System cost minimization in cloud RAN with limited fronthaul capacity

Tang, Jianhua; Tay, Wee Peng; Quek, Tony Q. S.; Liang, Ben

2017

Tang, J., Tay, W. P., Quek, T. Q. S., & Liang, B. (2017). System cost minimization in cloud RAN with limited fronthaul capacity. IEEE Transactions on Wireless Communications, 16(5), 3371-3384. doi:10.1109/TWC.2017.2682079

https://hdl.handle.net/10356/102695

https://doi.org/10.1109/TWC.2017.2682079

System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity

Jianhua Tang, *Member, IEEE,* Wee Peng Tay, *Senior Member, IEEE,* Tony Q.S. Quek, *Senior Member, IEEE* and Ben Liang, *Senior Member, IEEE*

***Abstract*—Cloud radio access network (C-RAN) is emerging as a potential alternative for the next generation RAN by merging RAN and cloud computing together. In this paper, we consider the baseband unit (BBU) pool of C-RAN as a collection of virtual machines (VMs). We allow each user equipment (UE) to associate with multiple VMs in the BBU pool, and each remote radio head (RRH) can only serve a limited number of UEs. Under this model, we jointly optimize the VM activation in the BBU pool and sparse beamforming in the coordinated RRH cluster, which is constrained by limited fronthaul capacity, to minimize the system cost of C-RAN. We formulate this problem as a mixed-integer nonlinear programming (MINLP) problem, and then propose efficient methods to optimize the number of active VMs, as well as the sparse beamforming vectors. Moreover, we derive closed-form solution for the beamforming vectors. Simulation results suggest that our proposed algorithms have better performance than the benchmark algorithms in terms of both system cost and robustness.**

***Index Terms*—C-RAN, VM activation, limited fronthaul capacity, computation capacity**

I. Introduction

The evolution of radio access network (RAN) over the past decade has been driven by fast data proliferation. Cisco Systems predicts that mobile data traffic will increase 8-fold from 2015 to 2020, and the number of mobile devices per capita will reach 1.5 by the year 2020 [1]. To maintain a high quality-of-service (QoS), the principal solution for RAN service providers is enhancing RANs' capacity and coverage. However, they are facing many challenges when adopting this solution approach [2]: Firstly, the explosive increase in network capacity demand (especially busy-hour demand) triggers an exponential increase in the number of base stations (BSs), which leads to a significantly higher power consumption.

Part of this paper was presented by the Asilomar Conference on Signals, Systems, and Computers (ASILOMAR), Pacific Grove, CA, USA, November 2015.

This work was supported in part by the Startup Funds of Chongqing University of Posts and Telecommunications under Grant A2016-114, Singapore MOE ARF Tier 2 under grant MOE2014-T2-1-028, the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2014, and the MOE ARF Tier 2 under Grant MOE2014-T2-2-002.

J. Tang is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China. He was with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. e-mail: jtang4@e.ntu.edu.sg

W. P. Tay is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. e-mail: wptay@ntu.edu.sg.

T. Q. S. Quek is with Singapore University of Technology and Design. He is also with Institute for Infocomm Research, A*STAR, Singapore. e-mail: tonyquek@sutd.edu.sg.

B. Liang is with Department of Electrical and Computer Engineering, University of Toronto, Canada. e-mail: liang@comm.utoronto.ca.

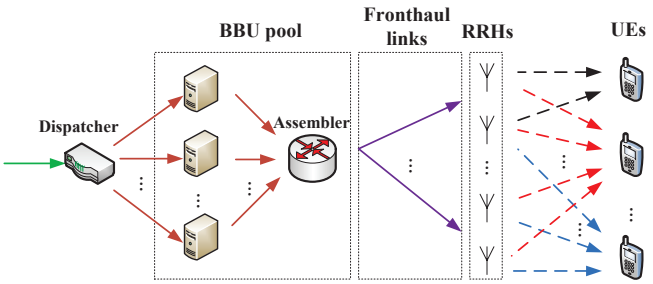


Fig. 1. A typical structure of C-RAN.

Secondly, costly capital and operating expenditure leads to falling average revenue per user. Moreover, with the dynamic nature of mobile traffic, the utilization of some BSs is actually quite low during non-peak hours.

Along with RANs' evolution, cloud computing has emerged as a popular computing paradigm, since cloud computing has its attractive characteristics like resource pooling and rapid elasticity. By introducing the merits of cloud computing into RANs, cloud radio access network (C-RAN) has been proposed as a prospective architecture to overcome the aforementioned challenges [3]. A typical C-RAN consists of three components (cf. Figure 1): baseband unit (BBU) pool, fronthaul links and remote radio heads (RRHs). The most significant innovation of C-RAN is utilizing a centralized cloud-based BBU pool instead of the conventional distributed baseband processing devices co-located with the BSs. That means, in C-RAN, baseband signal processing functionalities are decoupled from the RRHs, and RRHs just need to keep basic signal transmission and reception functionalities.

C-RAN possesses several advantages compared to the conventional RAN: Firstly, utilizing centralized signal processing in the BBU pool instead of the distributed BSs in the conventional RAN can significantly save the capital and operating expenditure. Secondly, joint processing in the BBU pool and cooperative radio techniques over RRHs, which are interconnected via the BBU pool, improves the spectrum efficiency, link reliability and the communication quality of the cell edge users. Thirdly, BBU pool consists of many general purpose servers. Applying cloud computing as the computing paradigm of the centralized BBU pool can reduce the power consumption and improve hardware utilization, through resource sharing and virtualization, i.e., a server can be further virtualized into many virtual machines (VMs). However, several challenges in C-RAN remain to be addressed, and

I summarize these challenges as a "*limited versus unlimited*" problem:

- Due to the high amount of data transfers (especially when joint transmission techniques are adopted in the C-RAN downlink) in the fronthaul, whose capacity is actually *limited*, efficient data transfer algorithms need to be developed.
- As all the computation resources are migrated from the BSs into a centralized BBU pool, the amount of computation resources in BBU pool is relatively *unlimited* (compared to these in BSs). We need to effectively manage and dispatch those computation resources in C-RAN. In particular, with the ability to elastically scale service capacities in the cloud-based BBU pool, many problems well studied in the conventional RANs have to be relooked at in C-RAN. For example, resource allocation schemes for conventional RANs are typically oblivious to computation capatities/costs since they are fixed. In C-RAN, however, the computation capatities/costs at the cloud BBU pool can be dynamically scaled, e.g., turning on or off VMs, according to system demands.

In this paper, we jointly optimize the VM activation in the BBU pool and sparse beamforming vector in the RRHs, which have limited fronthaul capacity, to minimize the system cost of C-RAN, including cloud processing cost and wireless transmission cost.

A. Related Work

C-RAN [4]–[6] has attracted increasing research interest over the past three years. C-RAN provides a centralized BBU pool to improve resource utilization, such as the hardware and energy utilization, and enables centralized processing of the receive and transmit signals at the RRHs. However, the main concern for this centralized processing structure is the high amount of data transfer in the capacity-limited fronthaul [7]–[20]. In fact, there are two main different definitions for the fronthaul capacity in the literature:

1) Fronthaul capacity was defined as the maximum sum data rate transmitted on each fronthaul, such as in [11] and [18].
2) Fronthaul capacity was defined as the maximum number of user equipments (UEs) can be served on each fronthaul, such as in [9], [10] and [21].

The authors always implicitly assume that each fronthaul can serve unlimited number of users when they adopt the first definition. However, due to the signaling and coordination overhead, in the real system, this assumption can not hold. Thus, we adopt the second definition in this work (See the detailed mathematical definition in Section II-C). Moreover, the second definition is also applied in the simulation part of [22].

With respect to the problem formulations, the works [7], [8] aim to minimize the number of active UE-RRH pairs to mitigate the amount of data transfer in the fronthaul, while in [9], [11], [13], [23], the authors consider the fronthaul capacity as a constraint in their optimization formulation. The references [14]–[17], [24] develop efficient signal compression/quantization algorithms to downsize the load in the C-RAN fronthaul. Admission control in heterogeneous networks with wireless backhaul is studied in [25]. In addition, to reduce the amount of data transfer in C-RAN, caching at access points is a promising approach [26]–[28], and users device level caching is also applicable [29].

Instead of focusing on the wireless transmission part of C-RAN, several works also investigate the problems introduced by the BBU pool. Computationally aware strategies are proposed to reduce computational outage in [30], and to maximize sum-rate in [31]. The reference [32] uses task assignment to minimize the power consumption in the BBU pool of C-RAN. However, most of these works fail to consider the interaction between cloud processing and wireless transmission in C-RAN. In [33], we jointly optimize the elastic service scaling in the BBU pool, RRH selection, and the beamformer design at the RRHs. However, the system model studied in [33] is somewhat idealized. This work considers a more practical system, which differs from our previous work in the following aspects:

1) In this paper, to fully utilize the computation capacity of VMs, we consider the case where each UE can be associated with multiple VMs in the BBU pool, and we need to optimize the number of active VMs in the BBU pool. However, in [33], we assume each UE can only associate with one VM, and one VM can only server one UE. Hence, the number of active VMs is just simply equal to the number of UEs, while this is not practical.
2) We consider the limited fronthaul capacity in this paper, which is assumed to be unlimited in [33]. As a consequence, in this paper, each UE can only access a limited number of RRHs in the active RRH cluster. In [33], we assume each UE can access every RRH in the active RRH cluster, which is not practical.

B. Main Contributions

In this paper, we jointly optimize the VM activation and sparse beamforming in order to minimize the overall cost for a C-RAN with limited fronthaul capacity. Our main contributions are as follows:

- We formulate the "*unlimited versus limited*" problem as a mixed-integer nonlinear programming (MINLP) by minimizing the overall cost, which consists of two parts: the cloud processing cost (in the BBU pool) with respect to (w.r.t.) the number of active VMs, and the wireless transmission cost (in the fronthaul and RRHs) w.r.t. the transmit beamformers.
- To avoid the feasibility problem caused by relaxing the $l_{2,0}$-norm constraint directly, we reformulate the original MINLP into an equivalent problem, which introduces a price vector. This equivalent problem can then be solved by adjusting the value of the price vector, and reducing to a subproblem. We propose two different approaches to solve this subproblem: an integer search (IS) approach and a joint optimization with integer recovery (JR) approach. Moreover, we derive the closed-form solution for the JR approach.

- We provide simulation results that suggest that our proposed approach can provide better feasibility guarantee and obtain lower system cost than the benchmark algorithms, for example, the recently proposed static clustering algorithm in [11].

