Joint User Selection and Energy Minimization for Ultra-Dense Multi-channel C-RAN with Incomplete CSI

Cunhua Pan, Huiling Zhu, Nathan J. Gomes and Jiangzhou Wang, *Fellow*, *IEEE*

Abstract

This paper provides a unified framework to deal with the challenges arising in dense cloud radio access networks (C-RAN), which include huge power consumption, limited fronthaul capacity, heavy computational complexity, unavailability of full channel state information (CSI), etc. Specifically, we aim to jointly optimize the remote radio head (RRH) selection, user equipment (UE)-RRH associations and beam-vectors to minimize the total network power consumption (NPC) for dense multi-channel downlink C-RAN with incomplete CSI subject to per-RRH power constraints, each UE's total rate requirement, and fronthaul link capacity constraints. This optimization problem is NP-hard. In addition, due to the incomplete CSI, the exact expression of UEs' rate expression is intractable. We first conservatively replace UEs' rate expression with its lower-bound. Then, based on the successive convex approximation (SCA) technique and the relationship between the data rate and the mean square error (MSE), we propose a single-layer iterative algorithm to solve the NPC minimization problem with convergence guarantee. In each iteration of the algorithm, the Lagrange dual decomposition method is used to derive the structure of the optimal beam-vectors, which facilitates the parallel computations at the Baseband unit (BBU) pool. Furthermore, a bisection UE selection algorithm is proposed to guarantee the feasibility of the problem. Simulation results show the benefits of the proposed algorithms and the fact that a limited amount of CSI is sufficient to achieve performance close to that obtained when perfect CSI is possessed.

Index Terms

Cloud radio access network (C-RAN), 5G ultra-dense networks, limited fronthaul capacity, incomplete CSI.

This work is supported by European Commission Horizon2020 project iCIRRUS under grant agreement No 644526.

C. Pan, H. Zhu, N. Gomes and J. Wang are with the School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NZ, U.K. C.Pan is also with the Queen Mary University of London, London E1 4NS, U.K. (Email:{C.Pan, H.Zhu, N.J.Gomes, J.Z.Wang}@kent.ac.uk).

I. Introduction

The fifth-generation (5G) wireless system is expected to offer a thousand times the throughput [1] of the current fourth-generation (4G) [2]–[4] and provide ubiquitous service access for a large number of user equipments (UEs) in hot spots such as shopping malls, stadia, etc. To achieve this goal, heterogeneous and small cell network (HetSNet) is regarded as one of the most promising techniques by exploiting spatial degrees of freedom through deploying more and more access points (APs) [5]. However, since all APs reuse the same frequency, the interference among the APs is a limiting factor [5], which should be carefully managed. Dense cloud radio access network (C-RAN) was proposed in [6] as one promising architecture to conquer this issue. In dense C-RAN, all the base-band processing is performed at the BBU pool through the recent development of cloud computing techniques [7], while the RRHs are only responsible for simple radio transmission or reception [8], [9]. Due to their simple functionality, RRHs can be densely deployed in the network with low hardware cost. Due to the centralized architecture of dense C-RAN, the multi-UE interference can be efficiently handled through joint signal processing techniques such as coordinated multi-point (CoMP), leading to significant performance gains. Although C-RAN has been introduced in 4G, it is usually deployed in a large geographical area by connecting macrocell base stations to the BBU pool through fronthaul links. This conventional C-RAN incurs large delays on the fronthaul links due to long transmission distance between RRHs and BBU pool [10], which will violate the stringent latency requirement in 5G [1], i.e., a roundtrip latency within 1 ms. In contrast, dense C-RAN studied in this paper is aimed to cover hot spots with much smaller geographical area. Hence, delays can be significantly reduced.

However, there are many technical and deployment issues associated with dense C-RAN. First, dense deployment of RRHs will require high power consumption if all RRHs are activated even when the network traffic load is low. In addition, if each RRH serves all UEs, significant power will be used on the fronthaul links. As a result, how to activate the RRHs and select the RRHs for serving each UE to minimize the total network power consumption (NPC) is a critical issue. Second, in a dense C-RAN there will be a need for a large number of fronthaul links, requiring them to be low cost. There may also be a need to use millimeter wave (mmWave) technology for flexible and low cost deployment. These cost considerations lead to the likelihood of a capacity constraint on the fronthaul. Third, in dense C-RAN, the BBU pool will support large number

of RRHs and the number of optimization variables for beam-vectors will become very large, which will incur high computational complexity and will become unaffordable. Finally, the dense C-RAN requires more CSI for the facilitation of CoMP transmission design, which will cause a heavy training overhead. The amount of training overhead will increase with the number of RRHs and UEs, and may counteract the cooperative gains provided by CoMP transmission [11]. The most promising way to deal with this issue is to restrict the number of RRHs that each UE should measure CSI to. The remaining CSI values can be regarded as zeros, or only long term channel statistics of the remaining CSI, such as path loss and shadowing, are considered. How to design transmission strategies for this incomplete CSI case becomes an imperative task.

Most of current work only deals with parts of the above challenges. For example, [12]–[15] considered the joint RRH selection and beamforming design to minimize the total NPC subject to UEs' quality of service (QoS) targets and per-RRH power constraints. These papers ignored the capacity constraints on the fronthaul links and assumed that the fronthaul capacity is unlimited. To address the fronthaul capacity constraints issue, [16] investigated the problem of minimizing the number of data transfers on the aggregated fronthaul links with UEs' QoS constraints and power constraints on each RRH. However, [16] did not explicitly impose the fronthaul capacity constraints in the optimization problem. Recently, several papers have addressed the case when the fronthaul capacity constraints are explicitly imposed [17]–[19]. The case when the optimization problem is infeasible was not considered. Then, some UEs can be removed to make the optimization problem feasible again. The UE admission control and total NPC minimization were jointly optimized in [20], where a single-stage optimization problem was formulated by introducing a weighting factor in the admission control part. Recently, [21] extended the work in [20] to multi-channel heterogeneous C-RAN where the C-RAN is overlaid by a macro-cell. However, for the admission control designs considered in [20] and [21], one has to carefully choose the weighting factor associated with the admission control part to ensure that the selected UEs can satisfy the QoS constraints, which is not easy.

However, the algorithms proposed in [12]–[21] were based on the assumption of full CSI at the BBU pool, which is not practical as explained. Unfortunately, the algorithms designed for perfect CSI cannot be directly extended to the case of incomplete CSI. To the best of our knowledge, only a few papers have considered the incomplete CSI case [22]–[25]. [22] proposed a CSI reduction scheme named compressive CSI acquisition, that can obtain the instantaneous CSIs for a subset

of channel links and the large scale fading gains of the others. Based on the incomplete CSI, [22] solved a transmit power minimization problem while guaranteeing UEs' QoS requirements by using a stochastic coordinated beamforming technique. However, the method needs to solve a high-dimension semi-definite programming (SDP) problem for each sample, and the number of samples increases with the size of the network, which incurs an unacceptable complexity for dense C-RAN. [23] focused on the beamforming algorithm to maximize the sum-rate for arbitrary UE-centric clustering C-RAN. The "C-cluster method" was introduced in [23] to reduce channel estimation overhead where only subsets of CSIs for each UE are measured, and the other unavailable CSIs are regarded as zeros. Recently, [24] proposed a conservative precoder design with the objective of maximizing the weighted sum-rate of UEs for arbitrary UE-centric clustering method with incomplete CSIs, where the long term channel statistic was incorporated into the optimization. Finally, [25] designed a clustering scheme maximizing the average net throughput of the dense C-RAN by taking the training overhead into account. The scheme is based on a hybrid CoMP transmission mode and operates under a long time duration that may be performed at the medium access control (MAC) layer since only large-scale CSIs are required. However, both the beam directions and power allocations were not optimized in [25]. None of the papers [22]–[25] considered the fronthaul capacity constraints and were mainly focused on sum-rate maximization problems without incorporating QoS requirements.

The aim of this paper is to provide a complete framework to jointly tackle the above-mentioned challenges together. Specifically, we investigate the joint optimization of RRH selection, RRH-UE associations and transmit beamforming to minimize the NPC for downlink multi-channel C-RAN with incomplete CSI, subject to fronthaul link capacity constraints, all UEs' rate requirements and per-RRH power constraints. The NPC is modeled as the sum of the RRH power consumption and the fronthaul link power consumption. The low-power sleep mode is considered in the RRH power consumption model, and the fronthaul link power consumption is modeled as a linear function of fronthaul traffic. To reduce the computational complexity, each UE is restricted to be served by its nearby RRHs since only nearby RRHs contribute significantly to the UE's signals. Moreover, to reduce the channel measurement overhead, we introduce the subset of RRHs that each UE should estimate the CSIs to, while the large-scale fading (such as path-loss and shadowing) is assumed to be known for the other unavailable CSI. In general, the candidate set of RRHs for serving UEs and the CSI estimation set of RRHs for each UE are

determined based on UEs' locations that may be the task of the upper-layer, which is beyond the scope of this paper. The NPC minimization problem is an NP-hard mixed-integer non-linear programming (MINLP) problem due to the indicator functions introduced in both objective function and fronthaul capacity constraints, whose optimal solution is intractable. In addition, due to the sum rate constraints and incomplete CSI, the QoS constraints are non-convex and difficult to handle. Furthermore, due to the conflicting constraints, the NPC minimization problem may be infeasible and the initialization solution should be carefully selected. As a result, the contributions of this paper can be summarized as follows:

1) Due to the incomplete CSI, it is intractable to derive the exact closed-form expression of the data rate for each UE, and thus stringent QoS requirements for each UE are difficult to be guaranteed. To alleviate this difficulty, we conservatively replace the data rate of each UE with its lower-bound expression derived by using the Jensen's inequality.
2) To resolve the feasibility issue, we provide a low-complexity UE selection algorithm based on bisection search method to maximize the number of admitted UEs that can achieve their QoS targets, and its complexity only increases logarithmically with the number of UEs. Simulation results show that this algorithm can achieve marginal performance loss with respect to (w.r.t.) that obtained by the exhaustive UE search algorithm with an exponential computational complexity over the number of admitted UEs.
3) Given the feasible set of UEs from the UE selection algorithm, we provide a low-complexity single-layer iterative algorithm (i.e., Algorithm 1) to solve the NPC minimization problem. Specifically, the non-smooth indicator function is approximated as a non-convex function and the successive convex approximation (SCA) technique [26] is adopted to approximate the non-convex function as a series of convex functions. To deal with the non-convex QoS constraints, we translate the technique in [27] that aimed at rate maximization problem to the NPC minimization problem with rate expressions in the constraints and incomplete CSI. The convergence of the iterative algorithm is strictly proved.
4) In each iteration of Algorithm 1, there is a subproblem that the beam-vectors should be optimized. We derive the structure of the optimal beam-vectors by employing the Lagrange dual decomposition method. Then, each beam-vector can be obtained in parallel for each sub-channel (SC), which facilitates the application of the cloud computing technique in

BBU pool.

