseudo code of the HERM algorithm is shown in Alg. 2. \( q^* \) and \( \beta_m \) are introduced in the problem transformation and \( l_0 \)-norm approximation, respectively. According to the Dinkelbach algorithm [14] and the \( l_0 \)-norm approximation method [15], the values of \( q^* \) and \( \beta_m \) should be derived via iterative processes. Therefore, there are two loops in the HERM algorithm. The inner loop aims to calculate the value of \( \beta_m \) while the outer loop is to obtain the value of \( q^* \). Given the values of \( q^* \) and \( \beta_m \), the RB assignment and power allocation are optimized by using Alg. 1.

**Lemma 3.** The HERM algorithm converges to the optimal solution.

**Proof:** The proof is presented in Appendix C. ■

**E. Computational Complexity Analysis**

The computational complexity of the HERM algorithm is calculated as \( N_q \cdot N_B \cdot N_{\mu,\gamma,v} \cdot O(KMN) \). \( N_q \) and \( N_B \) are the numbers of iterations required for obtaining \( q^* \) and \( \beta \), respectively. Based on the Dinkelbach algorithm [14] and the \( l_0 \)-norm approximation method [15], \( N_q \) and \( N_B \) should be fairly small numbers. \( N_{\mu,\gamma,v} \) is the number of iterations required for converging Lagrange multipliers \( \mu, \gamma, v \) in Alg. 1. The value of \( N_{\mu,\gamma,v} \) depends on multiple factors such as the step size and the distance between initial and optimal values. We analyze the convergence performance in terms of the value of \( N_{\mu,\gamma,v} \) in Appendix D. \( O(KMN) \) is the number of operations required in calculating the RB assignment and power allocation. The convergence of the HERM algorithm will be evaluated through extensive simulations presented in Section VI.

**V. THE PRACTICALITY OF THE HERM ALGORITHM**

In this section, we discuss the implementation of the HERM algorithm, and the timescales of the radio resource management and the fronthaul sleep mode.

**A. Implementation and Communication Protocol**

Fig. 3 illustrates the implementation of the HERM algorithm in H-CRAN. The HERM algorithm is implemented on the cloud computing platform and collects CSI from BBUs.
Meanwhile, the HERM algorithm also learns UEs’ data rate requirements and fronthaul capacity. With this information, the HERM algorithm optimizes the RB assignment and power allocation. Based on the RB assignment, H-CRAN determines the UE-RAP association and turns the fronthaul without any allocation. Based on the RB assignment, H-CRAN determines the HERM algorithm optimizes the RB assignment and power requirements and fronthaul capacity. With this information, the HERM algorithm also learns UEs’ data rate requirements and fronthaul capacity. Based on the RB assignment, H-CRAN determines the HERM algorithm optimizes the RB assignment and power requirements and fronthaul capacity. With this information, the HERM algorithm also learns UEs’ data rate requirements and fronthaul capacity.

With the HERM algorithm, the communication procedures can be briefly described in the following steps:

1) UEs attach to the network via a RAP and send their CSI and data rate requirements to the BBU pool.
2) The HERM algorithm learns the fronthaul capacity, collects the UE information including CSI and data rate requirements from the BBU pool, and optimizes the RB assignment and power allocation.
3) Based on the RB assignment, H-CRAN informs UEs about the updated UE-RAP association.
4) UEs switch their carrier frequencies based on the updated UE-RAP association. In H-CRAN, a UE may receive signals on two different carrier frequencies because the RRH and eRRH occupy different frequency bands. If the updated UE-RAP association requires a UE to handover from a RRH (eRRH) to an eRRH (RRH), the UE should switch its carrier frequency accordingly. After updating their carrier frequencies, UEs are ready to receive downlink data.
5) The HERM algorithm, based on the RB assignment, informs the fronthaul controller to turn the fronthaul without any traffic load into the sleep mode.

