Article:
Bsebsu, Ashraf, Zheng, Gan, Lambotharan, Sangarapillai et al. (2020) Joint Beamforming and Admission Control for Cache-Enabled Cloud-RAN with Limited Fronthaul Capacity. Iet signal processing. ISSN: 1751-9675

https://doi.org/10.1049/iet-spr.2019.0247

Joint Beamforming and Admission Control for Cache-Enabled Cloud-RAN with Limited Fronthaul Capacity

Ashraf Bsebsu*, Gan Zheng*, Sangarapillai Lambotharan*, Kanapathippillai Cumanan† and Basil AsSadhan ‡,
*Signal Processing and Networks Research Group, Loughborough University, Loughborough, LE11 3TU, UK
(E-mail: {a.n.bsebsu, g.zheng, s.lambotharan}@lboro.ac.uk)
†Department of Electronic Engineering, University of York, YO10 5DD, UK
(E-mail: kanapathippillai.cumanan@york.ac.uk)
‡Electrical Engineering Department, King Saud University, Saudi Arabia
(E-mail: bsadhan@ksu.edu.sa).

***Abstract*—Caching is a promising solution for the cloud radio access network (Cloud-RAN) to mitigate the traffic load problem in the fronthaul links. Multiuser downlink beamforming plays an important role for efficient utilization of spectrum and transmission power while satisfying the user's quality of service (QoS) requirements. When the number of users exceeds the serving capacity of the network, certain users will have to be dropped or re-scheduled. This is normally achieved by appropriate admission control mechanisms. Introducing local storage or cache at the remote radio heads (RRHs) where some popular contents are cached, we propose beamforming and admission control technique for cache-enabled Cloud-RAN in the downlink. This minimizes the total network cost including power and fronthaul cost while admitting as many users as possible. We formulate this multi-objective optimization problem as a single objective optimization problem. The original problem which is mixed-integer non-linear program (MINLP) is first converted to the mixed-integer second order cone programming form (MI-SOCP). Branch and Bound (BnB) algorithm is then used to determine the optimal and suboptimal solutions. Simulation study has been conducted to assess the performance of both methods.**

***Index Terms*—Cloud-RAN, downlink beamforming, admission control, fronthaul limitation, caching, second order cone programming.**

I. Introduction

Due to emergence of smart devices and high volume data applications, wireless networks are becoming very congested. This has resulted in a need for a heterogenous network architecture with more densely populated access points rather than the traditional single-layer network architecture [1]. However, increasing the number of access points will introduce major challenges in wireless networks, operational cost and management of network interference [2].

The cloud radio access network (Cloud-RAN) is a promising solution to tackle these challenges [3]. In the Cloud-RAN, the remote radio heads (RRHs) are attached to a centralized base band unit (BBU) through the fronthaul links. Base band data processing and acquisition of channel state information (CSI) are performed at the BBU which enables the RRHs to deliver data to users in a cooperative manner. This will allow multiple RRHs to jointly process data and design precoders. However, carrying out fully joint processing requires a significant amount of overhead for the RRHs to share data. This introduces a heavy load on fronthaul links [4]. Using wide bandwidth millimeter-wave frequency for fronthaul connectivity may not necessarily relieve fronthaul load due to continuous increase in the number of users and demand on high data rate applications. Hence, the capacity of the link between BBU and RRH will still be an issue, particularly for long distance and in non-line of sight environment.

One way to mitigate the fronthaul capacity requirements, is to serve each user by a cluster of RRHs cooperatively through joint precoding. In this case, the data for each user will only be transferred to a particular RRH from the BBU rather than that being distributed to all of the RRHs. As a result, a significant reduction in fronthaul load can be achieved [5].

Another way to reduce the fronthaul capacity requirement, is to use the local storage at the RRHs, where some popular contents can be cached. Getting the contents closer to the user or keeping them at the user terminal directly through broadcasting will enhance users' perceived experience in addition to minimizing the network congestion [6], [7], [8]. Caching some popular contents at the local storage in RRHs or user terminals during off peak time [9], [10], will significantly improve the network throughput and delay. Hence, caching and RRH clustering are promising techniques to reduce the fronthaul load in the Cloud-RAN. As the number of deployed RRHs increases in the Cloud-RAN, employing downlink beamforming will further enhance spectral and energy efficiency.

For the works related to energy efficient techniques in the literature, a weighted mixed norm minimization was proposed to enhance energy efficiency of the downlink beamforming in a Cloud-RAN in [11]. The authors in [12] investigated

The authors extend their appreciation to the Engineering and Physical Sciences Research Council for the support of this work through grants EP/R006385/1 and EP/N007840/1 and the International Scientific Partnership Program (ISPP) at King Saud University through grant ISPP134.

multiuser-access point (MU-AP) association and beamforming design for both the downlink and uplink power optimization. A multi-stage algorithm based on the group-sparsity was proposed to minimize the power consumption for multicast Cloud-RANs in [13]. The authors in [14] proposed an energy-saving mechanism for Cloud-RANs based on the formation of virtual base station. However, these works have not considered the fronthaul limitation. The work in [15] proposed low-complexity algorithms to minimize transmission power under the fronthaul link constraint. In [16], the authors proposed low-complexity algorithms to jointly optimize the data assignment and the transmit power minimization for beamforming in backhaul limited caching networks. The authors in [17] investigated the issues of user scheduling and beamforming for energy efficient Fog-RAN. A cache-enabled Cloud-RAN was introduced in [18] to enhance the network performance while reducing the fronthaul cost through content-centric base station clustering and beamforming approach.

The works discussed above do not consider admission control (AC) which is crucial for enhancing spectral efficiency in multi-antenna aided Cloud-RANs. In practice, the number of users in a particular frequency band may exceed the number of antennas which will force some users to be dropped or re-scheduled. The work in [19] proposed to maximize admitted users under the power constraint by jointly designing beamformers and AC. This problem is Nondeterministic Polynomial (NP) hard. However, it has been converted to a convex problem based on semidifinite relaxation and approximations. The authors in [20] proposed a holistic sparse optimization framework which considered power minimization and user AC for the multicast Cloud-RAN.
The authors in [21] developed a two-stage algorithm aiming to maximize the power efficiency by jointly optimizing the admitted users and cooperative beamformers. Three main approaches were presented in [22] to jointly optimize the admitted users and cooperative beamformers in heterogenous network based on the level of coordinations between macro and femto base stations. The authors in [23] studied the joint coordinated beamforming and AC for fronthaul constrained Cloud-RANs by formulating the optimization problem as a single-stage semidifinite program (SDP). However, joint multiuser downlink beamforming and AC with cache-enabled Cloud-RAN has not been considered in the literature.

The main focus of this paper is the network cost minimization which includes both the cost of transmission power and fronthaul capacity, by taking into account caching in Cloud-RAN design. The issue of users seeking access exceeding the limited network resources is tackled through our proposed joint downlink beamforming and AC (JBAC) technique. The contributions of this work are summarized as follows:

- We propose joint beamforming and AC for the cache-enabled Cloud-RAN with limited fronthaul capacity. In particular network cost under multiple constraints such as quality of service (QoS) requirement, fronthaul limitation and transmission power is optimized while admitting as many users as possible.
- The JBAC is a combinatorial problem which is NP hard and non-convex. To reduce the complexity of the problem, we first formulate the problem into a mixed-integer second order cone programming (MI-SOCP) with constraint relaxations [24] which makes our proposed problem formulation different from that of [23].
- We then develop a branch and bound (BnB) method [25] to obtain the optimal solution of the MI-SOCP. The proposed mixed-integer programming optimally selects the RRHs and designs the corresponding beamformers. To reduce the complexity further, a suboptimal BnB method is proposed.

We organize this paper as follows. Section II presents the network model and assumptions. The JBAC problem is formulated in Section III. In Section IV, we convert the original JBAC problem into MI-SOCP and propose the BnB algorithm to obtain the optimal solution. A low complexity suboptimal BnB algorithm is introduced in Section V. Simulation results are provided in Section VI, followed by conclusions in Section VII.

$Notations$: Boldface upper-case and boldface lower-case letters denote matrices and vectors respectively. $\mathbb{R}$ and $\mathbb{C}$ denote respectively the sets of real and complex numbers. $E[.]$ denotes the expected value of a random variable. $\mathcal{CN}(\mu, \sigma^2)$ represents the complex Gaussian distribution with mean μ and variance σ^2. The conjugate transpose of a vector is denoted as $(.)^H$. $\mathbf{0}_l$ and $\mathbf{1}_l$ represent the l-long all zeros and l-long all ones vectors respectively. $Re\{.\}$ and $Im\{.\}$ represent the real and imaginary parts of a complex variable, respectively.

