QoS-aware Joint BBU-RRH Mapping and User Association in Cloud-RANs

Citation:
J. Yao and N. Ansari, "QoS-aware Joint BBU-RRH Mapping and User Association in Cloud-RANs," in *IEEE Transactions on Green Communications and Networking*, DOI: 10.1109/TGCN.2018.2837867, early access.

URL:
https://ieeexplore.ieee.org/document/8360457/

QoS-aware Joint BBU-RRH Mapping and User Association in Cloud-RANs

Jingjing Yao, *Student Member, IEEE*, and Nirwan Ansari, *Fellow, IEEE*

***Abstract*—Cloud radio access network (C-RAN) is a promising wireless network architecture that can reduce the energy consumption by the centralized cloud architecture and subsequently decrease the number of required traditional base station (BS) sites and the site support equipments. C-RAN consists of the baseband units (BBUs) and the remote radio heads (RRHs). BBUs are pooled in a central cloud, i.e., the BBU pool to provide powerful computation and storage resources while RRHs are distributed across multiple sites to provide coverage and interact with user equipments (UEs). In order to exploit the benefits of C-RAN, each BBU can be actualized by a virtual machine (VM), i.e., virtual BBU (VB). VBs can be initiated and shut down as needed to serve clusters of RRHs (i.e., many-to-one mapping between RRHs and BBUs). RRHs can be turned into the sleep mode to reduce the energy consumption. In our work, we jointly optimize BBU-RRH mapping and user association with the objective to minimize the system cost incurred by the energy bill from RRHs and VB rentals under the constraint of user quality of service (QoS), which is formulated as an integer linear programming (ILP) problem. Furthermore, we decompose the joint problem into two subproblems and design a time-efficient algorithm to solve the problem. Simulation results demonstrate that our proposed algorithm performs close to the optimal solutions obtained from CPLEX.**

***Index Terms*—Cloud radio access network (Cloud-RAN), BBU-RRH mapping, user association, power consumption.**

I. Introduction

Nowadays, mobile data have grown exponentially owing to the advances of wireless technologies and the proliferation of mobile devices like smart phones, laptops, wearable devices, and Internet of Things (IoT) devices. Bandwidth hungry wireless Internet applications, such as video conferencing, video streaming and online games have generated drastically increased mobile communication demands. Cisco Systems predicts that the global mobile data traffic will increase sevenfold between 2016 and 2021, reaching 49.0 exabytes per month by 2021 [1]. The explosive increase in mobile traffic leads to the increasing number of base stations (BSs), thus incurring significantly higher power consumption as well as costly capital and operating expenditure. Furthermore, due to the non-uniform nature of mobile traffic, many BSs are under utilization during non-peak hours. Therefore, a novel and intelligent wireless network architecture is called for.

Cloud radio access network (C-RAN) has been proposed as a prospective architecture to satisfy the fast growing mobile traffic [2]. A typical C-RAN architecture (Fig. 1) consists of three major parts: remote radio heads (RRHs), fronthaul links and baseband units (BBUs). BBUs are aggregated in a BBU pool to provide powerful computing capacities for baseband processing. RRHs are distributed across multiple sites to provide basic signal transmission and reception functionalities. RRHs are connected to the BBU pool through fronthaul links. By utilizing cloud computing technology and virtualization, the BBU pool can reduce the power consumption and improve hardware utilization. Specifically, several virtual machines (VMs) are turned on and off according to the traffic load.

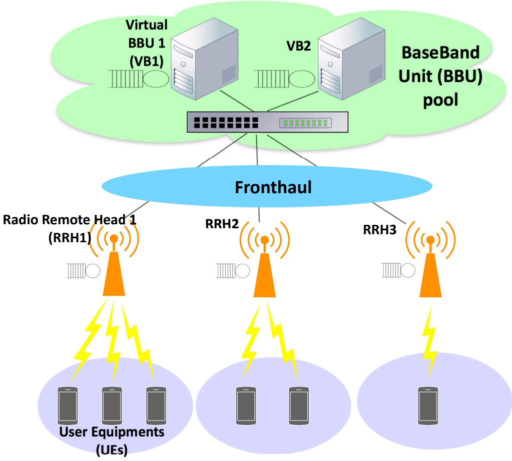


Fig. 1. QoS-aware C-RAN architecture

IoT is forming a pervasive network by connecting various kinds of devices which can communicate with Internet for various services or exchange information with other devices. The rapid development of IoT has driven an enormous amount of traffic with different qualify of service (QoS) [3]. However, the traditional mobile network cannot accommodate such diverse mobile services and fluctuating traffic patterns efficiently [4]. C-RAN, by utilizing the cloud computing technology to enable network flexibility, is considered as a promising architecture to address this challenge [5]. Specifically, all UEs in IoT have easier access to the mobile core network due to the densely deployed RRHs and various traffic with different QoS requirements will be processed in the centralized BBU pool.

Explosive mobile data demands are driving a significant growth in energy consumption in mobile networks. From the network operators' perspective, reducing energy consumption can potentially reduce a great amount of expenditures on their energy bills. C-RAN helps decrease the energy consumption by reducing the cooling infrastructure due to the deployment of lightweight RRHs and sharing of computing resources in the BBU pool. However, energy consumption is still consumed by densely deployed RRHs and for processing huge large data traffic in the BBU pool. Therefore, several RRHs could be turned into the sleep mode to save energy. Furthermore, the virtual machines, functioning as the virtual BBUs (VBs), can also be turned off to save the system cost when there are no traffic demands [6].

Selecting the serving RRH for each user, referred to as the user association problem, plays a pivotal role in enhancing the load balancing and energy efficiency of wireless networks. Unfortunately, it is a combinatorial problem of high complexity [7]. The BBU-RRH mapping problem addresses the many-to-one mapping between RRHs and BBUs. It is also known as the RRH clustering problem because one BBU can serve a cluster of RRHs. Decisions for each VB to serve which RRH is important for resource scheduling and allocation. Furthermore, since data traffic goes through both RRHs and BBUs before being served, any long delays caused in either RRHs or BBUs may degrade the user QoS. Hence, it is critical to jointly consider both the user association problem and BBU-RRH mapping problem.

The above facts motivate us to study the QoS-aware joint BBU-RRH mapping and user association problem in C-RAN with the objective to minimize the system cost of both RRHs and the BBU pool. The rest of this paper is organized as follows. We present related works in Section II. The problem formulation is described in Section IV. The problem analysis is discussed in Section V. Simulation results are given in Section VI. The paper is concluded in Section VII.

