On Joint BBU/RRH Resource Allocation in Heterogeneous Cloud-RANs

Kaiwei Wang, Wuyang Zhou, *Member, IEEE*, and Shiwen Mao, *Senior Member, IEEE*

***Abstract*—Cloud radio access network (Cloud-RAN) is a promising wireless network architecture that can satisfy the fast growing mobile data traffic and improve the performance of Internet of Things. In this paper, we propose an energy-efficient resource allocation scheme based on heterogeneous Cloud-RAN jointly considering the remote radio head (RRH) antenna resource with baseband unit (BBU) computation resource. We formulate our joint resource allocation problem and decompose it into two subproblems. The first subproblem is a network-wide beamforming vectors optimization problem, and it is solved by weighted minimum mean square error approach. Based on the optimized beamforming vector, we propose an algorithm to get the RRH-user equipment clusters. The second subproblem is a BBU scheduling problem, and we reformulate it as a bin packing problem which aims to minimize the number of BBUs in working model to save more energy. Compared to some existed works which form the BBU scheduling problem as a bin packing problem, we propose a bin packing algorithm based on the best-fit-decreasing method, which has better performance. With simulation results and detailed analysis, the system performance of our proposed joint resource allocation scheme is verified, which is more energy-efficient than other existing schemes.**

***Index Terms*—Bin packing, heterogeneous cloud radio access network (Cloud-RAN), Internet of Things (IoT), resource allocation, weighted minimum mean square error (WMMSE).**

I. Introduction

THE compelling developed wireless Internet applications, such as high-definition video streaming, mobile cloud computing, and so on, have generated drastically increased wireless communication demands [1]. As wireless data traffic is explosively increasing, the capacity of existing and future wireless networks will be greatly stressed [2]. The definition of next generation networks is under a global discussion [3], and the evolution will be defined as an increasing number of wireless devices with ubiquitous access requirement of mobile services [4]. The change of mobile communication services and applications raise a requirement for more flexible, cost-efficient and powerful network structure and deployment. In the meantime, the fast development of Internet of Things (IoT) has driven the enormous amount of traffic with different quality of service (QoS) requirements and ubiquitous, reliable, and high-data-rate communications [5]. For example, in a future IoT-based smart city, the wireless network for all devices should provide flexible and high quality service as well as low operation cost [6]. However, the current mobile networks are not able to support the diversity mobile services and fluctuating traffic patterns efficiently but are designed for peak-provisioning and typical Internet traffic [7].

To satisfy the future demand of growing mobile data traffic and high-speed data applications in wireless communication system, many advanced technologies are being developed, such as millimeter wave communications [8], [9] and massive MIMO [10], [11]. In addition, cloud technologies, which can provide the flexibility for the radio access network, have already received increasing attention for the deployment of mobile core network functionalities [12]–[14]. As a promising application for cloud concept in mobile wireless communication, a new cellular structure named cloud radio access network (Cloud-RAN) has been proposed by the industry and studied by several researchers for the next generation networks. Unlike the existing cellular networks, the computation resource for baseband process in each baseband unit (BBU) is centralized in baseband resource pool (BRP). The remote radio heads (RRHs) where the radio function locate are connected with the BRP through high reliable optical fibers, and all RRHs can dynamically share the baseband resource provided by any BBU in BRP [15]. The capital expenditure and operational expenditure can be significantly reduced by the distributed deployment of RRHs, while such a centralized processing structure enables several cooperative communication techniques, such as coordinated multipoint (CoMP) transmission and joint beamforming with efficient interference suppression. As a promising technique, cooperative communication can exploit the broadcasting nature of wireless channels and achieve spatial diversity gains, rate improvement, and energy efficiency [16].

Based on all the benefits mentioned above, cloud-based wireless networks also have been recognized as important deployment methods for future IoT networks. Deng *et al.* [17]

Manuscript received December 6, 2016; revised January 27, 2017; accepted February 3, 2017. Date of publication February 7, 2017; date of current version June 15, 2017. This work was supported in part by the Natural Science Foundation of China under Grant 61461136002, in part by the Key Program of National Natural Science Foundation of China under Grant 61631018, in part by the National Programs for High Technology Research and Development under Grant 2014AA01A707, in part by the Fundamental Research Funds for the Central Universities, in part by the Huawei Technology Innovative Research, in part by the U.S. National Science Foundation under Grant CNS-1247955 and Grant CNS-1320664, and in part by the Wireless Engineering Research and Education Center at Auburn University.

K. Wang and W. Zhou are with the Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, University of Science and Technology of China, Hefei 230027, China (e-mail: wangkw@mail.ustc.edu.cn; wyzhou@ustc.edu.cn).

S. Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: smao@ieee.org).

Digital Object Identifier 10.1109/JIOT.2017.2665550

discussed the workload allocation issues in a cloud-based IoT, while a centralized interference mitigation algorithm was presented in [18] to improve the QoS performance in a Cloud-RAN-based D2D communication. Zhang *et al.* [19] proposed a cellular partition zooming mechanism in Cloud-RAN-based IoT to achieve higher energy efficiency. Owing to the distributed RRH deployment, all user equipments (UEs) in IoT would have easier access to the core network with densely deployed RRHs. The centralized BBU structure in Cloud-RAN could facilitate cross-cell cooperation, improve spectrum efficiency, and improve the QoS for all UEs in IoT. However, there are still many new challenges to be addressed, such as the resource scheduling problems, the limited fronthaul and backhaul capacity, the virtualization of computation resource in BBU, and so on.

There have been some interesting works focused on resource allocation and scheduling in Cloud-RAN. Some existing works are on bandwidth resource allocation among users in OFDM-based Cloud-RAN [20], [21], while many other works have addressed CoMP transmission among different RRHs to capitalize the advantages brought by the centralized architecture of Cloud-RAN [22]–[24]. Among them, [22] jointly solved the RRH selection and beamforming vectors optimization problem. Dai and Yu [23] proposed a user-centric clustering scheme to maximize the network utility based on data joint transmission. A grouping scheme of users and RRHs is proposed by [24] to achieve high network performance in Cloud-RAN. However, all these works focused mainly on the antenna resource of RRHs to serve each UE and have overlooked the resource scheduling in the BRP. In a cloud-based network system like Cloud-RAN, the data processing center provides computation resource as well as consumes a significant amount of power [25]. Therefore, the resource allocation and power consumption of BRP should be important considerations in the resource scheduling problems in Cloud-RAN.

Tang *et al.* [26] proposed a cross-layer resource allocation scheme that jointly considers resource from BBUs and RRH antennas, while this paper only focused on the power consumption and rate allocation of virtual machines (VMs) generated by BBUs in the BRP and ignored the scheduling scheme among BBUs. In a Cloud-RAN system, the baseband processing ability is located on BBUs which are centralized in the BRP. This kind of structure brings a large amount of energy savings, such as the energy consumption of cooling system. At the same time, further energy cost reduction can be achieved through efficient BBU scheduling, which means to allocate the baseband processing resource according to different network loads dynamically. A graph-based dynamic frequency reuse scheme is proposed in [27] to minimize the number of BBUs in working model in BRP as well as alleviate the intercell interference in Cloud-RAN. Boulos *et al.* [28] and Sigwele *et al.* [29] formulated a similar problem as a bin packing problem to schedule the baseband resource provided by BBUs based on different network loads. While all these works failed to take the RRH antenna resource scheduling into consideration, and only focused on the resource and power of BBUs in the Cloud-RAN.

In this paper, we jointly consider the RRH antenna resource with BBU computation resource, and propose an energy-efficient resource allocation scheme. We formulate our problem in our proposed heterogeneous Cloud-RAN from our previous work [27]. Different from other existed heterogeneous Cloud-RAN [30], the macro cell coverage and pico cell coverage are all served by RRHs which are connected with the BRP to improve the flexibility of the network. We decompose our problem into two subproblems. The first subproblem is a network-wide beamforming vectors optimization, which aims to get approach to the optimized data processing rate and transmitting rate with the beamforming vectors. We use a weighted minimum mean square error (WMMSE) approach to solve it. Based on the optimized beamforming vector, we propose an algorithm to get the RRH-UE clusters. The second subproblem is a BBU scheduling problem, which aims to minimize the number of working BBUs to save more energy, and we reformulate it as a bin packing problem. Compared to the existed works which formed the BBU scheduling problem as a bin packing problem [28], [29], we propose a bin packing algorithm based on the best-fit-decreasing (BFD) method, which has better performance without increasing the computation complexity. To the best of our knowledge, it is the first resource allocation scheme in heterogeneous Cloud-RAN jointly consider the RRH antenna resource and the BBU scheduling.