This paper is organized as follows. Section II presents the system model, and Section III formulates the UE selection problem and NPC minimization problem along with the complexity analysis. The single-layer iterative algorithm to solve the NPC algorithm is given in Section IV when the UEs are selected to be admitted. Then, in Section V, the low-complexity UE selection algorithm is provided. Simulation results are presented in Section V to evaluate the performance of the proposed algorithms. Finally, conclusions are drawn in Section VII.

Notations: For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, while for a complex number x, $|x|$ denotes the magnitude of x. $\mathbf{1}$ denotes a vector with all elements equal to ones. 's.t.' is short for 'subject to'. $\mathbb{E}_{\{x\}}\{y\}$ means the expectation of y over x. The complex Gaussian distribution is denoted as $\mathcal{CN}(\cdot,\cdot)$. We use $\mathbb{C}$ to represent the complex set. The lower-case bold letters denote vectors and upper-case bold letters denote matrices. $\mathrm{blkdiag}(\cdot)$ denotes the block diagonalization operation.

II. System Model

A. System model

Consider a downlink ultra-dense C-RAN, as shown in Fig. 1, consisting of I RRHs and K UEs[1], where each RRH is equipped with M transmit antennas, and each UE has a single antenna. Denote the set of RRHs and UEs as $\mathcal{I} = \{1, \cdots, I\}$ and $\bar{\mathcal{U}} = \{1, \cdots, K\}$, respectively. Each RRH is connected to the BBU pool through wireless (e.g. mmWave communication) fronthaul links. The fronthaul links are represented by dark solid arrows in Fig. 1. The BBU pool is assumed to have all UEs' data and distributes each UE's data to a carefully selected set of RRHs through the fronthaul links. It is assumed that all the RRHs send their received data using the Orthogonal Frequency Division Multiple Access (OFDMA) technique and then cooperatively transmit to the UEs.

Denote $\mathcal{U} \subseteq \bar{\mathcal{U}}$ as the subset of UEs that are admitted in the C-RAN. To reduce the computational complexity of the large network, it is assumed that each UE $k \in \mathcal{U}$ can only be served

[1]Due to the simple functionalities of RRHs, the RRHs can be densely deployed with low hardware cost, wherein the number of RRHs may even be larger than that of UEs. Hence, the average distance between RRHs and UEs can be significantly reduced. As a result, the transmission power of the RRHs can also be reduced due to the decreased path loss.

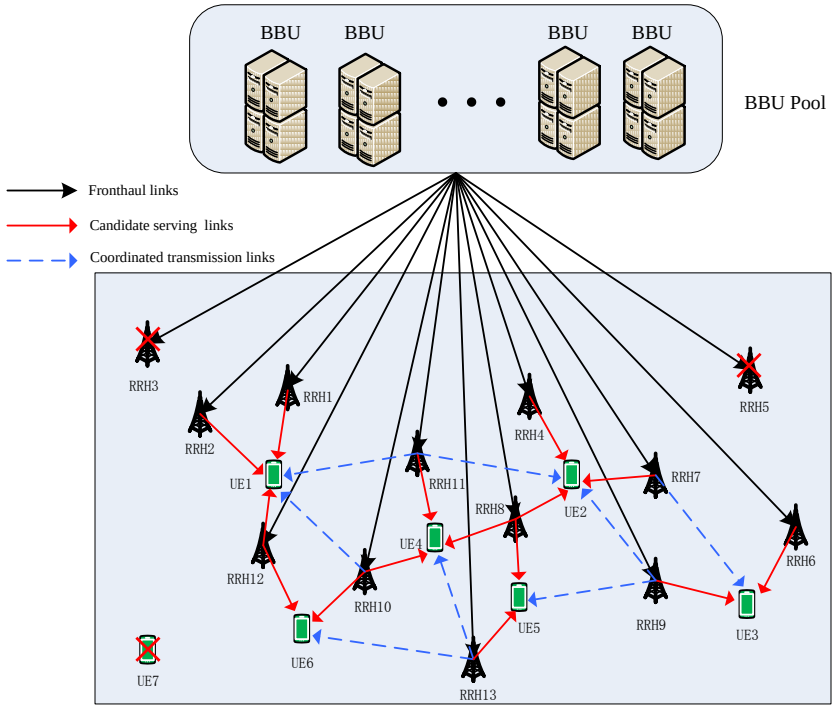

Fig. 1. Illustration of a C-RAN with thirteen RRHs and seven UEs. The RRHs are connected to a BBU pool through wireless fronthaul links. In this scenario, UE 7 is not selected for serving and the candidate sets of RRHs for the selected UEs are given by $\mathcal{I}_1 = \{1, 2, 12\}$, $\mathcal{I}_2 = \{4, 7, 8\}$, $\mathcal{I}_3 = \{6, 9\}$, $\mathcal{I}_4 = \{8, 10, 11\}$, $\mathcal{I}_5 = \{8, 13\}$ and $\mathcal{I}_6 = \{10, 12\}$, respectively. The sets of UEs that are potentially served by the RRHs are given by $\mathcal{U}_1 = \{1\}$, $\mathcal{U}_2 = \{1\}$, $\mathcal{U}_3 = \{\emptyset\}$, $\mathcal{U}_4 = \{2\}$, $\mathcal{U}_5 = \{\emptyset\}$, $\mathcal{U}_6 = \{3\}$, $\mathcal{U}_7 = \{2\}$, $\mathcal{U}_8 = \{2, 4, 5\}$, $\mathcal{U}_9 = \{3\}$, $\mathcal{U}_{10} = \{4, 6\}$, $\mathcal{U}_{11} = \{4\}$, $\mathcal{U}_{12} = \{1, 6\}$ and $\mathcal{U}_{13} = \{5\}$, respectively. The sets of RRHs for coordinating the interference for the selected UEs are given by $\mathcal{C}_1 = \{10, 11\}$, $\mathcal{C}_2 = \{9, 11\}$, $\mathcal{C}_3 = \{7\}$, $\mathcal{C}_4 = \{13\}$, $\mathcal{C}_5 = \{9\}$ and $\mathcal{C}_6 = \{13\}$, respectively. The sets of coordinated UEs by the RRHs are given by $\mathcal{T}_7 = \{3\}$, $\mathcal{T}_9 = \{2, 5\}$, $\mathcal{T}_{10} = \{1\}$, $\mathcal{T}_{11} = \{1, 2\}$ and $\mathcal{T}_{13} = \{4, 6\}$, respectively.

by its nearby RRHs since only nearby RRHs contribute significantly to the UE's signal quality due to the severe path loss. Denote $\mathcal{I}_k \subseteq \mathcal{I}$ and $\mathcal{U}_i \subseteq \mathcal{U}$ as the candidate set of RRHs that potentially serve UE k and the set of UEs that can be potentially served by RRH i, respectively. The transmission links from the RRHs in $\mathcal{I}_k$ to UE k are called the candidate serving links, which are represented in red solid arrows in Fig. 1. In this paper, it is assumed that $\mathcal{I}_k$ and $\mathcal{U}_i$ are predetermined by some well-known user-centric cluster methods [25], [28], [29] determined

by the MAC layer[2]. Please refer to [30] for a survey on user-centric cluster methods. Note that since no restrictions are placed on $\mathcal{I}_k$, they can overlap with each other, i.e., there may exist two different UEs k and k' that $\mathcal{I}_k \cap \mathcal{I}_{k'} \neq \emptyset$, for $\forall k, k' \in \mathcal{U}$. Moreover, the other-cluster interference due to overlapping coverage can be effectively handled under this user-centric cluster method. For example, UE 4 and UE 5 have one common serving RRH 8. Hence, RRH 8 will transmit useful signals to both UE 4 and UE 5, rather than only interference signals. In addition, the BBU pool has the CSI knowledge from RRH 3 to UE 4. Thus, the interference from RRH 3 to UE 4 will be carefully controlled when RRH 3 is serving UE 5. In contrast to the non-cooperative optimization where each cluster selfishly optimizes its own performance without considering its impact on the other clusters, in dense C-RAN all the signal processing operation is performed at the BBU pool, where the interference among different clusters can be centrally mitigated by resorting to the powerful cloud computing tool.