### Algorithm 2: The HERM Algorithm

1. Initialize the convergence conditions $\varepsilon_1$ and $\varepsilon_2$;
2. Initialize $q^*$ and $\beta$;
3. Set iteration indexes $i = 1$;
4. while ($q^*$ does not converge) do
   5. Set $j = 1$;
   6. while ($\beta$ does not converge) do
      7. Solve Problem $\mathcal{P}0$ to obtain $A^{(j)}, P^{(j)}$ using Alg. 1;
      8. if $\left| \beta_m^{(j)} - \beta_m^{(i-1)} \right| / \beta_m^{(i-1)} \leq \varepsilon_2, \forall m \in \mathcal{M}$ then
         9. Return: $\beta^* = \beta^{(j)}$ and $\{A^{(j)}, P^{(j)}\}$;
      else
         10. Update $\beta$ based on Eq. (17);
         11. Set $j = j + 1$;
   12. if $(R^T (A^{(i)}, P^{(i)}) - q^* P^T A^{(i)}, P^{(i)}) \leq \varepsilon_2$ then
      13. Return: $\{A^*, P^*\} = \{A^{(i)}, P^{(i)}\}$;
      else
         14. Update $q^* = R^T (A^{(i)}, P^{(i)}) / P^T A^{(i)}, P^{(i)}$;
         15. Set $i = i + 1$;
   16. Return: $\{A^*, P^*\}$ and $I_m$, $\forall m \in \mathcal{M}$;

6) The BBU pool processes the downlink data transmission according to the RB assignment and power allocation.

### B. Timescale Discussion

In the HERM algorithm, we dynamically optimize the radio resource management and the fronthaul sleep mode. One major concern may be the different operation timescales of the radio resource management and the fronthaul sleep mode. In current mobile networks, e.g., LTE, the transmission time interval (TTI) is 1 ms [50]. Therefore, the RB assignment and power allocation can be managed on a timescale of 1 ms. However, managing radio resource on such a small timescale may lead to excessive computing and communication overheads. Moreover, the channel coherence time may be much larger than one TTI. Therefore, the TTI building technique, which enables the combination and scheduling of multiple consecutive subframes, is supported in the LTE system [50]. As a result, the practical operation timescale of radio resource management can be several milliseconds.

Regarding the fronthaul, the passive optical network (PON) emerges as a key fronthaul solution for future cloud-based radio access network [51]. In PON, the optical line terminal (OLT) is responsible for receiving the upstream data and broadcasting downstream data to optical network units (ONUs). The ONU can be switched into the sleep for energy savings when it does not carry any traffic load [52]. The sleep mode in ONU operates at a timescale of several milliseconds [52]. Therefore, it is possible to operate the dynamic radio resource management and the fronthaul (ONU) sleep mode using the same timescale.

### VI. Simulation Results and Analysis

In this section, we set up simulations to evaluate the performance of the HERM algorithm. In the simulations, we consider one eRRH and multiple RRHs. The system bandwidth $B_T$ is 20 MHz. RRHs and UEs are randomly distributed in the network. Denote $d$ as the distance between a RAP (eRRH or RRH) and a UE in kilometers. The path loss between an eRRH and a UE is modeled as $128.1 + 37.6 \log_{10}(d)$. The path loss between a RRH and a UE is modeled as $140.7 + 36.7 \log_{10}(d)$. The standard deviation of the shadow fading of eRRHs and RRHs are 8 dB and 10 dB, respectively [53]. Since eRRHs support baseband signal processing, we assume that the fronthaul capacity demand of eRRHs is 10 times less than that of RRHs according to the typical baseband signal sampling ratio of LTE networks [4].

In the simulations, the maximum transmission power $P_m^{\text{max}}$ of eRRHs and RRHs are 43 dBm and 29 dBm, respectively [53]. The static power consumption $P_m^{R,S}$ of eRRHs

![Fig. 3. The practical implementation of the HERM algorithm](image-url)
The energy consumption of the BBU pool is closely related with the increment of the number of UEs. This indicates that the energy consumption of the cloud platform, which carries the baseband signal processing, a large number of CRAN with different number of UEs. Since the functionality of the RRH is simplified to carry only the RF signal processing, the energy consumption of the RRH is much less than that of the cloud platform and fronthaul. As the energy consumption of the BBU pool dominates the energy consumption of H-CRAN, optimizing the cloud resource management is essential in achieving energy efficient H-CRAN.