This paper is organized as follows. Section II introduces the heterogeneous Cloud-RAN structure and formulates the energy-efficient joint resource allocation problem. Based on WMMSE approach, Section III analyzes and solves the network-wide beamforming vectors optimization problem. We formulate the BBU scheduling problem as a bin packing problem in Section IV and propose a heuristic algorithm to solve it. Section V shows the simulation results and finally, Section VI provides the conclusion.

II. System Model and Problem Formulation

In this section, we introduce the heterogeneous Cloud-RAN system and formulate our energy-efficient joint resource allocation problem.

A. System Scenario of Heterogeneous Cloud-RAN

We consider a heterogeneous Cloud-RAN system with two kinds of RRHs: 1) the macro RRH and 2) the pico RRH. The macro RRHs are regularly deployed to form ordinary hexagonal macro cells and provide wide area coverage. Each macro cell is equipped with one macro RRH. Several pico RRHs are located in each macro cell to serve some hot points or edge areas. We denote the set of macro cells as $\mathcal{I} = \{1, 2, \ldots, I\}$, so there would be I macro RRHs to serve the whole network.

The heterogeneous Cloud-RAN proposed in this paper is shown in Fig. 1, where all the RRHs are connected to the BRP through a switch of 10 Gb/s Ethernet [31]. All the UEs of IoT can access to several RRHs based on different channel states and QoS requirements. The switch can dynamically change the connections between BBUs and RRHs under the control of a center management unit (CMU). The BRP consists of K

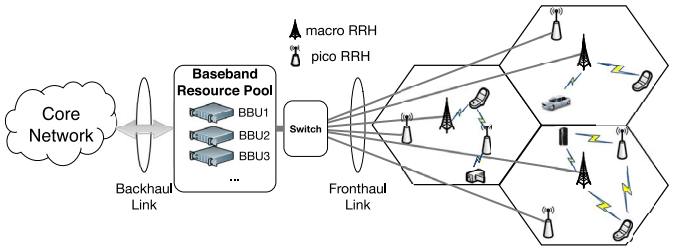


Fig. 1. Heterogeneous Cloud-RAN architecture.

BBUs, and we use $\mathcal{K} = \{1, 2, \ldots, K\}$ to denote the set of all BBUs. The BBUs could serve each UE by generating a VM to provide computation resource as common data center do in a cloud-based system [32]. Each UE has its own corresponding VM located in the BRP. Due to limited processing ability of each BBU [15], the number of VMs generated by one BBU is limited, which means that one BBU can only support a limit number of UEs. And we also assume that each UE can only be supported by one BBU. Similar assumption about the BBU processing ability can be found in [27], [33], and [34]. A BBU can be turned to sleep model when no UE need to be supported by it to save energy. All BBUs in the BRP is assumed to be the same with each other, which means different BBUs have the same processing ability.

To capitalize the advantages brought by the centralized architecture of Cloud-RAN, we consider a CoMP transmission scheme among all the RRHs in this paper. Among several kinds of downlink CoMP techniques, we choose joint processing scheme to formulate our system. In joint processing transmission scheme, each UE's data will be shared among all the RRHs in the CoMP cluster [35]. Considering the channel conditions and power consumption, there is no need for all RRHs in the whole network to joint serve one UE. So similar to the static base station clustering in [23] and the virtual base station clustering scheme in [24], we limit the maximum cooperative range of RRHs into one macro cell, i.e., each UE can be jointly served by the macro RRH and all pico RRHs located in the cell where the UE located in. The signal from RRHs located in other macro cells will be treated as interference. All RRHs in macro cell i forms a set $\mathcal{L}_i = \{1, 2, \ldots, |\mathcal{L}_i|\}$, where $|\mathcal{L}_i|$ denote the cardinality of set $\mathcal{L}_i$. Also, we denote the UE set in cell i as $\mathcal{M}_i = \{1, 2, \ldots |\mathcal{M}_i|\}$.

The arrival data of UE $m \in \mathcal{M}_i$ which comes from the core network will first be processed in VM generated by one BBU in the BRP, and then transmitted to the RRHs located in cell i which jointly serve UE $m \in \mathcal{M}_i$ through the fronthaul link. We assume that each RRH is equipped with N transmission antennas and each UE is equipped with one receiving antenna. We use $\mathbf{h}_{lm}^{ij} \in \mathbb{C}^{1\times N}$ to denote the channel gain from RRH $l \in \mathcal{L}_i$ to UE $m \in \mathcal{M}_j$. Since UE $m \in \mathcal{M}_i$ only receives useful signal from the RRHs located in the same cell, we use $\mathbf{w}_{lm}^{i} \in \mathbb{C}^{N\times 1}$ to denote the beamforming vector for UE $m \in \mathcal{M}_i$ from RRH $l \in \mathcal{L}_i$. We use $\mathbf{w}_{m}^{i} \in \mathbb{C}^{|\mathcal{L}_i|N\times 1}$ to denote the beamforming vector of UE $m \in \mathcal{M}_i$ based on all the RRHs in $\mathcal{L}_i$ and $\mathbf{w} = \{\mathbf{w}_m^i | i \in \mathcal{I}, m \in \mathcal{M}_i\}$ to denote all the beamforming vectors. Correspondingly, $\mathbf{h}_{m}^{ij} \in \mathbb{C}^{1\times|\mathcal{L}_i|N}$ is used to denote the channel gain from the RRHs in $\mathcal{L}_i$ to UE $m \in \mathcal{M}_j$ in $\mathcal{M}_j$. Given that $\mathbf{D}_{il} = \{\mathbf{0}_N^1, \ldots, \mathbf{I}_N^l, \ldots, \mathbf{0}_N^{|\mathcal{L}_i|}, l \in \mathcal{L}_i\}_{N\times|\mathcal{L}_i|N}$, we can represent $\mathbf{w}_{lm}^i$ with $\mathbf{w}_m^i$ through

$$\mathbf{w}_{lm}^i = \mathbf{D}_l^i \mathbf{w}_m^i. \tag{1}$$

Note that the beamforming vectors can also demonstrate the transmission relationship between UEs and RRHs. UE $m \in \mathcal{M}_i$ is served by RRH $l \in \mathcal{L}_i$ only when $||\mathbf{w}_{lm}^i|| \neq \mathbf{0}$.

Let x_m^i denote the data symbol for UE $m \in \mathcal{M}_i$ with $E[|x_m^i|^2] = 1$ and x_m^i's are independent with each other. Then the received signal at UE $m \in \mathcal{M}_i$ is given by

$$y_m^i = \mathbf{h}_m^{ii}\mathbf{w}_m^i x_m^i + \sum_{\substack{m'\in\mathcal{M}_i,\\ m'\neq m}} \mathbf{h}_m^{ii}(t)\mathbf{w}_{m'}^i x_{m'}^i + \sum_{i'\neq i}\sum_{m''\in\mathcal{M}_{i'}} \mathbf{h}_m^{i'i}\mathbf{w}_{m''}^{i'} x_{m''}^{i'} + z_m^i, m \in \mathcal{M}_i, i \in \mathcal{I} \tag{2}$$

where the first term on the right hand side is the useful signal for UE $m \in \mathcal{M}_i$, the second term is the intracell interference signal from other RRHs in cell i, and the third term is the intercell interference from the RRHs in other cells. $z_m^i \sim \mathcal{CN}(0, (\sigma_m^i)^2)$ represents the additive Gaussian noise. Consequently, the transmission rate for UE $m \in \mathcal{M}_i$ can be formulated as

$$c_m^i = B_0 \log_2\left(1 + \frac{\left|\mathbf{h}_m^{ii}\mathbf{w}_m^i\right|^2}{\left|\sum_{(j,n)\neq(i,m)} \mathbf{h}_m^{ji}\mathbf{w}_n^j\right|^2 + \left(\sigma_m^i\right)^2}\right) \tag{3}$$

where B_0 is the total system bandwidth. We can see clearly from (3) that the transmission rate of each UE depends not only on the beamforming vector of itself, but also on the transmission beamforming vectors of other UEs. So the coordination of RRHs in the same macro cell and among different macro cells is needed to suppress the intracell and intercell interference. And according to the time varying channel status and QoS requirements of UEs, the coordination of different access points need to be updated quickly. Such kind of coordination is hard to realized in traditional heterogeneous cellular network due to the distributed network structure. While in a centralized heterogeneous Cloud-RAN, it is much easier to be achieved. To fulfill the coordination of all RRHs in the network, the CMU can decide the network-wide beamforming vectors according to the channel state information and different UEs' QoS requirements, and allocate the BBU processing resource to each UE via generating VMs. So in this paper, we combine the optimization of beamforming vectors and BBU resource scheduling and propose an energy-efficient method to achieve better system performance.