Denote the set of available sub-channels (SCs) as $\mathcal{N} = \{1, 2, \cdots, N\}$, where N is the total number of SCs. To maximize the spectral efficiency, it is assumed that universal frequency reuse is adopted and the multiuser interference can be efficiently handled by the beamforming technique. Denoting $\mathbf{w}_{i,k}^{(n)} \in \mathbb{C}^{M\times 1}$ as the beam-vector at RRH i for UE k on SC n, the transmitted signal of RRH i on SC n is

$$\mathbf{x}_i^{(n)} = \sum\nolimits_{k\in\mathcal{U}_i} \mathbf{w}_{i,k}^{(n)} s_k^{(n)}, \tag{1}$$

where $s_k^{(n)}$ is the data symbol for UE k on SC n. Without loss of generality, it is assumed that $\mathbb{E}\{|s_k^{(n)}|^2\} = 1$ and $\mathbb{E}\{s_{k_1}^{(n_1)} s_{k_2}^{(n_2)}\} = 0$ for $(n_1, k_1) \neq (n_2, k_2), \forall n_1, n_2 \in \mathcal{N}, \forall k_1, k_2 \in \mathcal{U}$. The baseband received signal at UE k on SC n is given by

$$y_k^{(n)} = \underbrace{\sum\nolimits_{i\in\mathcal{I}_k} \mathbf{h}_{i,k}^{(n)} \mathbf{w}_{i,k}^{(n)} s_k^{(n)}}_{\text{desired signal}} + \underbrace{\sum\nolimits_{l\neq k, l\in\mathcal{U}} \sum\nolimits_{i\in\mathcal{I}_l} \mathbf{h}_{i,k}^{(n)} \mathbf{w}_{i,l}^{(n)} s_l^{(n)}}_{\text{interference}} + z_k^{(n)}, \tag{2}$$

where $\mathbf{h}_{i,k}^{(n)} \in \mathbb{C}^{1\times M}$ is the channel vector from RRH i to UE k on SC n, and $z_k^{(n)}$ is the additive complex white Gaussian noise following the distribution of $\mathcal{CN}(0, \sigma_k^2)$. The channel vector $\mathbf{h}_{i,k}^{(n)}$ can be written as $\mathbf{h}_{i,k}^{(n)} = \alpha_{i,k}^{(n)} \tilde{\mathbf{h}}_{i,k}^{(n)}$, where $\alpha_{i,k}^{(n)}$ denotes the large-scale channel gain that includes

[2]In general, the cluster method is mainly determined based on the large-scale CSI, which is usually performed in the upper layer such as MAC layer. In some hot spots such as stadia and shopping malls, the users move slowly. Hence, the cluster can be kept fixed for a long time compared with the instantaneous CSI. This paper only focuses on the beam-vectors at the physical layer, and how to design the optimal cluster method is beyond the scope of this paper.

the path loss and shadowing, and $\tilde{\mathbf{h}}_{i,k}^{(n)}$ denotes the small-scale fading vector, where all elements are dependent of each other and each one has zero mean and unit variance.

For the sake of reduced complexity of decoding at the receivers, we do not consider the joint decoding of the interfering signals and the multiuser interference is simply regarded as noise at the receivers. In addition, coherent joint transmission[3] is assumed as in most of existing papers [12]–[21]. Then, the SINR at UE k on SC n can be obtained from (2) as

$$\gamma_k^{(n)}(\mathbf{w}) = \frac{\left|\sum_{i\in\mathcal{I}_k} \mathbf{h}_{i,k}^{(n)}\mathbf{w}_{i,k}^{(n)}\right|^2}{\sum_{l\neq k, l\in\mathcal{U}}\left|\sum_{i\in\mathcal{I}_l} \mathbf{h}_{i,k}^{(n)}\mathbf{w}_{i,l}^{(n)}\right|^2 + \sigma_k^2}. \tag{3}$$

where $\mathbf{w}$ denotes the collection of all beam-vectors.

As seen in (3), to design the beam-vectors for all UEs, the overall CSI of all UEs is required. However, it is a formidable task to obtain all CSI for the dense C-RAN due to the limited training resources. To handle this difficulty, we introduce the set $\widetilde{\mathcal{I}}_k \supseteq \mathcal{I}_k$ for each UE k that is defined as the set of RRHs that UE k needs to measure CSI from. Also, we define $\widetilde{\mathcal{U}}_i \supseteq \mathcal{U}_i$ for each RRH i as the set of UEs that each RRH i knows the CSI to. In general, $\widetilde{\mathcal{I}}_k$ are the set of UE k's nearby RRHs and $\widetilde{\mathcal{U}}_i$ are the set of RRH i's nearby UEs. Note that at least the CSI from all RRHs in $\mathcal{I}_k$ is required for cooperative transmission design. The other CSI from RRHs in $\mathcal{C}_k = \widetilde{\mathcal{I}}_k \backslash \mathcal{I}_k$ to UE k is used to coordinate the interference, and the links from RRHs in $\mathcal{C}_k$ are called coordinated interference links, which are shown by blue dashed arrows in Fig. 1. Also, the UEs in $\mathcal{T}_i = \widetilde{\mathcal{U}}_i \backslash \mathcal{U}_i$ are called RRH i's coordinated UEs. For the CSI from RRHs in $\mathcal{I}\backslash\widetilde{\mathcal{I}}_k$ to UE k, it is assumed that the BBU pool only knows the large scale gains $\{\alpha_{i,k}^{(n)}, \forall i \in \mathcal{I}\backslash\widetilde{\mathcal{I}}_k, k \in \mathcal{U}, n \in \mathcal{N}\}$. This is possible because the large scale gains change much more slowly than the small-scale fading.

Since the CSI in $\mathcal{I}\backslash\widetilde{\mathcal{I}}_k$ is unknown, we consider the following data rate for UE k on SC n (bit/s/Hz) [31]

$$\bar{r}_k^{(n)}(\mathbf{w}) = \mathbb{E}_{\left\{\mathbf{h}_{i,k}^{(n)}, i\in\mathcal{I}\backslash\widetilde{\mathcal{I}}_k\right\}}\left\{\log_2(1+\gamma_k^{(n)}(\mathbf{w}))\right\}. \tag{4}$$

where the expectation operator is performed over the fast fading of the unknown CSI in $\mathcal{I}\backslash\widetilde{\mathcal{I}}_k$.

[3]This assumption is valid for dense C-RAN. The reason is that dense C-RAN is usually deployed in hot spots with smaller coverage area compared with that of the conventional C-RAN that covers multiple macrocells [10]. Hence, different transmission delays due to different transmission distances between RRHs and BBU pool can be ignored. Then, both the synchronization and coherent joint transmission are possible.

Each UE k's total data rate should be larger than the minimum rate requirement $R_{k,\min}$:

$$\mathrm{C1}: \bar{r}_{k,\mathrm{tot}}(\mathbf{w}) = \sum\nolimits_{n\in\mathcal{N}} \bar{r}_k^{(n)}(\mathbf{w}) \geq R_{k,\min}, \forall k \in \mathcal{U}. \tag{5}$$

In each fronthaul link, the maximum capacity that can be supported is limited. Hence, the following fronthaul capacity constraint follows:

$$\mathrm{C2}: \sum\nolimits_{k\in\mathcal{U}_i} \varepsilon\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) \bar{r}_{k,\mathrm{tot}}(\mathbf{w}) \leq C_{i,\max}, \forall i \in \mathcal{I}, \tag{6}$$

where $\varepsilon(\cdot)$ is an indicator function, defined as

$$\varepsilon(x) = \begin{cases} 1, \text{ if } x \neq 0, \\ 0, \text{ otherwise}, \end{cases} \tag{7}$$

$P_{i,k}^{\mathrm{tr}}(\mathbf{w}) = \sum_{n\in\mathcal{N}} \left\|\mathbf{w}_{i,k}^{(n)}\right\|^2$ denotes the total transmission power from RRH i to UE k, $C_{i,\max}$ is the maximum capacity that can be supported by the ith fronthaul link.

B. Network power consumption model

In this subsection, a practical NPC model is provided that consists of two parts: power consumption at the RRHs and power consumption on the fronthaul links.

As in [32], the power consumption of RRH i can be modeled as a piecewise linear function of the transmit power at RRH i:

$$P_i^{\mathrm{rrh}}(\mathbf{w}) = \begin{cases} \eta_i P_i^{\mathrm{tr}}(\mathbf{w}) + P_i^{\mathrm{active}}, & \text{if } P_i^{\mathrm{tr}}(\mathbf{w}) > 0 \\ P_i^{\mathrm{sleep}}, & \text{if } P_i^{\mathrm{tr}}(\mathbf{w}) = 0 \end{cases} \tag{8}$$

where $\eta_i > 1$ is the constant accounting for the efficiency of the power amplifier of RRH i, $P_i^{\mathrm{tr}}(\mathbf{w})$ is the total transmit power at RRH i that should be no larger than $P_{i,\max}$, i.e.,

$$\mathrm{C3}: \ P_i^{\mathrm{tr}}(\mathbf{w}) = \sum\nolimits_{k\in\mathcal{U}_i} P_{i,k}^{\mathrm{tr}}(\mathbf{w}) \leq P_{i,\max}, i \in \mathcal{I}, \tag{9}$$

P_i^{active} and P_i^{sleep} represent the circuit power consumption when RRH i is in active mode and sleep mode, respectively. In general, P_i^{active} is much larger than P_i^{sleep}, which motivates us strategically to switch off the RRHs to save power in case of very low traffic.

Fronthaul power consumption model is critical for the optimization of NPC. In [12] and [13], the fronthaul power consumption was simply modeled as a step function, with a larger constant value for active mode and smaller one for sleep mode. In [33], the fronthaul power consumption

is modeled to be proportional to the number of UEs that each one supports. However, these papers did not take into account the effect of data rate transmitting on each fronthaul link. Intuitively, to support high fronthaul transmit data rate, more power should be consumed on the fronthaul links. Compared with [12], [13], [33], we go one step further by modeling the power consumption of each fronthaul link to be proportional to the total fronthaul transmit data rate as in [34]:

$$P_i^{\mathrm{fr}}(\mathbf{w}) = \rho_i \sum\nolimits_{k\in\mathcal{U}_i} \varepsilon\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) \bar{r}_{k,\mathrm{tot}}(\mathbf{w}), \tag{10}$$

where ρ_i is a constant scaling factor [4].

Based on the above analysis and with some simple manipulations, the NPC is modeled as

$$\begin{aligned} P_{\mathrm{NPC}}(\mathbf{w}) &= \sum_{i\in\mathcal{I}} \left\{P_i^{\mathrm{rrh}}(\mathbf{w}) + P_i^{\mathrm{fr}}(\mathbf{w})\right\} \\ &= \sum_{i\in\mathcal{I}} \left\{\eta_i P_i^{\mathrm{tr}}(\mathbf{w}) + \varepsilon\left(P_i^{\mathrm{tr}}(\mathbf{w})\right) P_i^{\mathrm{c}} + \rho_i \sum_{k\in\mathcal{U}_i} \varepsilon\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) \bar{r}_{k,\mathrm{tot}}(\mathbf{w})\right\} + \sum_{i\in\mathcal{I}} P_i^{\mathrm{sleep}}, \end{aligned} \tag{11}$$

where $P_i^{\mathrm{tr}}(\mathbf{w})$ is given in (9), $P_i^{\mathrm{c}} = P_i^{\mathrm{active}} - P_i^{\mathrm{sleep}}, \forall i \in \mathcal{I}$.