### B. Impact of the fronthaul capacity

Fig. 6 shows the NEE of different algorithms versus the fronthaul capacity. In the simulation, we assume that there are high-capacity fronthaul, e.g., optical fiber links, and low-capacity fronthaul, e.g., wireless links. In order to quantify the impact of the fronthaul capacity on the NEE, we set the capacity of the high-capacity fronthaul to be 4 times larger than that of low-capacity fronthaul. In Fig. 6, x-axis is the capacity of high-capacity fronthaul and y-axis is the network energy efficiency. The figure shows that the NEE improves as the fronthaul capacity increases. When the fronthaul capacity is sufficiently large, e.g., 0.5 Gbps in the simulation, the NEE reaches its maximum value. This is because a small fronthaul capacity constrains the radio resource management from achieving the optimal NEE. As the fronthaul capacity increases, the constraint is slacking, thus leading to a better NEE. The fronthaul capacity increase will eventually eliminate the fronthaul capacity from the constraints in optimizing the NEE. Therefore, after reaching a certain value, the fronthaul capacity does not impact the NEE.

Fig. 6 also shows the performance comparison between the HERM and ERA algorithms. Since the HERM algorithm optimizes the energy consumption in the BBU pool, it improves the NEE by 52% as compared with the ERA algorithm. Both the HERM and ERA algorithms achieve better NEE when there are more UEs in the network. However, the performance gap between the HERM and ERA algorithms shrinks. This is because an increasing number of UEs may gradually saturate the network, which leaves a smaller room for improving the NEE.

### C. Impact of the maximum transmission power

Fig. 7 evaluates the NEE versus the maximum transmission power, $P_{m}^{\max}$, of RRHs with different algorithms. In this simulation, the maximum transmission power of the eRRH is 20 Watts. As shown in the figure, the NEE with the HERM and ERA algorithms improves with the increase of the maximum transmission power in the RRH. However, when $P_{m}^{\max}$ reaches a certain value, e.g., 9 Watts in the simulation, the NEE does not improve any more. When $P_{m}^{\max}$ is small, the data rates of UEs are constrained by $P_{m}^{\max}$. Therefore, increasing $P_{m}^{\max}$ enhances the throughput of the network and thus improves the NEE. When $P_{m}^{\max}$ is large, additional energy consumption caused by allocating more transmission power to UEs is not compensated by the throughput improvement in terms of the NEE. Therefore, the HERM and ERA algorithms do not allocate additional power to UEs. Hence, the NEE keeps constant after $P_{m}^{\max}$ reaches a certain value. As compared with the ERA algorithm, the NEE of the HERM algorithm improves 15% and 7% under the small and large $P_{m}^{\max}$, respectively.

### Table I

<table>
<thead>
<tr>
<th>Variable</th>
<th>Default value</th>
<th>Variable</th>
<th>Default value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>20</td>
<td>$M$</td>
<td>10</td>
</tr>
<tr>
<td>$N$</td>
<td>20</td>
<td>$\sigma$</td>
<td>-154 dBm</td>
</tr>
<tr>
<td>$P_{m,S}^{F,A}$</td>
<td>37 Watt</td>
<td>$P_{m,S}^{F,S}$</td>
<td>22.2 Watt</td>
</tr>
<tr>
<td>$P_{m,S}^{F,S}$</td>
<td>12 Watt</td>
<td>$P_{m,S}^{F,S}$</td>
<td>5 Mbp</td>
</tr>
</tbody>
</table>

and RRHs are 84 Watts and 3.5 Watts, respectively [45]. The static power consumption of the eRRH is larger than that of the RRH because the eRRH supports the baseband signal processing functions. The parameters $\rho_{m}$ and $\Delta_{m}$ are 5.676 and 4.75, respectively [10]. Table I summarizes the default values of important parameters in the simulations.

In the simulations, we compare the proposed HERM algorithm with the following algorithms:

- **Baseline Algorithm [23]**: In the baseline algorithm, UEs are associated with the nearest RAP. The RBs are evenly assigned to all UEs with the same transmission power. The fronthaul link is switched to the sleep mode if it does not carry any traffic load.

- **Energy-efficient Radio resource Allocation (ERA) Algorithm [34]**: The ERA algorithm does not optimize the energy consumption of the BBU pool. Hence, the energy consumption of the BBU pool for the $m$th RAP is $P_{m,S}^{F,S} + \rho_{m}N$ as shown in Eq. 8.