Base on the beamforming vectors of each UE, the total power consumption of RRH l in cell i can be expressed as

$$P_l^i = \sum_{m\in\mathcal{M}_i} \left\|\mathbf{D}_l^i \mathbf{w}_m^i\right\|_2^2 + P^C \tag{4}$$

where $\sum_{m\in\mathcal{M}_i} ||\mathbf{D}_l^i\mathbf{w}_m^i||_2^2$ represents the transmit power of RRH $l \in \mathcal{L}_i$ and P^C represents the circuit and fronthaul link power consumption of each RRH. Since each RRH is equipped

with no air conditioner and the fronthaul link power consumption is rather small, P^C could be neglected [36]. Then, the total power consumption of RRH $l \in \mathcal{L}_i$ only consists of transmit power, which can be expressed as

$$P_l^i = \sum_{m \in \mathcal{M}_i} \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2. \tag{5}$$

The transmit power of RRH $l \in \mathcal{L}_i$ is limited to $(P_l^i)^{\max}$, which can be given by

$$\sum_{m \in \mathcal{M}_i} \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2 \le \left(P_l^i\right)^{\max}. \tag{6}$$

We use a_{km}^i to indicate whether UE $m \in \mathcal{M}_i$ is served by the VM generated by BBU k, which can be expressed as

$$a_{km}^i = \begin{cases} 1, & \text{if UE } m \text{ is served by BBU } k, m \in \mathcal{M}_i \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

With the indicator variable a_{km}^i, the constraints of BBU processing ability can be formulated as

$$\sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \le U \tag{8}$$

which constrains the processing ability of BBU k to U, and

$$\sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} a_{km}^i \le 1 \tag{9}$$

which limits the data of each UE $m \in \mathcal{M}_i$ can only be processed by one BBU.

We use P_k^B to denote the power consumption of BBU k. Based on the BBU power consumption model given by [37], we have

$$P_k^B = \begin{cases} P_{\text{act}} + \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \varphi\left(\mu_m^i\right), & \text{if } k \in \mathcal{B} \\ P_{\text{sleep}}, & \text{otherwise.} \end{cases} \tag{10}$$

In P_k^B, $\mathcal{B}$ represents the set of BBUs in working model. The parameter P_{act} here indicates the statistic part of power consumption of BBU in working model, which includes the power consumption of backhaul transmission equipment, the air conditioner, and so on. We use P_{sleep} to denote the power consumption of BBU in sleeping model. μ_m^i here indicates the VM processing rate for the data of UE $m \in \mathcal{M}_i$. The power consumption model of VM is often assumed to be a convex and increasing function with the processing rate of VM, and it forms the dynamic power consumption part of BBUs in working model. We use $\varphi(\mu_m^i)$ to indicate the power consumption of VM according to the data processing rate it provides for UE $m \in \mathcal{M}_i$, which is linear with the data processing rate μ_m^i [38], where we have

$$\varphi\left(\mu_m^i\right) = \alpha \mu_m^i \tag{11}$$

α represents the variation coefficient of $\varphi(\mu_m^i)$ as a function of μ_m^i.

We assume that the BBU data processing rate should be equivalent to the RRH data transmission rate to satisfy the basic UE QoS requirement [28], [38], so usually we have $\mu_m^i = c_m^i$. Based on the energy consumption of each BBU, the energy consumption of the entire BRP can be expressed as

$$\begin{aligned} P^B &= \sum_{k \in \mathcal{K}} P_k^B \\ &= (K - |\mathcal{B}|) P_{\text{sleep}} + |\mathcal{B}| P_{\text{act}} + \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} a_{km}^i \varphi\left(\mu_m^i\right) \\ &= \left(K - \sum_{k \in \mathcal{K}} \mathbb{1}\left\{ \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \right\} \right) P_{\text{sleep}} \\ &\quad + \sum_{k \in \mathcal{K}} \mathbb{1}\left\{ \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \right\} P_{\text{act}} \\ &\quad + \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} a_{km}^i \varphi\left(\mu_m^i\right) \end{aligned} \tag{12}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function, which is denoted by

$$\mathbb{1}\left\{ \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} a_{km}^i \varphi\left(\mu_m^i\right) \right\} = \begin{cases} 0, & \text{if } \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} a_{km}^i \alpha \mu_m^i = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{13}$$

Note that $P_{\text{sleep}} \ll P_{\text{act}}$, it is reasonable to turn as many BBUs as we can to sleep model on the premise of meeting UEs' QoS requirements to save more energy.

B. Energy Efficient Joint Resource Scheduling Problem Formulation

To quantitatively describe the impact on system performance brought by transmission rate and power consumption, we use an energy efficiency weighted utility function $f(c_m^i, P_k^B, P_l^i)$ to denote the system energy efficiency, which is a widely used method adopted by many other works [39], [40]. The energy efficiency utility function is defined as

$$f\left(c_m^i, P_l^i, P_k^B\right) = \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} c_m^i - \kappa \left(\sum_{k \in \mathcal{K}} P_k^B + \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} P_l^i \right) \tag{14}$$

where κ represents the system power consumption weight. Our purpose of this paper is to maximize the energy efficiency utility function of the whole heterogeneous Cloud-RAN system, including the transmit power of RRHs and data processing power of BBUs. We jointly consider the RRH antenna resource and BBU processing resource scheduling, and combine the network-wide beamforming vector optimization with BBU processing rate allocation.

Note that the fronthaul links between RRHs and BBUs in the BRP are capacity-limited, which means the sum rate of all UEs access to one RRH is limited. Through the indicator function $\mathbb{1}\{\cdot\}$, the fronthaul capacity constraint of RRH $l \in \mathcal{L}_i$ can be expressed as

$$\sum_{m \in \mathcal{M}_i} \mathbb{1}\left\{ \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2 \right\} c_m^i \le R_l^i \tag{15}$$

where R_l^i is the maximum capacity of the fronthaul link connect between RRH $l \in \mathcal{L}_i$ and the BRP.

Considering all the constraints mentioned above, the energy efficiency maximization problem can be mathematically formulated as

$$\begin{aligned}
\text{(P0)} \quad & \max_{\mathbf{a},\mathbf{w},\boldsymbol{\mu}} \ f\left(c_m^i, P_l^i, P_k^B\right) \\
& \text{s.t.} \ \ \text{C1}: c_m^i \geq \left(c_m^i\right)^{\text{req}} \ \forall i \in \mathcal{I}, m \in \mathcal{M}_i \\
& \qquad \text{C2}: P_l^i \leq \left(P_l^i\right)^{\max} \ \forall i \in \mathcal{I}, l \in \mathcal{L}_i \\
& \qquad \text{C3}: \sum_{m \in \mathcal{M}_i} \mathbb{1}\left\{ \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2 \right\} c_m^i \leq R_l^i \\
& \qquad\qquad \forall i \in \mathcal{I}, m \in \mathcal{M}_i, l \in \mathcal{L}_i \\
& \qquad \text{C4}: \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \leq U \ \forall k \in \mathcal{K} \\
& \qquad \text{C5}: \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} a_{km}^i \leq 1 \ \forall m \in \mathcal{M}_i. \qquad (16)
\end{aligned}$$

In problem (P0), C1 is the constraint to ensure the QoS requirement of each UE jointly consider the BBU processing rate and RRH transmitting rate. C2 is the transmit power constraint of each RRH. C3 is the fronthaul capacity constraint. C4 and C5 are the BBU processing ability constraints, which limit the total number of UEs one BBU can process and that one UE can only be served by one BBU separately.