II. Related Works

C-RAN is a promising paradigm to reduce both capital and operating expenditures as well as to provide high spectral efficiency (SE) and energy efficiency (EE) [8], and has hence been advocated by both the industry and research community. According to [8], C-RAN reduces the power consumption by 41% and achieves 20-50% throughput gain as compared with traditional cellular networks. Energy efficiency plays an important role in the performance of C-RAN [2], [9]. In order to save energy, UEs should be well scheduled to be served by their optimal serving RRHs and RRHs should also be appropriately assigned to respective BBUs [10]. Opadere *et al.* [11] utilized C-RAN virtualization to enable inter-operator traffic offloading and explored the energy saving potential of the sleep mode scheme. Guo *et al.* [12] investigated a joint RRH-BBU association and energy sharing problem to minimize the brown energy usage in green energy powered C-RAN. Zeng *et al.* [13] proposed an energy-efficient Re-CRAN architecture which incorporates distributed renewable energy resources into C-RAN and verified the advantages of their architecture by investigating a renewable-energy-aware RRH activation problem.

The EE-based user association problem can be formulated into two forms including minimizing the power consumption and minimizing the overall energy efficiency (e.g., the ratio of the sum rate to the total energy consumption) [7]. Most of the existing works focus on the former one. A common algorithm is the nearest RRH association scheme, where the nearest RRH is chosen to serve each UE to minimize the RRHs' power consumption [14]. Zuo *et al.* [15] considered the EE-based user association problem in massive MIMO empowered C-RAN. They proposed three algorithms including nearest-RRH based user association, single-candidate RRH user association, and multi-candidate RRHs user association to solve this problem. Wang and Sun [16] proposed an effective user association strategy under the constraints of power budget, and solved this problem by an approximation algorithm. Han and Ansari [17] proposed the network utility aware traffic load balancing scheme to adapt the user association and investigated the tradeoff between the brown power consumption and the traffic delivery latency.

From the perspective of the BBU-RRH mapping problem, also known as the RRH clustering problem, several works [18]–[20] have formulated it as a bin packing problem with the objective to reduce the number of active BBUs. In [18], UE's baseband tasks are first assimilated to objects of different volumes and then packed into bins (e.g., BBUs). Boulos *et al.* [19] minimized the network power consumption by reducing the number of active BBUs and minimized the handover frequency by clustering neighboring RRHs. Some known heuristics, such as the first fit decreasing and the net fit decreasing, are designed to find the acceptable solutions for the BBU-RRH mapping problem. Qian *et al.* [20] proposed a heuristic simulated annealing method by combining the bin packing algorithm with a simulated annealing algorithm. They utilized a two-layer algorithm which first maps one or many RRHs to a single BBU and then maps each unmapped RRH to another BBU with additional power consumption.

Although the above works attempt to minimize the energy consumption in C-RAN by optimizing either the user association or BBU-RRH mapping, they do not take user QoS into consideration. In order to characterize user QoS, Tang *et al.* [6] proposed a queueing system to study the joint VM activation and sparse beamforming problem in C-RAN. However, they assumed that data traffic is homogenous, which is not practical in the IoT environment. They also assumed that all traffic from RRHs is distributed evenly among different BBUs, which may result in underutilization of several BBUs. Soliman *et al.* [21] investigated the joint RRH clustering and RRH activation problem under the QoS constraint with the objective to minimize the RRHs' power consumption. However, their work does not consider the BBU-RRH mapping problem and their QoS model is only related to the signal to interference and noise ratio (SINR) without the delays in the BBU pool. Khan *et al.* [22] formulated the QoS as a weighted combination of the number of blocked users and handovers in C-RAN and provided load balancing solutions.

Different from the above works, we consider the QoS-aware joint BBU-RRH mapping and user association problem in C-RAN where our user QoS requirement is modeled as delays of two queues in tandem, including the BBU processing queue and the RRH transmission queue with heterogenous data traffic. Our objective is to minimize the system cost caused by both RRHs and the BBU pool. In order to reduce the system cost, RRHs can be turned into the sleep mode and VBs can be turned off in order to save energy when there is no traffic.

III. System Model

A. System Description

In our proposed QoS-aware C-RAN architecture, we assume there are N UEs and R RRHs and each UE is served by one RRH. The Boolean variable x_{ij} indicates whether UE i is served by RRH j. BBUs are actualized by VMs. As VBs can be initialized and shut down according to the traffic load, the number of VBs varies. However, the maximum number of VBs is equivalent to the number of RRH R. Hence, we assume there are R VBs among which several turned-off VBs do not contribute any cost to the system. Each VB has a fixed computing capacity C_k. We define the Boolean variable y_{jk} to denote whether RRH j is mapped to VB k. We use $I = \{1, 2, ..., N\}, J = \{1, 2, ..., R\}$, and $K = \{1, 2, ..., R\}$ to denote the sets of UEs, RRHs and VBs respectively. We summarize all notations in Table I.

Table I. Summary of notations.

Notation	Definition
N	Number of UEs.
R	Number of RRHs.
B	Number of BBUs
C_k	Computing capacity of BBU k.
r_{ij}	Wireless data rate between UE i and RRH j.
p_j^{sleep}	Power consumption of RRH j in sleep mode.
p_j^{static}	Static power consumption of RRH j.
β	Load-power coefficient which reflects the relationship between the traffic and dynamic power consumption of RRHs.
ρ_j	Traffic load of RRH j.
$\tilde{\rho}_k$	Traffic load of VB k.
η	Electricity cost per milliwatt.
f	Fixed cost of renting a VB.
l_i	Average packet length of UE i.
λ_i	Request arrival rate of UE i.
Q_{th}	QoS requirement of RRHs.
$\tilde{Q}_{th}$	QoS requirement of VBs.
x_{ij}	Boolean variable which equals 1 when UE i is served by RRH j; otherwise, 0.
y_{jk}	Boolean variable which equals 1 when RRH j is associated with VB k; otherwise, 0.
t_j	Boolean variable which equals 1 when RRH j is active; otherwise, 0.
$\tilde{t}_k$	Boolean variable which equals 1 when VB k is active; otherwise, 0.

In the downlink, a UE's traffic is first processed by the VB which is mapped to its serving RRH and routed to the serving RRH through fronthaul links. We assume that the links between the BBU pool and RRHs are high-bandwidth, low-latency optimal fiber links with negligible transmission delay. Then, the serving RRH transmits the traffic to the UE via the wireless network. We also assume that the BBU pool can schedule the spectral resources well and hence the interferences between UEs can be neglected.