### A. Impact of the number of UEs

Figs. 4 and 5 show the NEE under different algorithms versus the number of UEs. It can be seen from Fig. 4 that the network achieves a higher energy efficiency when there are a larger number of UEs. When the number of UEs is small, the traffic load in the network is light, and the static power consumption dominates the overall power consumption of the network. Therefore, the network has a low energy efficiency. When the number of users increase, the network is saturated with the traffic loads. Thus, the network efficiency reaches its maximum value. However, as shown in the figure, the HERM algorithm improves the NEE by 51% as compared to the ERA algorithm when the traffic load is light in the network. The energy savings of the HERM algorithm come from the optimization of the BBU pool. That is, the HERM algorithm is able to dynamically switch the VMs in the BBU pool into the sleep mode when the traffic load is light in the network. As compared to the baseline algorithm, the HERM algorithm achieves about 59% improvement in the NEE when the network traffic load is heavy. The energy efficiency improvement is mainly a result of the optimization of the RRH transmission and fronthaul operation.

Fig. 5 shows a breakdown of the energy consumption in H-CRAN with different number of UEs. Since the BBU pool carries the baseband signal processing, a large number of UEs lead to heavy computing workloads in the BBU pool. Therefore, the energy consumption of the cloud platform, where the BBU pool is implemented, increases significantly with the increment of the number of UEs. This indicates that the energy consumption of the BBU pool is closely related to the traffic load in the network. Since the functionality of the RRH is simplified to carry only the RF signal processing, the energy consumption of the RRH is much less than that of the cloud platform and fronthaul. As the energy consumption of the BBU pool dominates the energy consumption of H-CRAN, optimizing the cloud resource management is essential in achieving energy efficient H-CRAN.
The performance gap between the HERM and ERA algorithms shrinks with the increase of $P_m^{max}$. This because a larger $P_m^{max}$ leads to a higher throughput which generates more workloads of baseband signal processing in the BBU pool. The heavy-loaded BBU pool leaves a small room to the HERM algorithm for further improving the energy efficiency.

Both the HERM and ERA algorithms show significantly better performance than the baseline algorithm. Besides, the NEE with the baseline algorithm decreases after $P_m^{max}$ reaches a certain value. This is because the baseline algorithm allocates as much power as available without any optimization on the energy efficiency.

D. Impact of energy factors

In the network energy consumption model, there are two energy factors, $\Delta_m$ and $\rho_m$. $\Delta_m$ characterizes the relationship between the dynamic power consumption and the traffic load in the $m$th RRH; $\rho_m$ defines the relationship between the VM energy consumption and the radio resource allocation. In this simulation, we evaluate how these energy factors impact the NEE. To quantify the impact, we assume that $\Delta_m = \Delta_n$ and $\rho_m = \rho_n, \forall m, n \in M$.

Fig. 8 shows the NEE under different algorithms versus the energy factor $\Delta_m$. It can be observed that the NEE decreases with the increase of $\Delta_m$. This because a larger $\Delta_m$ results in more energy consumption in transmitting the same amount of data traffic in a RAP. When $\Delta_m$ is small, e.g., $\Delta_m = 1$, the NEE with the HERM algorithm is 21% better than that with the ERA algorithm. When $\Delta_m$ increases to 80, the HERM algorithm is only 6% better than the ERA algorithm in terms of the NEE. That is, the advantage of the HERM algorithm is gradually neutralized by an increasing $\Delta_m$. The reason is that a large $\Delta_m$ makes RAPs the dominant energy consumer in the network. As a result, the optimization in the BBU pool provided by the HERM algorithm does not enable a significant performance improvement. Fig. 8 also shows that both the HERM and ERA algorithms can achieve much higher energy efficiency than the baseline algorithm.

Fig. 9 shows the impact of $\rho_m$ on the NEE with different algorithms. As shown in the figure, a larger $\rho_m$ leads to a lower energy efficiency because more energy is consumed for processing the same amount of baseband signals in the BBU pool. The advantage of the HERM algorithm in terms of optimizing the NEE becomes more significant with a larger $\rho_m$. For example, when $\rho_m = 9$ and $\rho_m = 27$, the NEE achieved by the HERM algorithm are 37% and 55% better than that with the ERA algorithm, respectively. This indicates that optimizing the BBU pool is more important in a less energy-efficient cloud platform.