Since there is no difference between the data processing ability of all BBUs in the BRP, the obvious way to save energy consumptions of the whole network is minimizing the number of working BBUs. Based on this idea, the energy efficiency utility function can be separated into two subfunctions as

$$f\left(c_m^i, P_l^i, P_k^B\right) = f_1\left(c_m^i, P_l^i\right) + f_2\left(a_{km}^i\right) \qquad (17)$$

where we have

$$\begin{aligned}
f_1\left(c_m^i, P_l^i\right) = & \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} c_m^i \\
& - \kappa \left(\sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} \varphi\left(c_m^i\right) + \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}_i} \sum_{m \in \mathcal{M}_i} \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2 \right) \qquad (18)
\end{aligned}$$

and

$$\begin{aligned}
f_2\left(a_{km}^i\right) = & \ \kappa \sum_{k \in \mathcal{K}} \mathbb{1}\left\{ \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \right\} P_{\text{act}} \\
& + \kappa \left(K - \sum_{k \in \mathcal{K}} \mathbb{1}\left\{ \sum_{i \in \mathcal{I}} \sum_{m \in \mathcal{M}_i} a_{km}^i \right\} \right) P_{\text{sleep}}. \qquad (19)
\end{aligned}$$

The first subutility function $f_1(c_m^i, P_l^i)$ includes the data transmission rate of each UE, the power consumption of each RRH and each VM generated by BBUs in the BRP, which are all decided by the network-wide beamforming vectors. The second subutility function $f_2(a_{km}^i)$ includes the statistic power consumption of all BBUs in working model and the power consumption of all BBUs in sleeping model.

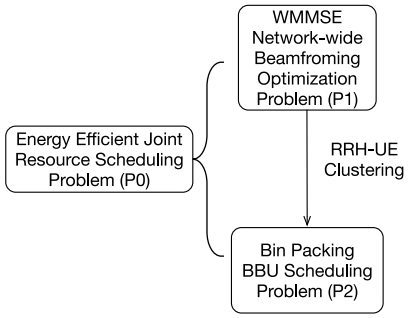


Fig. 2. Two-step approach to solve problem P1.

Based on the two subutility functions, we can separate the energy efficiency maximization problem into two subproblems. The first subproblem is to determine the optimized network-wide beamforming vectors of each UE by maximizing $f_1(c_m^i, P_l^i)$. Based on the beamforming vectors, we can get the RRH-UE cluster, and the energy-efficient data transmission rate of RRHs with data processing rate of BBUs. The first subproblem can be formulated as a WMMSE problem. Then the second subproblem is to decide the UEs in each RRH-UE cluster should be served by which BBU in the BRP, and to minimize the number of working BBUs. We will formulate the second subproblem as a bin packing problem. These two subproblems will be elaborately analyzed in the following parts of this paper.

As shown in Fig. 2, the energy-efficient joint resource allocation problem (P0) will be divided into two subproblems: 1) the network-wide beamforming optimization problem (P1) and 2) the BBU scheduling problem (P2). We need to solve problem (P1), and use a RRH-UE clustering algorithm to get the RRH-UE clusters, then solve the problem (P2) to obtain the final resource schedule scheme.

III. Energy Efficient Beamforming Strategy

In this section, we first formulate the network-wide beamforming optimization problem as problem P1 based on the subutility function $f_1(c_m^i, P_l^i)$ and solve ti with WMMSE approach, then we propose an RRH-UE clustering algorithm to get the RRH-UE clusters based on the optimized beamforming vectors.

A. Network-Wide Beamforming Optimization

Based on the subutility function $f_1(c_m^i, P_l^i)$ and the constraints related to beamforming vectors, the subproblem of network-wide beamforming optimization can be formulated as

$$\begin{aligned}
\text{(P1)} \quad & \max_{\mathbf{w}} \ f_1\left(c_m^i, P_l^i\right) \\
& \text{s.t.} \ \ \text{C1}: c_m^i \geq c^{\text{req}} \ \forall i, m \\
& \qquad \text{C2}: P_l^i \leq \left(P_l^i\right)^{\max}, \ \forall i \in \mathcal{I}, l \in \mathcal{L}_i \\
& \qquad \text{C3}: \sum_{m \in \mathcal{M}_i} \mathbb{1}\left\{ \left\| \mathbf{D}_l^i \mathbf{w}_m^i \right\|_2^2 \right\} c_m^i \leq R_l^i \ \forall l. \qquad (20)
\end{aligned}$$

Since the phase of $\mathbf{w}_m^i$ will not have impact on the optimization problem or change the constraints, so we can assume that

each term of $\mathbf{h}_m^{ii}\mathbf{w}_m^i$ has a zero imaginary part. Then, we can rewrite the constrain C1 as

$$\text{C1}: \sqrt{\left|\sum_{(j,n)\neq(i,m)} \mathbf{h}_m^{ji}\mathbf{w}_n^j\right|^2 + \left(\sigma_m^i\right)^2} \leq \frac{1}{\sqrt{\gamma_m^i}}\Re\left\{\mathbf{h}_m^{ii}\mathbf{w}_m^i\right\} \ \forall m \tag{21}$$

which is a second order cone (SOC) constraint and $\gamma_m^i = 2^{((c_m^i)^{\text{req}}/B_0)} - 1$ is the equivalent signal-to-interference-plus-noise ratio threshold for UE m.

With the concept that a non convex ℓ_0-norm optimization objective can be approximated by a convex ℓ_1 reweighted norm [41], we can approximate the fronthaul capacity constraint C3 as

$$\text{C3}: \sum_{m\in\mathcal{M}} \beta_{lm}^i \tilde{c}_m^i \left\|\mathbf{D}_l^i \mathbf{w}_m^i\right\|_2^2 \leq R_l \quad \forall i \in \mathcal{I}, l \in \mathcal{L}_i \tag{22}$$

where β_{lm}^i is a constraint weight, which is updated iteratively according to

$$\beta_{lm}^i = \frac{1}{\left\|\mathbf{D}_l^i \mathbf{w}_m^i\right\|_2^2 + \tau} \quad \forall i \in \mathcal{I}, l \in \mathcal{L}_i, m \in \mathcal{M}_i \tag{23}$$

and $\tilde{c}_m^i$ is the optimum RRH transmission rate for UE m obtained by the previous iteration.

After all these processing of constraints, the problem (20) is still not convex. So motivated by [42], we can use a WMMSE-based solution to deal with the problem. We can reformulate the target function as

$$f_1\left(c_m^i, P_l^i\right) = \sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}_i} g\left(c_m^i\right) + \kappa \sum_{i\in\mathcal{I}}\sum_{l\in\mathcal{L}_i}\sum_{m\in\mathcal{M}_i} \left\|\mathbf{D}_l^i \mathbf{w}_m^i\right\|_2^2 \tag{24}$$

where

$$g\left(c_m^i\right) = c_m^i - \kappa\varphi\left(c_m^i\right). \tag{25}$$

We denote $\theta(\cdot) = g(-B_0\log(\cdot))$, since $\varphi(\cdot)$ has the form of (11) which is liner with the data processing rate of each VM, $g(c_m^i)$ is strictly concave. With a carefully selected parameter κ, $g(c_m^i)$ could also be strictly increasing with c_m^i. So $\theta(\cdot)$ is also concave. Following a similar proof as that of [42, Th. 2], the problem of (20) can be equivalently transformed into

$$\begin{aligned}\min_{\mathbf{w}} \ & \sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}_i} \omega_m^i e_m^i + \theta\left(\phi\left(\omega_m^i\right)\right) - \omega_m^i\phi\left(\omega_m^i\right) \\ & + \kappa\sum_{i\in\mathcal{I}}\sum_{l\in\mathcal{L}_i}\sum_{m\in\mathcal{M}_i} \left\|\mathbf{D}_l^i\mathbf{w}_m^i\right\|_2^2 \\ \text{s.t.} \ & \text{C1} \sim \text{C3}\end{aligned} \tag{26}$$

where $\phi(\cdot)$ is the inverse mapping of the gradient map $\nabla\theta(e_m^i)$ and ω_m^i is the MSE weight of UE $m \in \mathcal{M}_i$. Under the independence assumption of s_m^i and z_m^i, the corresponding MSE e_m^i is defined as

$$\begin{aligned} e_m^i &\triangleq E\left\{\left(s_m^i - u_m^i y_m^i\right)\left(\left(s_m^i\right)^* - \left(u_m^i y_m^i\right)^*\right)\right\} \\ &= \left(u_m^i\right)^H\left(\sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{M}_i} \mathbf{h}_m^i\mathbf{w}_n^i\left(\mathbf{w}_n^i\right)^H\left(\mathbf{h}_m^i\right)^H + \sigma^2\right)u_m^i \\ &\quad - 2\Re\left\{\left(u_m^i\right)^H\mathbf{h}_m^i\mathbf{w}_m^i\right\} + 1 \end{aligned} \tag{27}$$