III. Problem Formulation and Analysis

Based on the above system model, we formulate the user selection problem and the NPC minimization problem in a two-stage form. Then, we provide the complexity analysis for the formulated problems.

A. Problem Formulation

Due to the limited fronthaul capacity constraints C2 in (6) and the power constraints C3 in (9), the system may not be able to support all UEs with their rate requirements of C1 in (5). Hence, some UEs may be dropped or rescheduled in other orthogonal time slots to make the optimization problem feasible. As a result, we may consider a two-stage optimization problem. In

[4]In general, this scaling factor may not be a constant, rather depend on the total transmit data rate on the fronthaul link. However, how to accurately model this relationship is still under investigation. To the best of our knowledge, only [34] provided the detailed study of this model that has been adopted by the existing work such as [14].

the first stage, one should find the largest subsets of UEs that can be supported by the system[5], while in the second stage, one should optimize the corresponding beam-vectors to minimize P_{NPC} with the selected subset of UEs obtained from the first stage.

As a result, the optimization problem at the first stage is formulated as

$$\begin{aligned} \mathcal{P}_1: \ & \max_{\mathbf{w}, \mathcal{U} \subseteq \bar{\mathcal{U}}} \quad |\mathcal{U}| \\ & \text{s.t.} \qquad \mathrm{C1}, \mathrm{C2}, \mathrm{C3}. \end{aligned} \tag{12}$$

Denote $\mathcal{U}^\star$ as the solution from Stage I and the corresponding $\mathcal{U}_i$ becomes $\mathcal{U}_i^\star$. Then, the optimization problem at the second stage is formulated as

$$\mathcal{P}_2: \ \min_{\mathbf{w}} \quad \sum\nolimits_{i \in \mathcal{I}} \left\{ \eta_i P_i^{\mathrm{tr}}(\mathbf{w}) + \varepsilon \left(P_i^{\mathrm{tr}}(\mathbf{w}) \right) P_i^{\mathrm{c}} + \rho_i \sum\nolimits_{k \in \mathcal{U}_i^\star} \varepsilon \left(P_{i,k}^{\mathrm{tr}}(\mathbf{w}) \right) \bar{r}_{k,\mathrm{tot}}(\mathbf{w}) \right\} \tag{13a}$$

$$\text{s.t.} \quad \mathrm{C1}, \mathrm{C2}, \mathrm{C3} \tag{13b}$$

In the constraints C1, C2, and C3, $\mathcal{U}$ and $\mathcal{U}_i$ are replaced by $\mathcal{U}^\star$ and $\mathcal{U}_i^\star$, respectively. Note that the constant term $\sum_{i \in \mathcal{I}} P_i^{\mathrm{sleep}}$ in (11) has been omitted in the objective function (13a).

We emphasize that the aim of Stage I is to find the maximum number of admitted UEs with feasible beam-vectors. These obtained beam-vectors are not guaranteed to be optimal in terms of NPC. Hence, we need to perform Stage II to optimize the beam-vectors to reduce the NPC. The beam-vectors obtained from Stage I will be a feasible initial input that is required by the algorithm developed in Stage II.

The incomplete CSI at the BBU pool makes the design of beam-vectors very difficult to solve and the expression for the data rate is difficult to derive. In the following, we consider its lower-bound and replace the data rate with its lower-bound, which makes the optimization problem more tractable.

We first simplify the SINR expression in (3). The beam-vectors for each UE on each SC n are merged into a single large-dimension vector $\bar{\mathbf{w}}_k^{(n)} = [\mathbf{w}_{i,k}^{(n)\mathrm{H}}, \forall i \in \mathcal{I}_k]^{\mathrm{H}} \in \mathbb{C}^{|\mathcal{I}_k| M \times 1}, \forall n \in \mathcal{N}$. Then, we define a set of new channel vectors $\bar{\mathbf{h}}_{l,k}^{(n)} = [\mathbf{h}_{i,k}^{(n)}, \forall i \in \mathcal{I}_l] \in \mathbb{C}^{1 \times |\mathcal{I}_l| M}$, representing the

[5]Dense C-RAN is usually deployed in hot spots (such as shopping mall, stadia, et al.) where the number of UEs is huge, and the amount of available communication resource is limited. Hence, maximizing the number of admitted users for each time slot should be placed in higher priority. In some other scenarios, where there are abundant wireless resource, maximizing the number of admitted UEs in each time slot may not be a good option in reducing NPC, and dynamically scheduling the UE in different time slots may further reduce NPC, which will be left for future work.

aggregated CSI from the RRHs in $\mathcal{I}_l$ to UE k on SC n. The SINR expression in (3) can be rewritten as

$$\gamma_k^{(n)}(\mathbf{w}) = \frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\sum_{l\neq k, l\in\mathcal{U}}\left|\bar{\mathbf{h}}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}\right|^2+\sigma_k^2}. \tag{14}$$

Note that $\bar{\mathbf{h}}_{k,k}^{(n)}$ is perfectly known in the BBU pool according to the previous assumption, and only the denominator in (14) contains the uncertain terms. However, it is difficult to obtain the accurate rate expression. To deal with this challenge, we consider its lower-bound with more tractable form. Specifically, since $\log_2(1+a/x)$ is a convex function for any positive a, by using Jensen's inequality [35], the lower bound of the data rate in (4) can be derived as

$$\begin{aligned}
& \bar{r}_k^{(n)}(\mathbf{w}) && (15)\\
\geq\ & \log_2\left(1+\frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\mathbb{E}_{\left\{\mathbf{h}_{i,k}^{(n)}, i\in\mathcal{I}\setminus\tilde{\mathcal{I}}_k\right\}}\left\{\sum_{l\neq k,l\in\mathcal{U}}\left|\bar{\mathbf{h}}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}\right|^2\right\}+\sigma_k^2}\right) && (16)\\
=\ & \log_2\left(1+\frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\sum_{l\neq k,l\in\mathcal{U}}\bar{\mathbf{w}}_l^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}+\sigma_k^2}\right) && (17)\\
\triangleq\ & \tilde{r}_k^{(n)}(\mathbf{w}) && (18)
\end{aligned}$$

where $\mathbf{A}_{l,k}^{(n)} = \mathbb{E}_{\left\{\mathbf{h}_{i,k}^{(n)}, i\in\mathcal{I}\setminus\tilde{\mathcal{I}}_k\right\}}\left\{\bar{\mathbf{h}}_{l,k}^{(n)\mathrm{H}}\bar{\mathbf{h}}_{l,k}^{(n)}\right\} \in \mathbb{C}^{M|\mathcal{I}_l|\times M|\mathcal{I}_l|}$. To obtain the closed-form expression of $\mathbf{A}_{l,k}^{(n)}$, we define the indices of $\mathcal{I}_l$ as $\mathcal{I}_l = \{s_1^l, \cdots, s_{|\mathcal{I}_l|}^l\}$. Then, we have

$$\mathbf{A}_{l,k}^{(n)} = \begin{bmatrix} \left(\mathbf{A}_{l,k}^{(n)}\right)_{1,1} & \cdots & \left(\mathbf{A}_{l,k}^{(n)}\right)_{1,|\mathcal{I}_l|} \\ \vdots & \ddots & \vdots \\ \left(\mathbf{A}_{l,k}^{(n)}\right)_{|\mathcal{I}_l|,1} & \cdots & \left(\mathbf{A}_{l,k}^{(n)}\right)_{|\mathcal{I}_l|,|\mathcal{I}_l|} \end{bmatrix}, l\neq k \tag{19}$$

where $\left(\mathbf{A}_{l,k}^{(n)}\right)_{i,j} \in \mathbb{C}^{M\times M}, i,j\in 1,\cdots,|\mathcal{I}_l|$ is the block matrix of $\mathbf{A}_{l,k}^{(n)}$ at the ith row and jth column, given by

$$\left(\mathbf{A}_{l,k}^{(n)}\right)_{i,j} = \begin{cases} \mathbf{h}_{s_i^l,k}^{(n)\mathrm{H}}\mathbf{h}_{s_j^l,k}^{(n)}, & \text{if } s_i^l, s_j^l \in \tilde{\mathcal{I}}_k, \\ \left|\alpha_{s_i^l,k}^{(n)}\right|^2\mathbf{I}_{M\times M}, & \text{if } s_i^l, s_j^l \notin \tilde{\mathcal{I}}_k, \text{and } i=j, \\ \mathbf{0}_{M\times M,} & \text{otherwise.} \end{cases} \tag{20}$$

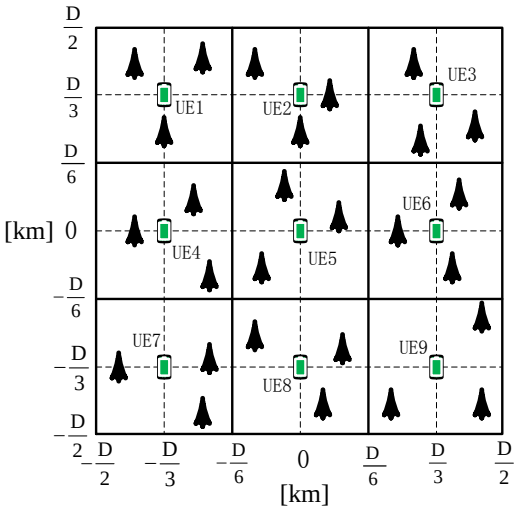

Fig. 2. Non-overlapped cluster topology.

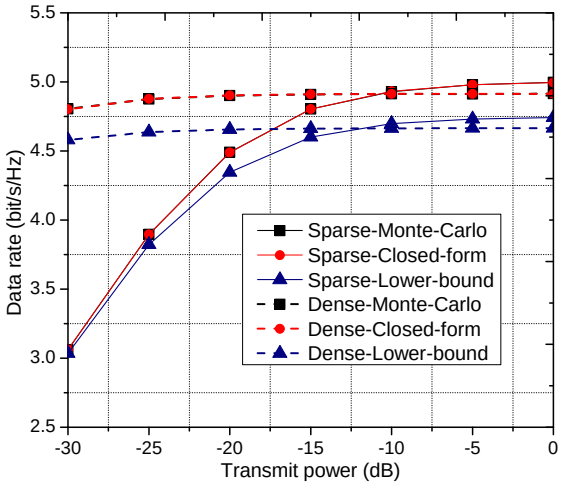

Fig. 3. Data rate versus the transmit power for the special case.