II. NETWORK MODEL AND ASSUMPTION

A. System Model

We consider a downlink transmission of a cache-enabled Cloud-RAN with L RRHs and K users. Each RRH consists of N antennas and each user is equipped with a single antenna as depicted in Figure 1. We consider a set of RRH $\mathcal{L} = \{1, 2, \cdots, L\}$ and a set of users $\mathcal{K} = \{1, 2, \cdots, K\}$. Each RRH is attached to the BBU by a capacity limited fronthaul link. The BBU can access the content server that contains F number of contents. It is also assumed that each RRH has a local storage with a limited size. Each user is served cooperatively by a cluster of RRHs during each time frame.

We define $\mathbf{w}_{l,k} \in \mathbb{C}^{N\times 1}$ as the beamforming vector at the RRH l for transmitting data to user k. The transmit signal of RRH l can be expressed as

$$x_l = \sum_{k=1}^{K} \mathbf{w}_{l,k} s_k, \forall l \in \mathcal{L}, \tag{1}$$

where $s_k \in \mathbb{C}$ denotes the data symbol for user k with unit power, i.e., $E[|s_k|^2] = 1$. The received signal at the user k can be written as:

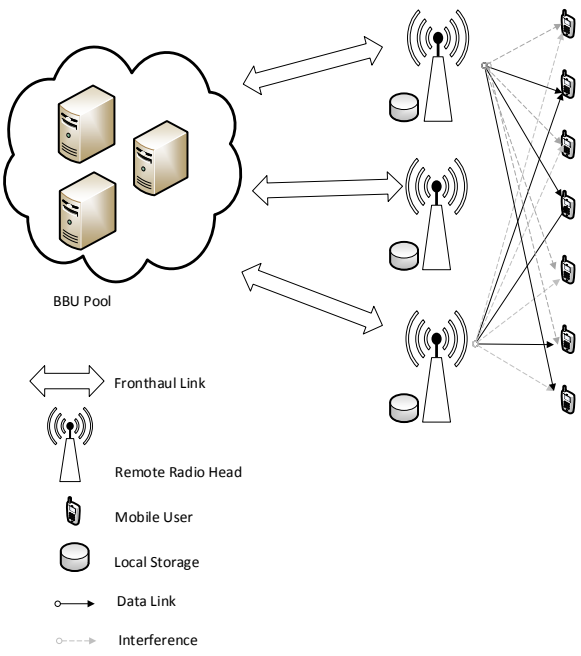

Figure 1: Paradigm of Cache-enabled downlink Cloud-RAN with limited fronthaul.

$$y_k = \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} s_k + \sum_{i=1, i\neq k}^{K} \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i} s_i + n_k, \forall k \in \mathcal{K}, \tag{2}$$

where the noise at the receiver $n_k \in \mathbb{C}$ is distributed according to $\mathcal{CN}(0, \sigma_k^2), k \in \mathcal{K}$, and $\mathbf{h}_{l,k} \in \mathbb{C}^{N\times 1}$ is the channel vector from RRH l to user k. The signal to interference plus noise ratio (SINR) at the receiver of user k is given by

$$SINR_k = \frac{|\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k}|^2}{\sum_{i=1, i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2}, \forall k \in \mathcal{K}. \tag{3}$$

B. Cache Model and Request Model

We consider $\mathcal{F} = \{1, 2, \cdots, F\}$ as a library of F contents. All the contents are of the same normalized size. We denote the local storage size of RRH l as Y_l ($Y_l < F$), which is also the total number of contents that can be stored. Let $\mathbf{C} \in \{0,1\}^{F\times L}$ denote the caching placement matrix where $c_{f,l} = 1$ indicates that the f-th content is stored in the l-th RRH and $c_{f,l} = 0$ otherwise. Due to storage limitation, we have to constrain $\sum_{f=1}^{F} c_{f,l} \leq Y_l$. There is a time scale difference between the long-term caching placement and the short-term transmission, so we consider that cache placement matrix $\mathbf{C}$ is known a priori and fixed.

We define $b_{l,k}$ as the RRH-MU association indicator, where $b_{l,k} = 1$ means that the user k is served by the RRH l and $b_{l,k} = 0$ otherwise. The relation between $b_{l,k}$ and $\mathbf{w}_{l,k}$ is described as:

$$\begin{cases} b_{l,k} = 0 \Leftrightarrow \mathbf{w}_{l,k} = \mathbf{0}, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}, \\ b_{l,k} = 1 \Leftrightarrow \mathbf{w}_{l,k} \neq \mathbf{0}, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}. \end{cases} \tag{4}$$

Define a user requests matrix as $\mathbf{R} \in \mathbb{R}^{F\times K}$, where $r_{f,k} = B_k \log_2(1+\gamma_k)$ if the k-th user requests for the f-th content and demands a target SINR γ_k. Otherwise, $r_{f,k} = 0$, where B_k is the user bandwidth. If the requested content of user k is available in a RRH l, the user can get the content directly from RRH l without relying on the fronthaul. Otherwise, RRH l needs to fetch this content from the BBU via the fronthaul link. Our aim is to maximize delivery of contents from RRH whenever possible and to minimize reliance on BBU. Hence, this has an indirect impact on the delay as the contents are aimed to be delivered by RRHs whenever possible. We assume that a user cannot request more than one content at each scheduled time slot.

C. Network Cost Model

we consider the network cost as the sum of the costs of fronthaul and transmission power. We denote the content file requested by the user k as f_k and model the fronthaul cost as [18]:

$$C_B = \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{f=1}^{F} b_{l,k}(1 - c_{f_k,l}) r_{f,k}. \tag{5}$$

The transmission power cost is modeled using the beamforming vectors of RRHs as

$$C_p = \sum_{l=1}^{L} \sum_{k=1}^{K} \|\mathbf{w}_{l,k}\|_2^2. \tag{6}$$

The total network cost can be expressed as:

$$C_N = C_p + \eta C_B, \tag{7}$$

where η is a weighting factor ($\eta > 0$).

For the purpose of joint processing, we assume that CSI, caching placement matrix and user request matrix are available at the BBU.

III. PROBLEM FORMULATION

We formulate the problem as minimizing the total network cost for the Cloud-RAN while admitting as many users as possible and satisfying transmission power, admitted users' QoS and the fronthaul capacity.

A. Network Cost Minimization

The QoS constraint is written as:

$$SINR_k \geq \gamma_k, \forall k \in \mathcal{K}, \tag{8}$$

where γ_k is the target SINR of user k. Since the rotation of each element of $\mathbf{w}_{l,k}$ by an arbitrary phase angle, i.e., $\mathbf{w}_{l,k}\exp(j\phi)$ for any ϕ, will not have any impact on the network power consumption and the QoS constraints, i.e., (6)

and (8), the SINR constraint can be formulated into a standard SOCP constraint as follows:

$$\sqrt{\sum_{i=1,i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2} \leq$$

$$\frac{1}{\sqrt{\gamma_k}} Re\Big\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} \Big\}, \forall k \in \mathcal{K}, \tag{9a}$$

$$Im\Big\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} \Big\} = 0, \forall k \in \mathcal{K}, . \tag{9b}$$

The transmit power of RRH l is equal to $\sum_{k=1}^{K} \|\mathbf{w}_{l,k}\|_2^2$, hence, the power budget constraint is:

$$\sum_{k=1}^{K} \|\mathbf{w}_{l,k}\|_2^2 \leq P_l^M, \forall l \in \mathcal{L}, \tag{10}$$

where P_l^M represents the power budget of RRH l.

The data transfer rate from database cloud to RRH l according to (5) is equal to $\sum_{k=1}^{K} \sum_{f=1}^{F} b_k(1 - c_{f_k,l}) r_{f,k}$. The fronthaul is constrained as follows:

$$\sum_{k=1}^{K} \sum_{f=1}^{F} b_{l,k}(1 - c_{f_k,l}) r_{f,k} \leq R_l^M, \tag{11}$$

where R_l^M represents the maximum fronthaul capacity link of the RRH l.