Algorithm 1 Energy Efficient Beamforming Strategy With QoS Constrains

Initialization: Choose network-wide beamforming vector $\mathbf{w}_m^i$, then based on $\mathbf{w}_m^i$, initialize β_{lm}^i, $\tilde{c}_m^i$, $\forall i \in \mathcal{I}, l \in \mathcal{L}_i, m \in \mathcal{M}_i$.

repeat

With fixed beamforming vector $\mathbf{w}_m^i$, calculating the MSE weight ω_m^i and the optimum receiver u_m^i according to (27), (29) and (28) in turn, $\forall i \in \mathcal{I}, m \in \mathcal{M}_i$;

Solve the problem (30) with fixed ω_m^i and u_m^i to get the optimized energy-efficient beamforming vector $(\mathbf{w}_m^i)^*$;

Compute the UE data rate c_m^i according to (3) with $(\mathbf{w}_m^i)^*$;

Update $\tilde{c}_m^i = c_m^i$, $\mathbf{w}_m^i = (\mathbf{w}_m^i)^*$, and update β_{lm}^i according to (23), $\forall i \in \mathcal{I}, l \in \mathcal{L}_i, m \in \mathcal{M}_i$;

until convergence

where u_m^i is the optimum receiver under given beamforming vector $\mathbf{w}_m^i$ and can be given by

$$u_m^i = \frac{\mathbf{h}_m^i\mathbf{w}_m^i}{\sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{M}_i}\mathbf{h}_m^i\mathbf{w}_n^i\left(\mathbf{w}_n^i\right)^H\left(\mathbf{h}_m^i\right)^H + \sigma^2}. \tag{28}$$

By fixing the beamforming vector $\mathbf{w}_m^i$ and the optimum receiver u_m^i, the optimum MSE weight ω_m^i for UE m in cell i can be given by

$$\omega_m^i = \nabla\theta\left(e_m^i\right). \tag{29}$$

Note that the optimization problem (26) is convex with respect to each variables $\mathbf{w}_m^i$, under given ω_m^i and u_m^i, the optimum beamforming vector can be obtained by solving the following problem:

$$\begin{aligned}\text{(P1-1)} \quad \min_{\mathbf{w}} \ & \sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}_i} \omega_m^i e_m^i + \kappa\sum_{i\in\mathcal{I}}\sum_{l\in\mathcal{L}_i}\sum_{m\in\mathcal{M}_i} \left\|\mathbf{D}_l^i\mathbf{w}_m^i\right\|_2^2 \\ \text{s.t.} \ & \text{C1} \sim \text{C3}\end{aligned}$$

which is a quadratically constrained quadratic programming with quadratic constraints and SOC constraints. The problem can be easily solved by using standard convex optimization solver such as CVX [43]. And the complete method to solve the energy-efficient beamforming optimization problem (25) can be elaborated as Algorithm 1.

Suppose Algorithm 1 needs T total number of iterations to converge or the maximum number of iterations is set to T, then the computational complexity can be approximately given as $O(T\cdot(NML)^{3.5})$ [44].

B. RRH-UE Clustering

Based on the network-wide beamforming vector of each UE, we can get the connection between RRHs and UEs. In this paper, for any UE $m \in \mathcal{M}_i$, all RRHs located in cell i just provide a range of CoMP RRH cluster options. Limited by other constraints, such as the channel states and fronthaul capacity, the optimum results of $\mathbf{w}_m^i$ may have some zero terms, which indicate that the UE does not receive useful signals from certain RRHs. We use $\mathcal{L}_m^i$ to denote the set of RRHs which serve the UE $m \in \mathcal{M}_i$. Based on the optimum results of beamforming vectors, we can get the RRH-UE clusters as indicated in

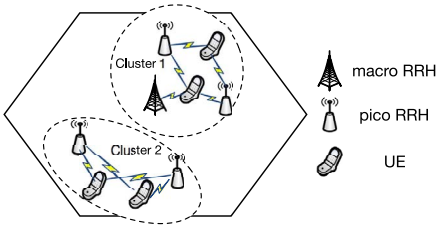


Fig. 3. RRH-UE clusters.

Algorithm 2 RRH-UE Clustering

Initialize:$p = 1$;
for $i \in \mathcal{I}$ **do**
  repeat
    choose $m \in \mathcal{M}_i$, $u_p = 1$, let $\mathbf{v}_p = \mathbf{w}_m^i$, $\mathcal{L}^p = \mathcal{L}_m^i$;
    update $\mathcal{M}_i$ as $\mathcal{M}_i = \mathcal{M}_i \backslash \{m\}$;
    repeat
      choose $n \in \mathcal{M}_i$
      if $\mathbf{v}_p \bigcap \mathbf{w}_n^i \neq \mathbf{0}$ **then**
        update $u_p = u_p + 1$
        $\mathbf{v}_p = \mathbf{v}_p \bigcap \mathbf{w}_n^i$
        update $\mathcal{M}_i$ as $\mathcal{M}_i = \mathcal{M}_i \backslash \{n\}$;
        update $\mathcal{L}^p = \mathcal{L}^p \bigcap \mathcal{L}_n^i$
      end if
    until can not find a UE $n \in \mathcal{M}_i$ meet the condition $\mathbf{v}_p \bigcap \mathbf{w}_n^i \neq \emptyset$
    update $p = p + 1$;
  until $\mathcal{M}_i = \emptyset$
end for

Fig. 3. All UEs in one RRH-UE cluster will be jointly served by all RRHs in that cluster. According to the processing limit of BBUs, all RRHs in one cluster need to be assigned to one BBU, which means all data for UE in that cluster will be processed by that BBU. In this paper, our purpose is to find the RRH-UE clusters based on the optimum beamforming vectors first, then assigned these clusters to certain BBUs.

To obtain the RRH-UE clusters according to the network-wide beamforming vectors, we propose an algorithm as is described as Algorithm 2, which has a computation complexity of $O(M^2)$. We use $\mathcal{U} = \{u_1, u_2, \ldots, u_P\}$ to denote the set of all formed P RRH-UE clusters, where u_p indicates the number of UEs cluster p has. And we also use $\mathcal{L}^p$ to denote the set of RRHs which serves all the UEs in cluster p, where we have $\mathcal{L}^p \in \mathcal{L}, p = 1, 2, \ldots, P$.

IV. BBU Scheduling Scheme

In this section, we formulate the BBU scheduling subproblem as a bin packing problem, and propose an algorithm to solve it.