It can be easily verified that $\mathbf{A}_{l,k}^{(n)}$ is a positive definite matrix. Note that the derivations of matrix $\mathbf{A}_{l,k}^{(n)}$ place no restrictions on the channel distributions and only large-scale channel gains are required. Hence, the following developed algorithms are applicable for any channel distributions, such as Rayleigh fading, Ricean channels, Nakagami-m fading channels, et al.

We now start to check the tightness of this rate lower-bound. It is difficult to derive the accurate data rate expression for general case. Instead, in Appendix A, we derive the accurate closed-form expression of data rate for one special case under three assumptions: 1) The RRH serving cluster is the same as the CSI cluster for each UE: $\mathcal{I}_k = \widetilde{\mathcal{I}}_k$; 2) The RRH serving cluster for each UE is non-overlapped with each other: $\mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset, \forall k, k' \in \mathcal{U}$; 3) The small-scale fading vector $\tilde{\mathbf{h}}_{i,k}$ follows the distribution of $\mathcal{CN}(\mathbf{0}, \mathbf{I})$ for $\forall i, k$. We consider one non-overlapped C-RAN scenario deployed within a square area of coordinates $[-D/2, D/2] \times [-D/2, D/2]$ km as shown in Fig. 2. This network area is divided into nine $D/3$ km $\times$ $D/3$ km squares. In each square, one UE is located at the center point and three RRHs are randomly generated in this square to exclusively serve this UE. For simplicity, only one SC is considered. The other simulation parameters are the same as in the simulation Section. It is assumed that each RRH transmits at their maximum power and the beam direction is chosen to be channel direction. The values of $D = 3$ and $D = 1$ are tested, which correspond to sparse and dense scenarios, respectively.

Only UE 5 is considered. Fig. 3 plots three kinds of curves for comparison: one is the lower bound of data rate derived in (18), one is the accurate closed-form data rate expression derived in (A.4) in Appendix A, and the last one is the Monte-Carlo simulations. It is seen from Fig. 3 that the curve of the closed-form expression coincides with that of the Monte-Carlo simulations, which verifies the correctness of the derivations. Furthermore, for the sparse scenarios, when the transmit power is low, $P_{\max} < -20\mathrm{dB}$, the lower-bound is quite tight. With the increase of transmit power, the gap increases and becomes a constant in the high transmit power regime. Note that only roughly 3% data rate loss will be incurred when using the lower-bound compared with the accurate data rate, which is negligible. On the other hand, for the dense scenario, the C-RAN becomes interference limited and the data rate remains fixed for all ranges of the transmit power as expected. It is again observed that the gap between the lower-bound and the exact value is small. Hence, considering the complicated data rate expression in (A.4) in Appendix A, our derived lower-bound expression in (18) is much easier to handle and more suitable for algorithm design.

By replacing the data rate $\bar{r}_k^{(n)}$ in Problems $\mathcal{P}_1$ and $\mathcal{P}_2$ with its lower-bound $\tilde{r}_k^{(n)}$ given in (18) and considering the fact that the minimum rate constraints are met with equality at the optimal point, Problems $\mathcal{P}_1$ and $\mathcal{P}_2$ can be transformed as

$$\mathcal{P}_3: \max_{\mathbf{w}, \mathcal{U} \subseteq \overline{\mathcal{U}}} \quad |\mathcal{U}| \tag{21a}$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C4}: \sum\nolimits_{n \in \mathcal{N}} \tilde{r}_k^{(n)}(\mathbf{w}) \geq R_{k,\min}, \forall k \in \mathcal{U}, \tag{21b}$$

$$\mathrm{C5}: \sum\nolimits_{k \in \mathcal{U}_i} \varepsilon\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) R_{k,\min} \leq C_{i,\max}, \forall i \in \mathcal{I}, \tag{21c}$$

and

$$\mathcal{P}_4: \min_{\mathbf{w}} \quad \tilde{P}_{\mathrm{tot}}(\mathbf{w}) \triangleq \sum_{i \in \mathcal{I}} \left\{ \eta_i P_i^{\mathrm{tr}}(\mathbf{w}) + \varepsilon\left(P_i^{\mathrm{tr}}(\mathbf{w})\right) P_i^{\mathrm{c}} + \rho_i \sum_{k \in \mathcal{U}_i^\star} \varepsilon\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) R_{k,\min} \right\} \tag{22a}$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C4}, \mathrm{C5},$$

respectively.

In the following, we focus on Problems $\mathcal{P}_3$ and $\mathcal{P}_4$.

B. Problem Analysis

By adopting the user-centric clustering method in Section II, the number of optimization variables in Problems $\mathcal{P}_3$ and $\mathcal{P}_4$ has been reduced from $NMI\left|\mathcal{U}\right|$ in fully cooperative transmission scheme to $NM\sum_{k\in\mathcal{U}}\left|\mathcal{I}_k\right|$ here. By appropriately setting the cluster sizes, the reduced number of variables $\left(NM\left(I\left|\mathcal{U}\right|-\sum_{k\in\mathcal{U}}\left|\mathcal{I}_k\right|\right)\right)$ may be very large, which significantly reduces the computational complexity. In addition, some redundant constraints can be removed, which can additionally reduce the computational complexity. For example, in Fig. 1, RRH 3 and RRH 5 are not in any UE's candidate serving set, and thus the power constraints associated with RRH 3 and RRH 5 in C3 can be removed. Moreover, if each link supports at most two UEs, then only link 8 (i.e., RRH 8) should be imposed with the fronthaul capacity constraints. Hence, by employing the user-centric clustering with limited cooperation, the computational complexity can be reduced significantly.

However, Problems $\mathcal{P}_3$ and $\mathcal{P}_4$ are still difficult to solve due to the following reasons. Both the objective functions and constraint C5 contain the non-smooth and non-differential indicator function or (and) continuous variables, which are usually named as an MINLP problem. Although the generalized Benders decomposition method [18], [36] is effective in solving this kind of problems, it is very difficult to directly apply this method to Problems $\mathcal{P}_3$ and $\mathcal{P}_4$ due to the non-convex sum data rate constraints over all multiple SCs. An exhaustive search method can be applied to solve Problems $\mathcal{P}_3$ and $\mathcal{P}_4$. Specifically, to solve Problem $\mathcal{P}_3$, one should check whether Problem $\mathcal{P}_3$ is feasible or not for each given user set $\mathcal{U}$ and each given set of UE-RRH associations. This requires $O\left(2^{|\mathcal{U}|+|\mathcal{U}|I}\right)$ operations, which will become prohibitive for large values of $|\mathcal{U}|$ and I. In addition, even given the selected UE set $\mathcal{U}$ and the set of UE-RRH associations, it is still difficult to check the feasibility since constraint C4 is non-convex. Moreover, for dense C-RAN, the complexity associated with the exhaustive search method is unaffordable for BBU pool. Similar difficulties hold for Problem $\mathcal{P}_4$.

In the next section, we first deal with NPC minimization Problem $\mathcal{P}_4$ by assuming that the UEs have been selected with feasible beam-vectors, then one low-complexity UE selection algorithm to deal with Problem $\mathcal{P}_3$ is provided in Section V.

IV. Low-complexity Algorithm to deal with Problem $\mathcal{P}_4$

In this section, we propose a low-complexity algorithm to solve Problem $\mathcal{P}_4$ when UEs have been selected by using the UE selection algorithms in Section V, and denote the selected subset of UEs as $\mathcal{U}$. As analyzed in Section III-B, there are two difficulties to solve Problem $\mathcal{P}_4$: one is the non-convex sum data rate constraint C4 and the other one is the non-smooth indicator function.

To deal with the first difficulty, we resort to the relationship between the data rate and weighted mean square error (MSE). In [27], the authors considered the sum rate maximization problem by showing that maximizing the sum rate is equivalent to minimizing the weighted MSE. Unfortunately, there are two hurdles that preclude the direct application of the technique in [27]: First, [27] considered the multiple-antenna UEs with perfect CSI. When each UE has only one antenna with perfect CSI, the rank of channel covariance matrices will be equal to one, i.e., $\mathrm{rank}\left(\bar{\mathbf{h}}_{l,k}^{(n)\mathrm{H}}\bar{\mathbf{h}}_{l,k}^{(n)}\right) = 1, \forall l,k,n$. However, for the incomplete CSI considered in this paper, the rank of channel covariance matrix may be larger than 1 according to (19), i.e., $\mathrm{rank}\left(\mathbf{A}_{l,k}^{(n)}\right) > 1, \forall n,l,k$. Second, in [27], the rate expression is in the objective function, while the rate expressions are in the constraints here.