The remainder of this paper is organized as follows. We present the C-RAN system model in Section II, and propose the problem formulation and its equivalent formulation in Section III. In Section IV and V, we propose approaches to solve the problem step-by-step. And in Section VI, the numerical results are presented. We conclude the paper in Section VII.

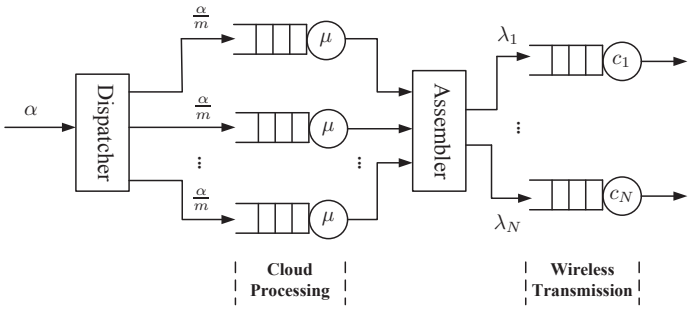


Fig. 2. Queueing network model representation of a C-RAN cloud processing and wireless transmission.

Notations

We use calligraphy letters to represent sets, boldface lower case letters to denote vectors, and boldface upper case letters to denote matrices. The notation $\|\cdot\|_2$ stands for the Euclidean norm, while $(\cdot)^T$ and $(\cdot)^H$ are the transpose and the conjugate transpose, respectively. We use $\mathbb{N}$, $\mathbb{C}$ and $\mathbb{R}^+$ to represent the natural numbers, complex numbers and non-negative real numbers, respectively. The notation $\mathcal{A}\backslash\mathcal{B}$ denotes the set $\mathcal{A}$ with its subset $\mathcal{B}$ removed. We also use the notation $x^+ = \max(0, x)$. $\lfloor x \rfloor$ stands for the largest integer smaller than or equals x and $\lceil x \rceil$ stands for the smallest integer larger than or equals x.

II. SYSTEM MODEL

In this section, we present our C-RAN system model and its practical constraints.

A. System description

Suppose that there are N single-antenna UEs and L RRHs, each with K antennas, in a C-RAN cluster. We denote the set of all UEs and all RRHs as $\mathcal{N} = \{1, \cdots, N\}$ and $\mathcal{L} = \{1, \cdots, L\}$, respectively. There are M homogenous VMs in the BBU pool. Each has computation capacity μ and incurs a *VM cost* $\varphi > 0$ when it is active. We denote the number of active VMs as $m \in \mathbb{N}$, where $m \leq M$. This model reflects the popular commercial cloud service models, e.g., Amazon Elastic Compute Cloud (EC2). In EC2, there are thousands of instances (VMs) in the data center, and each instance has a fixed computation capacity. Cloud users just need to decide how many instances they need to rent.

In the downlink of a C-RAN (cf. Figure 1), all UEs' incoming traffic is first processed by a dispatcher. We assume that the mean arrival data rate of UE i to the dispatcher is λ_i, $\forall i \in \mathcal{N}$, and let $\alpha = \sum_{i\in\mathcal{N}} \lambda_i$. Then, each transport block [34] (or even a code block within each transport block) in the data flow to UE i can be routed to one of m active VMs for processing (e.g., turbo coding) with probability $1/m$ by the dispatcher. Therefore, the mean incoming traffic rate routed to each active VM is α/m.

In the wireless transmission part, we consider joint transmission as the CoMP technique in C-RAN, i.e., each UE's data can be shared among all the coordinated/associated RRHs, while the RRHs have limited fronthaul link capacity (the fronthaul links between the BBU pool and the RRHs are heterogeneous, and can be fiber links, copper cables or wireless channels). After processing by the VMs, each UE's data is forwarded to the UE via at most L RRHs (since the data is shared among the limited fronthaul RRHs). Let the achievable wireless transmission rate to UE i be c_i.

B. Queueing system model

Each active VM in the BBU pool can be modeled as a queue. Specifically, for each queue, the mean arrival rate is α/m and the mean service rate is μ. Throughout the paper, we assume the tasks within each queue is served in a first in first out (FIFO) manner and the buffer length is infinite. We note that the use of queuing models, where the wireless transmission rate is the queue's service rate, is not new, and has been widely used to characterize wireless communication systems [35].

We consider a double-layer queueing network to represent each UE's data processing and transmitting behavior in the C-RAN downlink (cf. Figure 2). Specifically, in the BBU pool, the transport blocks to each UE is processed (e.g., encoded) by m parallel active VMs, each of which is abstracted as a queue with mean service rate μ. Then, the processed data is transmitted to UE i via RRHs over wireless channels, which are modeled by a wireless transmission queue with mean service rate c_i.

We denote the mean processing delay for data to UE i in the BBU pool as b_i. Let d_i be the mean transmission delay of the data to UE i in the wireless transmission queue (i.e., the expected delay incurred at the queue before the data is completely transmitted). We assume that UE i's packet arrival process to the dispatcher is a Poisson process with mean rate λ_i. Hence, the arrival process to each VM also forms Poisson process with mean arrival rate α/m. Suppose that the service time of each data packet in each VM queue follows an exponential distribution with mean $1/\mu$, for $\mu > \alpha/m$. Then, for each UE's data, the arrival rate to the wireless transmission queue is the same as the one to the dispatcher [36], [37] [1]. We assume that the service time of each data packet in the wireless transmission queue follows an exponential distribution with

[1]Note that, we do not consider the impact, introduced by baseband processing, on the arrival data rate to the wireless transmission queue. Specifically, we assume that, after being processed by the VM, the size of the transport blocks and the inter-arrival times to the assembler still remain the same as those to the dispatcher.

mean $1/c_i$. Therefore, the data processing and transmission in our C-RAN model can be treated as two layers of M/M/1 queues in tandem. In addition, from queueing theory [38], we have

$$b_i = \frac{m}{m\mu - \alpha}, \tag{1}$$

$$d_i = \frac{1}{c_i - \lambda_i} \tag{2}$$

where $\mu > \alpha/m, c_i > \lambda_i$, $\forall i \in \mathcal{N}$.

Let u_i denote the data symbol for UE i with $\mathrm{E}[|u_i|^2] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^{K\times 1}$ denote the transmit beamformer for UE i from RRH j. We assume block fading wireless transmission channels from the RRHs to the UEs. We define the $l_{2,0}$-norm and $l_{2,1}$-norm of vector $\mathbf{w}_{ij}$ as $\|\mathbf{w}_{ij}\|_{2,0} \triangleq \left\|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\right\|_0$ and $\|\mathbf{w}_{ij}\|_{2,1} \triangleq \mathbf{w}_{ij}^H \mathbf{w}_{ij}$ respectively. Then, the associations between UEs and RRHs can be represented by $\|\mathbf{w}_{ij}\|_{2,0}$, i.e., $\|\mathbf{w}_{ij}\|_{2,0} = 1$ if and only if UE i is connected to j.

The block fading channel from RRH j to UE i is denoted as $\mathbf{h}_{ij}^H$, where $\mathbf{h}_{ij} \in \mathbb{C}^{K\times 1}$, for $i \in \mathcal{N}$ and $j \in \mathcal{L}$. The received signal at UE i is then given by

$$\hat{u}_i = \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} u_i + \sum_{k\neq i}^{N} \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj} u_k + \delta_i,$$

where the first term is the useful signal for UE i, the second term is the interference to UE i, and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) at UE i. The signal-to-interference-plus-noise ratio (SINR) at UE i is

$$\mathrm{SINR}_i = \frac{|\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}|^2}{\sigma_i^2 + \sum_{k\neq i}^{N} |\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}|^2}. \tag{3}$$

The downlink achievable rate c_i to UE i satisfies

$$c_i \leq B_i \log(1 + \mathrm{SINR}_i),$$

where B_i is the wireless transmission bandwidth for UE i. Each RRH j's maximum transmit power is denoted as E_j, i.e.,

$$\sum_{i=1}^{N} \|\mathbf{w}_{ij}\|_{2,1} \leq E_j, \text{ for } j \in \mathcal{L}.$$

C. Practical constraints

In addition to the basic system model above, we include the following two practical constraints to capture the features of C-RAN:

1) **System delay constraint**: To couple cloud processing and wireless transmission in C-RAN, we propose the cross-layer system delay constraint:

$$b_i + d_i \leq \tau_i, \forall i \in \mathcal{N},$$

where τ_i is a predefined maximum system delay for UE i, which can be treated as its QoS requirement.

2) **Fronthaul capacity constraint**: We denote $S_j \in \mathbb{N}$ as the fronthaul link j's capacity, i.e., the maximum number of UEs that can be connected with this fronthaul link. In other words, at most S_j UEs can be associated with RRH j. We can cast this fronthaul capacity constraint as

$$\sum_{i\in\mathcal{N}} \|\mathbf{w}_{ij}\|_{2,0} \leq S_j, \forall j \in \mathcal{L},$$

where $\|\mathbf{w}_{ij}\|_{2,0} = 1$ if and only if RRH j is associated with UE i.

III. Problem Formulation

The system cost incurred by a C-RAN is the total VM cost in the BBU pool and the power consumption incurred by the RRHs. Our aim is to minimize the system cost by optimizing the number of active VMs and the beamformers at the RRHs. Specifically, (i) the cost for cloud processing is $m\varphi$; (ii) the cost incurred by wireless transmission is $\eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,1} + \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,0} P_j$, where $\eta > 0$ is a weight and P_j is a static cost when RRH j (and its fronthaul) is active. Our optimization problem is formulated as:

$$\text{(P0)} \quad \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,1} + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j \|\mathbf{w}_{ij}\|_{2,0}$$

$$\text{s.t.} \quad \frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \ \forall i \in \mathcal{N}, \tag{4}$$

$$\alpha < m\mu, \ \lambda_i < c_i, \ \forall i \in \mathcal{N}, \tag{5}$$

$$0 < m \leq M, \ m \in \mathbb{N}, \tag{6}$$

$$c_i \leq B_i \log\left(1 + \mathrm{SINR}_i\right), \ \forall i \in \mathcal{N}, \tag{7}$$

$$\sum_{i=1}^{N} \|\mathbf{w}_{ij}\|_{2,1} \leq E_j, \ \forall j \in \mathcal{L}, \tag{8}$$

$$\sum_{i=1}^{N} \|\mathbf{w}_{ij}\|_{2,0} \leq S_j, \ \forall j \in \mathcal{L}, \tag{9}$$

where "s.t." stands for "subject to" and SINR_i is given by (3). We assume the feasible region of problem (P0) is nonempty. Let the optimal solution for problem (P0) be $\{(m^*, c_i^*, \mathbf{w}_{ij}^*) : i \in \mathcal{N}, j \in \mathcal{L}\}$.