Fig. 10 shows the NEE achieved by the HERM algorithm versus $\rho_m$ under different numbers of UEs. As shown in the figure, when $\rho_m$ is large, e.g., $\rho_m = 24$, the NEE with different numbers of UEs tends to be the same. When $\rho_m$ is small, the static power consumption accounts for the largest parts of the

Fig. 4. The NEE versus the number of UEs.
Fig. 5. Network energy consumption breakdown.
Fig. 6. The NEE versus $C_m$.
Fig. 7. The NEE versus $P_m^{max}$.
Fig. 8. The NEE versus $\Delta_m$.
Fig. 9. The NEE versus $\rho_m$.
Fig. 10. The NEE versus $\rho_m$ with different numbers of UEs.
overall energy consumption in the network. As a result, when the number of UEs increases, the network may serve more traffic loads at the cost of slightly higher traffic-dependent energy consumption (dynamic energy consumption). When \( \rho_m \) is large, the profit (higher data rate) achieved by serving more UEs is neutralize by the energy consumption increment. Therefore, the network tends to have the same NEE under different number of UEs when \( \rho_m \) is large.

VII. CONCLUSION

In this paper, we have studied the impact of the radio resource management on the energy efficiency of H-CRAN. We have developed a network energy consumption model that captures the energy consumption of base stations (BSs), fronthaul, and the BBU pool in H-CRAN. Based on the network energy computing model, we have formulated the network energy efficiency optimization problem and designed H-CRAN energy-efficient radio resource management (HERM) algorithm to solve the problem. We have proved the convergence and optimality of the HERM algorithm and validated its performance through extensive simulations. The simulation results have revealed insights on optimizing the energy efficiency of H-CRAN under different network conditions.

APPENDIX A

PROOF OF LEMMA 1

Denote \((A, P)\) as any feasible solution to the problem in Eq. (12). Since \((A^*, P^*)\) is the optimal solution to the problem,

\[
R^T(A^*, P^*) - q^* P^T(A^*, P^*) \geq R^T(A, P) - q^* P^T(A, P). \tag{33}
\]

Because \(R^T(A^*, P^*) - q^* P^T(A^*, P^*) = 0\), \(R^T(A, P) - q^* P^T(A, P) \leq 0\), \(P^T(A, P)\) is the H-CRAN power consumption which is larger than zero. Therefore,

\[
\frac{R^T(A, P)}{P^T(A, P)} \leq q^*. \tag{34}
\]

Hence,

\[
\frac{R^T(A, P)}{P^T(A, P)} \leq \frac{R^T(A^*, P^*)}{P^T(A^*, P^*)}. \tag{35}
\]

Thus, \((A^*, P^*)\) maximizes \(\frac{R^T(A, P)}{P^T(A, P)}\) while satisfying all the constraints in the NEE optimization problem. That is, \((A^*, P^*)\) is the optimal solution to the NEE optimization problem. Then, the lemma is proved.

APPENDIX B

PROOF OF LEMMA 2

In order to prove Lemma 2, we first prove \(\{A^*, P^*\}\) is the optimal solution to Problem \(\mathcal{P}1\) and then prove that \(\{A^*, P^*\}\) is also the optimal solution to Problem \(\mathcal{P}0\).

A. The optimal solution to Problem \(\mathcal{P}1\):

To prove \(\{A^*, P^*\}\) is the optimal solution to Problem \(\mathcal{P}1\), we first show that \(\{A^*, P^*\}\) derived based on the HERM algorithm maximizes \(f(A, P)\), and then prove the convergence of Lagrange dual function \(g(\mu, \gamma, \nu)\).

Proposition 1. \(\{A^*, P^*\}\) maximizes \(f(A, P)\).

Proof: Denote \(\{A, P\}\) as any feasible RB assignment and power allocation. Let \(\{A^*, P^*\}\) be the RB assignment and power allocation obtained derived based on Eqs. (30) and (26), respectively. Since \(f(A, P)\) is convex with respect to \(P, P^*\) is derived based on the Karush-Kuhn-Tucker (KKT) condition [16]. Therefore, \(f(A, P^*) \geq f(A, P)\).