To resolve the first hurdle, we construct an auxiliary signal transmission model by decomposing each interfering UE into multiple interfering sources. Specifically, for each UE k on SC n, since $\mathbf{A}_{l,k}^{(n)}, \forall l \neq k$, are positive definite matrices, they can be decomposed as

$$\mathbf{A}_{l,k}^{(n)} = \mathbf{V}_{l,k}^{(n)}\mathbf{V}_{l,k}^{(n)\mathrm{H}}, \tag{23}$$

where $\mathbf{V}_{l,k}^{(n)} = \left[\mathbf{v}_{l,k,1}^{(n)}, \cdots, \mathbf{v}_{l,k,d_{l,k}^{(n)}}^{(n)}\right], \forall l \neq k$, with $d_{l,k}^{(n)}$ being the rank of $\mathbf{A}_{l,k}^{(n)}$. Then, we construct the following auxiliary signal transmission model for UE k

$$\tilde{y}_k^{(n)} = \bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\tilde{s}_k^{(n)} + \sum\nolimits_{l\in\mathcal{U},l\neq k}\sum\nolimits_{d=1}^{d_{l,k}^{(n)}} \mathbf{v}_{l,k,d}^{(n)\mathrm{H}}\bar{\mathbf{w}}_l^{(n)}\tilde{s}_{l,d}^{(n)} + z_k^{(n)}, \tag{24}$$

where $d_{l,k}^{(n)}$ can be regarded as the number of interfering sources from UE l, $\mathbf{v}_{l,k,d}^{(n)\mathrm{H}}$ can be treated as the CSI from the dth interfering source of UE l to UE k, $\tilde{s}_{l,d}^{(n)}$ is the corresponding transmission data. Both $\tilde{s}_{l,d}^{(n)}$ and $\tilde{s}_k^{(n)}$ are assumed to obey the distribution of $\mathcal{CN}(0,1)$. The data from different interfering sources are mutually independent and independent of $\tilde{s}_k^{(n)}$. Note that all interfering sources from the same UE use the same beam-vector. By using the receive decoding $u_k^{(n)} \in \mathbb{C}$

to decode UE k's received signal on SC n [6], the estimated signal is given by

$$\hat{s}_k^{(n)} = u_k^{(n)\mathrm{H}} \tilde{y}_k^{(n)}, \tag{25}$$

Due to the independence of the transmit data and noise, the mean square error (MSE) matrix at UE k is given by

$$\begin{aligned}
&\epsilon_k^{(n)}\left(\mathbf{u},\mathbf{w}\right)\\
=&\mathbb{E}_{\{\tilde{\mathbf{s}},z_k^{(n)}\}}\left[\left(\hat{s}_k^{(n)}-\tilde{s}_k^{(n)}\right)\left(\hat{s}_k^{(n)}-\tilde{s}_k^{(n)}\right)^H\right]\\
=&\left(u_k^{(n)\mathrm{H}}\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}-1\right)\left(u_k^{(n)\mathrm{H}}\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}-1\right)^{\mathrm{H}}+\sum_{l\in\mathcal{U},l\neq k}\left|u_k^{(n)}\right|^2\bar{\mathbf{w}}_l^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}+\sigma_k^2\left|u_k^{(n)}\right|^2,
\end{aligned} \tag{26}$$

where $\mathbf{u}$ and $\tilde{\mathbf{s}}$ are the collections of decoding variables and data symbols, respectively, and (23) has been used to derive (26).

To deal with the second hurdle, we successfully find a lower bound of the sum rate for each UE and this lower bound is tight at certain point. Then, we replace the sum rate in constraints C4 with its lower bound and iteratively solve the beam-vectors by using the block coordinate decent method. Specifically, defining the following functions:

$$\Psi_k^{(n)}\left(\mathbf{w},q_k^{(n)},u_k^{(n)}\right)=\log_2 e\left(\ln(q_k^{(n)})-q_k^{(n)}\epsilon_k^{(n)}\left(\mathbf{u},\mathbf{w}\right)+1\right),\forall k\in\mathcal{U}, \tag{27}$$

where $q_k^{(n)}\geq 0$ is an introduced variable, we have the following lemma:

Lemma 1: Given the beam-vectors $\mathbf{w}$, function $\Psi_k^{(n)}\left(\mathbf{w},q_k^{(n)},u_k^{(n)}\right)$ is a lower bound for $\tilde{r}_k^{(n)}(\mathbf{w})$. In addition, the optimal $u_k^{(n)}$ and $q_k^{(n)}$ for $\Psi_k^{(n)}\left(\mathbf{w},q_k^{(n)},u_k^{(n)}\right)$ to achieve $\tilde{r}_k^{(n)}(\mathbf{w})$ are

$$u_k^{(n)\star} = \left(\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2+\sum\nolimits_{l\in\mathcal{U},l\neq k}\bar{\mathbf{w}}_l^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}+\sigma_k^2\right)^{-1}\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}, \tag{28}$$

$$q_k^{(n)\star} = \left(\epsilon_k^{(n)}\left(\mathbf{u}^\star,\mathbf{w}\right)\right)^{-1}, \tag{29}$$

where $\epsilon_k^{(n)}\left(\mathbf{u}^\star,\mathbf{w}\right)$ is given by

$$\epsilon_k^{(n)}\left(\mathbf{u}^\star,\mathbf{w}\right)=1-\frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2+\sum\limits_{l\in\mathcal{U},l\neq k}\bar{\mathbf{w}}_l^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_l^{(n)}+\sigma_k^2}. \tag{30}$$

[6]The decoding parameters $\{u_k^{(n)}\in\mathbb{C},\forall n,k\}$ can be iteratively calculated at the BBU pool by using the following iterative algorithm. Then BBU pool will send these parameters to the corresponding UEs for decoding their signals. Note that these parameters cannot be included in beam-vectors, otherwise they will affect the transmit power at each RRH.

Proof: Please see Appendix B. □

By replacing $\tilde{r}_k^{(n)}(\mathbf{w})$ in Problem $\mathcal{P}_4$ with its lower-bound $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$, Problem $\mathcal{P}_4$ can be transformed into the following optimization problem

$$\mathcal{P}_5: \min_{\mathbf{u},\mathbf{q},\mathbf{w}} \quad \tilde{P}_{\mathrm{tot}}(\mathbf{w}) \tag{31a}$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C5},$$

$$\mathrm{C6}: \sum\nolimits_{n\in\mathcal{N}} \Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right) \geq R_{k,\min}, \forall k \in \mathcal{U}, \tag{31b}$$

where $\mathbf{u}$ and $\mathbf{q}$ are the collection of variables $\left\{u_k^{(n)}, \forall n, k\right\}$ and $\left\{q_k^{(n)}, \forall n, k\right\}$, respectively. Note that given $\mathbf{u}$ and $\mathbf{q}$, constraint C6 is a convex set over beam-vectors, which is more tractable than Problem $\mathcal{P}_4$, wherein constraint C4 is non-convex. Hence, Problem $\mathcal{P}_5$ can be solved by using the block coordinate decent method: given $\mathbf{w}$, update $\mathbf{u}$ and $\mathbf{q}$ in (28) and (29), respectively; update $\{\alpha_k\}_{k\in\mathcal{U}}$ and $\mathbf{w}$ with fixed $\mathbf{u}$ and $\mathbf{q}$. We only need to deal with the latter one. Given $\mathbf{u}$ and $\mathbf{q}$, by inserting the MSE expression in (26) into C6, Problem $\mathcal{P}_5$ can be transformed as

$$\mathcal{P}_6: \min_{\mathbf{w}} \ \tilde{P}_{\mathrm{tot}}(\mathbf{w}) \tag{32a}$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C5},$$

$$\mathrm{C7}: \sum_{n\in\mathcal{N}} q_k^{(n)} \left(\left| u_k^{(n)} \right|^2 \bar{\mathbf{w}}_k^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)} \bar{\mathbf{w}}_k^{(n)} - u_k^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)} \bar{\mathbf{w}}_k^{(n)} - u_k^{(n)} \bar{\mathbf{w}}_k^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}} \right)$$

$$+ \sum_{n\in\mathcal{N}} \sum_{l\in\mathcal{U}, l\neq k} q_k^{(n)} \left| u_k^{(n)} \right|^2 \bar{\mathbf{w}}_l^{(n)\mathrm{H}} \mathbf{A}_{l,k}^{(n)} \bar{\mathbf{w}}_l^{(n)} \leq \omega_k, \forall k \in \mathcal{U}, \tag{32b}$$

where $\omega_k = \sum_{n\in\mathcal{N}} \left[\ln\left(q_k^{(n)}\right) - q_k^{(n)} \sigma_k^2 \left| u_k^{(n)} \right|^2 - q_k^{(n)} \right] + N - R_{k,\min} \ln 2$.

Now, we deal with the second difficulty: the non-smooth indicator function $\varepsilon(\cdot)$ in (7) in the objective function and C5 in Problem $\mathcal{P}_6$. The non-smooth indicator function is approximated as a fractional function $f_\theta(x) = \frac{x}{x+\theta}$, where θ is a very small positive value that controls the

smoothness of approximation[7]. Then, $\tilde{P}_{\mathrm{tot}}(\mathbf{w})$ can be approximated as

$$\tilde{P}_{\mathrm{tot}}(\mathbf{w}) \approx \sum\nolimits_{i\in\mathcal{I}}\left\{\eta_i P_i^{\mathrm{tr}}(\mathbf{w}) + f_\theta\left(P_i^{\mathrm{tr}}(\mathbf{w})\right)P_i^{\mathrm{c}} + \rho_i \sum\nolimits_{k\in\mathcal{U}_i} f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) R_{k,\min}\right\} \quad (33)$$

$$\triangleq \hat{P}_{\mathrm{tot},\theta}(\mathbf{w}). \quad (34)$$

Note that for any positive θ, the fractional function $f_\theta(x)$ is strictly smaller than one. Hence, $\hat{P}_{\mathrm{tot},\theta}(\mathbf{w})$ is actually the lower bound of $\tilde{P}_{\mathrm{tot}}(\mathbf{w})$. However, this gap is negligible when θ is very small and x is comparatively large. By replacing the indicator function in Problem $\mathcal{P}_6$ with $f_\theta(x)$, we have

$$\mathcal{P}_7: \min_{\mathbf{w}} \quad \hat{P}_{\mathrm{tot},\theta}(\mathbf{w}) \quad (35a)$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C7},$$

$$\mathrm{C8}: \sum\nolimits_{k\in\mathcal{U}_i} f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) R_{k,\min} \le C_{i,\max}, \forall i \in \mathcal{I}. \quad (35b)$$

Problem $\mathcal{P}_7$ is much more tractable than Problem $\mathcal{P}_6$ since both the objective function and constraints in Problem $\mathcal{P}_7$ are differentiable and continuous. Although Problem $\mathcal{P}_7$ is still nonconvex due to the concavity of $f_\theta(\cdot)$, it is a well-known difference of convex (d.c.) program, which can be efficiently solved by the SCA method [37]. The main idea of this method is to approximate the concave function as its first order Taylor expansion. Specifically, by using the concavity of $f_\theta(\cdot)$, one has

$$f_\theta\left(P_i^{\mathrm{tr}}(\mathbf{w})\right) \le f_\theta\left(P_i^{\mathrm{tr}}(\mathbf{w}(t))\right) + \beta_i(t)\left(P_i^{\mathrm{tr}}(\mathbf{w}) - P_i^{\mathrm{tr}}(\mathbf{w}(t))\right), \quad (36)$$

$$f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right) \le f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w}(t))\right) + \chi_{i,k}(t)\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w}) - P_{i,k}^{\mathrm{tr}}(\mathbf{w}(t))\right) \quad (37)$$

where $\mathbf{w}(t)$ is a collection of beam-vectors at the t^{th} iteration, $\beta_i(t)$ and $\chi_{i,k}(t)$ are given by

$$\beta_i(t) = f_\theta'\left(P_i^{\mathrm{tr}}(\mathbf{w}(t))\right), \chi_{i,k}(t) = f_\theta'\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w}(t))\right), \quad (38)$$