For a given set of users, the optimization of cost taking into consideration of QoS, power budget and fronthaul capacity can be formulated as,

$$\mathcal{P}_1 : \min_{\{\mathbf{w}_{l,k}, b_{l,k}\}} \quad C_N \tag{12a}$$

$$s.t. \sqrt{\sum_{i=1,i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2} \leq$$

$$\frac{1}{\sqrt{\gamma_k}} Re\Big\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} \Big\}, \forall k \in \mathcal{K}, \tag{12b}$$

$$Im\Big\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} \Big\} = 0, \forall k \in \mathcal{K}, \tag{12c}$$

$$\sum_{k=1}^{K} \|\mathbf{w}_{l,k}\|_2^2 \leq P_l^M, \forall l \in \mathcal{L}, \tag{12d}$$

$$\sum_{k=1}^{K} \sum_{f=1}^{F} b_{l,k}(1 - c_{f_k,l}) r_{f,k} \leq R_l^M, \tag{12e}$$

$$\|\mathbf{w}_{l,k}\|_2^2 \leq b_{l,k} P_l^M, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}, \tag{12f}$$

$$b_{l,k} \in \{0,1\}, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}. \tag{12g}$$

The QoS constraint (12b) ensures that user k achieves the target SINR γ_k. The total power budget constraint (12d) means that the total transmit power of RRH l is below the maximum transmit power. The fronthaul constraint (12e) indicates that the data transfer rate from database cloud to RRH l is not more than the link capacity. The constraint (12f) is to ensure that when $b_{l,k} = 0$, $\mathbf{w}_{l,k} = 0$. The constraint (12g) indicates that the value of $b_{l,k}$ can only be 0 or 1.

B. AC for Downlink Beamforming

The problem $\mathcal{P}_1$ is MI-SOCP [19], [20], [22], [26]. The problem will turn out to be infeasible if the number of users seeking access exceeds the number of antennas at each RRH or when the data rate requirement exceeds the fronthaul capacity. In this case, only a subset of users will be selected for transmission and the remaining users will be dropped temporarily. This is equivalent to the AC at the physical layer level which is the focus of this paper. However, a higher-level AC will also be required to monitor non-admitted users and allocate higher priority to those users that had been denied access for a longer period due to poor channel conditions etc. Hence, we incorporate the maximization of the number of admitted users to the total network cost minimization problem, and reach a two-stage problem. The first stage is to maximize the admitted users while satisfying the constraints as follows:

$$\mathcal{P}_2 : \max_{\{\mathbf{w}_{l,k}, b_{l,k}, a_k\}} \sum_{k=1}^{K} a_k \tag{13a}$$

$$s.t. \quad \text{(12d)} - \text{(12f)}, \tag{13b}$$

$$SINR_k \geq \gamma_k \frac{a_k + 1}{2}, \forall k \in \mathcal{K}, \tag{13c}$$

$$b_{l,k} \leq \frac{a_k + 1}{2}, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}, \tag{13d}$$

$$\sum_{l=1}^{L} b_{l,k} \geq \frac{a_k + 1}{2}, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}, \tag{13e}$$

$$b_{l,k} \in \{0,1\}, a_k\{-1,1\} \in, \forall l \in \mathcal{L}, \forall k \in \mathcal{K}, \tag{13f}$$

where a_k is the MU access indicator, whereas $a_k = 1$ means that the user k accesses the network and $a_k = -1$ otherwise. The constraints (12f) and (13d) guarantee that when $a_k = -1, \forall k \in \mathcal{K}$ both $b_{l,k} = 0$ and $\mathbf{w}_{l,k} = \mathbf{0}, \forall l \in \mathcal{L}$. i.e., this is formulated such as way that when a user k is not admitted, it sets automatically the corresponding beamformer to zero and the link of this user to RRH to be inactive. In addition, when a_k is 1, the constraint in (13c) will be transformed to $SINR_k \geq \gamma_k$, i.e., an admitted user k should satisfy the target SINR γ_k. However, if $a_k = -1$, the corresponding constraint will turn out to be $SINR_k \geq 0$, i.e., if a user k is not admitted, the corresponding QoS constraint will hold true always, i.e., it will be ignored. If $a_k = 1$, the constraint (13e) means that user k should be served by at least one RRH. However, if $a_k = -1$, i.e. when the user k is not admitted, the constraint (13d) will turn out to be $b_{l,k} \leq 0$ and the constraint (13e) will turn out to be $\sum_{l=1}^{L} b_{l,k} \geq 0$, i.e., it will automatically force $b_{l,k}$ to be zero for all values of l. i.e., none of the RRH will be attached to the user k. The second stage is to solve the

following problem with the selected set of users in the first stage.

$$\mathcal{P}_1 : \min_{\{\mathbf{w}_{l,k}, b_{l,k}, a_k\}} \quad C_N \tag{14a}$$

$$s.t. \qquad (12\text{d}) - (12\text{f}), (13\text{c}) - (13\text{f}). \tag{14b}$$

C. Joint Beamforming and AC (JBAC)

The principle aim of our optimization is to maximize as many admitted users as possible within the constraints of available transmission power (12d) and fronthaul capacity (12e) as formulated in $\mathcal{P}_2$. However, if there is a choice between selection of users, we wish to admit users that reduce demand on fronthaul traffic and RRH transmission power. Hence we have explicitly included cost C_N together with the user maximization in the optimization cost as described in Section C. In case if we do not include C_N, the problem formulation $\mathcal{P}_2$ will maximize admitted users within the constraints of (12d) and (12e). As the aim is to maximize users, it will allocate users until at least one of the constraints (12d) and (12e) is violated. It is very likely that admitted users will be restricted only by one of the constraints, for example, if the fronthaul capacity constraint is very tight compared to the power constraint. If there is a choice between users for example user j or user k, both requiring identical data rate, cost in $\mathcal{P}_2$, will not guarantee choosing the user that require less power assuming both users will satisfy the power limit. In this case, there are multiple solutions for the problem in $\mathcal{P}_2$. However, when both the cost C_N and the user admission are combined into one cost, it ensures that we maximize the number of users while allocating users that demand less fronthaul capacity and transmission power. This will also ensure a unique solution of the problem considered in this manuscript. Hence, by adopting the approach in [19], we convert the two-stage problem to JBAC problem by introducing two small positive constants α, β, which are used as penalty factors for user admission and feasibility guarantee, respectively. The JBAC problem is expressed as:

$$\mathcal{P}_3 : \min_{\{\mathbf{w}_{l,k}, b_{l,k}, a_k\}} \quad \alpha C_N + (1-\alpha)\sum_{k=1}^{k}(a_k - 1)^2 \tag{15a}$$

$$s.t. \qquad (10), (11), (12\text{f}), (13\text{d}) - (13\text{f}), \tag{15b}$$

$$\frac{|\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k}|^2 - \beta^{-1}(a_k - 1)}{\sum_{i=1, i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2} \geq \gamma_k, \forall k \in \mathcal{K}. \tag{15c}$$

The second part of the cost function $\sum_{k=1}^{k}(a_k - 1)^2$ ensures that admitting more users reduces the overall cost function value. When the user k is dropped, we have $a_k = -1$, which ensures that $b_{l,k} = 0, \mathbf{w}_{l,k} = \mathbf{0}, \forall l \in \mathcal{L}$. The QoS requirement constraint (15c) of user k will be satisfied when we use a proper feasibility guarantee factor β. Specifically, when $0 < \beta \leq \min_{k\in K} \frac{2}{\gamma_k(\sum_{l=1}^{L} P_l^M \sum_{l=1}^{L} \|h_{l,k}\|^2 + \sigma_k^2)}$ holds, the problem $\mathcal{P}_3$ is always feasible [19].

Although we reformulate the two-stage problem as a JBAC problem with a simple form $\mathcal{P}_3$, this problem is still non-convex due to the constraint (15c) and discrete values of constraint (13f). In the following, this problem will be converted into a mixed-integer programming problem (MIP).

We first define a vector $\mathbf{b}_k \in \mathbb{R}^{L\times 1}, k = 1, ..., K$, that could take one of the following combinations [27]:

$$\mathbf{b}_k \in \left\{ \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ . \\ . \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ 0 \end{bmatrix}, ..., \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{bmatrix} \right\}. \tag{16}$$

If the l^{th} element of $\mathbf{b}_k$ is one, it represents that the k^{th} user is served by the l^{th} RRH and $a_k = 1$. When all the elements of this vector are zeros, then the k^{th} user is dropped and this is only the case where $a_k = -1$.