Remark 1: Actually, the system cost in this paper is a wide concept, and its physical meaning can be varied. For instance,

1) It can be the monetary cost per unit time, if ϕ stands for the price of renting/turning on a VM per unit time, P_j denotes the unit time electricity price of turning on RRH j and η captures the electricity price per Watts per unit time.
2) It also can be the power consumption, if ϕ stands for the static power consumption of turning on a VM, P_j denotes the static power consumption of turning on RRH j and η captures the inefficiency coefficient of the amplifier of each RRH.

For problem (P0), we first note that there is no loss in optimality if we restrict the constraints (4) to be equalities.

Proposition 1: There exists an optimal solution $\{(m^*, c_i^*, \mathbf{w}_{ij}^*) : i \in \mathcal{N}, j \in \mathcal{L}\}$ such that constraints

(4) are active, i.e., (m^*, c_i^*) for problem (P0) satisfies the following equation

$$\frac{m^*}{m^*\mu - \alpha} + \frac{1}{c_i^* - \lambda_i} = \tau_i, \ \forall i \in \mathcal{N}. \tag{4'}$$

Proof: For any optimal solution, if there exists an $i \in \mathcal{N}$ such that (4) is a strict inequality, then we can decrease c_i^* until equality in (4) holds. This is still a feasible solution for (P0), and the proposition is proved. □

Proposition 1 shows the interaction between cloud processing and wireless communication. For example, if we need a lower processing delay b_i, which means more VMs should be added, then this will result in a higher cloud processing cost. However, on the other hand, based on Proposition 1, a lower processing delay b_i will lead to a higher transmission delay d_i in return. That means we can save some wireless transmission power and active RRHs. Then a lower wireless transmission cost can be achieved. This interaction reveals C-RAN as a coupled system.

Problem (P0) is difficult to solve, due to the following reasons: (i) it is an MINLP with two integer constraints (6) and (9); and (ii) the problem is nonconvex even if we assume $m \in \mathbb{R}^+$ and the fronthaul capacity constraint (9) is removed. In the following section, we propose a reformulation and some relaxation techniques to make problem (P0) tractable.

A. An equivalent formulation for problem (P0)

In problem (P0), one of the solution challenges is the $l_{2,0}$-norm constraint (9). Two commonly used approaches to deal with the $l_{2,0}$-norm are: smoothing function approximation [8], [39] and reweighted $l_{2,1}$-norm approximation [11], [40]. However, if we just relax the left hand side of constraint (9) with a smoothing function or $l_{2,1}$-norm approximations, and solve the relaxed problem directly, then we have no guarantee that the optimal solution derived from the relaxed problem is also feasible for the original problem (P0).

Let's consider the following problem, which considers the trade-off between the system cost and fronthaul capacity:

$$\begin{aligned}(\text{P1}) \quad \min_{m, c_i, \mathbf{w}_{ij}} \quad & m\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,1} \\ & + \sum_{i=1}^{N}\sum_{j=1}^{L} (P_j + \gamma_j) \|\mathbf{w}_{ij}\|_{2,0} \\ \text{s.t.} \quad & (4'),\ (5),\ (6),\ (7) \text{ and } (8),\end{aligned}$$

where $\gamma_j \geq 0$ is the price for RRH j. Let $\Gamma \triangleq [\gamma_1, \cdots, \gamma_L]^T$.

We denote $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ as the optimal solution for problem (P1) for a given price vector Γ. Define $\beta_j(\Gamma) = \sum_{i=1}^{N} \|\mathbf{w}_{ij}(\Gamma)\|_{2,0}$. The following theorem shows the relationship between problem (P0) and problem (P1).

Theorem 1: For problem (P1), if

$$\beta_j(\Gamma) = S_j, \ \forall j \in \mathcal{L}, \tag{10}$$

then the optimal solution to problem (P1) is also optimal for problem (P0).

Proof: See Appendix A. □

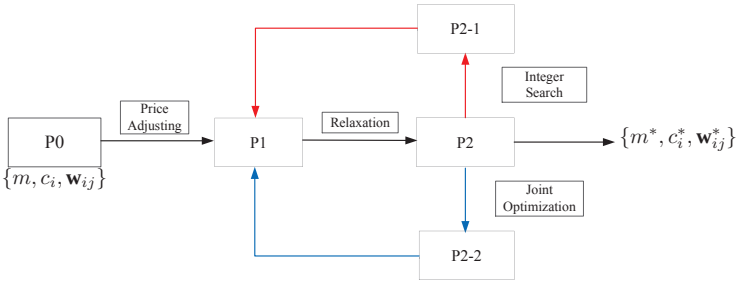


Fig. 3. An iterative approach to solve problem (P0).

Based on Theorem 1, if a price vector Γ can be found so that (10) holds, then we can solve (P1) instead of (P0). However, such a price vector may not exist, and in general, the solution of (P1) is sub-optimal for (P0) if

$$\beta_j(\Gamma) \leq S_j, \ \forall j \in \mathcal{L}. \tag{11}$$

Instead of solving problem (P0) directly, in what follows, we propose a step-by-step relaxation and reformulation approach to simplify the problem, and obtain a reasonable but sub-optimal solution:

1) In Section IV, we introduce some properties of the price vector Γ in problem (P1), and we propose a price adjusting algorithm to find a price vector Γ that satisfies equation (11). For a fixed price vector Γ, we apply reweighted l_1-norm relaxation on problem (P1) to simplify it into another problem (P2) in subsection IV-B.
2) In Section V, we propose two different approaches to solve problem (P2).

We show the logic flow for solving problem (P0) in Figure 3.

IV. Approximation for problem (P1)

In this section, we first present some properties that the price vector Γ satisfies. Then, we propose a *price adjusting* algorithm to obtain a sub-optimal solution for problem (P0).

A. Bisection search for price vector Γ

In the following proposition, we present results that allow us to iteratively adjust the price vector in problem (P1) such that equation (11) holds. The proposition is inspired by [9].

Proposition 2: Fix each γ_k as a constant $\bar{\gamma}_k$ for all $k \in \mathcal{L}\backslash j$, and let $\bar{\Gamma}_j = [\bar{\gamma}_1, \cdots, \bar{\gamma}_{j-1}, \gamma_j, \bar{\gamma}_{j+1}, \cdots, \bar{\gamma}_L]^T$. Then the following holds.

(i) $\beta_j(\bar{\Gamma}_j)$ is a non-increasing function w.r.t. γ_j.
(ii) There is a threshold price for RRH j, $\theta_j = \varphi M + \sum_{j\in\mathcal{L}}(\eta E_j + P_j S_j) + \sum_{k\in\mathcal{L}\backslash j} \bar{\gamma}_k S_k$, such that for $\gamma_j \geq \theta_j$, $\beta_j(\bar{\Gamma}_j) \leq S_j$.

Proof: The proof is similar with the one in [9]. We provide it for completeness in Appendix B. □

Recall that the feasible region of problem (P0) is nonempty, therefore, we can always satisfy equation (11) by iteratively searching over $\gamma_j \in [0, \theta_j]$. We elaborate the algorithm to solve problem (P1) by iteratively adjusting the price vector in Algorithm 1, in which $\gamma_j^{(l)}$ is the j-th component of $\Gamma^{(l)}$ in

the l-th iteration, and $\theta_{\tilde{j}}^{(l)} = \varphi M + \sum_{j\in\mathcal{L}}(\eta E_j + P_j S_j) + \sum_{k\in\mathcal{L}\backslash j}\gamma_k^{(l-1)} S_k$.

Algorithm 1 Price adjusting algorithm for problem (P1)

1: Initialize: Let $\Gamma^{(0)} = [0, \cdots, 0]^T$.
2: Iteration l: Solve problem (P1) with given $\Gamma^{(l-1)}$, obtaining $\beta_j(\Gamma^{(l-1)})$, for $j \in \mathcal{L}$.
3: **if** $\beta_j(\Gamma^{(l-1)}) \le S_j$, $\forall j \in \mathcal{L}$, **then**
4: break;
5: **else**
6: for those $\tilde{j} \in \mathcal{A}^{(l)} \triangleq \{j : \beta_j(\Gamma^{(l-1)}) > S_j, \forall j \in \mathcal{L}\}$, set $\gamma_{\tilde{j}}^{(l)} = \theta_{\tilde{j}}^{(l)}$. Fix $\gamma_k^{(l)} = \gamma_k^{(l-1)}$, $\forall k \in \mathcal{L}\backslash\mathcal{A}^{(l)}$.
7: **end if**
8: Let $l = l + 1$, go to step 2.

In Algorithm 1, the main iteration in Step 2 involves an algorithm to solve problem (P1). Although we avoid the feasibility problem by reformulating problem (P0) into (P1), the $l_{2,0}$-norm still remains unsolved in the objective function. In the next subsection, we introduce reweighted $l_{2,1}$-norm approximation for problem (P1).

B. Reweighted $l_{2,1}$-norm relaxation

In compressive sensing [41], reweighted l_1-norm is regarded as an effective way to deal with the l_0-norm in the objective function, since l_1-norm is the convex relaxation for l_0-norm [40]. In the same spirit, we adopt an iterative procedure to solve problem (P1), and the terms involving $l_{2,0}$-norm in the objective function of problem (P1) as:

$$\|\mathbf{w}_{ij}\|_{2,0} \approx \rho_{ij}^{(p)} \|\mathbf{w}_{ij}\|_{2,1}, \tag{12}$$

where $\rho_{ij}^{(p)} = \left(\left\|\mathbf{w}_{ij}^{(p-1)}\right\|_2^2 + \phi\right)^{-1}$ is the weight in the p-th iteration, $\mathbf{w}_{ij}^{(p-1)}$ is a constant vector obtained from previous iteration and ϕ is a small positive constant to guarantee the numerical stability. The intuition behind the weight $\rho_{ij}^{(p)}$ is that the beamformer vector that has smaller norm in iteration $p-1$ is allocated a larger weight $\rho_{ij}^{(p)}$ in iteration p, and hence, the norm is further reduced after solving the problem in iteration p.

With the $l_{2,1}$-norm relaxation, problem (P1) can be approximated as the following problem in the p-th iteration:

$$\text{(P2)} \quad \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij}^{(p)} \|\mathbf{w}_{ij}\|_{2,1}$$
$$\text{s.t.} \quad (4'),\ (5),\ (6),\ (7) \text{ and } (8),$$

where $z_{ij}^{(p)} = \eta + (P_j + \gamma_j)\rho_{ij}^{(p)}$.