B. System Cost Model

The system cost is attributed from both RRHs and the BBU pool. We characterize the cost from RRHs as the electricity bill and hence the power consumption of RRHs should be reduced. RRHs can be selectively turned into the sleep mode in which case the power consumption of RRH j is p_j^{sleep}. When RRH j is active, its power consumption includes the static power consumption and the dynamic power consumption [23]. The static power consumption p_j^{static} is generated without carrying any traffic load. The dynamic power consumption is incurred by traffic load and hence can be denoted as a linear function of the traffic load ρ_j. We denote β as the load-power coefficient that reflects the relationship between RRH j's traffic load and its dynamic power consumption. The power consumption of each RRH can be expressed as

$$p_j^{RRH} = \begin{cases} \beta\rho_j + p_j^{static}, & \text{if RRH } j \text{ is active} \\ p_j^{sleep}, & \text{otherwise.} \end{cases}$$

We assume all VBs have the same fixed cost, and so to reduce the cost is to reduce the number of VBs. This is the popular commercial cloud service model, e.g., Amazon Elastic Compute Cloud (EC2) [24]. In EC2, there are several VMs for rent. The mobile operator decides how many VMs they need to rent for a period of time. After this period, the mobile operators decide the number of VMs once again to adapt to the dynamic traffic. Within one period, we denote η as the electricity cost per milliwatt and f as the fixed cost for renting a VM. Hence, the total system cost can be described as $\eta P^{RRH} + fN^{VB}$, where P^{RRH} and N^{VB} denote the power consumption of RRHs and the number of VBs, respectively.

C. QoS Model

We model our downlink traffic demand and QoS model based on queuing models. As data of a UE are first processed in the VB and then transmitted by the RRH, we consider a two-layer queueing network to represent each UE's traffic processing and transmission in the downlink, including the VB processing queue and RRH transmission queue. Throughout the paper, we assume that each queue behaves in a first in first out (FIFO) manner.

1) VB Processing Queue: We assume that the traffic arrival of each UE i follows the Poisson process with the arrival rate λ_i and its packet size per arrival follows the exponential distribution with the average value of l_i. We also assume that the traffic arrivals of different UEs under a certain RRH are independent with each other. Hence, the traffic arrivals toward one VB is still a Poisson process. Furthermore, the computation capacity C_k of each VB k is considered constant, and so the average service time, which equals the packet size divided by the processing rate, is also an exponential distribution. Therefore, the traffic processing in each VB realizes an M/M/1 queuing model. The traffic load in each VB k can be expressed as

$$\tilde{\rho}_k = \sum_{j=1}^{R} \sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij} y_{jk}}{C_k}, \quad \forall k \in K,$$

where x_{ij} and y_{jk} are the user association and BBU-RRH mapping indicators, respectively. According to the properties of the M/M/1 queue [25], the average traffic delivery time in each VB k can be calculated as

$$\tilde{\tau}_{ik} = \frac{l_i}{C_k(1 - \tilde{\rho}_k)}.$$

Then, the waiting time for each UE is

$$\tilde{W}_{ik} = \tilde{\tau}_{ik} - \frac{l_i}{C_k} = \frac{l_i \tilde{\rho}_k}{C_k(1-\tilde{\rho}_k)}.$$

Since UE's traffic loads are diverse, different users may require different amounts of service time. We utilize the average waiting time per unit service time to reflect a queue's performance [25]. Specifically, we denote the latency ratio $\frac{\tilde{W}_{ik}}{\tilde{\tau}_{ik}} = \frac{\tilde{\rho}_k}{1-\tilde{\rho}_k}$ to reflect the QoS performance. Note that a larger latency ratio implies a longer waiting time, and thus leads to a worse QoS. We denote $\tilde{Q}_{th}$ as the threshold of the latency ratio that each VB cannot exceed. Hence, for each VB k, the QoS requirement is

$$\frac{\tilde{\rho}_k}{1-\tilde{\rho}_k} \le \tilde{Q}_{th}.$$

2) RRH Transmission Queue: According to Burke's Theorem, the traffic departure process of an M/M/1 queue is still a Poisson process with average departure rate equivalent to the average traffic arrival rate [25]. Hence, for each RRH transmission queue, the traffic arrival follows the Poisson process. Since the users transmission rate r_{ij} is generally distributed, the service time, which equals $\frac{l_i}{r_{ij}}$, follows a general distribution. Therefore, the RRH transmission queue realizes an M/G/1 processor sharing queue where multiple UEs share the RRH's downlink service. Then, the traffic load in RRH j can be calculated as

$$\rho_j = \sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}}.$$

Although there are various downlink scheduling algorithms to enable sharing of the limited radio resource, we adopt the round robin (RR) scheduling. For the M/G/1-RR queue [25] , the average traffic delivery time for UE i in RRH j is

$$\tau_{ij} = \frac{l_i}{r_{ij}(1-\rho_j)},$$

and the waiting time for each UE is

$$W_{ij} = \tau_{ij} - \frac{l_i}{r_{ij}} = \frac{l_i \rho_j}{r_{ij}(1-\rho_j)}.$$

Similar to the BBU processing queue, we denote the latency ratio for each RRH j as $\frac{W_{ij}}{\tau_{ij}} = \frac{\rho_j}{1-\rho_j}$ to reflect the QoS performance. Hence, for each RRH, the QoS requirement is

$$\frac{\rho_j}{1-\rho_j} \le Q_{th}.$$

IV. Problem Formulation

In our work, we jointly consider the QoS-aware joint BBU-RRH mapping and user association in C-RAN with the objective to minimize the system cost of both RRHs and the BBU pool. Our joint optimization problem is formulated as follows.

$$\textbf{P0:} \min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{t},\tilde{\boldsymbol{t}}} \eta[\sum_{j=1}^{R}(\beta\rho_j + p_j^{static})t_j + \sum_{j=1}^{R} p_j^{sleep}(1-t_j)] + f\sum_{k=1}^{R}\tilde{t}_k \quad (1)$$

s.t.