We transform the problem (15) to a convex problem by manipulating the constraint (15c) to a more attractive (SOCP) form. The key step is to rewrite (15c) into the following form:

$$\frac{|\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} - \beta^{-1}(a_k - 1)|^2}{\sum_{i=1, i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2} \geq \gamma_k, \forall k \in \mathcal{K}. \tag{17}$$

As stated before, since an arbitrary phase rotation of $\mathbf{w}_{l,k}$, will not affect the transmission power consumption and the QoS constraints, the problem can be written in the form of MI-SOCP as follows:

$$\min_{\{\mathbf{w}_{l,k}, b_{l,k}, a_k\}} \quad \alpha C_N + (1-\alpha)\sum_{k=1}^{k}(a_k - 1)^2 \tag{18a}$$

$$s.t. \qquad (12\text{d}) - (12\text{f}), (13\text{d}) - (13\text{f}), \tag{18b}$$

$$Re\left\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} - \beta^{-1}(a_k - 1) \right\} \geq \sqrt{\gamma_k \left\{ \sum_{i=1, i\neq k}^{K} |\sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,i}|^2 + \sigma_k^2 \right\}}, \tag{18c}$$

$$Im\left\{ \sum_{l=1}^{L} \mathbf{h}_{l,k}^H \mathbf{w}_{l,k} - \beta^{-1}(a_k - 1) \right\} = 0. \tag{18d}$$

The value of β can be determined as $0 < \beta \leq \min_{k\in K} \frac{2}{\sqrt{\gamma_k(\sum_{l=1}^{L} P_l^M \sum_{l=1}^{L} \|h_{l,k}\|^2 + \sigma_k^2)}}$. Furthermore, it can be shown that β satisfies the SINR constraint explicitly when $a_k = -1$. The main advantage of the proposed MI-SOCP formulation over the mixed-integer semidifinite program (MI-SDP) is that it significantly reduces computational complexity.

IV. THE OPTIMAL ALGORITHM BASED ON THE BRANCH AND BOUND METHOD

The JBAC formulation is a combinatorial optimization problem. To obtain the optimal solution for a such a problem, an exhaustive search is generally required. Due to this exhaustive search enumerations increase exponentially with the number of variables which needs more time and more storage for computation. Hence, we propose to use BnB method to solve it. In the sequel, we introduce the BnB method and our

algorithm to solve the MI-SOCP problem in (18) and to obtain beamformers for each RRH. Branching is the first step in the BnB method where the feasible set of the problem is divided into subsets according to various combinations of b_k in (16). The second is bounding step to evaluate the lower bounds of those subproblems. Subsets will be divided into smaller subsets which will create a tree structure. In this method, some of the branches will be removed or pruned according to the following conditions:

1. The branch corresponding to an infeasible subproblem.

2. The branch with an optimal objective value that is above the best-known global objective value of the minimization problem (18).

The global lower bound will be updated at each level only if the current global lower bound is greater than the minimum of the lower bounds of all subsets. We now develop an optimal RRH allocation technique based on the BnB method from the original problem in (18). We first relax the constraints in (13f) as follows:

$$\begin{aligned} &\mathbf{0}_{L\times 1} \leq \mathbf{b}_k \leq \mathbf{1}_{L\times 1}, \qquad k = 1, 2, 3, ..., K, \\ &a_k \in [-1, 1]. \end{aligned} \tag{19a}$$

After the above relaxation, the problem (18) becomes an SOCP problem and can be optimally solved. By solving the problem (18), the lower bound of the original problem will be obtained. If all the elements of $\mathbf{b}_k$ and a_k are integer values, then the problem is deemed to have been solved with the optimal solution. If the problem (18) is infeasible, then the original problem is also infeasible and the algorithm will be terminated. The branching step will start when solving the problem (18) results in non-integer values. The number of branches will be all combinations of values that $\mathbf{b}_k$ and a_k can take. We generate the branches in the first level by allocating the first user to the RRHs and solving the problem (18) to obtain the objective value at each branch. Then we sort all the objective values in descending order and choose the last branch (the branch with the minimum objective value) to proceed to the next level. The rest of the branches will be stored. This process will be repeated until the last level is reached. At this level, the branch with the minimum objective value will be chosen and the objective value of this branch will be designated to a variable $Globalobjectivevalue$. Then we remove all the branches with objective values higher than $Globalobjectivevalue$ at all levels from the algorithm. The branch with the next least objective value in the previous level will be picked up and proceed to the next level until the last level is reached. The objective values of all branches at the last level will be compared to the $Globalobjectivevalue$. If they are higher than $Globalobjectivevalue$, then they will be removed from the algorithm, otherwise, the branch with the minimum objective value at the last level will be assigned to be $Globalobjectivevalue$. We repeat the above procedure until all branches are pruned. The solution will be the variable vector $\mathbf{b}_k$ and a_k of the branches of the paths which are traced back from the minimum objective value of the last level to

Algorithm 1: OPTIMAL ALGORITHM BASED ON BnB METHOD TO SOLVE MI-SOCP

Step 1: Set $Globalobjectivevalue = \infty, level = 0, \mathcal{K} = \{1, 2, ..., K\}, Result = \emptyset, Node = 0, FinalSolution = infeasible$
Step 2: Solve SOCP the problem in (18) with the relaxed integer constraints and obtain the objective value.
if all solution ($\mathbf{b}_k$ and a_k) consist of integer elements **then**
    $FinalSolution \leftarrow [a_1, ..., a_k], \begin{bmatrix} b_{11} & . & . & b_{1k} \\ . & . & . & . \\ . & . & . & . \\ b_{l1} & . & . & b_{lk} \end{bmatrix}$ and go to step 10.
else if Objective value$< Globalobjectivevalue$ **then**
    go to step 3.
else
    go to step 10.
end if
Step 3:
if objective value$\leq Globalobjectivevalue$ **then**
    $m \leftarrow 1, \mathbf{b}_{level} = \mathbf{0}, a_{level} = -1$ and go to step 4.
else
    go to step 9.
end if
Step 4: Update $level = level + 1, levelTemp = \emptyset$ and go to step 5.
Step 5: Generate the branch.
if $m = 1$ **then**
    $\mathbf{b}_{level} = \mathbf{0}, a_{level} = -1$
else
    $\mathbf{b}_{level}$ =Decimal to binary $(m-1), a_{level} = 1$
end if
$Node \leftarrow Node + 1$
$\mathbf{\Gamma}(Node) \leftarrow [\mathbf{\Gamma}(parentNode) \quad a_{level}]$
$\mathbf{\Gamma}_1(Node) \leftarrow [\mathbf{\Gamma}_1(ParentNode) \quad \mathbf{b}_{level}]$ store $\mathbf{\Gamma}(Node)$ and $\mathbf{\Gamma}_1(Node)$ to this branch and got to step 6.
Step 6: Solve SOCP relaxation problem (18) with the values saved at this branch.
if subproblem is feasible **then**

    if $level \neq K$ **then**
        save Objective value to this branch and attach this branch $levelResult$ and go to step 7.
    else
        if objective value$< Globalobjectivevalue$ **then**
            $Globalobjectivevalue \leftarrow$ objective value
            $FinalSolution \leftarrow [a_1, ..., a_k], \begin{bmatrix} b_{11} & . & . & b_{1k} \\ . & . & . & . \\ . & . & . & . \\ b_{l1} & . & . & b_{lk} \end{bmatrix}$
        end if
    end if
else
    remove this branch and go to step 7.
end if
Step 7: Update $q = q + 1$.
if $q \leq 2^L$ **then**
    go to step 5.
else if $level \neq K$ **then**
    go to step 8.
else
    go to step 9.
end if
Step 8: sort $levelResult$ in decreasing order and add them to $Result$. Empty $levelResult$.
Step 9:
if $Result \neq \emptyset$ **then**
    collect the last branch from $Result$ and recall the saved values with this branch and go to step 3.
else
    go to step 10
end if
Step 10: terminate and show the $FinalSolution$.

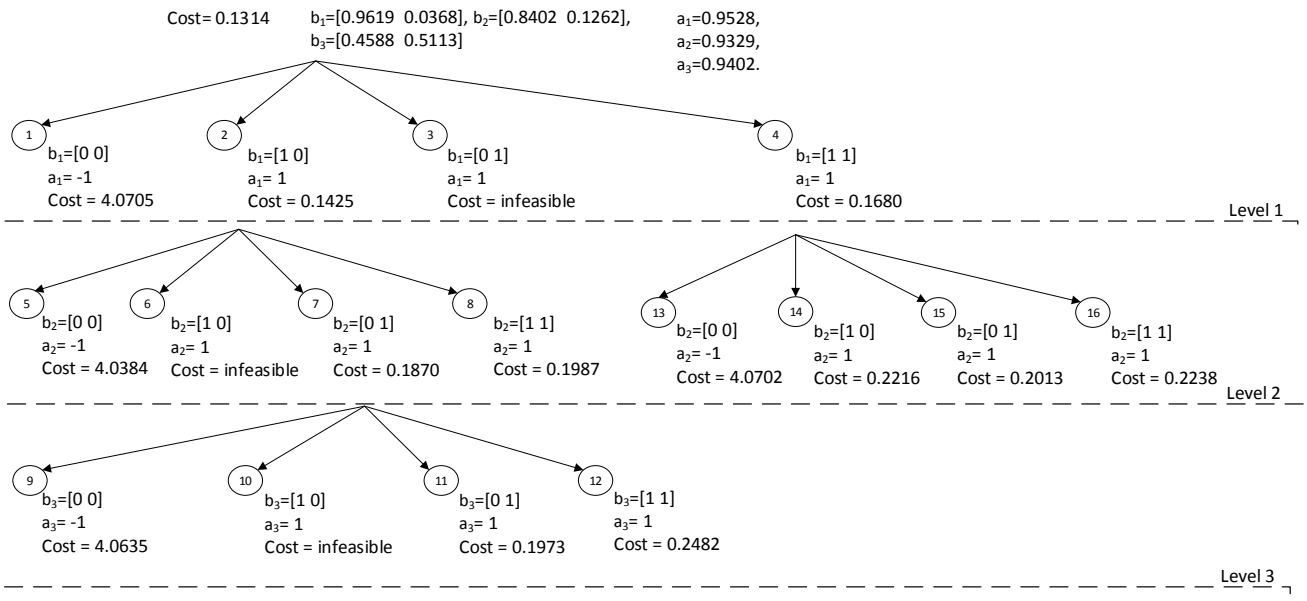

Figure 2: Illustrated model for the proposed BnB based optimal algorithm with two RRHs and three users.