After the reweighted $l_{2,1}$-norm relaxation, a sub-optimal solution can be obtained for problem (P1). However, in each iteration of the reweighted $l_{2,1}$-norm relaxation, problem (P2) is required to be solved, which is still an MINLP. In the next section, we discuss two different approaches to solve problem (P2).

V. Two optimization approaches for problem (P2)

In this section, we propose two different approaches to solve problem (P2), which obtain its global and local optimal solution respectively.

First of all, constraints (4′) and (5) imply that

$$m = \alpha\left(\frac{\tau_i}{n_i} + \frac{1}{n_i(n_i(c_i - \lambda_i) - \mu)}\right) \ge \alpha\left(\frac{\tau_i}{n_i} + \frac{1}{n_i(n_i(\check{c}_i - \lambda_i) - \mu)}\right) \triangleq m_i,\ \forall i \in \mathcal{N}, \tag{13}$$

and

$$c_i = \lambda_i + \frac{\mu_i}{n_i} + \frac{\alpha}{n_i(n_i m - \alpha\tau_i)} \triangleq g_i(m),\ \forall i \in \mathcal{N}, \tag{14}$$

where $n_i = \tau_i\mu - 1 > 0$, and $\check{c}_i$ is an upper bound of c_i, which can be derived by applying the Cauchy-Schwarz inequality on (7) as follows:

$$c_i \le B_i \log\left(1 + \frac{1}{\sigma_i^2}\sum_{j=1}^{L}\|\mathbf{h}_{ij}\|_2^2 \sum_{j=1}^{L}\|\mathbf{w}_{ij}\|_2^2\right) \le B_i \log\left(1 + \frac{1}{\sigma_i^2}\sum_{j=1}^{L}\|\mathbf{h}_{ij}\|_2^2 E_j\right) \triangleq \check{c}_i,\ \forall i \in \mathcal{N}. \tag{15}$$

Based on (13) and (14), in what follows, we discuss two different approaches to solve problem (P2).

A. Integer search (IS) approach

Once m is fixed as an integer $\bar{m}$, problem (P2) is reduced into the following weighted sum-power minimization (WSPM) problem:

$$\text{(P2-1)} \quad \min_{\mathbf{w}_{ij}} \quad \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij}^{(p)} \|\mathbf{w}_{ij}\|_{2,1}$$
$$\text{s.t.} \quad \bar{c}_i \le B_i \log(1 + \text{SINR}_i),\ \forall i \in \mathcal{N}, \tag{16}$$
$$\sum_{i=1}^{N}\|\mathbf{w}_{ij}\|_{2,1} \le E_j,\ \forall j \in \mathcal{L},$$

where $\bar{c}_i = g_i(\bar{m})$ is a constant. Since phase rotation of $\mathbf{w}_{ij}$ does not affect problem (P2-1), we can recast constraint (16) as the following second-order cone (SOC) [42], [43]:

$$\|\mathbf{r}_i\|_2 \le \sqrt{1 + 1/(2^{\frac{\bar{c}_i}{B_i}} - 1)}\,\mathfrak{Re}[R_{ii}],\ \forall i \in \mathcal{N}, \tag{17}$$

where $R_{ik} = \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}$, $\mathbf{r}_i = [R_{i1}, \cdots, R_{iN}, \sigma_i]^T$ and $\mathfrak{Re}(\cdot)$ stands for the real part of a complex number. Thus, problem (P2-1) can be reformulated as a second-order cone programming (SOCP), which can be easily solved by interior point method with standard optimization tool boxes like CVX [44].

Since $\bar{c}_i = g_i(\bar{m})$ is decreasing in $\bar{m}$, the optimal objective function value in (P2-1) is non-increasing in $\bar{m}$. Therefore, a straightforward approach to obtain an optimal solution for problem (P2) is to first perform a search for the optimal $m^* \in \mathbb{N}$ in the following interval

$$\left[\max_{i\in\mathcal{N}} \lceil m_i \rceil,\ M\right], \tag{18}$$

which minimizes the optimal objective function value in problem (P2). This IS approach can obtain the global optimal solution for problem (P2).

B. Joint optimization with integer recovery (JR) approach

If the number of available VMs M is very large, the aforedescribed integer search algorithm may not be applicable. For the case with large M, we can relax m from nature numbers to non-negative real numbers, i.e., $m \in \mathbb{R}^+$.

For the received signal $\hat{u}_i$ at UE i, let $y_i \in \mathbb{C}$ be the receive beamformer. Then, the mean square error (MSE) e_i is defined as [11]:

$$\begin{aligned} e_i &\triangleq \mathbb{E}\left[\left\|y_i^H \hat{u}_i - u_i\right\|_2^2\right] \\ &= \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} - 1\right|^2 + \sum_{l\neq i}^{N} \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj}\right|^2 + \sigma_i^2 \left|y_i\right|^2 \\ &= \sum_{l=1}^{N} \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj}\right|^2 - 2\mathfrak{Re}\left[y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}\right] + \sigma_i^2 \left|y_i\right|^2 \\ &\quad + 1. \end{aligned} \tag{19}$$

Hence, for a given transmit beamformer $\mathbf{w}_{ij}$, the minimum mean square error (MMSE) is

$$e_i^{\text{mmse}} = 1 - \sum_{j\in\mathcal{L}} \mathbf{w}_{ij}^H \mathbf{h}_{ij} y_i^{\text{mmse}}, \tag{20}$$

where y_i^{mmse} is the well-known MMSE receive beamformer given by

$$y_i^{\text{mmse}} = \frac{\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}}{\sum_{k\in\mathcal{N}} \left(\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}\right)\left(\sum_{j\in\mathcal{L}} \mathbf{w}_{kj}^H \mathbf{h}_{ij}\right) + \sigma_i^2}. \tag{21}$$

Lemma 1: Each UE i's achievable rate $B_i \log\left(1 + \text{SINR}_i\right)$ satisfies the following equation [45], [46]:

$$B_i \log\left(1 + \text{SINR}_i\right) = \max_{x_i, y_i} (\log x_i - x_i e_i + 1) B_i, \tag{22}$$

where $x_i \in \mathbb{R}^+$ is the *MSE weight*.

With Lemma 1, we have the following Proposition.

Proposition 3: For $m \in \mathbb{R}^+$, problem (P2) can be represented as:

$$\begin{aligned} \text{(P2-2)} \quad \min_{x_i, y_i, m, \mathbf{w}_{ij}} \quad & \left(m\varphi + \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij}^{(p)} \left\|\mathbf{w}_{ij}\right\|_{2,1}\right) \\ \text{s.t.} \quad & g_i(m) \leq (\log x_i - x_i e_i + 1) B_i, \\ & \qquad \forall i \in \mathcal{N}, \qquad (23) \\ & \max_{i\in\mathcal{N}} \lceil m_i \rceil \leq m \leq M, \ m \in \mathbb{R}^+, \qquad (24) \\ & \sum_{i=1}^{N} \left\|\mathbf{w}_{ij}\right\|_{2,1} \leq E_j, \ \forall j \in \mathcal{L}, \end{aligned}$$

where $g_i(m)$ and e_i are given by (14) and (19) respectively. Let the optimal solution for problem (P2-2) be $\{(\tilde{m}, \tilde{c}_i, \tilde{\mathbf{w}}_{ij}) : i \in \mathcal{N}, j \in \mathcal{L}\}$.

We partition the variables in problem (P2-2) into three groups, i.e., x_i, y_i and $\{m, \mathbf{w}_{ij}\}$. The reasons that we introduce two new groups of variables x_i and y_i in problem (P2-2) are:

- If we fix x_i and y_i, then problem (P2-2) can be easily recast as a convex optimization problem w.r.t. $\{m, \mathbf{w}_{ij}\}$, since $g_i(m)$ is a convex function.
- For the right hand side of (23), by checking the first order optimality condition, optimal receive beamformers y_i can be obtained using (21).
- The optimal MSE weight x_i in the right hand side of (23) is given by

$$x_i = (e_i^{\text{mmse}})^{-1}, \tag{25}$$

for fixed $\{m, \mathbf{w}_{ij}\}$ and y_i.

Problem (P2-2) is convex w.r.t. each variable group while keeping other variable groups fixed. Therefore problem (P2-2) is much easier to solve using an alternating optimization approach than problem (P2). Specifically, in problem (P2-2), the optimal y_i and x_i can be obtained with closed-form solutions if we fix the other variable groups as constants, and $\{m, \mathbf{w}_{ij}\}$ can be obtained by solving the following convex optimization problem, for fixed x_i and y_i:

$$\begin{aligned} \text{(P2-2.1)} \quad \min_{m, \mathbf{w}_{ij}} \quad & m\varphi + \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij}^{(p)} \left\|\mathbf{w}_{ij}\right\|_{2,1} \\ \text{s.t.} \quad & g_i(m) + B_i x_i e_i \leq B_i(\log x_i + 1), \\ & \qquad \forall i \in \mathcal{N}, \qquad (26) \\ & (24) \text{ and } (8). \end{aligned}$$

A local optimal solution for problem (P2-2) can be achieved by this alternating optimization procedure.

Problem (P2-2.1) can be solved by applying the interior point method. However, to reduce the complexity further, we propose the following dual decomposition approach, which obtains the closed-form solution for Problem (P2-2.1).

*1) **Dual decomposition approach**:* The Lagrangian associated with problem (P2-2.1) is (we drop the superscript (p) in this subsubsection)

$$\begin{aligned} \mathscr{L}(m, \mathbf{w}_{ij}, \varepsilon_i, \nu_j) = {} & m\varphi + \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij} \left\|\mathbf{w}_{ij}\right\|_{2,1} \\ & + \sum_{i=1}^{N} \varepsilon_i \left(g_i(m) + B_i x_i e_i - B_i(\log x_i + 1)\right) \\ & + \sum_{j=1}^{L} \nu_j \left(\sum_{i=1}^{N} \left\|\mathbf{w}_{ij}\right\|_{2,1} - E_j\right), \end{aligned} \tag{27}$$

where $\varepsilon_i \geq 0$, $\forall i \in \mathcal{N}$ and $\nu_j \geq 0$, $\forall j \in \mathcal{L}$ are the Lagrange multipliers associated with constraint (26) and (8) respectively[2]. Then, the Lagrange dual function can be written as

$$f(\varepsilon_i, \nu_j) = \min_{m, \mathbf{w}_{ij}} \mathscr{L}(m, \mathbf{w}_{ij}, \varepsilon_i, \nu_j)$$

[2]Constraint (24) can be simply omitted in the Lagrangian, since it is just a constant bound for m.

$$= \min_{m,\mathbf{w}_{ij}} m\varphi + \sum_{i=1}^{N} \varepsilon_i g_i(m) + \sum_{i=1}^{N}\sum_{j=1}^{L} (z_{ij}+\nu_j) \left\|\mathbf{w}_{ij}\right\|_{2,1} + \sum_{i=1}^{N} B_i\varepsilon_i x_i e_i - \sum_{i=1}^{N} \varepsilon_i B_i(\log x_i + 1) - \sum_{j=1}^{L} \nu_j E_j, \quad (28)$$

where m should satisfy (24).