$$\sum_{j=1}^{R} x_{ij} = 1, \quad \forall i \in I, \quad (2)$$

$$\sum_{k=1}^{R} y_{jk} = 1, \quad \forall j \in J, \quad (3)$$

$$x_{ij} \le t_j, \quad \forall i \in I, j \in J, \quad (4)$$

$$y_{jk} \le \tilde{t}_k, \quad \forall j \in J, k \in K, \quad (5)$$

$$\rho_j = \sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}}, \quad \forall j \in J, \quad (6)$$

$$\tilde{\rho}_k = \sum_{j=1}^{R}\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij} y_{jk}}{C_k}, \quad \forall k \in K, \quad (7)$$

$$0 \le \rho_j < 1, \quad \forall j \in J, \quad (8)$$

$$0 \le \tilde{\rho}_k < 1, \quad \forall k \in K, \quad (9)$$

$$\frac{\rho_j}{1-\rho_j} \le Q_{th}, \quad \forall j \in J, \quad (10)$$

$$\frac{\tilde{\rho}_k}{1-\tilde{\rho}_k} \le \tilde{Q}_{th}, \quad \forall k \in K, \quad (11)$$

$$x_{ij} \in \{0,1\}, \quad \forall i \in I, j \in J, \quad (12)$$

$$y_{jk} \in \{0,1\}, \quad \forall j \in J, k \in K, \quad (13)$$

$$t_j \in \{0,1\}, \quad \forall j \in J, \quad (14)$$

$$\tilde{t}_k \in \{0,1\}, \quad \forall k \in K. \quad (15)$$

Eq. (2) indicates that each UE can only be associated with one RRH. Eq. (3) implies that each RRH is only mapped to one VB. In Eqs. (4) and (5), t_j and $\tilde{t}_k$ are Boolean variables to indicate whether RRH j is active ($t_j = 1$ if affirmative) and VB k is active ($\tilde{t}_k = 1$ if affirmative), respectively. Eqs. (4) and (5) ensure that only active RRHs and VBs can be connected. Eqs. (6) and (7) compute ρ_j and $\tilde{\rho}_k$; Eqs. (8) and (9) are the constraints of the traffic loads of queues; Eqs. (10) and (11) are the QoS requirements for RRHs and VBs, respectively. Eqs. (12)-(15) indicate that x_{ij}, y_{jk}, t_j and $\tilde{t}_k$ are Boolean variables.

Note that Eqs. (10) and (11) can be transformed into $\rho_j \le \frac{Q_{th}}{1+Q_{th}}$ and $\tilde{\rho}_k \le \frac{\tilde{Q}_{th}}{1+\tilde{Q}_{th}}$, respectively. These two provide tighter bounds for ρ_j and $\tilde{\rho}_k$ than Eqs. (8) and (9), and so we can combine them. In Eq. (1), when $t_j = 0$, we must have $\rho_j = 0$. On the other hand, if $\rho_j > 0$, we have $t_j = 1$. Hence, we can deduce that $\rho_j t_j = \rho_j$. For ease of readability, we denote $\hat{p_j}^s = p_j^{static} - p_j^{sleep}$. In the actual wireless system, p_j^{static} is usually greater than p_j^{sleep}, and so we can assume $\hat{p_j}^s > 0$.

Unfortunately, this problem is non-linear due to the product $x_{ij}y_{jk}$ in Eq. (7). Therefore, we introduce another variable z_{ijk}, which is also a Boolean variable, and assign $z_{ijk} = x_{ij}y_{jk}$. To guarantee the transformed problem is equivalent to the original one, the following additional inequality constraints should be satisfied: 1) $z_{ijk} \leq x_{ij}$; 2) $z_{ijk} \leq y_{jk}$; 3) $z_{ijk} \geq x_{ij} + y_{jk} - 1$.

Our transformed formulation becomes:

$$\begin{aligned}\textbf{P1:}\quad \min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{t},\tilde{\boldsymbol{t}}} \quad & \eta \sum_{j=1}^{R}\sum_{i=1}^{N} \frac{\beta\lambda_i l_i x_{ij}}{r_{ij}} + \eta \sum_{j=1}^{R} \hat{p_j}^{s} t_j \\ & + \eta \sum_{j=1}^{R} p_j^{sleep} + f \sum_{k=1}^{R} \tilde{t}_k \end{aligned} \tag{16}$$

s.t. (2), (3), (4), (5), (12), (13), (14), (15),

$$z_{ijk} \leq x_{ij}, \quad \forall i \in I, j \in J, k \in K, \tag{17}$$

$$z_{ijk} \leq y_{jk}, \quad \forall i \in I, j \in J, k \in K, \tag{18}$$

$$z_{ijk} \geq x_{ij} + y_{jk} - 1, \quad \forall i \in I, j \in J, k \in K, \tag{19}$$

$$z_{ijk} \in \{0, 1\}, \quad \forall i \in I, j \in J, k \in K, \tag{20}$$

$$\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}} \leq \frac{Q_{th}}{1 + Q_{th}}, \quad \forall j \in J, \tag{21}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{R} \frac{\lambda_i l_i z_{ijk}}{C_k} \leq \frac{\tilde{Q}_{th}}{1 + \tilde{Q}_{th}}, \quad \forall k \in K. \tag{22}$$

This transformed problem **P1**, which is equivalent to problem **P0**, is an integer linear programming (ILP) problem. It can be addressed via exhaustive search and can also be solved with CPLEX by the branch-and-bound scheme. However, those two approaches are both computationally expensive (exponential). Hence, we design suboptimal algorithms to solve this joint problem and compare their performances with the optimal solutions obtained by CPLEX in Section VI.

V. Problem Analysis

We decompose this joint optimization problem into two subproblems including the user association problem and the BBU-RRH mapping problem. We try to solve the user association problem first and then utilize its optimal solutions to address the BBU-RRH mapping problem. The total system cost $P = \eta P_1 + f P_2$, where P_1 and P_2 are optimal values from the first problem (i.e., the minimum power consumption of all RRHs) and the second one (i.e., the minimum number of VBs), respectively. We next discuss these two subproblems.

A. User Association Problem

In the user association problem, UEs are scheduled to be associated with their optimal RRHs to minimize RRHs' power consumption with consideration of wireless channel conditions and QoS requirement. Hence, the user association problem can be formulated as

$$\begin{aligned}\textbf{P2:}\quad \min_{\boldsymbol{x}} \quad & \sum_{j=1}^{R}\sum_{i=1}^{N} \frac{\beta\lambda_i l_i x_{ij}}{r_{ij}} + \sum_{j=1}^{R} \hat{p_j}^{s} t_j + \sum_{j=1}^{R} p_j^{sleep} \\ s.t. \quad & (2), (4), (12), (14), (21). \end{aligned}$$