the first level. The computational complexity depends on the number of branches and bounds. The proposed BnB based algorithm is summarized in Algorithm 1.

The algorithm is illustrated using an example in Figure 2 which shows the key steps of the BnB method. There are two RRHs and three users. According to the reference of Algorithm 1, the variables $Globalobjectivevalue, Cost, FinalSolution, Level, \mathcal{K}$, $Result$ and $Node$ are initialized with suitable values. The branch numbers and all levels are depicted in Figure 2. In step 2 of Algorithm 1, the problem (18) is solved with relaxed integer constraints. If the solutions ($\mathbf{b}_k$ and a_k) obtained with the relaxed constraints are integer values, then the algorithm will be terminated at this stage. Otherwise, the algorithm will proceed to the next step (Level 1 in Figure 2). The relaxed problem in (18) will be solved for all possibilities so that the first user can take by settings $\mathbf{b}_1 = [0 \quad 0], a_1 = -1$, i.e., the user 1 is served by neither RRH1 nor RRH2; and $\mathbf{b}_1 = [1 \quad 0], a_1 = 1$, i.e., the user 1 is served by the RRH1; and $\mathbf{b}_1 = [0 \quad 1], a_1 = 1$, i.e., the user 1 is served by the RRH2; and $\mathbf{b}_1 = [1 \quad 1], a_1 = 1$, i.e., the user 1 is served by both RRH1 and RRH2. This generates branch 1, branch 2, branch 3 and branch 4, respectively, as shown in Figure 2 which are steps 4, 5, 6 and 7 of the Algorithm 1. We pick the branch with the minimum objective value (branch 2 in Figure 2) at Level 1 to proceed to the next level (step 8 and 9 in the Algorithm 1). We keep repeating this process at each level until we reach the last level. At the last level (Level 3 in Figure 2), from all the branches created at this level (Level 3), we select the branch with the minimum objective value (branch 11 in Figure 2) and assign the objective value of this branch to the variable $Globalobjectivevalue$.

Next, all the branches at every level with objective values higher than the $Globalobjectivevalue$ will be removed from the algorithm (branch 12, branch 9, branch 8, branch 5 and branch 1 in Figure 2). The next least objective value in the previous levels will be chosen to proceed to the next level (branch 4 in Figure 2). The objective values of branches which are generated from branch 4 at Level 2, i.e., branch 13, branch 14, branch 15 and branch 16 in Figure 2 will be compared to the $Globalobjectivevalue$. If they are higher than the $Globalobjectivevalue$, then they will be removed and the algorithm will be terminated. Otherwise, the minimum objective value branch will be taken to the next level and more branches will be generated until the last level is reached. The procedure will be repeated until all the branches are removed from the algorithm. The solution will be the values of $\mathbf{b}_k, a_k$ of the branches of the path which are traced back from the minimum objective value at the last level to the first level (branch 11, branch 7 and branch 2 in Figure 2). If the last level cannot be reached, i.e., all the branches in a specific level turn out to be infeasible, then the relaxed problem will be solved for $\mathbf{b}_k = [0 \quad 0], a_k = -1$. This means that the k^{th} user will be dropped from the network. Then the algorithm will continue until the last level is reached and the solution will be obtained from other branches as explained above.

V. THE SUBOPTIMAL ALGORITHM BASED ON THE BRANCH AND BOUND METHOD

To reduce complexity, we propose a suboptimal BnB algorithm to solve the relaxed problem in (18). The design of this algorithm is based on the first feasible solution achieved by the optimal algorithm using the BnB method. The SOCP problem (18) with the relaxed integer constrains in (19a) will be solved. If all the elements of $\mathbf{b}_k$ and a_k at the solution are integers, then we consider that the problem has been solved and the algorithm will be ended. If the relaxed problem is infeasible, then the original problem is also infeasible. Branching step will be carried out to generate branches according to all possibilities of vector $\mathbf{b}_k$ and a_k as mentioned in Algorithm 1. The branches in the first level will be generated by allocating the first user to the RRHs and the problem will be solved to obtain the objective value at each branch of every level. Then all objective values are sorted in the descending order and the last branch (the branch with the minimum objective value) will be chosen to proceed to the next level. The rest of the branches will be removed. This algorithm does not need to keep the objective values of all other branches after choosing the branch with the minimum value. This feature is an advantage in terms of reducing the memory requirement during the process. This procedure will be carried out until the last level is reached. At the last level, the branch with the minimum objective value will be chosen and all other branches will be removed. The solution will be the variable vector $\mathbf{b}_k$ and a_k of the branches of the paths which are traced back from the minimum objective value of the last level to the first level. Algorithm 2 summaries the proposed suboptimal BnB algorithm.

The major steps of the suboptimal algorithm can be illustrated using the same example as in Algorithm 1. The branch numbers and all the levels are indicated in Figure 3. The problem (18) is solved with the relaxed integer constraints. If the solutions obtained by setting the relaxed values of constraints are non integer values, then the Algorithm 2 will proceed to the next level (Level 1 in our example). Branches will also be generated considering all the possibilities of $\mathbf{b}_1$ and a_1 (branch 1, branch 2, branch 3 and branch 4 in Figure 3). The branch with the minimum objective value (branch 2 in our example) will be chosen to proceed to the next level. The remaining branches will be removed from the algorithm. This process will be repeated until the last level is reached (Level 3 in Figure 3). At the last level, the branch with the minimum objective value will be chosen and the solution will be obtained from the value of $\mathbf{b}_k$ and a_k of the branches of the path which are traced back from the minimum objective value at the last level to the first level (branch 11, branch 7 and branch 2 in our example).

The complexity of the proposed algorithms is mainly determined by two most important parameters: the convergence speed of the algorithm and the number of arithmetic operations in each iteration. The interior point method proposed in [28] is used to solve the subproblems as the binary variables are relaxed. In the worst-case, the number of subproblems for the

Algorithm 2: SUBOPTIMAL ALGORITHM BASED ON BnB METHOD TO SOLVE MI-SOCP

1: **Step 1:** Set $Globalobjectivevalue = \infty, Level = 0, \mathcal{K} = \{1, 2, ..., K\}, \Gamma = \emptyset, \Gamma_1 = \emptyset, FinalSolution = infeasible.$
2: **Step 2:** Solve SOCP the problem in (18) with the relaxed integer constraints and obtain the objective value.
3: **if** all solution ($\mathbf{b}_k$ and a_k) consist of integer elements **then**
4: $FinalSolution \leftarrow [a_1, ..., a_k], \begin{bmatrix} b_{11} & . & . & b_{1k} \\ . & . & . & . \\ . & . & . & . \\ b_{l1} & . & . & b_{lk} \end{bmatrix}$ and go to step 7.
5: **else if** Objective value$<$ $Globalobjectivevalue$ **then**
6: go to step 3.
7: **else**
8: go to step 7.
9: **end if**
10: **Step 3:** Update $Level = Level + 1, ResultTemp = \emptyset, m = 1, \mathbf{b}_{level} = \mathbf{0}, a_{level} = -1.$
11: **Step 4:** Generate the branch.
12: **if** $m = 1$ **then**
13: $\mathbf{b}_{level} = \mathbf{0}, a_{level} = -1$
14: **else**
15: $\mathbf{b}_{level}$ = Decimal to binary $(m-1), a_{level} = 1$
16: **end if**
17: **Step 5:** Solve SOCP relaxation problem in (18) with the values $\mathbf{\Gamma}, a_{level}$ and $\mathbf{\Gamma}_1, \mathbf{b}_{level}$ saved at this branch.
18: **if** Objective value$<$ $Globalobjectivevalue$ **then**
19: $Globalobjectivevalue \leftarrow$ Objective value
20: **if** $level = K$ **then**
21: $FinalSolution \leftarrow [a_1, ..., a_k], \begin{bmatrix} b_{11} & . & . & b_{1k} \\ . & . & . & . \\ . & . & . & . \\ b_{l1} & . & . & b_{lk} \end{bmatrix}$
22: **else**
23: $\mathbf{\Gamma} \leftarrow a_{level}$
24: $\mathbf{\Gamma}_1 \leftarrow \mathbf{b}_{level}$
25: **end if**
26: **else**
27: remove this branch.
28: **end if**
29: **Step 6:** Update $q = q + 1$.
30: **if** $q \leq 2^L$ **then**
31: go to step 4.
32: **else**
33: go to step 3.
34: **end if**
35: **Step 7:** terminate and show the $FinalSolution$.