[7]Smaller value of θ will result in more accurate approximation but leads to less smoothness in the function, while larger value of θ leads to high approximation error. From simulations, we find that $\theta = 10^{-5}$ can achieve a good balance between smoothness and approximation accuracy. In the simulations, when transmit power for each link is smaller than 10^{-8} watt, the transmit power is set to be zero. The effect on the rate of each user can be negligible. In addition, for practical analog to digital conversion (ADC) or digital to analog conversion (DAC), there is a minimum required power to activate it. Hence, when the transmit power is very small, it can be ignored.

where $f_\theta^{'}(x)$ denotes the first-order derivative of x. By replacing $f_\theta\left(P_i^{\mathrm{tr}}(\mathbf{w})\right)$ and $f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w})\right)$ in Problem $\mathcal{P}_7$ with the right hand side (RHS) of (36) and (37), respectively, one can solve the following optimization problem in the $(t+1)^{\mathrm{th}}$ iteration

$$\mathcal{P}_8: \min_{\mathbf{w}} \quad \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}\mathbf{G}_k(t)\bar{\mathbf{w}}_k^{(n)} \tag{39a}$$

$$\text{s.t.} \quad \mathrm{C3}, \mathrm{C7},$$

$$\mathrm{C9}: \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}_i} \tau_{i,k}(t)\left\|\mathbf{w}_{i,k}^{(n)}\right\|^2 \le \tilde{C}_{i,\max}(t), \forall i\in\mathcal{I}. \tag{39b}$$

where $\mathbf{G}_k(t)$ is given by $\mathbf{G}_k(t) = \mathrm{blkdiag}\left(\kappa_{s_1^k}(t)\mathbf{I}_{M\times M}, \cdots, \kappa_{s_{|\mathcal{I}_k|}^k}(t)\mathbf{I}_{M\times M}\right)$ with $\kappa_{s_i^k}(t) = \left(\eta_{s_i^k} + \beta_{s_i^k}(t)P_{s_i^k}^{\mathrm{c}} + \rho_{s_i^k}\chi_{s_i^k,k}(t)R_{k,\min}\right), i = 1,\cdots,|\mathcal{I}_k|$, $\tau_{i,k}(t) = \chi_{i,k}(t)R_{k,\min}$, and $\tilde{C}_{i,\max}(t) = C_{i,\max} - \sum_{k\in\mathcal{U}_i}\left(f_\theta\left(P_{i,k}^{\mathrm{tr}}(\mathbf{w}(t))\right) - \chi_{i,k}(t)P_i^{\mathrm{tr}}(\mathbf{w}(t))\right)R_{k,\min}$. Note that some constant terms in the RHS of (36) and (37) are omitted in (39a). Obviously, $\mathbf{G}_k(t)$ is a positive definite matrix and all constraints form a convex set. Then Problem $\mathcal{P}_8$ is a convex problem. The details to solve it will be given in the next subsection.

Based on the above analysis, an iterative algorithm is given to solve Problem $\mathcal{P}_4$. A straightforward way to solve Problem $\mathcal{P}_4$ would involve two layers: the inner layer to solve Problem $\mathcal{P}_8$ by using the SCA method given $\mathbf{u}$ and $\mathbf{q}$; the outer layer to update $\mathbf{u}$ and $\mathbf{q}$ by using (28) and (29) given $\mathbf{w}$. Although the inner layer is guaranteed to converge to a Karush-Kuhn-Tucker (KKT) point of Problem $\mathcal{P}_7$ as proved in [38], this two-layer algorithm will incur high computational complexity. Instead, we merge these two layers into one layer and update $\{\beta_i(t), \tilde{C}_{i,\max}(t), \forall i\}$, $\{\chi_{i,k}(t), \tau_{i,k}(t), \forall i,k\}$, $\mathbf{u}(t)$ and $\mathbf{q}(t)$ at the same layer, as given in Algorithm 1. Fortunately, Algorithm 1 is guaranteed to converge, as proved in Theorem 1.

Theorem 1: Given the feasible initial input $\mathbf{w}(0)$, Algorithm 1 is guaranteed to converge both in objective value and variables.

Proof: Please see Appendix C. □

A. Lagrange dual decomposition method to solve Problem $\mathcal{P}_8$

In Step 2 of Algorithm 1, Problem $\mathcal{P}_8$ should be solved. Since both the maximum power limit and fronthaul capacity limit are positive, i.e., $C_{i,\max} > 0, P_{i,\max} > 0, \forall i \in \mathcal{I}$, the Slater's condition of Problem $\mathcal{P}_8$ is satisfied and the duality gap between Problem $\mathcal{P}_8$ and its dual problem is zero

Algorithm 1 Iterative Algorithm to Solve Problem $\mathcal{P}_4$

1: Initialize the iterative number $t = 1$, error tolerance δ. Initialize $\mathbf{w}(0)$ with the output from the UE selection algorithm in Section V, calculate $\{\beta_i(0), \chi_{i,k}(0), \mathbf{G}_k(0), \tau_{i,k}(0), \tilde{C}_{i,\max}(0), \forall i,k\}$, calculate $\mathbf{u}(0)$ and $\mathbf{q}(0)$ by using (28) and (29) with $\mathbf{w}(0)$, calculate the objective value of Problem $\mathcal{P}_7$, denoted as $\mathrm{Obj}(\mathbf{w}(0))$.
2: Solve Problem $\mathcal{P}_8$ to get $\mathbf{w}(t)$ with $\{\beta_i(t-1), \chi_{i,k}(t-1), \mathbf{G}_k(t-1), \tau_{i,k}(t-1), \tilde{C}_{i,\max}(t-1), \forall i,k\}$, $\mathbf{u}(t-1)$ and $\mathbf{q}(t-1)$;
3: Update $\{\beta_i(t), \chi_{i,k}(t), \mathbf{G}_k(t), \tau_{i,k}(t), \tilde{C}_{i,\max}(t), \forall i,k\}$ with $\mathbf{w}(t)$;
4: Update $\mathbf{u}(t)$ and $\mathbf{q}(t)$ by using (28) and (29) with $\mathbf{w}(t)$;
5: If $|\mathrm{Obj}(\mathbf{w}(t-1)) - \mathrm{Obj}(\mathbf{w}(t))|/\mathrm{Obj}(\mathbf{w}(t)) < \delta$, terminate. Otherwise, set $t \leftarrow t+1$ and go to step 2.

[35]. Hence, Problem $\mathcal{P}_8$ can be equivalently solved by solving its dual problem. In the following, we derive the optimal form of $\mathbf{w}$ for Problem $\mathcal{P}_8$ by using the Lagrange dual decomposition method. For notation simplicity, we omit the iteration index t in the following derivations.

Define the following block diagonal matrices

$$\mathbf{B}_{i,k} = \mathrm{diag}\left\{\overbrace{\mathbf{0}_{1\times M}}^{s_1^k}, \cdots, \overbrace{\mathbf{1}_{1\times M}}^{s_j^k}, \overbrace{\mathbf{0}_{1\times M}}^{s_{j+1}^k}, \cdots, \overbrace{\mathbf{0}_{1\times M}}^{s_{|\mathcal{I}_k|}^k}\right\}, \text{ if } s_j^k = i, \forall i \in \mathcal{I}, k \in \mathcal{U}, \tag{40}$$

then $\left\|\mathbf{w}_{i,k}^{(n)}\right\|^2 = \bar{\mathbf{w}}_k^{(n)\mathrm{H}} \mathbf{B}_{i,k} \bar{\mathbf{w}}_k^{(n)}$. With some manipulations, the Lagrangian function of Problem $\mathcal{P}_8$ is given by

$$\begin{aligned}
&\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\
=\ & \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{U}} \bar{\mathbf{w}}_k^{(n)\mathrm{H}} \mathbf{J}_k^{(n)} \bar{\mathbf{w}}_k^{(n)} - \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{U}} \nu_k q_k^{(n)} \left(u_k^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)} \bar{\mathbf{w}}_k^{(n)} + u_k^{(n)} \bar{\mathbf{w}}_k^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}} \right) \\
& + \ln 2 \sum_{k\in\mathcal{U}} \nu_k R_{k,\min} - \sum_{i\in\mathcal{I}} \lambda_i P_{i,\max} - \sum_{i\in\mathcal{I}} \mu_i \tilde{C}_{i,\max} - \sum_{k\in\mathcal{U}} \nu_k \omega_k,
\end{aligned} \tag{41}$$

where $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}$ are the collections of non-negative Lagrangian multipliers corresponding to C3 in (9), C9 in (39b) and C7 in (32b), respectively, $\mathbf{J}_k^{(n)}$ is given by

$$\mathbf{J}_k^{(n)} = \mathbf{G}_k + \sum_{i\in\mathcal{I}_k} (\lambda_i + \mu_i \tau_{i,k}) \mathbf{B}_{i,k} + \nu_k q_k^{(n)} \left| u_k^{(n)} \right|^2 \bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}} \bar{\mathbf{h}}_{k,k}^{(n)} + \sum_{l\in\mathcal{U}, l\neq k} \nu_l q_l^{(n)} \left| u_l^{(n)} \right|^2 \mathbf{A}_{k,l}^{(n)}. \tag{42}$$

Then, the dual function is given by

$$g\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}\right) \tag{43}$$

$$= \min_{\mathbf{w}} \mathcal{L}\left(\mathbf{w},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}\right) \tag{44}$$

$$= \min_{\mathbf{w}} \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}\mathbf{J}_k^{(n)}\bar{\mathbf{w}}_k^{(n)} - \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}} \nu_k q_k^{(n)}\left(u_k^{(n)\mathrm{H}}\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)} + u_k^{(n)}\bar{\mathbf{w}}_k^{(n)\mathrm{H}}\bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}}\right)$$
$$+ \ln 2\sum_{k\in\mathcal{U}}\nu_k R_{k,\min} - \sum_{i\in\mathcal{I}}\lambda_i P_{i,\max} - \sum_{i\in\mathcal{I}}\mu_i\tilde{C}_{i,\max} - \sum_{k\in\mathcal{U}}\nu_k\omega_k. \tag{45}$$