The dual problem is then formulated as

$$\begin{aligned} \max_{\varepsilon_i,\nu_j} \quad & f(\varepsilon_i,\nu_j) \\ \text{s.t.} \quad & \varepsilon_i \geq 0,\ \forall i \in \mathcal{N} \\ & \nu_j \geq 0,\ \forall j \in \mathcal{L}. \end{aligned} \quad (29)$$

To solve the dual problem, we first observe that the Lagrangian (27) is separable over m and $\mathbf{w}_{ij}$. Hence, in the Lagrange dual function, the minimization can be achieved by the following two subproblems:

$$\begin{aligned} \min_{m} \quad & m\varphi + \sum_{i=1}^{N} \varepsilon_i g_i(m) \\ \text{s.t.} \quad & \max_{i\in\mathcal{N}} \lceil m_i \rceil \leq m \leq M,\ m \in \mathbb{R}^+, \end{aligned} \quad (30)$$

and

$$\min_{\mathbf{w}_{ij}} \quad \sum_{i=1}^{N}\sum_{j=1}^{L} (\nu_j + z_{ij}) \left\|\mathbf{w}_{ij}\right\|_{2,1} + \sum_{i=1}^{N} B_i\varepsilon_i x_i e_i, \quad (31)$$

where e_i is given by (19).

Problem (30) can be easily solved numerically. And the closed-form solution for problem (31) can be derived as follows.

Let $\mathbf{w}_i = [\mathbf{w}_{i1}, \mathbf{w}_{i2}, \cdots, \mathbf{w}_{iL}] \in \mathbb{C}^{KL\times 1}$ and $\mathbf{h}_i = [\mathbf{h}_{i1}, \mathbf{h}_{i2}, \cdots, \mathbf{h}_{iL}] \in \mathbb{C}^{KL\times 1}$ be the network-wide beamformer and channel for UE i respectively. Define $\mathbf{A}_j = \{\underbrace{\mathbf{0}_K, \cdots, \mathbf{0}_K}_{j-1}, \mathbf{I}_K, \underbrace{\mathbf{0}_K, \cdots, \mathbf{0}_K}_{L-j}\} \in \mathbb{R}^{K\times KL}$, where $\mathbf{0}_K$ is the $K \times K$ zero matrix and $\mathbf{I}_K$ is the $K \times K$ identity matrix. Then, we have

$$\mathbf{w}_{ij} = \mathbf{A}_j \mathbf{w}_i. \quad (32)$$

Thus, problem (31) is equivalent to

$$\min_{\mathbf{w}_i} \quad \sum_{i=1}^{N} \mathbf{w}_i^H \mathbf{Q_i} \mathbf{w}_i - \sum_{i=1}^{N} 2B_i\varepsilon_i x_i \mathfrak{Re}\left[y_i^H \mathbf{w}_i^H \mathbf{h}_i\right], \quad (33)$$

where $\mathbf{Q_i} = \sum_{j\in\mathcal{L}} (\nu_j + z_{ij}) \mathbf{A}_j^H \mathbf{A}_j + (\sum_{l\in\mathcal{N}} y_l^H \mathbf{h}_l \mathbf{h}_l^H y_l) B_i\varepsilon_i x_i$. And $\mathbf{Q_i}$ can be easily proven as a positive definite matrix. Then, the closed-form solution for problem (33) can be derived as

$$\mathbf{w}_i^* = \mathbf{Q_i}^\dagger \mathfrak{Re}\left[y_i \mathbf{h}_i\right] B_i\varepsilon_i x_i, \quad (34)$$

where $\mathbf{Q_i}^\dagger$ is the pseudo-inverse of $\mathbf{Q_i}$.

Therefore, the dual problem (29) can be solve via the following gradient projection algorithm:

$$\varepsilon_i(t+1) = [\varepsilon_i(t) + \pi_1(t)(g_i(m(t)) + B_i x_i e_i(t) - B_i(\log x_i + 1))]^+, \quad (35)$$

and

$$\nu_j(t+1) = \left[\nu_j(t) + \pi_2(t)\left(\sum_{i=1}^{N} \left\|\mathbf{w}_{ij}(t)\right\|_{2,1} - E_j\right)\right]^+, \quad (36)$$

where $\pi_1(t) > 0$ and $\pi_2(t) > 0$ are step sizes, $m(t)$ denotes the number of active VMs calculated by solving problem (30) in the t-th iteration, $\mathbf{w}_{ij}(t)$ is the beamformer derived from (32) and (34) in the t-th iteration, and

$$e_i(t) = \sum_{l=1}^{N} \left| y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj}(t) \right|^2 - 2\mathfrak{Re}\left[y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}(t) \right] + \sigma_i^2 \left|y_i\right|^2 + 1. \quad (37)$$

*2) **Implementation**:* With dual decomposition and gradient projection, problem (P2-2.1) is ready to be tackled in parallel. Specifically, optimal $m(t)$ and dual variable $\varepsilon_i(t)$ can be calculated by one certain computation resource block in BBU pool, and the optimal $\mathbf{w}_{ij}(t)$ and dual variable $\nu_j(t)$ can be calculated by another computation resource block. Moreover, the parallel computing property of the cloud BBU pool provides a nature environment to implement this parallel computing, and the enormous computation resource in the cloud BBU pool can help solve the problem quickly.

From subsection IV-B and subsection V-B, we see that one approach to solve problem (P2) includes two nested loops: an outer loop to update the weights $z_{ij}^{(p)}$ and an inner loop to solve problem (P2-2) by the iteratively weighted minimum mean square error (WMMSE) method [47], [48]. To reduce the complexity, we can combine the two nested loops together with in a single loop [11], as elaborated in Algorithm 2, in which

$$O^{(p)} = m^{(p)}\varphi + \sum_{i=1}^{N}\sum_{j=1}^{L} z_{ij}^{(p)} \left\|\mathbf{w}_{ij}^{(p)}\right\|_2^2.$$

VI. Numerical results

In this section, we conduct simulations to verify the performance of our proposed algorithms, and compare them with current benchmark algorithms in the literature.

A. Simulation setup

In our simulation, we define the VM cost as its power consumption (in Watts), i.e., $\varphi = k\mu^3$, which is measured by [49], and adopted by [50] and [51]. Here k is a parameter determined by the processor structure, and μ (cycles/s) is the computation capacity of the VM. We choose $k = 10^{-26}$ and $\mu = 10^9$ in our simulation, which is consistent with the measurements in [52]. Moreover, we assume that, for each data byte arrives at the BBU pool, 1900 processor cycles are needed to finish its baseband (cloud) processing [53], and mean packet size is 1000 bytes.

We consider a C-RAN system of 3 RRHs, where RRH 1 to 3 are located on a circle with radius 0.5 km. The 3 RRHs

Algorithm 2 Joint reweighted $l_{2,1}$-norm relaxation and iteratively WMMSE approach for problem (P2)

1: Initialize: $\mathbf{w}_{ij}^{(0)}$ and $p = 1$.
2: **while** $\left|O^{(p)} - O^{(p-1)}\right| > \xi$ **do**
3: Given $\mathbf{w}_{ij}^{(p-1)}$, obtain receive beamformer $y_i^{(p)}$ by (21);
4: Fix $\mathbf{w}_{ij}^{(p-1)}$ and $y_i^{(p)}$, obtain the MSE weight $x_i^{(p)}$ from (20) and (25);
5: Given $x_i^{(p)}$, $y_i^{(p)}$ and $z_{ij}^{(p)}$, utilize the proposed low-complexity dual decomposition approach to solve the convex optimization problem (P2-2.1), obtaining the number of active VMs $m^{(p)}$ and transmit beamformer $\mathbf{w}_{ij}^{(p)}$;
6: Update $z_{ij}^{(p)}$;
7: Let $p = p + 1$.
8: **end while**
9: Integer recovery: Set $\tilde{m} = m^{(p)}$. Therefore, m^* is chosen from $\{\lfloor \tilde{m} \rfloor, \lceil \tilde{m} \rceil\}$ to minimize the optimal objective function value of (P2). Then $\mathbf{w}_{ij}^*$ can be obtained by solving problem (P2-1) with $\bar{c}_i = g_i(m^*)$.
10: Output: $\{(m^*, c_i^*, \mathbf{w}_{ij}^*) : i \in \mathcal{N}, j \in \mathcal{L}\}$.

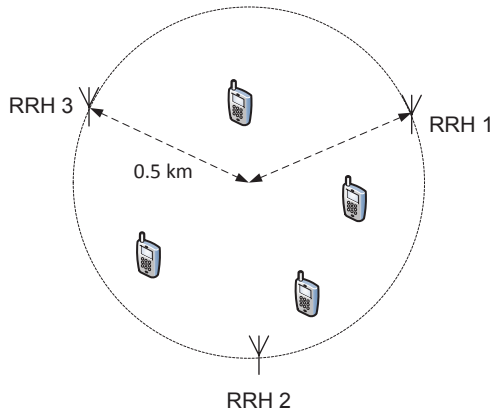


Fig. 4. Simulation topology.

are placed at equal distances apart, as shown in Figure 4. UEs are randomly, uniformly and independently distributed within this disk. The wireless transmission bandwidth is 10 MHz for each UE. We adopt the path loss model used by the 3GPP specification for Evolved Universal Terrestrial Radio Access in [54], where the received power at a UE d km from a RRH is given by

$$p \text{ (dB)} = 128.1 + 37.6 \log_{10} d.$$

The transmit antenna gain at each RRH is ϑ. The lognormal shadowing standard deviation is s.

In our simulations, we consider homogeneous RRHs with $E_1 = E_2 = \cdots = E_L = E$, $P_1 = P_2 = \cdots = P_L = P$, and $S_1 = S_2 = \cdots = S_L = S$. We also consider homogeneous UEs with $\sigma_1 = \sigma_2 = \cdots = \sigma_N = \sigma$, $\lambda_1 = \cdots = \lambda_N = \lambda$, and $\tau_1 = \tau_2 = \cdots = \tau_N = \tau$. We summarize our default simulation parameters in Table I, if not specified.