K number of users and L number of RRH is $\sum_{k=1}^{K}(2^L)^k$ in case of the optimal solution. Each of these subproblem is an SOCP and consists of KL linear constraints. $O[\sqrt{KL}\log(\frac{1}{\epsilon})]$ iterations are required to converge with ϵ solution accuracy at the termination of the algorithm using interior point method. Each iteration requires at most $O[K^3L^3 + K^2L^2]$ arithmetic operations [29] in the worst-case. For the proposed suboptimal algorithm, only $K(2^L)$ number of subproblems is required to be solved, therefore the complexity is reduced substantially.

VI. SIMULATION RESULTS

A. Network Performance versus Target SINR

For the simulation results, we consider a Cloud-RAN network with three RRHs, (i.e., L = 3), each equipped with N = 2 antennas and K = 6 users. The target SINRs requirement for all the users is identical. The channels between the RRH and the users have been generated using a Rayleigh fading model. Each entry of the channel vector is independently and

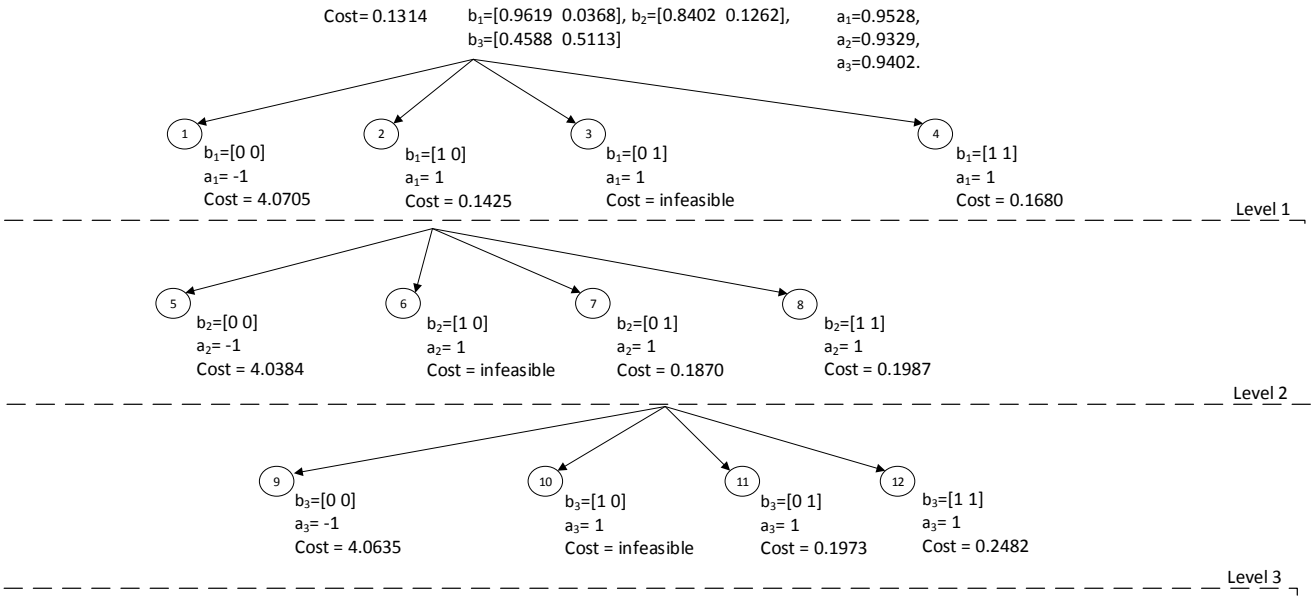

Figure 3: Illustrated model for the proposed BnB based suboptimal algorithm with two RRHs and three users.

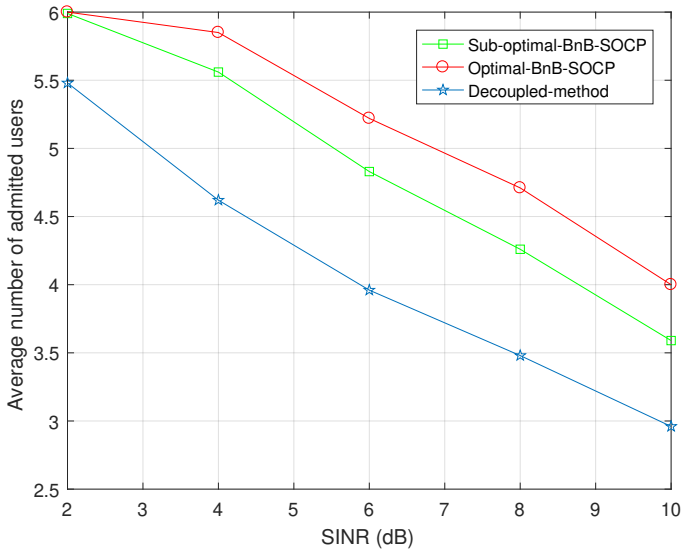

Figure 4: Average number of admitted users versus target SINR for three different schemes, optimal method, sub-optimal method and a simple decoupled design method.

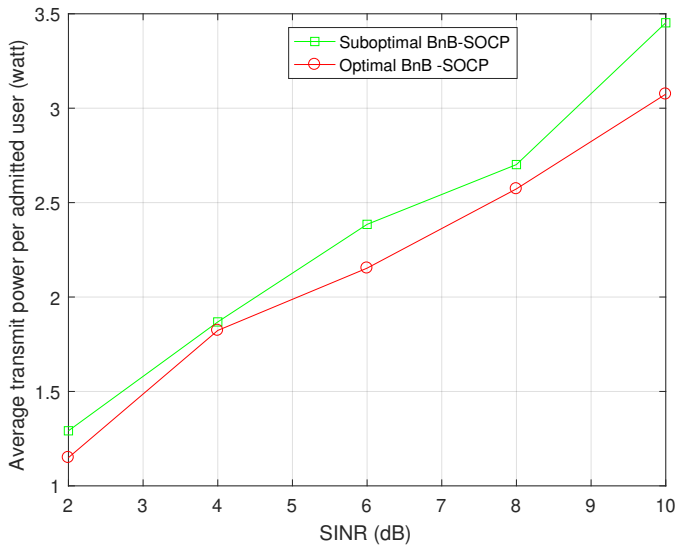

Figure 5: Average transmit power per admitted user versus target SINR.

identically distributed symmetric zero-mean complex Gaussian random variable with variance one. The maximum transmit power at the RRH (P_l^M) is 5 Watt. The available channel bandwidth is 5 MHz and the parameter α is set 0.05. The maximum fronthaul link capacity of RRH l link (R_l^M) is 100 Mbps. We also assume that all the RRHs are equipped with a local storage of equal size (cache) of Y $(Y_l = Y)$. Each user submits contents request independently. Different users cannot request the same content simultaneously. Cache placement matrix is known at RRHs. The constant η in the problem (7) is set to 1. The simulation results are based on 100 Monte Carlo experiments. We evaluate the performance of the two proposed algorithms using MI-SOCP in terms of the average number of admitted users and the average transmission power per admitted user versus the target SINR as shown in Figure 4 and Figure 5, respectively. As expected, the number of admitted users declines as the data rate for the admitted

user increases. Similarly, the transmission power per admitted user increases as the target SINR increases. Although the optimal algorithm is able to allocate more users and consumes less power, the sub-optimal algorithm have the advantage of requiring lower complexity while achieving a comparable performance. The optimal algorithm is able to achieve the best performance because it searches for all possible combinations of $\mathbf{b}_k$ and a_k, where as the suboptimal algorithm searches for the solution through only K iterations. Furthermore, we assessed the performance of our joint optimization algorithms with a decoupled optimization algorithm. We have chosen a decoupled method based on BnB method. In this method, we use maximum ratio transmission (MRT) beamformers for each possible admitted user, then we solve the problem P_2 of user maximization and user association to obtain the optimal number of admitted users and RRHs allocation. Then we perform beamformer design and power allocation only for those set of admitted users. We compared the results using MRT method with our proposed algorithms. The results shows that our proposed algorithms outperform MRT method in terms of the average number of admitted users versus the target SINR as shown in Figure 4. We also evaluated the cost function of the proposed algorithms and MRT method as shown Figure 6. Again our proposed algorithms perform better than MRT method with low cost function. There is a trade-off between user admission and fronthaul cost plus transmission power through the parameter α in the cost function (18). As α increases, the number of admitted user decreases as shown in Figure 7. We have studied the advantage of introducing caching contents at RRH from the simulation results provided in Figure 8. We have studied the impact of varying the fronthaul capacity limit for a given target SINR of 10dB and computing the average number of admitted users. As seen, the average number of admitted users is more than that without caching contents at RRH as the fronthaul capacity has significant impact on the overall system performance limits. Furthermore, as observed in other simulations, we also expect the same trend for the case of optimal BnB SOCP.