As the system know the network-wide optimum beamforming vector of each UE, the system could know all the new formed RRH-UE clusters through Algorithm 2. According to the energy consumption minimization problem of (16) and the subutility function $f_2(a_{km}^i)$ with BBU processing limit constraints, the BBU scheduling problem can be formulated as

$$\begin{aligned} \text{(P2)} \quad & \min_{\mathbf{a}} \ f_2\left(a_{km}^i\right) \\ & \text{s.t. C4}: \sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}_i} a_{km}^i \leq U \\ & \qquad \forall i \in \mathcal{I}, m \in \mathcal{M}_i, k \in \mathcal{K} \\ & \qquad \text{C5}: \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{I}} a_{km}^i \leq 1 \ \forall k \in \mathcal{K}, i \in \mathcal{I}, m \in \mathcal{M}_i. \end{aligned} \tag{30}$$

Note that the statistic power consumption P_{act} of BBU in working model is much larger than the power consumption P_{sleep} in sleeping model, so we can equivalently transfer problem (P2) to a BBU number in working model minimization problem as (P2-1)

$$\begin{aligned} \text{(P2-1)} \quad & \min_{\mathbf{a}} \ \sum_{k\in\mathcal{K}} \mathbb{1}\left\{\sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}_i} a_{km}^i\right\} \\ & \text{s.t. C4}: \sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}_i} a_{km}^i \leq U \\ & \qquad \forall i \in \mathcal{I}, m \in \mathcal{M}_i, k \in \mathcal{K} \\ & \qquad \text{C5}: \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{I}} a_{km}^i \leq 1 \ \forall k \in \mathcal{K}, i \in \mathcal{I}, m \in \mathcal{M}_i. \end{aligned} \tag{31}$$

Since the BBU processing ability limit that the data from one UE can only be processed in one BBU, all RRHs in one cluster need to establish connections with one BBU, and all UEs in that cluster need to be assigned to the certain BBU. Based on the RRH-UE clustering results, we can transfer the corresponding relationship between UEs and BBUs into the relationship between RRH-UE clusters and BBUs. We use b_k^p as the indicator which can be expressed as

$$b_k^p = \begin{cases} 1, & \text{if RRH-UE cluster } p \text{ is served by BBU } k \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

and based on b_k^p, we can easily get the related indicate value a_{km}^i. If $b_k^p = 1$, then all UEs in RRH-UE clusters p will be served by BBU k. We also use y_k to indicate whether BBU k is in working model to provide computation resource, where we have

$$y_k = \begin{cases} 1, & \text{if BBU } k \text{ is in working model} \\ 0, & \text{if BBU } k \text{ is in sleep model.} \end{cases} \tag{33}$$

Based on the indicator b_k^p and y_k, the BBU processing ability constraints can be rewrote as

$$\begin{aligned} & \text{C4}: \sum_{p=1}^{P} u_p b_k^p \leq U y_k \ \forall k \in \mathcal{K} \\ & \text{C5}: \sum_{p=1}^{P} b_k^p \leq 1 \ \forall k \in \mathcal{K} \end{aligned} \tag{34}$$

separately, where constraint C4 limits the maximum number of UEs one BBU can support, and constraint C5 limits that

Algorithm 3 Bin Packing-Based BBU Scheduling

Initialization: Rearrange $u_p \in \mathcal{U}$ by descending order, $\mathcal{L} = \{\mathcal{L}^p | p \in \{1, 2, ..., P\}\}$, for all BBUs,$C_k = U, k = 1, ..., K$.
$i = 1, j = 1$;
Choose u_j and its associate RRH set $\mathcal{L}^j$;
Assign $\mathcal{L}^j$ to BBU i;
Update $\mathcal{L} = \mathcal{L} \backslash \mathcal{L}^j$;
Update the remaining capacity of BBU i as $C_i = C_i - u_j$;
repeat
  Update $j = j + 1$, choose u_j;
  if there is BBU k fulfill $C_k - u_j \geq 0$ **then**
    Choose $k^* = \arg\min_k (C_k - u_j)$,
      $\forall k$ fulfill $C_k - u_j \geq 0, k \in [1, ..., i]$;
    Assign $\mathcal{L}^j$ to BBU k^*;
    Update the remaining capacity of BBU k^* as $C_{k^*} = C_{k^*} - u_j$;
    Update $\mathcal{L} = \mathcal{L} \backslash \mathcal{L}^j$;
  else
    Update $i = i + 1$;
    Assign $\mathcal{L}^j$ to BBU i;
    Update the remaining capacity of BBU i as $C_i = C_i - u_j$;
    Update $\mathcal{L} = \mathcal{L} \backslash \mathcal{L}^j$;
  end if
until $\mathcal{L} = \emptyset$

the data for one UE can only be processed by one BBU. Thus, the subutility function $f_2(a_{km}^i)$ can be replaced by

$$\tilde{f}_2(b_k^p) = \kappa \sum_{k \in \mathcal{K}} b_k^p P_{\text{act}} + \kappa \left(K - \sum_{k \in \mathcal{K}} b_k^p \right) P_{\text{sleep}}. \quad (35)$$

Thus, we can reformulated the problem (P2-1) as

$$\text{(P2-2)} \quad \min_{\mathbf{a}} \sum_{k \in \mathcal{K}} y_k \quad \text{s.t. C4, C5} \quad (36)$$

which is a typical 1-D bin packing problem.

Bin packing problem is a kind of problem aimed to assign a set of items in different sizes into the minimum number of bins. Each bin has a fixed bin capacity, so that the sum size of all items assigned into one bin can not exceeds the capacity. In this paper, each BBU is regarded as a bin, and RRH-UE clusters are regarded formed as different items which have different numbers of UEs as sizes. And now we need to assign these RRH-UE clusters to different BBUs in BRP to set up the physical fronthaul link between BBUs and RRHs, and minimize the number of BBUs in working model to save more energy.

Since bin packing problem is one of the classical NP-hard problems, many heuristic algorithms were proposed to find better packing results. Boulos *et al.* [28] formulated the BBU resource allocation problem in a frequency reuse system as a bin packing problem, and proposed a heuristic algorithm to assign each RRH to BBUs. It tried to assign the RRH with largest resource demand and the RRH with the least resource demand to one BBU in one turn. Sigwele *et al.* [29] proposed

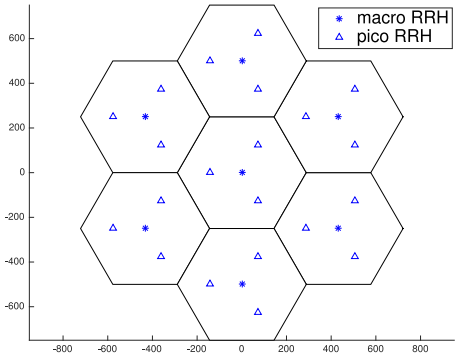


Fig. 4. Simulation scene of heterogeneous Cloud-RAN.

three kinds of heuristic algorithms to allocate the computing resource of BBU based on next-fit (NF), first-fit, and first-fit-decreasing (FFD), those are three widely used approximate algorithms for bin packing problem. In this paper, we propose a heuristic algorithm based on BFD, which is another bin packing approximate algorithm and has better performance without increasing the complexity.

We first sort all RRH-UE clusters in descending order by the number of UE in each cluster, then assign these RRH-UE clusters to each BBU in order. Every time we check the clusters in the list, then we will try to put it into the most full BBU where it fits, or open a new BBU to serve it when no existed BBU in working model has enough spare processing ability, until all the clusters are assigned to BBUs. The BFD-based heuristic BBU scheduling algorithm is elaborated in Algorithm 3. The complexity of solving Algorithm 3 is same with BFD bin packing solution, which is $O(P \log P)$ and P represents the total number of RRH-UE clusters here.

In summary, we first get the optimum data transmission rate and VM processing rate of each UE through the optimum network-wide beamforming vectors obtained by solving problem (P1). Based on beamforming vectors, the RRH-UE clusters can be obtained through Algorithm 2. Then, we can get the scheduling results between RRH-UE clusters and BBUs through Algorithm 3, which is the final resource allocation result of problem (P0).

V. Simulation Results

In this section, we evaluate the performance of our proposed joint resource allocation scheme through numerical simulation results.

We consider a heterogeneous Cloud-RAN with seven macro RRHs formulate seven macro cell as shown in Fig. 4. Each macro cell has three pico RRHs to provide hot-spot coverage. The network simulation parameters mainly reference to [45] are listed in Table I. The UEs are randomly located in the whole area, and we simulate our scheme in different numbers of total UEs. We also assume the UEs are homogeneous, which means different UE has the same QoS requirement c^{req}.