To solve problem **P2**, we design a Lagrangian relaxation algorithm where we relax Eq. (2) and Eq. (21), i.e., the constraint that guarantees that each UE is only served by one RRH and the QoS constraint for each RRH. The Lagrangian relaxation problem can be formulated as

$$\begin{aligned}\textbf{P3:}\quad \max_{\boldsymbol{u},\boldsymbol{v}} \min_{\boldsymbol{x},\boldsymbol{t}} \quad & \sum_{j=1}^{R}\sum_{i=1}^{N} \frac{\beta\lambda_i l_i x_{ij}}{r_{ij}} + \sum_{j=1}^{R} \hat{p_j}^{s} t_j + \sum_{j=1}^{R} p_j^{sleep} \\ & + \sum_{i=1}^{N} u_i (1 - \sum_{j=1}^{R} x_{ij}) \\ & + \sum_{j=1}^{R} v_j (\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}} - \frac{Q_{th}}{1 + Q_{th}}) \\ & = \sum_{j=1}^{R}\sum_{i=1}^{N} [\frac{(\beta + v_j)\lambda_i l_i}{r_{ij}} - u_i] x_{ij} + \sum_{j=1}^{R} \hat{p_j}^{s} t_j \\ & + \sum_{i=1}^{N} u_i - \frac{Q_{th}}{1 + Q_{th}} \sum_{j=1}^{R} v_j + \sum_{j=1}^{R} p_j^{sleep} \end{aligned} \tag{23}$$

$$s.t. \quad (4), (12), (14),$$

$$v_j \geq 0, \quad \forall j \in J, \tag{24}$$

where u_i and v_j are the Lagrangian multipliers. For fixed values of the Lagrangian multipliers, the above relaxed problem **P3** will yield an optimal objective value that provides the lower bound (LB) of the original user association problem (i.e., problem **P2**).

Lemma 1. *The solutions of problem **P3** with fixed **u** and **v** are*

$$x_{ij} = \begin{cases} 1, & if \ \frac{(\beta + v_j)\lambda_i l_i}{r_{ij}} - u_i < 0 \ \& \ t_j = 1, \\ 0, & otherwise. \end{cases}$$

$$t_j = \begin{cases} 1, & if \ \hat{p}_j + \sum_{i=1}^{N} \min\{0, \frac{(\beta + v_j)\lambda_i l_i}{r_{ij}} - u_i\} < 0, \\ 0, & otherwise. \end{cases}$$

Proof: For fixed multipliers $\boldsymbol{u}$ and $\boldsymbol{v}$, in order to minimize the objective function, it is preferable to set $x_{ij} = 1$, if its coefficient $\frac{(\beta + v_j)\lambda_i l_i}{r_{ij}} - u_i < 0$, and 0 otherwise. However, setting $x_{ij} = 1$ means that we must also set $t_j = 1$ under the constraint of Eq. (4) which stipulates $x_{ij} \leq t_j, \forall i, j$. If we set $t_j = 1$, the value we add to the objective function is $\Delta V = \hat{p}_j + \sum_{i=1}^{N} \min\{0, \frac{(\beta + v_j)\lambda_i l_i}{r_{ij}} - u_i\}$. If only ΔV is negative, the objective value is reduced. Hence, we set $t_j = 1$ when $\Delta V < 0$. Therefore, the lemma is proved. ■

We can obtain the optimal solutions for problem **P3** by Lemma 1. However, the solutions can only provide the LB and hence may not be feasible because Eq. (2) and Eq. (21) are relaxed, i.e., several UEs may be served by more than one RRH and some RRHs' QoS requirement may be violated. Hence, we need to find the feasible solution. The feasible solution acts as the upper bound (UB) for our original problem **P2** because the feasible solution cannot guarantee optimality, and so we can always find other feasible solutions with performances not worse than the UB.

In order to attain the UB, we utilize $\boldsymbol{t}$ obtained from Lemma 1 and then substitute it into problem **P2**. We can observe that Eq. (4) and Eq. (21) can be combined as $\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}} \leq \frac{Q_{th}}{1+Q_{th}} t_j$, because when $t_j = 0$, we have $x_{ij} = 0$. Contrarily, if $t_j = 1$, x_{ij} can be either 0 or 1; this equation imposes the QoS requirement. Then, problem **P2** can be transformed as

$$\textbf{P4:} \quad \min_{\boldsymbol{x}} \quad \sum_{j=1}^{R} \sum_{i=1}^{N} \frac{\beta \lambda_i l_i x_{ij}}{r_{ij}} + \sum_{j=1}^{R} \hat{p}_j{}^s t_j^* + \sum_{j=1}^{R} p_j^{sleep}$$

$$s.t. \quad (2), (12),$$

$$\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}}{r_{ij}} \leq \frac{Q_{th}}{1+Q_{th}} t_j^*, j \in J.$$

Problem **P4** has the similar form of the generalized assignment problem (GAP), which has been demonstrated to be an NP-hard problem [26]. We design a heuristic UB searching algorithm to solve this problem. In order to minimize the objective value, each UE prefers to connect with the RRH with the minimum $\frac{\beta \lambda_i l_i}{r_{ij}}$ if the QoS requirement is not considered. Our idea is to define the reassignment gain weight $\Delta w_i = \frac{\beta \lambda_i l_i}{r_{ij_{min2}}} - \frac{\beta \lambda_i l_i}{r_{ij_{min}}}$ for each UE, where j_{min} and j_{min2} indicates the RRHs with the minimum and second minimum value of $\frac{\beta \lambda_i l_i}{r_{ij}}$, respectively. The reassignment gain weight measures how much we can add to the objective value if we move the UE i from RRH j_{min} to RRH j_{min2}. We prefer to assign the UE with the maximum Δw_i to its RRH j_{min} to avoid the case that if we assign it to RRH j_{min2}, a large reassignment gain weight Δw_i will be added to the objective value. The specific process of the UB searching algorithm is shown in Alg. 2. Lines 4-23 iteratively assign UE i with maximum value Δw_i until all UEs find their RRHs. The Loop in lines 5-16 calculates Δw_i for each UE. Two special cases are considered in lines 7-12. If one UE cannot find any RRH to connect (e.g., $\boldsymbol{t}^* = 0$), this problem has no feasible solution. Hence, we set $P_{ub} = +\infty$. Another case is when a UE can only find one RRH, then the UE can only be assigned to the unique RRH.