In order to assess the running time of the algorithms, we investigated the running time for each algorithm at a certain target SINR of 2dB as shown in Table I. Even though the optimal algorithm provides an attractive user admission solution, it requires very long time for processing, hence sub optimal algorithm provides a compromise between optimal solution and computational complexity.

Table I: Average running time of the algorithms

Method	Time (sec)
MRT-optimal	2.85×10^4
Proposed-optimal	2.26×10^3
Proposed-sub-optimal	1.41×10^2

B. Network Performance versus Total Number of Users

For the simulation results of the proposed optimal BnB and suboptimal algorithm, we consider a Cloud-RAN with two RRHs, (i.e, $L = 2$), each equipped with $N = 2$ antennas. The target SINR requirement for all the users is 6 dB. The rest of the system assumptions have been assumed the same as that in the previous subsection.

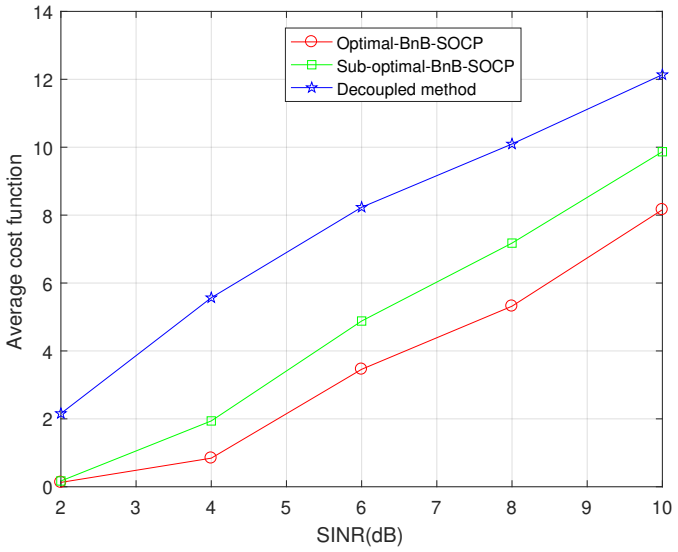

Figure 6: Average cost function versus SINR.

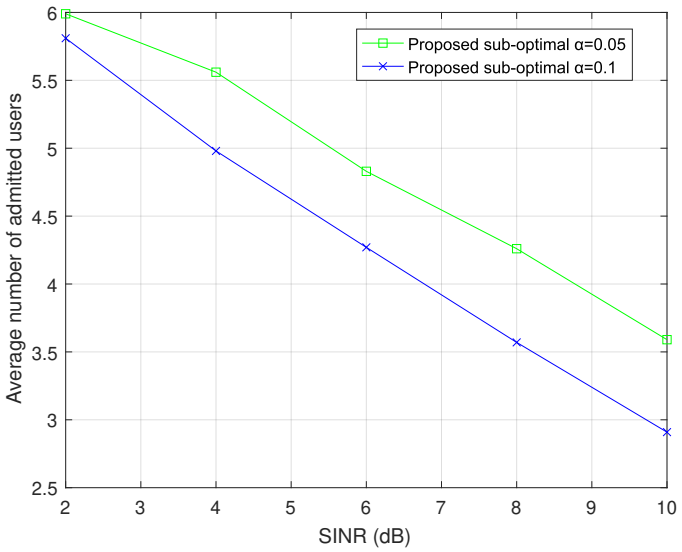

Figure 7: Average number of admitted users versus target SINR of the proposed sub-optimal method for two different network cost - user admission cost combinations.

In Figure 9, we study the trade-off in terms of the average network cost (C_n) and the total number of admitted users. We evaluate the performance of the two proposed algorithms using MI-SOCP in terms of the average number of admitted users and the average transmission power versus the total number of users as shown in Figure 10 and Figure 11 respectively. As seen in Figure 10, with more users deployed in the network, the growth of the average number of admitted users becomes slow, since some users have to be temporarily dropped. Figure

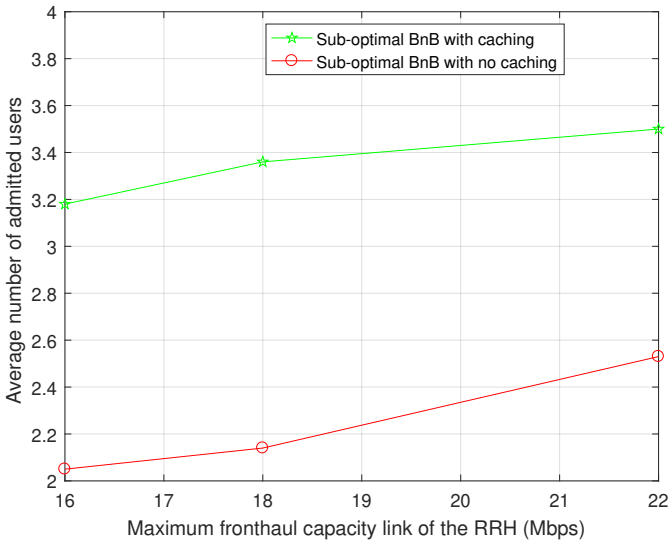

Figure 8: Average number of admitted users versus fronthaul capacity constraint of the RRH when the target SINR is 10dB.

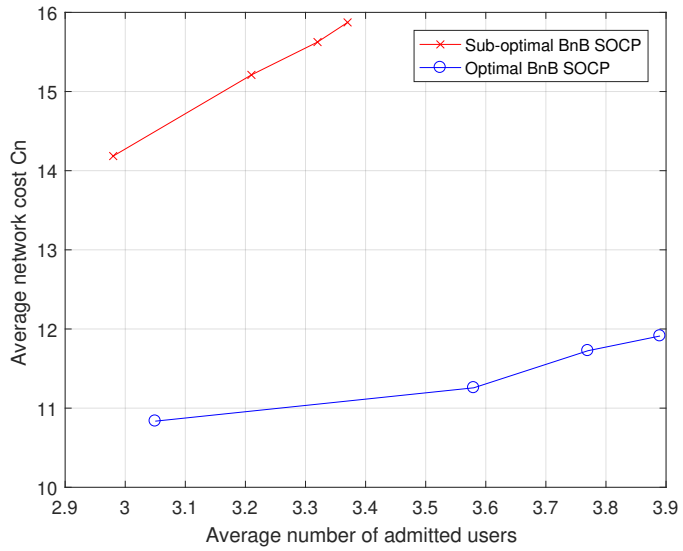

Figure 9: Average network cost versus average number of admitted users when the target SINR is 6dB.

11 depicts the average transmit power versus the total number of users. As seen, the average transmission power increases as the number of users increases.