Let \(\tilde{\alpha}_{k,m,n}\) be the RB assignment derived using the HERM algorithm. Then, \(f(A^*, P^*) = \sum_{k \in K} \sum_{m \in M} \tilde{\alpha}_{k,m,n} H_{k,m,n}\), and

\[
f(A^*, P^*) - f(A, P^*) = \sum_{k \in K} \sum_{m \in M} (\tilde{\alpha}_{k,m,n} H_{k,m,n} - \alpha_{k,m,n} H_{k,m,n}) \tag{36}
\]

Since \(\sum_{k \in K} \sum_{m \in M} \tilde{\alpha}_{k,m,n} = 1\), \(\sum_{k \in K} \sum_{m \in M} \alpha_{k,m,n} = 1\) and

\[
\tilde{\alpha}_{k,m,n} = \begin{cases} 1, & k = \arg \max_{k \in K} H_{k,m,n} \\ 0, & \text{otherwise} \end{cases} \tag{37}
\]

\(f(A^*, P^*) - f(A, P^*) \geq 0. \tag{38}\)

Therefore, the power allocation and RB assignment derived using the HERM algorithm maximizes \(f(A, P)\).

After deriving \(\{A^*, P^*\}\), we obtain \(g(\mu, \gamma, \nu)\) based on Eqs. (22), (23), and (24) in each iteration.

Proposition 2. \(g(\mu, \gamma, \nu)\) converges if the step sizes in updating the Lagrangian multipliers are properly selected.

Proof: The convergence of the Lagrangian multipliers are well proved in Proposition 8.2.4 in [17]. For the sake of brevity, we omit the detailed proof.

Based on the above analysis, we can conclude that the power allocation and RB assignment derived by the HERM algorithm optimize Problem \(\mathcal{P}0\).

B. The optimal solution to Problem \(\mathcal{P}0\)

According to Eq. (30), \(\tilde{\alpha}_{k,m,n}\) equals either 0 or 1. Therefore, the RB assignment derived by the HERM algorithm is the binary RB assignment that satisfies the constraints of Problem \(\mathcal{P}0\). Therefore, the RB assignment derived by the HERM algorithm is the optimal solution to Problem \(\mathcal{P}0\). Since the problem relaxation does not impact the power allocation. Therefore, the optimal RB assignment and power allocation, \(\{A^*, P^*\}\), derived by the HERM algorithm optimizes \(\mathcal{P}0\).

APPENDIX C

PROOF OF LEMMA 3

We prove the convergence of the HERM algorithm by showing the convergence of \(\beta\) and \(q^*\), respectively.

A. The convergence of \(\beta\)

The HERM algorithm relies on the iterative weight updates in Eq. (17) to deactivate fronthaul for energy savings. It is challenging to prove the convergence of \(\beta\) under arbitrary reweighting function. However, we show that if the reweighting function is chosen as

\[
h(I_m, \tau) = \frac{1}{(I_m + \tau) \ln(1 + \tau^{-1})}. \tag{39}
\]
where $\xi = \frac{1}{\|1+\gamma\|^2}$. The $l_0$-norm approximation in the HERM algorithm can be seen as a special case of the majorization-minimization (MM) algorithms [54] [55] and is guaranteed to converge.

**B. The convergence of $q^*$**

Define $F(q) = \max_{(A,P)} \{R^T(A,P) - qP^T(A,P)\}$.

**Proposition 3.** $F(q)$ is a strictly monotonic decreasing function with respect to $q$.

**Proof:** Denote $\{A^1, P^1\}$ and $\{A^2, P^2\}$ as the solutions to $F(q^1)$ and $F(q^2)$, respectively. Then,

$$F(q^1) = \max_{(A,P)} \{R^T(A,P) - q^1P^T(A,P)\}$$

$$= R^T(A^1,P^1) - q^1P^T(A^1,P^1) \geq R^T(A^2,P^2) - q^1P^T(A^2,P^2)$$

Assume $q^2 > q^1$. Since $P^T(A^2,P^2) \geq 0$

$$R^T(A^2,P^2) - q^1P^T(A^2,P^2) \geq R^T(A^2,P^2) - q^2P^T(A^2,P^2)$$

$$= F(q^2)$$

Therefore, $F(q^1) > F(q^2)$. Hence, $F(q)$ is a strictly monotonic decreasing function with respect to $q$. ■

**Proposition 4.** $F(q)$ is a non-negative function when $p$ is determined by any feasible RB assignement and power allocation.