B. The optimality of price adjusting

It is noted that once the sparsity of the beamforming vector is obtained, the association relationship between the RRHs and the UEs is also determined. Specifically, $\left\|\mathbf{w}_{ij}^*\right\|_2 > 0$ if and only if UE i is associated with RRH j. In Section IV, we obtain the number of active VMs and the sparse beamforming vectors by solving problem (P1). Hence, the sparsity of the beamforming vectors, or the RRH-UE associations, can be achieved by our proposed price adjusting (PA) algorithm. To show the optimality of our proposed PA algorithm, we utilize the following benchmark algorithms:

- Static Clustering (SC). This is proposed by the authors in [11], which obtains the sparse beamforming vectors in two steps:
  1) The heuristic Algorithm 3 in [11] is used to obtain the UE set $\mathcal{N}_j$ that associates with RRH j, such that $\left\|\mathbf{w}_{ij}^*\right\|_2 > 0$ for $i \in \mathcal{N}_j$ and $\left\|\mathbf{w}_{ij}^*\right\|_2 = 0$ for $i \in \mathcal{N}_j{}^c$, where $\mathcal{N}_j{}^c$ is the complementary set of $\mathcal{N}_j$.
  2) Based on the given user association set $\mathcal{N}_j$, a solution to the following problem is found:

$$\begin{aligned} \text{(P0-S)} \quad \min_{m, c_i, \mathbf{w}_{ij}} \quad & m\varphi + \eta \sum_{i=1}^{N} \sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,1} \\ & + \sum_{i=1}^{N} \sum_{j=1}^{L} P_j \|\mathbf{w}_{ij}\|_{2,0} \\ \text{s.t.} \quad & (4'),\ (5),\ (6),\ (7) \text{ and } (8), \\ & \|\mathbf{w}_{ij}\|_2 = 0,\ \forall i \in \mathcal{N}_j{}^c, j \in \mathcal{L}. \end{aligned}$$

  Problem (P0-S) is very similar with problem (P2), and, hence, can be solved by either IS or JR approach.
- Exhaustive Search (ES). This algorithm aims to find out the best association by searching all possible RRH-UE associations. It has an extremely high complexity. In the worst case, the number of possible RRH-UE association relationship can be $\prod_{j=1}^{L} \left(\sum_{i=0}^{S_j} \binom{N}{i} \right)$, where $\binom{N}{i}$ stands for the number of i-combinations of a set with N elements. For each possible RRH-UE association case, we denote $\mathcal{L}_i$ as the RRH set that serves UE i, and use $\mathcal{L}_i{}^c$ as the complementary set of $\mathcal{L}_i$. The following problem (similar with problem (P0-S)) is needed to be solved (by IS or JR):

$$\begin{aligned} \text{(P0-E)} \quad \min_{m, c_i, \mathbf{w}_{ij}} \quad & m\varphi + \eta \sum_{i=1}^{N} \sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_{2,1} \\ & + \sum_{i=1}^{N} \sum_{j=1}^{L} P_j \|\mathbf{w}_{ij}\|_{2,0} \\ \text{s.t.} \quad & (4'),\ (5),\ (6),\ (7) \text{ and } (8), \\ & \|\mathbf{w}_{ij}\|_2 = 0,\ \forall j \in \mathcal{L}_i{}^c, i \in \mathcal{N}. \end{aligned}$$

- Closest Clustering (CC). This algorithm simply assumes each UE only associates with one RRH which provides the best channel gain. Specifically, the RRH associates with UE i is determined by

$$\mathcal{L}_i = \arg\max_j \|\mathbf{h}_{ij}\|_2 . \tag{38}$$

TABLE I
SIMULATION PARAMETERS

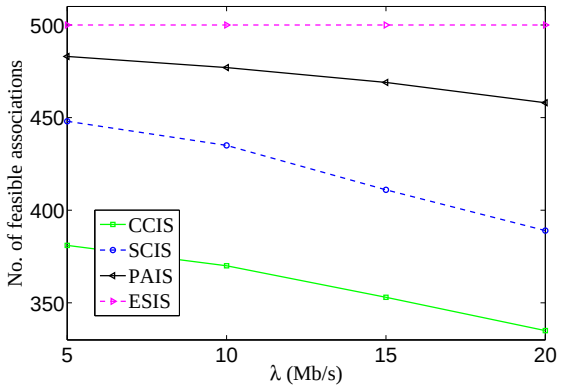
Parameter	Value	Parameter	Value
Computation capacity μ	1×10^9 cycles/s	Weight η	5×10^4
Static cost P	5	Number of UEs N	4
Maximum number of VMs M	25	VM cost φ	10
Fronthaul capacity S	2	Number of antennas K	2
Maximum transmitting power E	1 W	QoS requirement τ	50 ms
White noise power density σ^2	-174 dBm/Hz	Mean arrival rate λ	10 Mb/s
Transmit antenna gain ϑ	15 dBi	Lognormal shadowing s	10 dB

However, this algorithm may easily obtain *infeasible associations*, i.e., the number of UEs associated with certain RRH is more than its fronthaul capacity S. For any feasible RRH-UE association case, problem (P0-E) is also needed to be solved, where $\mathcal{L}_i$ is calculated by (38).

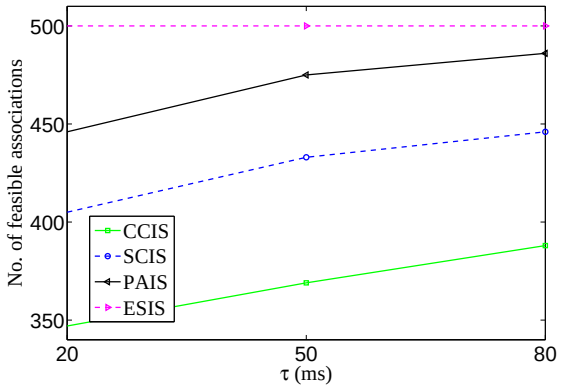
As we can learn from Section V, IS always performs better than JR (since IS obtains the global optimal for problem (P2), while JR only obtains the local optimal). In this subsection, to identify the performance gap between PA and its benchmark algorithms above, we only utilize IS to solve problem (P2) (and its similar problems (P0-S) and (P0-E)).

We show the number of feasible associations in Figure 5 under 500 channel realizations. Specifically, we fix $\tau = 50$ (ms) in Figure 5(a) and increase λ gradually. While in Figure 5(b), λ is set as 10 Mb/s and τ is varied. We can conclude that CC and SC algorithms are unable to guarantee feasibility when λ becomes high and τ becomes low because that the user association sets obtained from CC and SC algorithms are oblivious to UEs' incoming traffic rates and their QoS requirements.

In Figure 6, we present the system cost with different traffic rate, different QoS requirements, and different η values respectively, under different algorithms. We observe that, firstly, PA algorithm outperforms the CC and SC algorithms. Secondly, PA algorithm has much lower complexity than ES algorithm, but still has close performance with ES algorithm.



(a) Different arrival rates.



(b) Different system delay constraints.

Fig. 5. Feasibility under different user association algorithms.

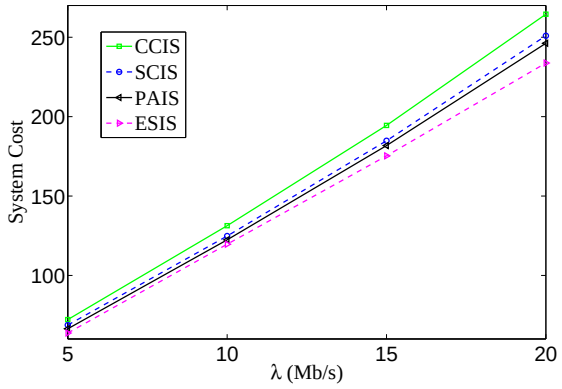
C. Allocations by PAIS

With the "cross-layer" resource allocation in both BBU pool and RRHs, an interesting question is how the allocation works in the system by applying our proposed algorithms. In Figure 7, we present the cost and delay allocations under the QoS requirement as 20 ms, 50 ms and 80 ms respectively (with fixed $\lambda = 10$ (Mb/s)). Leveraging our proposed PAIS algorithm, we can obtain the number of active VMs under these three different QoS requirements as 11, 10 and 10 respectively. And the optimal achievable rate for the UE are 11.33, 10.67 and 10.19 (Mb/s) respectively.

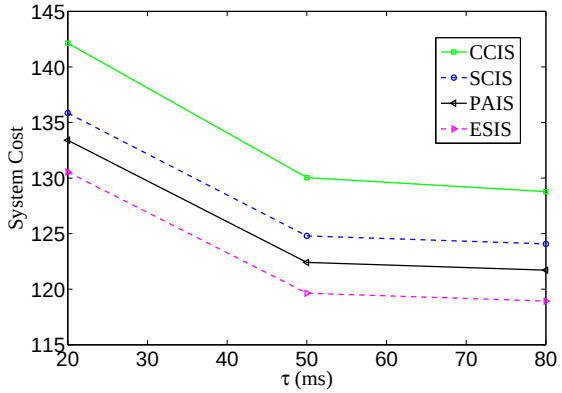
In Figure 7(a), we show the cost of cloud processing (in the BBU pool, i.e., the first term in problem (P0)) and the cost of wireless transmission (in the fronthauls and RRHs, i.e., the second and third terms in problem (P0)). The first interesting observation is, when τ increases from 50 to 80 (ms), the cost of cloud processing remains the same. That because the optimal numbers of VMs for $\tau = 50$ and $\tau = 80$ are identical. However, the cost of wireless transmission decreases strictly with τ. Secondly, the cost of wireless transmission is much lower than the cost of cloud processing (under the parameters in Table I). In Figure 7(b), we show the delay in cloud processing and the delay in wireless transmission. It can be learnt that, interestingly, the delay in wireless transmission strictly increases with τ, while the delay in cloud processing may not vary during some increments of τ.