To evaluate the BBU scheduling algorithm in this paper, we choose several existed bin-packing-based BBU scheduling algorithms as comparison, which are listed below.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
System Bandwidth B	10MHz
Macrocell Inter-site Distance	500m
Distance-dependent Path Loss From Macro to User	$128.1 + 37.6 * \lg(R)$, R in kilometers
Distance-dependent Path Loss From Pico to User	$140.7 + 36.7 * \lg(R)$, R in kilometers
Multipath Fading	3GPP TU channel
Log-normal Shadowing	-8dB
Penetration Loss	-20dB
Maximum macro RRH Transmit Power	46dBm
Maximum pico RRH Transmit Power	30dBm
Noise Power Density	-174dBm/Hz
α	50
β	0.6
κ	1

1) Optimum Algorithm: The optimum bin packing-based BBU scheduling algorithm needs to traverse all feasible solutions, and then chooses the solution with minimum number of BBUs in working model. Since bin packing problem is a typical NP-hard problem, the complexity of optimum algorithm is $O(2^P)$.

2) Next-Fit-Based Algorithm: The NF-based BBU scheduling algorithm is proposed in [29]. Different from our proposed algorithm, the RRH-UE clusters do not need to be sorted before the assignment. And the RRH-UE clusters are just assigned to the next feasible BBUs in the list. The complexity of NF-based BBU scheduling algorithm is $O(P \log P)$.

3) First-Fit-Decreasing-Based Algorithm: The FFD-based BBU scheduling algorithm is proposed in [29]. It also needs to sort the RRH-UE clusters before assignment, but for each RRH-UE cluster in the list, it just chooses the first feasible BBU instead of the best feasible one comparing to our algorithm in this paper. The complexity of this algorithm is also $O(P \log P)$.

4) Max–Min Algorithm: The max–min algorithm is proposed in [28]. Compared to the previous several BBU scheduling algorithms, it lowers the complexity with the loss of performance. In each iteration, the algorithm tries to assign the cluster with maximum number of UEs and the cluster with minimum number of UEs to one BBU. The complexity of this algorithm is $O(P)$.

We first evaluate the performance of our BBU scheduling scheme. Fig. 5 shows the number of BBU in working model under different traffic loads of the whole network. When the number of UE in the area is small, we can see that different algorithms have the same performance. The reason is that one BBU can support all UEs when the number of UEs is equivalent to or smaller than the processing limit of a single BBU. As the number of UEs grows, the performance difference among different BBU scheduling algorithms will become more obvious. It is clear that except the optimum algorithm, our proposed BFD-based BBU scheduling algorithm outperforms others, for the reason that it chooses the best fit BBU for each cluster with least capacity loss.

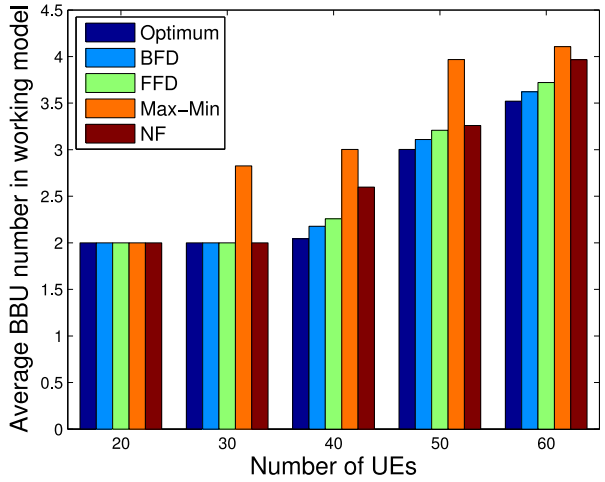


Fig. 5. BBU number in working model under different UE number.

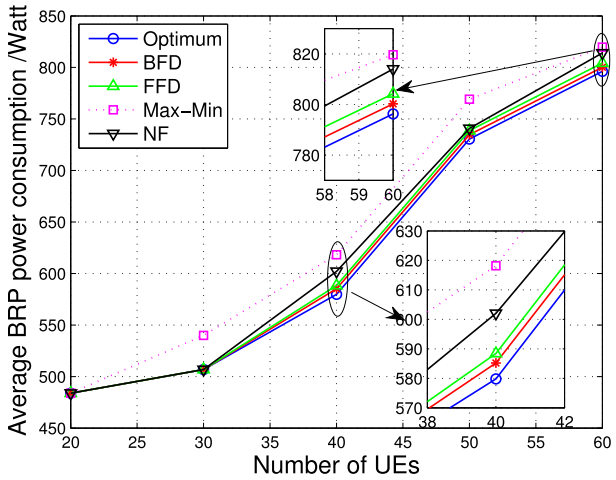


Fig. 6. BRP power consumption under different UE number.

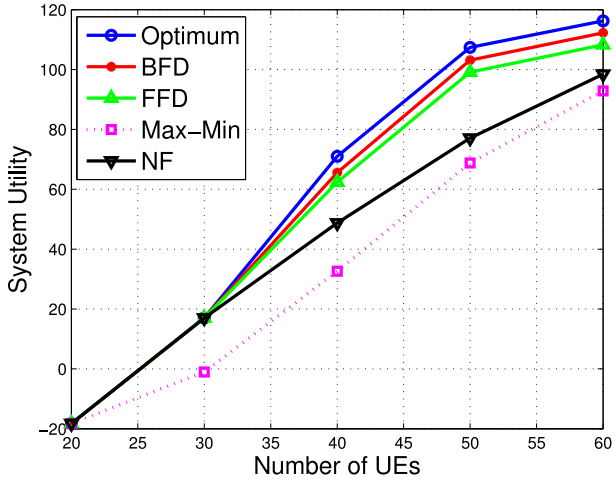


Fig. 7. System utility under different UE number.

Fig. 6 illustrates the energy consumption of BRP under different UE numbers. According to the BBU power model introduced in Section II, the power consumption of BRP is

mainly consisted of the power consumption of BBUs in working model and sleeping model. We set the QoS requirement of each UE as $c^{\text{req}} = 10$ Mb/s, and simulate the performance of each BBU scheduling algorithm. Based on the network-wide beamforming vectors approached by Algorithm 1, we can get the power consumption of each VM provided by BBUs in working model, and then obtain the BRP power consumption. From Fig. 6, we can see that our proposed BFD-based BBU scheduling algorithm outperforms all other heuristic algorithms, and the performance of the algorithm is closest to the optimum result.

As two simple basic bin-packing algorithms with lower complexity, the NF algorithm and the max–min algorithm do not sort the elements before packing, but with loss of performance as Fig. 6 shows. Compared to a better algorithm FFD, BFD also sorts the elements in decreasing size before packing, but it puts each element in the fullest bin which it fits, rather than the first fit one. So it has slightly better performance than all the other algorithms except the optimum one, which means it requires the least number of bins to pack all the numbers. So based on BFD, our proposed BBU scheduling scheme in this paper requires the least number of BBUs in working model. Furthermore, it can fulfill all the UEs' requirements to get the best performance with the lowest energy consumption as shown in Fig. 6.

Fig. 7 illustrates the system utility under different UE numbers. We set the QoS requirement of each UE as $c^{\text{req}} = 10$ Mb/s. As the number of UE grows in the network, the system utility increases, for the reason that more UEs bring more throughput, while the energy consumption does not increase much. Clearly that different BBU scheduling algorithms may result in different system utility performances. From the simulation results, Compared to other BBU scheduling algorithms, we can see that besides the optimum algorithm, our proposed BFD-based algorithm has the best performance comparing to other BBU scheduling algorithms.

VI. Conclusion

We jointly consider the RRH antenna resource and BBU computation resource and proposed an energy-efficient resource allocation scheme based on heterogeneous Cloud-RAN in this paper. We decompose our problem into two subproblems. The first subproblem is a network-wide beamforming vectors optimization, and we use a WMMSE approach to solve it. Based on the optimized beamforming vector, we propose an algorithm to get the RRH-UE clusters. The second subproblem is the BBU scheduling problem, and we reformulate it as a bin packing problem which means to minimize the number of working BBUs to save more energy. Compared to some existed works which form the BBU scheduling problem as a bin packing problem, we propose a bin packing algorithm based on the BFD method, which has better performance. With the detailed theoretical analysis and simulation results, the performance of our proposed BBU scheduling algorithm is better than other heuristic algorithms, which will lead to a higher system energy efficiency.

References

[1] Y. Wang, H. Ji, and H. Zhang, "Spectrum-efficiency enhancement in small cell networks with biasing cell association and eICIC: An analytical framework," *Int. J. Commun. Syst.*, vol. 29, no. 2, pp. 362–377, Jan. 2016.