Note that the original problem **P2** always chooses its UB as its objective value because the UB can guarantee the feasibility. Different values of Lagrange multipliers lead to different values of UBs and LBs. Thus, by applying the subgradient method [27], we adjust the values of u_i and v_j iteratively. The iteration terminates when the UB and LB are close to each other or reaching the maximum number of iterations. We denote P^{opt} as the optimal UB in previous iterations. In the n-th iteration, we denote P_{lb} and P_{ub} as the values of LB and UB, respectively. The values of u_i and v_j are calculated as follows.

$$u_i^{n+1} = u_i^n + \theta^n (1 - \sum_{j=1}^{R} x_{ij}^n), \forall i \in I, \tag{25}$$

$$v_j^{n+1} = \max\{0, v_j^n + \theta^n (\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}^n}{r_{ij}} - \frac{Q_{th}}{1+Q_{th}})\}, \forall j \in J, \tag{26}$$

Algorithm 1: Lagrangian Relaxation Algorithm

```
Input  : $R, N, \beta, \lambda_i, l_i, r_{ij}, \hat{p}_j^s, Q_{th}$
Output: user association matrix $\boldsymbol{x}$, RRH activation
        vector $\boldsymbol{t}$ and optimal value $P^{opt}$
1  Initialize the Lagrangian multipliers $u_i, v_j$;
2  Initialize $P_{lb} = 0, P_{ub} = +\infty, P^{opt} = +\infty, n = 1$;
3  while $P_{lb} \not\approx P_{ub}$ and $n < n_{max}$ do
4      Calculate the solution of problem P3, $\boldsymbol{x}^{lb}$ and
         $\boldsymbol{t}^{lb}$, by Lemma 1;
5      Calculate the LB $P_{lb}$ by Eq. (23);
6      $\boldsymbol{t}^{ub} = \boldsymbol{t}^{lb}$;
7      Calculate the UB $P_{ub}$ and its solution $\boldsymbol{x}^{ub}, \boldsymbol{t}^{ub}$ by
         Algorithm 2 ;
8      if $P_{ub} < P^{opt}$ then
9          $\boldsymbol{x} = \boldsymbol{x}^{ub}$;
10         $\boldsymbol{t} = \boldsymbol{t}^{ub}$;
11         $P^{opt} = P_{ub}$;
12     end
13     Update step size $\theta$ according to Eq. (27);
14     Update Lagrangian multipliers $u_j$ according to
         Eq. (25) and Eq. (26);
15     $n = n + 1$;
16 end
17 return $\boldsymbol{x}, \boldsymbol{t}, P^{opt}$;
```

where θ^n is the step size in the n-th iteration, which can be calculated as

$$\theta^n = \frac{\delta (P^{opt} - P_{lb}^n)}{\sum_{i=1}^{N} (1 - \sum_{j=1}^{R} x_{ij}^{lbn})^2 + \sum_{j=1}^{R} (\sum_{i=1}^{N} \frac{\lambda_i l_i x_{ij}^{lbn}}{r_{ij}} - \frac{Q_{th}}{1+Q_{th}})^2}, \tag{27}$$

where δ is the decreasing adaptation parameter. Usually, δ can be a constant or is set to 2 and then halved if P_{lb} does not change for several iterations. The Lagrangian relaxation algorithm is summarized in Algorithm 1. Lines 1-2 initialize the Lagrangian multipliers, LB and the optimal value. Lines 3-16 iteratively change the Lagrangian multipliers as well as update the LB and UB to find the optimal value. The iteration terminates when the LB is close to the UB or the maximum number of iterations n_{max} is reached.

B. BBU-RRH Mapping Problem

The BBU-RRH mapping problem determines which RRH is served by which VB to minimize the number of VBs, on the condition that the user association decisions have been decided. Hence, the BBU-RRH mapping problem can be formulated as

$$\textbf{P5:} \quad \min_{\boldsymbol{y}, \tilde{\boldsymbol{t}}} \quad \sum_{k=1}^{R} \tilde{t}_k$$

$$s.t. \quad (3), (13), (15),$$

$$\sum_{i=1}^{N} \sum_{j=1}^{R} \frac{\lambda_i l_i x_{ij}^* y_{jk}}{C_k} \leq \frac{\tilde{Q}_{th}}{1+\tilde{Q}_{th}} \tilde{t}_k, \quad \forall k \in K.$$

```
Algorithm 2: UB Searching Algorithm
    Input  : $R, N, \beta, \lambda_i, l_i, r_{ij}, \hat{p}_j^s, Q_{th}, t_j^*$
    Output: user association matrix $\boldsymbol{x}^{ub}$ and UB $P_{ub}$
    Residential capacity $R_j^{cap} = \frac{Qth}{1+Qth}, \forall j \in J$;
    Unassigned UE set $\phi = \{1, ..., N\}$;
    Boolean variable $feasible = 1$, $\boldsymbol{x}^{ub} = 0$;
    while $\phi \neq \emptyset$ do
        for each UE $\in \phi$ do
            Find the RRH set
              $\mathcal{A} = \{j | \frac{\lambda_i l_i}{r_{ij}} \leq R_j^{cap}, t_j = 1\}$;
            if $|\mathcal{A}| = 0$ then
                $feasible = 0$;
                break;
            else if $|\mathcal{A}| = 1$ then
                Assign UE $i$ to RRH $j \in \mathcal{A}$, $x_{ij}^{ub} = 1$;
                Update $R_j^{cap}$ and $\phi$;
            else
                Calculate $\Delta w_i$;
            end
        end
        if $feasible = 0$ then
            break;
        end
        Choose UE $i = \arg\max_i \{\Delta w_i\}$;
        Assign UE $i$ to RRH $j_{min}$, $x_{ij_{min}}^{ub} = 1$;
        Update $R_j^{cap}$ and $\phi$;
    end
    if $feasible = 0$ then
        return $\boldsymbol{x}_{ub} = 0, P_{ub} = +\infty$ ;
    end
    return $\boldsymbol{x}_{ub}, P_{ub}$ ;
```

Problem **P5** has the similar form of the bin packing problem [28], where VBs are "backpacks" and RRHs are "objects" to be put in the backpacks. We utilize the best-fit-decreasing algorithm [28] to solve the problem. We consider $\frac{\tilde{Q}_{th}}{1+\tilde{Q}_{th}}$ as the capacity of VB k if $\tilde{t}_k = 1$, while $\frac{\lambda_i l_i x_{ij}^*}{C_k}$ indicates the weight added to VB k by RRH j if they are connected. The idea is to connect RRH j to VB k (i.e., put object RRH j to backpack VB k) which will have the minimum remaining capacity after adding the object RRH (i.e., "best fit"). The weights of RRHs are sorted in descending order and RRHs with larger weights are handled preferentially (i.e., "decreasing"). If no VB can accommodate the RRH, a new VB is added to serve the RRH.