D. Performance gain by cross-layer design

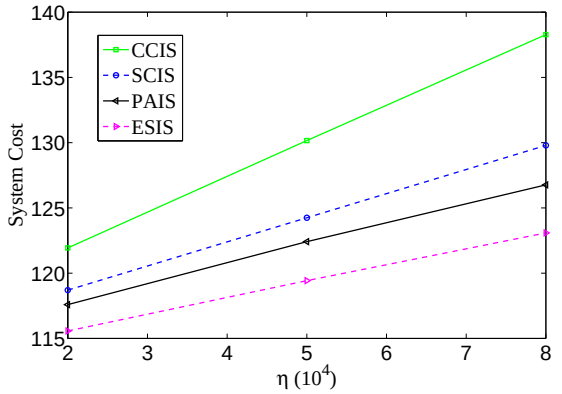
Most of the previous work in C-RAN just optimizes the cost in wireless transmission, without any considerations of the cost in cloud processing, for instance, [11] and [42]. We call those algorithms which consider the cost in wireless transmission and cloud processing independently as decoupled-layer (DL)



(a) Different arrival rates.

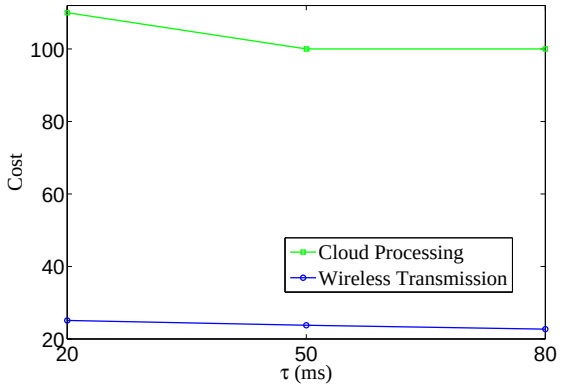


(b) Different QoS requirements.

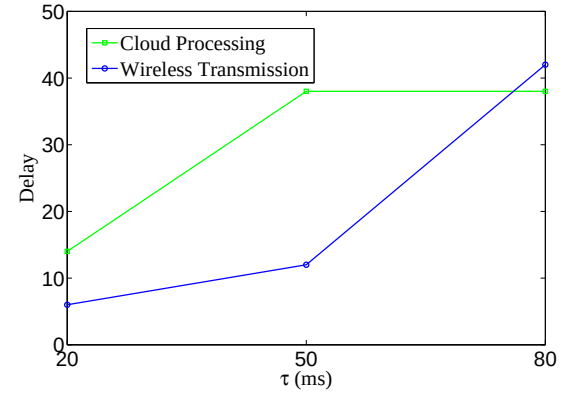


(c) Different values of the weight.

Fig. 6. System cost under different user association algorithms.



(a) Cost allocation.



(b) Delay allocation.

Fig. 7. Allocations in the system.

algorithms. We assume that, for the DL algorithm in our simulations, the delay in the cloud processing queue b_i and the delay in the wireless transmission queue d_i satisfy $b_i \leq \tau_i/2$ and $d_i \leq \tau_i/2$, respectively. Specifically, in our simulations, we use the following DL algorithm:

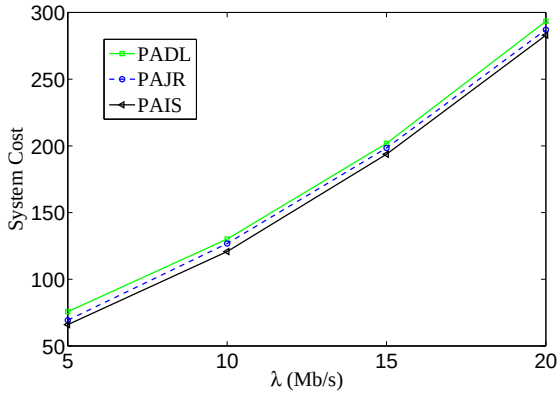
1) we obtain the optimal number of VMs $m^* = \max_{i\in\mathcal{N}} \left\lceil \frac{\tau_i \alpha}{\tau_i \mu - 2} \right\rceil$ from (1);
2) we obtain the optimal beamformers $\mathbf{w}_{ij}^*$ by solving problem (P2-1) with $\bar{c}_i = \lambda_i + \frac{2}{\tau_i}$.

This DL algorithm can be used in tandem with the PA algorithm, in place of the IS and JR algorithms. Hence, in this subsection, we have three algorithms: PADL, PAJR and PAIS.

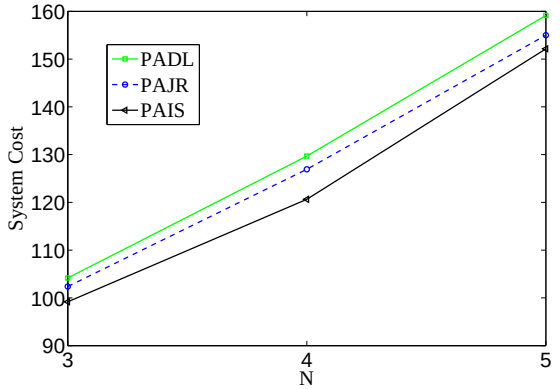
In Figure 8, we present the system cost with different traffic rate and number of UEs under different algorithms. In particular, we show the relationship between UEs' mean arrival rate and system cost for $N = 4$ in Figure 8(a), and the performance of system cost versus the number of UEs is depicted in Figure 8(b) when $\lambda = 10$ Mb/s. We observe that, firstly, JR algorithm have very close performance with IS algorithm. In addition, IS and JR algorithms have lower system cost than DL algorithm, since the optimal delay allocation for cloud processing and wireless transmission is not trivial (as we can learn from Section VI-C). However, DL algorithm always trivially allocates this two delays.

VII. Conclusion

In this paper, we considered the joint VM activation and sparse beamforming problem in C-RAN, which has limited fronthaul capacity. We aim to minimize the system cost of C-RAN, including VM cost (w.r.t. the number of active VMs) in the BBU pool and RRH cost (w.r.t. the beamformer vectors). To tackle the limited fronthaul capacity constraint, we propose a price adjusting algorithm. To find out the the optimal number of VMs, we proposed two different algorithms: integer search and joint optimization. Simulation results suggest that our



(a) Different arrival rates.



(b) Different number of UEs.

Fig. 8. System cost under different algorithms.

proposed algorithms have more robust performance and lower system cost than the benchmark algorithms.

With dense RRH placement in C-RAN, the huge amount of channel state information (CSI) exchange will lead to additional overhead and even cause potential problems for C-RAN. Therefore, in future work, it would be of interest to study effective techniques to reduce CSI overhead. Besides, more practically, we will examine both maximum sum data rate and maximum number of associated UEs as the fronthual capacity constraint.

Appendix A
Proof of Theorem 1

Firstly, if $\beta_j(\Gamma) = S_j$, which implies that $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ is also a feasible solution for problem (P0).

Then, based on $\{m^*, \mathbf{w}_{ij}^*\}$ and $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ are the optimal solutions for problem (P0) and (P1) respectively, we have

$$\begin{aligned}
&m(\Gamma)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}(\Gamma)\|_2^2 + \sum_{j=1}^{L}(P_j+\gamma_j)\beta_j(\Gamma) \\
&\le m^*\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}^*\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L}(P_j+\gamma_j)\left\|\mathbf{w}_{ij}^*\right\|_{2,0} \\
&\le m^*\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}^*\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}^*\right\|_{2,0} + \sum_{j=1}^{L}\gamma_j S_j \\
&\le m(\Gamma)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}(\Gamma)\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\|\mathbf{w}_{ij}(\Gamma)\|_{2,0} \\
&\quad + \sum_{j=1}^{L}\gamma_j S_j,
\end{aligned}$$

where the first inequality is based on that $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ is the optimal solution for problem (P1), the second inequality in based on constraint (9) in problem (P0), the third inequality is based on that $\{c_i^*, \mathbf{w}_{ij}^*\}$ is the optimal solution for problem (P0) and $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ is a feasible solution for problem (P0). Then, let's substitute the equation $\beta_j(\Gamma) = S_j$ into the right hand side of the third inequality above, we can have

$$\begin{aligned}
&m(\Gamma)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\mathbf{w}_{ij}(\Gamma)\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\|\mathbf{w}_{ij}(\Gamma)\|_{2,0} \\
&= m^*\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}^*\right\|_2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}^*\right\|_{2,0} \qquad (39)
\end{aligned}$$

Therefore, the theorem is now proved.

Appendix B
Proof of Proposition 2

Let $\bar{\Gamma}_j' \triangleq [\bar{\gamma}_1', \cdots, \gamma_j', \cdots, \bar{\gamma}_L']^T$ be a different price vector from $\bar{\Gamma}_j$, such that $\gamma_j' > \gamma_j$ and $\bar{\gamma}_k' = \bar{\gamma}_k$, for $k \in \mathcal{L}\backslash j$. We have

$$\begin{aligned}
&m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_{2,0} \\
&\quad + \gamma_j\beta_j(\bar{\Gamma}_j) + \sum_{k\in\mathcal{L}\backslash j}\bar{\gamma}_k\beta_k(\bar{\Gamma}_k) \\
&\le m(\bar{\Gamma}_j')\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j')\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}(\bar{\Gamma}_j')\right\|_{2,0} \\
&\quad + \gamma_j\beta_j(\bar{\Gamma}_j') + \sum_{k\in\mathcal{L}\backslash j}\bar{\gamma}_k\beta_k(\bar{\Gamma}_k'),
\end{aligned}$$

and

$$\begin{aligned}
&m(\bar{\Gamma}_j')\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j')\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}(\bar{\Gamma}_j')\right\|_{2,0} \\
&\quad + \gamma_j'\beta_j(\bar{\Gamma}_j') + \sum_{k\in\mathcal{L}\backslash j}\bar{\gamma}_k'\beta_k(\bar{\Gamma}_k') \\
&\le m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j\left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_{2,0} \\
&\quad + \gamma_j'\beta_j(\bar{\Gamma}_j) + \sum_{k\in\mathcal{L}\backslash j}\bar{\gamma}_k'\beta_k(\bar{\Gamma}_k),
\end{aligned}$$

where the first inequality is based on the assumption that $\{m(\bar{\Gamma}_j), \mathbf{w}_{ij}(\bar{\Gamma}_j)\}$ is the optimal solution for problem (P1), and the second inequality is based on the assumption that $\{m(\bar{\Gamma}_j'), \mathbf{w}_{ij}(\bar{\Gamma}_j')\}$ is the optimal solution for problem (P1).

Adding up both sides of the two inequalities above and simplifying it, we have

$$(\bar{\gamma}_j' - \bar{\gamma}_j)\beta_j(\bar{\Gamma}_j') \leq (\bar{\gamma}_j' - \bar{\gamma}_j)\beta_j(\bar{\Gamma}_j).$$

Hence, the first statement is now proved.