[2] Y. Xu, S. Mao, and X. Su, "Interference alignment improves the capacity of OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 756–767, Feb. 2016.

[3] G. P. Fettweis, "A 5G wireless communications vision," *Microw. J.*, vol. 55, no. 12, pp. 24–36, Dec. 2012.

[4] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[5] R. Zhang, M. Wang, X. Shen, and L.-L. Xie, "Probabilistic analysis on QoS provisioning for Internet of Things in LTE-A heterogeneous networks with partial spectrum usage," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 354–365, Jun. 2016.

[6] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[7] D. Wubben *et al.*, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[8] M. X. Gong, R. J. Stacey, D. Akhmetov, and S. Mao, "A directional CSMA/CA protocol for mmWave wireless PANs," in *Proc. IEEE WCNC*, Sydney, NSW, Australia, Apr. 2010, pp. 1–6.

[9] M. S. Dastgahian and H. Khoshbin, "Rank-defective millimeter-wave channel estimation based on subspace-compressive sensing," *Elsevier Digit. Commun. Netw.*, vol. 2, no. 4, pp. 206–217, Nov. 2016.

[10] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access J.*, vol. 2, no. 1, pp. 947–959, Sep. 2014.

[11] M. Khumalo, W.-T. Shi, and C.-K. Wen, "Fixed-point implementation of approximate message passing (AMP) algorithm in massive MIMO systems," *Elsevier Digit. Commun. Netw.*, vol. 2, no. 4, pp. 218–224, Nov. 2016.

[12] L.-W. Chen, Y.-F. Ho, W.-T. Kuo, and M.-F. Tsai, "Intelligent file transfer for smart handheld devices based on mobile cloud computing," *Int. J. Commun. Syst.*, vol. 30, no. 1, pp. 1–12, Feb. 2015.

[13] J. Park, H. Kim, Y.-S. Jeong, and E. Lee, "Two-phase grouping-based resource management for big data processing in mobile cloud computing," *Int. J. Commun. Syst.*, vol. 27, no. 6, pp. 839–851, Jun. 2014.

[14] Y.-S. Jeong, J. S. Park, and J. H. Park, "An efficient authentication system of smart device using multi factors in mobile cloud service architecture," *Int. J. Commun. Syst.*, vol. 28, no. 4, pp. 659–674, Mar. 2015.

[15] "C-RAN: The road towards green RAN," China Mobile Res. Inst., Beijing, China, White Paper, ver. 2.5, Oct. 2011.

[16] M. W. Baidas and M. M. Afghah, "Energy-efficient partner selection in cooperative wireless networks: A matching-theoretic approach," *Int. J. Commun. Syst.*, vol. 29, no. 8, pp. 1451–1470, May 2016.

[17] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1171–1181, Dec. 2016.

[18] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for D2D communications underlaying cloud-RAN-based LTE-A networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 428–438, Jun. 2016.

[19] D. Zhang, Z. Zhou, S. Mumtaz, J. Rodriguez, and T. Sato, "One integrated energy efficiency proposal for 5G IoT communications," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1346–1354, Dec. 2016.

[20] 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.

[21] Y. He, E. Dutkiewicz, G. Fang, and M. D. Mueck, "Fractional frequency reuse in distributed antenna systems in cloud radio access networks," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, U.K., 2015, pp. 907–912.

[22] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[23] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

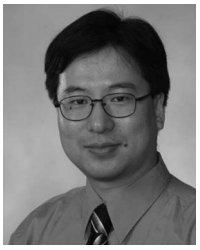
[24] X. Huang, G. Xue, R. Yu, and S. Leng, "Joint scheduling and beamforming coordination in cloud radio access networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5449–5460, Jul. 2016.
[25] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, Jan. 2009.
[26] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.
[27] K. Wang, M. Zhao, and W. Zhou, "Traffic-aware graph-based dynamic frequency reuse for heterogeneous Cloud-RAN," in *Proc. IEEE Glob. Commun. Conf.*, Austin, TX, USA, 2014, pp. 2308–2313.
[28] K. Boulos, M. El Helou, and S. Lahoud, "RRH clustering in cloud radio access networks," in *Proc. Int. Conf. Appl. Res. Comput. Sci. Eng. (ICAR)*, Beirut, Lebanon, 2015, pp. 1–6.
[29] T. Sigwele, A. S. Alam, P. Pillai, and Y. F. Hu, "Evaluating energy-efficient cloud radio access networks for 5G," in *Proc. IEEE Int. Conf. Data Sci. Data Intensive Syst.*, Sydney, NSW, Australia, 2015, pp. 362–367.
[30] 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.
[31] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-RAN architecture for future cellular network," in *Proc. Future Netw. Mobile Summit (FutureNetw)*, Berlin, Germany, 2012, pp. 1–8.
[32] R. Urgaonkar, U. C. Kozat, K. Igarashi, and M. J. Neely, "Dynamic resource allocation and power management in virtualized data centers," in *Proc. IEEE Netw. Oper. Manag. Symp. (NOMS)*, Osaka, Japan, 2010, pp. 479–486.
[33] W. Ni and I. B. Collings, "A new adaptive small-cell architecture," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 829–839, May 2013.
[34] V. N. Ha, L. B. Le, and N. D. ðào, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Istanbul, Turkey, 2014, pp. 1862–1867.
[35] D. Gesbert *et al.*, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.
[36] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.
[37] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, "Reducing energy consumption by dynamic resource allocation in C-RAN," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Paris, France, 2015, pp. 169–174.
[38] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-delay tradeoffs of virtual base stations with a computational-resource-aware energy consumption model," in *Proc. IEEE Int. Conf. Commun. Syst.*, Macau, China, 2014, pp. 26–30.
[39] Z. Zhou, M. Dong, K. Ota, J. Wu, and T. Sato, "Energy efficiency and spectral efficiency tradeoff in device-to-device (D2D) communications," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 485–488, Oct. 2014.
[40] C. He, B. Sheng, P. Zhu, X. You, and G. Y. Li, "Energy and spectral-efficiency tradeoff for distributed antenna systems with proportional fairness," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 894–902, May 2013.
[41] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[42] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.
[43] M. Grant and S. Boyd. (Sep. 2013). *CVX: Matlab Software for Disciplined Convex Programming, Version 2.0 Beta*. [Online]. Available: http://cvxr.com/cvx
[44] Y. Ye, *Interior Point Algorithms: Theory and Analysis*. New York, NY, USA: Wiley, 1997.
[45] 3GPP, "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRAN physical layer aspects (Release 9)," 3GPP TR 36.814, v9.0.0, Mar. 2010.

Kaiwei Wang received the B.S. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2012, where he is currently pursuing the Ph.D. degree in electronic engineering at the Department of Electronic Engineering and Information Science, School of Information Science and Technology.

Since 2015, he has been a visiting student with Prof. Shiwen Maos' group at Auburn University, Auburn, AL, USA. His current research interests include heterogeneous and cellular networks, cloud radio access network, green communications, and cooperative communications.

Wuyang Zhou (M'06) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2000.

He is currently a Professor of wireless communication networks with the Department of Electronic Engineering and Information Science, USTC. He participated in the National 863 Research Project Beyond Third Generation of Mobile System in China (FUTURE Plan) and has been a Task Director for many projects, including Innovative Wireless Campus Experimental Networks Research on High Frequency Networking Technologies and Research on Transmission and Networking Technologies in Satellite Mobile Communications. His current research interests include green technologies for communication systems, satellite mobile communications, and underwater acoustic communications.



Shiwen Mao (S'99–M'04–SM'09) received the Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA.

He is currently the Samuel Ginn Distinguished Professor with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. His current research interests include wireless networks and multimedia communications.

Dr. Mao was a recipient of the 2015 IEEE ComSoC TC-CSR Distinguished Service Award, the 2013 IEEE ComSoc MMTC Outstanding Leadership Award, and the NSF CAREER Award in 2010. He was a co-recipient of Best Paper Awards of IEEE GLOBECOM 2016, IEEE GLOBECOM 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the field of communications systems. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is on the Editorial Board of the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE INTERNET OF THINGS JOURNAL, IEEE MULTIMEDIA, among others. He is the Chair of the IEEE ComSoc Multimedia Communications Technical Committee.