C. Computational Complexity Analysis

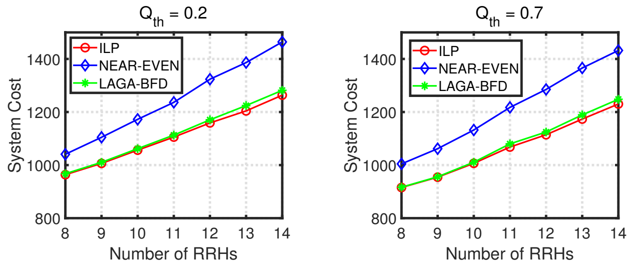
In the user association problem, we design a Lagrangian relaxation algorithm. In Alg. 1, the while loop in lines 3-16 is executed n_{max} times in the worst case. In each iteration, lines 4-5 yield an asymptotical factor of $O(NR)$ to calculate the lower bound. To obtain the upper bound, we design the UB searching algorithm, where we have to enumerate all UEs and RRHs to obtain the user association solutions in the worst case and hence yields an asymptotical factor of $O(NR)$. Updating Lagrangian multipliers in line 14 produces an asymptotical factor of $O(N + R)$. Therefore, the computational complexity of Alg. 1 is $O(n_{max}(NR + NR + N + R)) = O(n_{max}NR)$. In the BBU-RRH Mapping problem, the best fit decreasing scheme is utilized to solve the problem, which yields an asymptotical factor of $O(NR)$ [28]. Hence, the overall computational complexity of the joint problem is $O(n_{max}NR + NR) = O(n_{max}NR)$, which can be solved in polynomial time.

VI. Simulation Results

In this section, we set up simulations to investigate the performance of our proposed joint BBU-RRH mapping and user association algorithm, which is solved by LAGrAngian relaxation algorithm and Best Fit Decreasing algorithm (LAGA-BFD). To evaluate the performance of this algorithm, we choose the commonly used NEARest-first user association scheme [14], and the BBU-RRH mapping policy in [6] where user requests from each RRH are distributed EVENly among the VBs (NEAR-EVEN). In addition, both LAGA-BFD and NEAR-EVEN will be compared with the optimal solutions obtained from the ILP by CPLEX.

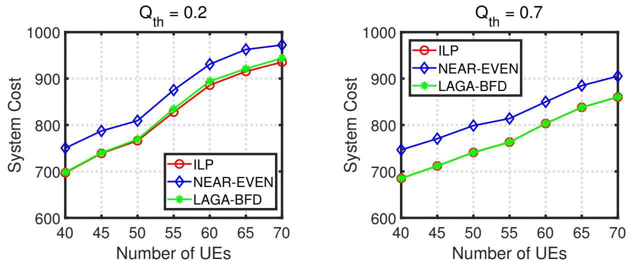
In our simulation, 6 macro RRHs and 60 UEs are randomly deployed in a $3000m \times 3000m$ area. All of RRHs' and UEs' x-coordinates and y-coordinates follow the uniform distribution ranging from 0 *m* to 3000 *m*. The system bandwidth is 10 *MHz* and the frequency reuse factor is one. We adopt the path loss model $128.1 + 37.6 * \log_{10}(D)$ (D in kilometers) based on the 3GPP specification. The transmit power for each RRH is 43 *dBm* and the noise power density is -174 *dBm/Hz*. The static and sleep power consumption of a RRH are 84 *W* and 56 *W*, respectively [23]. The load-power coefficient of RRH $\beta = 500$ *W/Mb*. The average traffic arrival rate for each UE is 1.0 *request/s* and the average traffic size for each UE is 1 *Mb*. The system cost coefficient (i.e., the system cost incurred by RRH power consumption) for RRHs $\eta = 1$ per watt and the one (i.e., the system cost incurred by renting each VM) for VBs $f = 30$ per VB. Throughout the simulation, we assume the QoS latency ratio for all RRHs and VBs are the same (i.e., $Q_{th} = \tilde{Q}_{th}$).

We first evaluate the performance of LAGA-BFD with different numbers of RRHs ranging from 6 to 14, shown in Fig. 2. We also conduct the simulation under two different QoS requirements of latency ratio 0.2 and 0.7, respectively. Fig. 2(a) and Fig. 2(b) depict the comparisons with the stricter requirement and the looser requirement, respectively. The general trends of the system cost in Fig. 2 go up with the increasing number of RRHs because building more RRHs potentially increase power consumption of all RRHs and hence increase the system cost. For example, an additional RRH consumes the least power in the sleep mode and even more if it is active. We can observe from Fig. 2 that LAGA-BFD performs close to the ILP optimal solution and much better than NEAR-EVEN. NEAR-EVEN exhibits a much steeper slope than those of LAGA-BFD and ILP because NEAR-EVEN always connects the UEs with the nearest RRHs; this will likely activate most RRHs if UEs are distributed evenly and hence draws more



(a) (b)

Fig. 2. System cost vs number of RRHs.



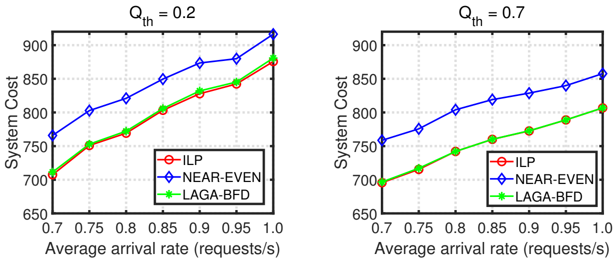
(a) (b)

Fig. 3. System cost vs number of UEs.

power consumption. Moreover, NEAR-EVEN incurs a higher system cost because it evenly distributes the requests from each RRH among all VBs without considering different traffic loads of all VBs; this may cause underutilization of some VBs and hence more VBs are activated as compared to LAGA-BFD.

In comparing Fig. 2(a) with Fig. 2(b), we can observe that for a certain number of RRHs, a stricter QoS requirement incurs a higher system cost. The reason is that with a looser QoS requirement, each RRH and VB can serve more UEs' requests and hence the active RRHs and VBs can be reduced. In both figures, LAGA-BFD performs close to ILP and better than NEAR-EVEN. ILP always provides the lowest system cost because ILP from CPLEX uses the branch and bound method to derive the exact optimal solutions.

We then compare the performances of the three algorithms with different numbers of UEs. Fig. 3 illustrates the system cost under different UE numbers from 30 to 70, and the comparisons of different QoS requirements of latency ratios are shown in Fig. 3(a) and Fig. 3(b). When the number of UEs increases, the system cost rises because more RRHs and VBs are needed to serve these UEs. In both Fig. 3(a) and Fig. 3(b), LAGA-BFD performs close to ILP and better than NEAR-EVEN. In comparing Fig. 3(a) with Fig. 3(b), a stricter QoS requirement (Fig. 3(a)) introduces a higher system cost. In addition, the solutions obtained by LAGA-BFD are closer to those of ILP in Fig. 3(a) as compared to Fig. 3(b). A stricter QoS requirement (Fig. 3(a)) implies that user association and BBU-RRH mapping strategies should be better scheduled and hence LAGA-BFD, the suboptimal solution, incurs a larger deviation from the optimal value.



(a) (b)

Fig. 4. System cost vs average arrival rate.