We denote $\{\hat{m}, \hat{\mathbf{w}}_{ij}\}$ as a feasible solution for problem (P0), whose feasible region is nonempty. Then, we have

$$\begin{aligned}
&m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j \left\|\mathbf{w}_{ij}(\bar{\Gamma}_j)\right\|_{2,0} \\
&+ \gamma_j\beta_j(\bar{\Gamma}_j) + \sum_{k\in\mathcal{L}/j} \bar{\gamma}_k\beta_k(\bar{\Gamma}_j) \\
&\leq \hat{m}\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\hat{\mathbf{w}}_{ij}\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j \|\hat{\mathbf{w}}_{ij}\|_{2,0} \\
&+ \gamma_j \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ij}\|_{2,0} + \sum_{k\in\mathcal{L}/j} \bar{\gamma}_k \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ik}\|_{2,0},
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&\beta_j(\bar{\Gamma}_j) - \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ij}\|_{2,0} \\
&< (\hat{m}\varphi + \eta \sum_{i=1}^{N}\sum_{j=1}^{L} \|\hat{\mathbf{w}}_{ij}\|_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{L} P_j \|\hat{\mathbf{w}}_{ij}\|_{2,0} \\
&\quad + \sum_{k\in\mathcal{L}/j} \bar{\gamma}_k \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ik}\|_{2,0})/\gamma_j \\
&\leq (M\varphi + \eta \sum_{j=1}^{L} E_j + \sum_{j=1}^{L} P_j S_j + \sum_{k\in\mathcal{L}/j} \bar{\gamma}_k S_k)/\gamma_j.
\end{aligned}$$

Therefore, if $\gamma_j \geq \theta_j$, then $\beta_j(\bar{\Gamma}_j) - \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ij}\|_{2,0} < 1$. Since $\beta_j(\bar{\Gamma}_j)$ and $\sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ij}\|_{2,0}$ are both integers, then we have $\beta_j(\bar{\Gamma}_j) = \sum_{i=1}^{N} \|\hat{\mathbf{w}}_{ij}\|_{2,0} \leq S_j$.

This completes the proof.

REFERENCES

[1] Cisco Systems, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," Cisco Systems, Technical Report, Feb. 2016.

[2] China Mobile Research Institute, "C-RAN: The road towards green RAN," China Mobile Research Institute, White Paper, V2.5, Oct. 2011.

[3] 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.

[4] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, First Quarter 2015.

[5] 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 Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, Third Quarter 2016.

[6] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge, U.K.: Cambridge University Press, 2016.

[7] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[8] F. Zhuang and V. K. N. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb. 2014.

[9] V. N. Ha, L. B. Le, and N.-D. Dao, "Energy-efficient coordinated transmission for Cloud-RANs: Algorithm design and trade-off," in *Proc. 48th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[10] ——, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints," in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014, pp. 1862–1867.

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

[12] M. Peng, C. Wang, V. K. N. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[13] B. Dai and W. Yu, "Backhaul-aware multicell beamforming for downlink cloud radio access network," in *Proc. IEEE ICC Workshop*, London, U.K., Jun. 2015, pp. 2689–2694.

[14] L. Liu, S. Bi, and R. Zhang, "Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4097–4110, Nov. 2015.

[15] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[16] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[17] T. Vu, H. Nguyen, and T. Q. S. Quek, "Adaptive compression and joint detection for fronthaul uplinks in cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 63, no. 11, pp. 4565–4575, Nov. 2015.

[18] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2063–2077, Apr. 2016.

[19] V. N. Ha, L. B. Le, and N. D. Dao, "Coordinated multipoint transmission design for cloud-RANs with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sep. 2016.

[20] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2063–2077, Apr. 2016.

[21] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4063–4076, Dec. 2016.

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

[23] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-rans," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 4397–4402.

[24] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[25] J. Zhao, T. Q. S. Quek, and Z. Lei, "Heterogeneous cellular networks using wireless backhaul: Fast admission control and large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2128–2143, Oct. 2015.

[26] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE PIMRC*, Washington, DC, USA, Sep. 2014, pp. 1370–1374.

[27] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, p. 907922, Apr. 2016.

[28] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[29] J. Tang and T. Q. S. Quek, "The role of cloud computing in content-centric mobile networking," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 52–59, Aug. 2016.

[30] P. Rost, S. Talarico, and M. C. Valenti, "The complexity-rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164–6176, Nov. 2015.

[31] P. Rost, A. Maeder, M. C. Valenti, and S. Talarico, "Computationally aware sum-rate optimal scheduling for centralized radio access networks," in *Proc. IEEE GLOBECOM*, San Diego, CA, USA, Dec. 2015.

[32] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 189–192, Apr. 2015.

[33] 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.

[34] M. C. Valenti, S. Talarico, and P. Rost, "The role of computational outage in dense cloud-based centralized radio access networks," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 1489–1495.

[35] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[36] P. J. Burke, "The output of a queuing system," *Operations Research*, vol. 4, no. 6, pp. 699–704, Dec. 1956.

[37] E. Reich, "Waiting times when queues are in tandem," *The Annals of Mathematical Statistics*, vol. 28, no. 3, pp. 768–773, Sep. 1957.

[38] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. New Jersey, U.S.: Prentice Hall, 1992.

[39] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l^0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.

[40] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[41] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[42] 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.

[43] A. Wiesel, Y. C. Eldar, and S. S. (Shitz), "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[44] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0," http://cvxr.com/cvx, Aug. 2012.

[45] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max-min fairness," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, Nov. 2011, pp. 1309–1313.

[46] W.-C. Liao, M. Hong, H. Farmanbar, X. Li, Z.-Q. Luo, and H. Zhang, "Min flow rate maximization for software defined radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1282–1294, Jun. 2014.

[47] 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.

[48] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "Towards system cost minimization in cloud radio access network," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2015, pp. 1460–1464.

[49] S. Kaxiras and M. Martonosi, *Computer Architecture Techniques for Power-Efficiency*. Morgan & Claypool, 2008.

[50] J. Tang, W. P. Tay, and Y. Wen, "Dynamic request redirection and elastic service scaling in cloud-centric media networks," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1434–1445, Aug. 2014.

[51] M. H. Chen, B. Liang, and M. Dong, "A semidefinite relaxation approach to mobile cloud offloading with computing access point," in *Proc. IEEE SPAWC*, Stockholm, Sweden, Jun. 2015, pp. 186–190.

[52] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. USENIX HotCloud*, Boston, MA, USA, Jun. 2010, pp. 1–7.

[53] L. Chen and N. Li, "On the interaction between load balancing and speed scaling," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2567–2578, Dec. 2015.

[54] 3GPP, "LTE; Evolved universal terrestrial radio access (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B (release 9)," 3rd Generation Partnership Project (3GPP), TS 36.931, May 2011, v9.0.0.

Jianhua Tang (S'11-M'15) received the B.E. degree in communication engineering from Northeastern University, China, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2015. He was a Post-Doctoral Research Fellow with the Singapore University of Technology and Design from 2015 to 2016. He is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China. Currently, he is a Research Assistant Professor with the Department of Electrical and Computer Engineering, Seoul National University, Korea. His research interests include cloud computing, content-centric network, and cloud radio access network.

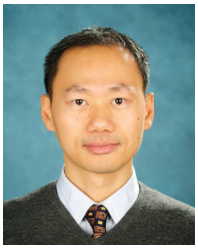
Wee Peng Tay (S'06-M'08-SM'14) received the B.S. degree in Electrical Engineering and Mathematics, and the M.S. degree in Electrical Engineering from Stanford University, Stanford, CA, USA, in 2002. He received the Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently an Assistant Professor in the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore. His research interests include distributed inference and signal processing, sensor networks, social networks, information theory, and applied probability.

Dr. Tay received the Singapore Technologies Scholarship in 1998, the Stanford University President's Award in 1999, the Frederick Emmons Terman Engineering Scholastic Award in 2002, and the Tan Chin Tuan Exchange Fellowship in 2015. He is the coauthor of the best student paper award at the Asilomar conference on Signals, Systems, and Computers in 2012, and coauthor for the IEEE Signal Processing Society Young Author Best Paper Award in 2016. He is currently an Associate Editor for the IEEE Transactions on Signal Processing, serves on the MLSP TC of the IEEE Signal Processing Society, and is the chair of DSNIG in IEEE MMTC. He has also served as a technical program committee member for various international conferences.

Tony Q.S. Quek (S'98-M'08-SM'12) received the B.E. and M.E. degrees in Electrical and Electronics Engineering from Tokyo Institute of Technology, respectively. At MIT, he earned the Ph.D. in Electrical Engineering and Computer Science. Currently, he is a tenured Associate Professor with the Singapore University of Technology and Design (SUTD). He also serves as the Associate Head of ISTD Pillar and the Deputy Director of the SUTD-ZJU IDEA. His main research interests are the application of mathematical, optimization, and statistical theories to communication, networking, signal processing, and resource allocation problems. Specific current research topics include heterogeneous networks, wireless security, internet-of-things, and big data processing.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is serving as the Workshop Chair for IEEE Globecom in 2017, the Tutorial Chair for the IEEE ICCC in 2017, and the Special Session Chair for IEEE SPAWC in 2017. He is currently an elected member of IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS. He is a co-author of the book "Small Cell Networks: Deployment, PHY Techniques, and Resource Allocation" published by Cambridge University Press in 2013 and the book "Cloud Radio Access Networks: Principles, Technologies, and Applications" by Cambridge University Press in 2017.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE Globecom 2010 Best Paper Award, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, the IEEE WCSP 2014 Best Paper Award, the 2015 SUTD Outstanding Education Awards – Excellence in Research, the 2016 Thomson Reuters Highly Cited Researcher, and the 2016 IEEE Signal Processing Society Young Author Best Paper Award.



Ben Liang (S'94-M'01-SM'06) received honors-simultaneous B.Sc. (valedictorian) and M.Sc. degrees in Electrical Engineering from Polytechnic University in Brooklyn, New York, in 1997 and the Ph.D. degree in Electrical Engineering with a minor in Computer Science from Cornell University in Ithaca, New York, in 2001. In the 2001 - 2002 academic year, he was a visiting lecturer and post-doctoral research associate with Cornell University. He joined the Department of Electrical and Computer Engineering at the University of Toronto in 2002, where he is now a Professor. His current research interests are in networked systems and mobile communications. He has served as an editor for the IEEE Transactions on Communications since 2014, and he was an editor for the IEEE Transactions on Wireless Communications from 2008 to 2013 and an associate editor for Wiley Security and Communication Networks from 2007 to 2016. He regularly serves on the organizational and technical committees of a number of conferences. He is a senior member of IEEE and a member of ACM and Tau Beta Pi.