VII. CONCLUSION

In this paper, we developed a JBAC technique for a cache-enabled Cloud-RAN with limited fronthaul capacity. The one-stage objective function is able to minimize the total network cost including the power cost and the fronthaul cost while admitting as many users as possible under a number of constraints such as QoS for each user, transmit power at RRHs and fronthaul capacity. The problem in its original form is non-convex, and was converted into MI-SOCP by relaxing

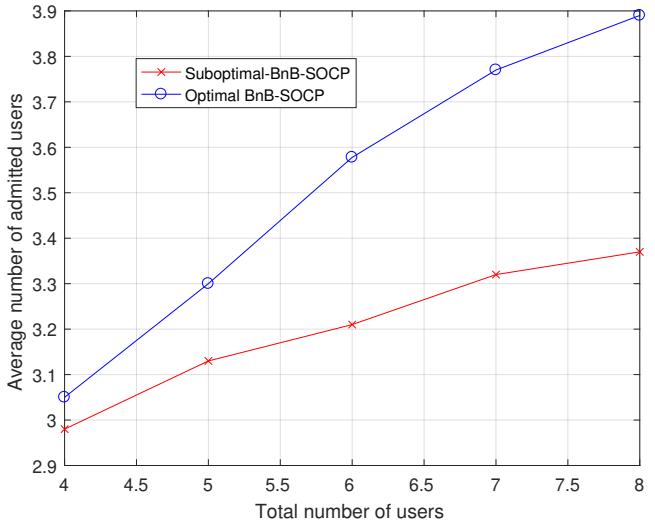

Figure 10: Average number of admitted users versus total number of users.

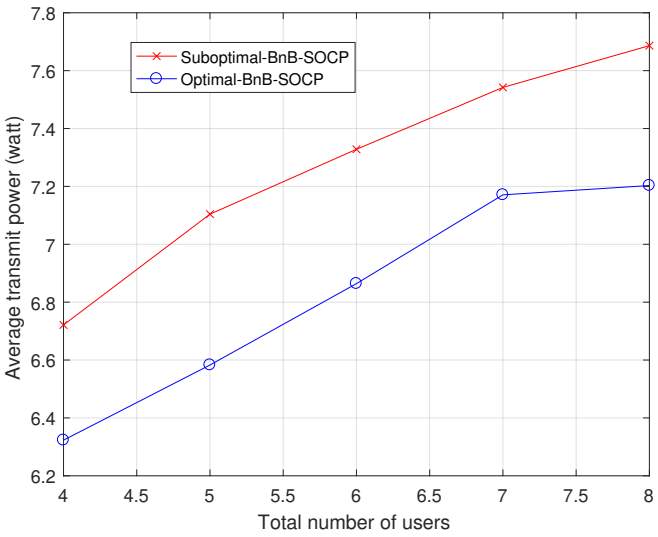

Figure 11: Average transmit power versus total number of users.

the integer variables. The optimal RRHs allocation and user maximization were solved using the BnB method. A suboptimal algorithm was also proposed to reduce computational complexity.

REFERENCES

[1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[2] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 84–91, June 2015.

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology Overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, First quarter 2015.

[4] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: a new look at interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, December 2010.

[5] M. Hong, R. Sun, H. Baligh, and Z. Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 226–240, February 2013.

[6] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, April 2013.

[7] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[8] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: the role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, August 2014.

[9] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, December 2013.

[10] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2894–2905, May 2014.

[11] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud radio access networks," in *2013 IEEE Global Communications Conference (GLOBECOM), Atlanta, GA, USA*, December 2013, pp. 4662–4667.

[12] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 494–508, January 2015.

[13] J. Cheng, Y. Shiy, B. Bai, W. Chen, J. Zhangy, and K. B. Letaief, "Group sparse beamforming for multicast green cloud-RAN via parallel semidefinite programming," in *2015 IEEE International Conference on Communications (ICC), London, UK*, June 2015, pp. 1886–1891.

[14] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1130–1139, 2016.

[15] V. N. Ha, L. B. Le *et al.*, "Cooperative transmission in cloud-RAN considering fronthaul capacity and cloud processing constraints," in *2014 IEEE Wireless Communications and Networking Conference (WCNC), Istanbul, Turkey*, April 2014, pp. 1862–1867.

[16] X. Peng, J. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), Washington, DC, USA*, Sep. 2014, pp. 1370–1374.

[17] T. H. L. Dinh, M. Kaneko, and L. Boukhatem, "Energy-efficient user association and beamforming for 5G fog radio access networks," in *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC), Las Vegas, NV, USA*, Jan 2019, pp. 1–6.

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

[19] E. Matskani, N. D. Sidiropoulos, Z. Q. Luo, and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2682–2693, July 2008.

[20] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green cloud-RAN with user admission control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1022–1036, April 2016.

[21] J. Lin, C. Jiang, and H. Shao, "Joint base station activation, user admission control and beamforming in a green downlink cooperative MISO network," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, China*, July 2015, pp. 913–917.

[22] D. H. Nguyen, L. L. Bao, and T. Le-Ngoc, "Multiuser admission control and beamforming optimization algorithms for MISO heterogeneous networks," *IEEE Access*, vol. 3, pp. 759–773, 2015.

[23] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-RANs," in *2014 IEEE Global Communications Conference (GLOBECOM), Austin, TX, USA*, Dec 2014, pp. 4054–4059.

[24] G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization*. Wiley, 1998.

[25] S. Boyd and J. Mattingley, "Branch and bound methods," https://see.stanford.edu/materials/lsocoee364b/17-bb_notes.pdf/, [Online; accessed 16-April-2020].

[26] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA cloud-RAN of small cells underlaying a macrocell," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2837–2850, November 2016.

[27] K. Cumanan, R. Krishna, L. Musavian, and S. Lambotharan, "Joint beamforming and user maximization techniques for cognitive radio networks based on branch and bound method," *IEEE Transactions on Wireless Communications*, vol. 9, no. 10, pp. 3082–3092, October 2010.

[28] Y. Zhang, *Interior point algorithms: Theory and analysis*. Wiley, 1997.

[29] T. Tsuchiya, "A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming," *Optimization methods and software*, vol. 11, no. 1-4, pp. 141–182, 1999.

Ashraf Bsebsu received his B.S. degree in Electrical and Electronic Engineering from Tripoli University, Libya 2006. In 2011 has awarded MSc in digital communication systems from Loughborough University. He is currently pursuing the Ph.D degree in signal processing and networking. His research interests including wireless communication, wireless caching and optimization.

Gan Zheng (S'05-M'09-SM'12) received the BEng and the MEng from Tianjin University, Tianjin, China, in 2002 and 2004, respectively, both in Electronic and Information Engineering, and the PhD degree in Electrical and Electronic Engineering from The University of Hong Kong in 2008. He is currently Reader of Signal Processing for Wireless Communications in the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK. His research interests include machine learning for communications, UAV communications, mobile edge caching, full-duplex radio, and wireless power transfer. He is the first recipient for the 2013 IEEE Signal Processing Letters Best Paper Award, and he also received 2015 GLOBECOM Best Paper Award, and 2018 IEEE Technical Committee on Green Communications & Computing Best Paper Award. He currently serves as an Associate Editor for IEEE Communications Letters.

Sangarapillai Lambotharan (SM'06) received the Ph.D. degree in signal processing in 1997 from Imperial College London, London, where he remained until 1999 as a postdoctoral research associate. He was a visiting scientist with the Engineering and Theory Centre of Cornell University, USA in 1996. He is a Professor of Digital Communications and the Head of Signal Processing and Networks Research Group with the Wolfson School Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, U.K. Between 1999 and 2002, he was with Motorola Applied Research Group, U.K. and investigated various projects including physical link layer modeling and performance characterization of GPRS, EGPRS, and UTRAN. He was with Kings College London and Cardiff University as a Lecturer and Senior Lecturer, respectively, from 2002 to 2007. His current research interests include 5G networks, MIMO, radars, smart grids, machine learning, network security, and blockchain technology. He has authored more than 200 journal and conference articles in these areas.

Kanapathippillai Cumanan received the BSc degree with first class honors in electrical and electronic engineering from the University of Peradeniya, Sri Lanka in 2006 and the PhD degree in signal processing for wireless communications from Loughborough University, Loughborough, UK, in 2009. He is currently a lecturer at the Department of Electronic Engineering, The University of York, UK. From March 2012 to November 2014, he was working as a research associate at School of Electrical and Electronic Engineering, Newcastle University, UK. Prior to this, he was with the School of Electronic, Electrical and System Engineering, Loughborough University, UK. In 2011, he was an academic visitor at Department of Electrical and Computer Engineering, National University of Singapore, Singapore. From January 2006 to August 2006, he was a teaching assistant with Department of Electrical and Electronic Engineering, University of Peradeniya, Sri Lanka. His research interests include non-orthogonal multiple access (NOMA), massive MIMO, physical layer security, cognitive radio networks, convex optimization techniques and resource allocation techniques. Dr. Cumanan was the recipient of an overseas research student award scheme (ORSAS) from Cardiff University, Wales, UK, where he was a research student between September 2006 and July 2007.

Basil AsSadhan received the M.S. degree in electrical and computer engineering from the University of Wisconsin and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University. He is currently an Assistant Professor with the Electrical Engineering Department, King Saud University. His research interests are in the areas of cybersecurity, network security, network traffic analysis, and anomaly detection.