**Proof:** Denote $\{A^1, P^1\}$ as any feasible RB assignment and power allocation. According to the Dinkelbach algorithm [14],

$$q^1 = \frac{R^T(A^1,P^1)}{P^T(A^1,P^1)}.$$  

Therefore,

$$F(q^1) = \max_{(A,P)} \{R^T(A,P) - q^1P^T(A,P)\}$$

$$\geq R^T(A^1,P^1) - q^1P^T(A^1,P^1) = 0.$$  

Hence, $F(q)$ is larger than zero when $p$ is determined by any feasible RB assignment and power allocation. ■

Given Propositions 3 and 4, the convergence of $q^*$ can be proved by showing that $q^*$ increases in each iteration. When $q^*$ keeps increasing, $F(q^*)$ will gradually decrease to zero. Once $F(q^*) = 0$, $q^*$ converges. Next, we show that $q^*$ increases in each iteration. In the HERM algorithm, the value of $q^*$ in the $(i+1)$th iteration is updated according to

$$q^{(i+1)} = R^T(A^{(i)}, P^{(i)})/P^T(A^{(i)}, P^{(i)})$$

Here, $\{A^{(i)}, P^{(i)}\}$ is the optimal solution to $F(q^{(i)})$. Therefore,

$$F(q^{(i)}) = R^T(A^{(i)}, P^{(i)}) - q^{(i)}P^T(A^{(i)}, P^{(i)})$$

Since $F(q^{(i)}) > 0$ until $q^*$ converges,

$$R^T(A^{(i)}, P^{(i)}) - q^{(i)}P^T(A^{(i)}, P^{(i)}) > 0$$

Hence,

$$q^{(i)} = R^T(A^{(i)}, P^{(i)})/P^T(A^{(i)}, P^{(i)}) = q^{(i+1)}.$$  

As $q^{(i+1)}$ increases, $F(q^{(i+1)})$ will eventually reach zero. Then, $q^{(i+1)}$ converges to $q^*$.

Therefore, with the proofs of the convergence $\beta$ and $q^*$, we can conclude that the proposed HERM algorithm converges. Since $F(q^*)$ reaches its maximum value when $q^*$ converges, the HERM algorithm converges to the optimal solution.

**APPENDIX D**

**The Value of $N_{\mu,\gamma,\nu}$**

Denote $\phi(i) = [\mu(i), \gamma(i), \nu(i)]$ and $\delta = [\delta_\mu, \delta_\gamma, \delta_\nu]$ as the Lagrangian multipliers in the $i$th iteration and the corresponding step sizes, respectively. Define $\phi(0)$ and $\phi^*$ as the initial and optimal Lagrangian multipliers, respectively. The sub-gradient of the dual function expressed in Eq. (31) is defined as $\nabla (\phi)$. Let

$$C_j = \sup_{i>0} \{\|f\| \in \nabla \phi(i) \cup \nabla \phi(i-1)\}.$$  

Then, $C = \sum_{j \in \phi} C_j$. The value of $N_{\mu,\gamma,\nu}$ can be derived based on the following proposition.

**Proposition 5.** For any positive scalar $\varepsilon$, we have

$$\min_{0 \leq \phi \leq N_{\mu,\gamma,\nu}} g(\phi(i)) \leq g^* + \frac{\delta \|C\| + \varepsilon}{2},$$

where $N_{\mu,\gamma,\nu}$ is given by

$$N_{\mu,\gamma,\nu} = \min_{\phi \in \frac{\delta \|C\| + \varepsilon}{2}}.$$  

**Proof:** The proposition can be proved by showing that $g(\phi(i))$ keeps decreasing in each iteration until reaching its minimum value if the step size $\delta$ is properly selected. $N_{\mu,\gamma,\nu}$ is derived when $g(\phi(i))$ is optimized. The detail proof can be found in Proposition 8.2.3 in [17]. ■

**REFERENCES**