We also investigate the impact of different traffic arrival rates on the system cost. Fig. 4 depicts the system cost under different average arrival rate λ from 0.06 to 1.0. The results with different QoS requirements are shown in Fig. 4(a) and Fig. 4(b). NEAR-EVEN always incurs the highest system cost for the same reason as explained in Fig. 2. A larger arrival rate implies more traffic from each UE, and so more active RRHs and VBs are required to serve the requests and hence a higher system cost is incurred, as observed in Fig. 4. We can also observe that LAGA-BFD performs close to ILP and even tigher when the QoS requirement is 0.7 in Fig. 4(b) for the same reason observed in Fig. 3. Similarly, the system cost incurred for a stricter QoS requirement ($Q_{th} = 0.2$ in Fig. 4(a)) is higher than that for a less strict QoS requirement ($Q_{th} = 0.7$ in Fig. 4(b)) because more active RRHs and VBs are required with a stricter QoS requirement. In summary, we can observe that LAGA-BFD incurs a lower system cost than the existing algorithm NEAR-EVEN, and achieves a performance very close to the optimal solution of ILP and performs even better when the QoS requirement is not strict.

VII. Conclusion

In this paper, we have investigated the QoS-aware joint BBU-RRH mapping and user association problem in C-RAN with the objective to minimize the system cost incurred by the power consumption of all RRHs and rentals of VBs. We have modeled our QoS requirement model as the delays of two queues in tandem, including the BBU processing queue and the RRH transmission queue. An ILP model has been proposed to address this joint optimization problem and to provide insights on which RRH should be activated, how to associate UEs to different RRHs, how many VBs are needed, and how to connect RRHs with different VBs. However, the ILP model incurs high computing complexity. Hence, we have further decomposed this joint problem into two subproblems: the user association problem and the BBU-RRH mapping problem. We have designed a Lagrangian relaxation algorithm for the user association problem and transformed the BBU-RRH mapping problem into a bin-packing problem which is solved by the BFD algorithm. Simulation results have demonstrated that our proposed algorithm LAGA-BFD performs very close to the optimal solution and performs even better when the QoS requirement is not strict.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," *White Paper*, Mar. 2017.

[2] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, “Recent advances in cloud radio access networks: System architectures, key techniques, and open issues,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308, Third Quarter 2016.

[3] X. Sun and N. Ansari, “EdgeIoT: Mobile edge computing for the internet of things,” *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, December 2016.

[4] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, “Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN,” *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov 2014.

[5] 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 of Things Journal*, vol. 3, no. 3, pp. 428–438, June 2016.

[6] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, “System cost minimization in cloud RAN with limited fronthaul capacity,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[7] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, “User association in 5G networks: A survey and an outlook,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second Quarter 2016.

[8] C. L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, “Recent progress on C-RAN centralization and cloudification,” *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[9] Q. Liu, T. Han, N. Ansari, and G. Wu, “On designing energy-efficient heterogeneous cloud radio access networks,” *IEEE Trans. on Green Communications and Networking*, DOI: 10.1109/TGCN.2018.2835451, early access.

[10] X. Huang and N. Ansari, “Joint spectrum and power allocation for multi-node cooperative wireless systems,” *IEEE Transactions on Mobile Computing*, vol. 14, no. 10, pp. 2034–2044, Oct. 2015.

[11] J. Opadere, Q. Liu, and T. Han, “Energy-efficient RRH sleep mode for virtual radio access networks,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec. 2017, pp. 1–6.

[12] S. Guo, D. Zeng, L. Gu, and J. Luo, “When green energy meets cloud radio access network: Joint optimization towards brown energy minimization,” *Mobile Networks and Applications*, Feb. 2018. [Online]. Available: https://doi.org/10.1007/s11036-018-1028-9

[13] D. Zeng, J. Zhang, S. Guo, L. Gu, and K. Wang, “Take renewable energy into CRAN toward green wireless access networks,” *IEEE Network*, vol. 31, no. 4, pp. 62–68, July 2017.

[14] M. Peng, S. Yan, and H. V. Poor, “Ergodic capacity analysis of remote radio head associations in cloud radio access networks,” *IEEE Wireless Communications Letters*, vol. 3, no. 4, pp. 365–368, Aug 2014.

[15] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo, “Energy efficient user association for cloud radio access networks,” *IEEE Access*, vol. 4, pp. 2429–2438, 2016.

[16] S. Wang and Y. Sun, “Enhancing performance of heterogeneous cloud radio access networks with efficient user association,” in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[17] T. Han and N. Ansari, “Network utility aware traffic load balancing in backhaul-constrained cache-enabled small cell networks with hybrid power supplies,” *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2819–2832, Oct 2017.

[18] T. Sigwele, A. S. Alam, P. Pillai, and Y. F. Hu, “Evaluating energy-efficient cloud radio access networks for 5G,” in *2015 IEEE International Conference on Data Science and Data Intensive Systems*, Dec 2015, pp. 362–367.

[19] K. Boulos, M. E. Helou, and S. Lahoud, “RRH clustering in cloud radio access networks,” in *2015 International Conference on Applied Research in Computer Science and Engineering (ICAR)*, Oct 2015, pp. 1–6.

[20] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, “Baseband processing units virtualization for cloud radio access networks,” *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 189–192, April 2015.

[21] H. M. Soliman and A. Leon-Garcia, “QoS-aware joint RRH activation and clustering in cloud-RANs,” in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[22] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, “QoS-aware dynamic RRH allocation in a self-optimized cloud radio access network with RRH proximity constraint,” *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 730–744, Sept 2017.

[23] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, “How much energy is needed to run a wireless network?” *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, October 2011.

[24] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, “A performance analysis of EC2 cloud computing services for scientific computing,” in *Cloud Computing*, D. R. Avresky, M. Diaz, A. Bode, B. Ciciani, and E. Dekel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 115–131.

[25] L. Kleinrock, *Queueing Systems: Computer Applications*. Hoboken, NJ, USA: Wiley-Interscience, 1976.

[26] G. T. Ross and R. M. Soland, “Modeling facility location problems as generalized assignment problems,” *Management Science*, vol. 24, no. 3, pp. 345–357, 1977. [Online]. Available: https://doi.org/10.1287/mnsc.24.3.345

[27] M. S. Daskin, *Network and Discrete Location : Models, Algorithms, and Applications*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.

[28] E. G. Coffman Jr., J. Csirik, G. Galambos, S. Martello, and D. Vigo, “Bin packing approximation algorithms: Survey and classification,” in *Handbook of Combinatorial Optimization*, P. M. Pardalos, D.-Z. Du, and R. L. Graham, Eds. New York, NY: Springer New York, 2013, pp. 455–531.