Obviously, Problem (45) is a strictly convex problem and the optimal solution can be easily obtained from its first-order optimality condition as:

$$\bar{\mathbf{w}}_k^{(n)\star} = \nu_k q_k^{(n)} u_k^{(n)}\left(\mathbf{J}_k^{(n)}\right)^{-1}\bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}}, \forall n\in\mathcal{N}, k\in\mathcal{U}. \tag{46}$$

By inserting the solution of $\{\mathbf{w}_k^\star, k\in\mathcal{U}\}$ in (46) into (45), the dual function can be rewritten as

$$\begin{aligned} g(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}) \;=\; & -\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}}\nu_k^2 q_k^{(n)2}\left|u_k^{(n)}\right|^2\bar{\mathbf{h}}_{k,k}^{(n)}\left(\mathbf{J}_k^{(n)}\right)^{-1}\bar{\mathbf{h}}_{k,k}^{(n)\mathrm{H}} \\ & + \ln 2\sum_{k\in\mathcal{U}}\nu_k R_{k,\min} - \sum_{i\in\mathcal{I}}\lambda_i P_{i,\max} - \sum_{i\in\mathcal{I}}\mu_i\tilde{C}_{i,\max} - \sum_{k\in\mathcal{U}}\nu_k\omega_k. \end{aligned} \tag{47}$$

Hence, the dual problem of Problem $\mathcal{P}_9$ is given by

$$\max_{\{\lambda_i\geq 0,\mu_i\geq 0,\nu_k\geq 0,\forall k,i\}} g(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}). \tag{48}$$

Fortunately, the objective function of the dual Problem (48) is differentiable and dual problem is a convex optimization problem as defined in [35]. Hence, the classic descent methods such as the gradient descent method can be applied to solve it as detailed in [35].

Remark 1 - Parallel Computations: Note that for given Lagrangian multipliers, the optimal beam-vectors $\{\bar{\mathbf{w}}_k^{(n)\star}, \forall k\}$ can be obtained in (46) in closed forms for each SC in parallel. In C-RAN, multiple-core processors or multiple virtual machines (VMs) are aggregated together in the BBU pool, which entails C-RAN to be capable of the parallel computation. Hence, the Lagrange dual decomposition method can run smoothly under the C-RAN architecture.

V. A Low-complexity UE Selection Algorithm

In this section, we propose a low-complexity UE selection algorithms to deal with Problem $\mathcal{P}_3$: the bisection UE selection algorithm, the complexity of which increases logarithmically with the number of UEs K.

Inspired by the UE selection problem formulations (28)-(30) in [39], we first construct the following alternative optimization problem by introducing a series of auxiliary variables $\{\varphi_k\}_{k\in\bar{\mathcal{U}}}$[8]:

$$\mathcal{P}_9: \min_{\{\varphi_k\}_{k\in\bar{\mathcal{U}}},\mathbf{w}} \quad \sum\nolimits_{k\in\mathcal{U}}(\varphi_k-1)^2 \tag{49a}$$

$$\text{s.t.} \quad \text{C3, C5,}$$

$$\text{C10}: \sum\nolimits_{n\in\mathcal{N}} \tilde{r}_k^{(n)}(\mathbf{w}) \geq \varphi_k^2 R_{k,\min}, \forall k\in\bar{\mathcal{U}}, \tag{49b}$$

Obviously, Problem $\mathcal{P}_9$ is always feasible since at least $\{\varphi_k=0, \mathbf{w}_k^{(n)}=\mathbf{0}, \forall k\in\bar{\mathcal{U}}, n\in\mathcal{N}\}$ is a feasible solution. In addition, it is easy to verify that the optimal $\{\varphi_k, k\in\bar{\mathcal{U}}\}$ should lie between zero and one, i.e., $0\leq\varphi_k\leq 1, \forall k\in\bar{\mathcal{U}}$. If UE k can be admitted, the optimal φ_k must be equal to one. This can be easily proved by contradiction. Denote the solution of $\{\varphi_k\}_{k\in\bar{\mathcal{U}}}$ as $\{\varphi_k^\star\}_{k\in\bar{\mathcal{U}}}$. If $\varphi_k^\star=1, \forall k\in\bar{\mathcal{U}}$, all UEs can be admitted in the network and output the corresponding optimal beam-vectors for the initial solution for Algorithm 1 in Section IV. Otherwise, some UEs should be removed. Intuitively, the UE with a smaller $\varphi_k^\star$ should have a higher priority to be removed since it has the largest gap away from its rate targets. Hence, we sort $\{\varphi_k^\star\}_{k\in\bar{\mathcal{U}}}$ in the ascending order: $\varphi_{\pi_1}^\star\leq\cdots\leq\varphi_{\pi_K}^\star$. Then admitting the maximum number of UEs is equivalent to finding a minimum L_0 such that all the users in $\mathcal{U}=\{\pi_{L_0+1},\cdots,\pi_K\}$ can be supported by C-RAN with $L_0=1,\cdots,K-1$. The bisection search procedure can be adopted to determine the minimum L_0. In each iteration of the bisection UE search algorithm, we only need to check whether the C-RAN can support all users in $\mathcal{U}$ or not. Hence, in each iteration, we need to solve the following

[8]The authors in [39] considered the single-channel case by introducing auxiliary variables $\{s_k, \forall k\}$ in each UE's useful signal power. When all the optimal $\{s_k, \forall k\}$ are no larger than zeros, all UEs can be admitted. The method in [39] cannot be directly extended to our work since we consider the multi-channel case. Instead, we introduce the auxiliary variables $\{\varphi_k\}_{k\in\bar{\mathcal{U}}}$ on the right hand side of constraint C10. When all the optimal $\{\varphi_k\}_{k\in\bar{\mathcal{U}}}$ are equal to one, all UEs can be admitted. Note that [39] optimized the UE selection and transmit power minimization problem simultaneously by introducing a large M. In our paper, we consider each problem individually. The reason is that we find from simulations that the big M is difficult to choose and unproperly chosen M may result in unexpected results.

optimization problem

$$\mathcal{P}_{10}: \min_{\varphi,\mathbf{w}} \quad (\varphi-1)^2 \tag{50a}$$

$$\text{s.t.} \quad \text{C3, C5,}$$

$$\text{C11}: \sum\nolimits_{n\in\mathcal{N}} \tilde{r}_k^{(n)}(\mathbf{w}) \geq \varphi^2 R_{k,\min}, \forall k \in \mathcal{U}, \tag{50b}$$

where φ is the introduced optimization variable. Obviously, when the optimal $\varphi^\star$ is equal to one, Problem $\mathcal{P}_{10}$ is feasible. Problem $\mathcal{P}_{10}$ can be similarly solved by using Algorithm 1. Note that all UEs' rate requirements in $\mathcal{P}_{10}$ use the same φ and thus Problem $\mathcal{P}_{10}$ has less variables than Problem $\mathcal{P}_9$.

Finally, the bisection search method is summarized in Algorithm 2. Notice that Problem $\mathcal{P}_{10}$ only needs to be solved no more than $\lceil \log_2(1+K) \rceil$ times.

Algorithm 2 Bisection UE Selection (BUES) Algorithm

1: Solve Problem $\mathcal{P}_9$.

   1) If $\varphi_k^\star = 1, \forall k \in \bar{\mathcal{U}}$, terminate and all UEs can be supported, output the corresponding optimal beam-vectors for the initial point for Algorithm 1;

   2) If there exists at least one UE k such that $\varphi_k^\star < 1$, sort all $\{\varphi_k^\star\}_{k\in\mathcal{U}}$ in the ascending order: $\varphi_{\pi_1}^\star \leq \cdots \leq \varphi_{\pi_K}^\star$, go to step 2;

2: Set $\mathcal{U} = \{\pi_K\}$, solve Problem $\mathcal{P}_{10}$:

   1) If $\varphi_{\pi_K}^\star = 1$, go to step 3;

   2) Otherwise, terminate and claim that no UE can be supported;

3: Initialize $L_{\text{low}} = 0, L_{\text{up}} = K$;

4: Repeat

   1) Set $l \leftarrow \left\lfloor \frac{L_{\text{low}}+L_{\text{up}}}{2} \right\rfloor$;

   2) Solve Problem $\mathcal{P}_{10}$ with $\mathcal{U} = \{\pi_{l+1}, \cdots, \pi_K\}$. If $\varphi^\star = 1$, set $L_{\text{up}} = l$; Otherwise, set $L_{\text{low}} = l$;

5: Until $L_{\text{up}} - L_{\text{low}} = 1$. Output the optimal active UE set $\mathcal{U} = \{\pi_{L_{\text{low}}+1}, \cdots, \pi_K\}$ and the corresponding optimal beam-vectors.

VI. Simulation Results

A. System parameters

In this section, we present simulation results to evaluate the performance of the proposed algorithms. The dense C-RAN is within a square area of coordinates $[-1000, 1000] \times [-1000, 1000]$ meters. Both the UEs and RRHs are assumed to be independently and uniformly distributed in this square area. The channel model consists of three parts: 1) the channel path-loss modeled as $PL_{i,k} = 148.1 + 37.6\log_{10} d_{i,k}$ (dB) [40], where $d_{i,k}$ (in km) is the distance between the ith RRH to the kth UE; 2) the log-normal shadowing with zero mean and 8 dB standard derivation; 3) small-scale Rayleigh fading with zero mean and unit variance. All the UEs are assumed to have the same rate requirements, i.e., $R_{k,\min} = R_{\min}, \forall k$. For ease of exposition, each fronthaul link is assumed to have the same capacity constraints, i.e., $C_{i,\max} = C_{\max}, \forall i$. Then, normalized fronthaul capacity is considered, i.e., $\tilde{C}_{\max} = C_{\max}/R_{\min}$, which represents the maximum number of UEs that can be supported on each fronthaul link is the same. It is assumed that each UE is potentially served by its nearest X RRHs, i.e., $|\mathcal{I}_k| = X, \forall k$. Also, each UE is assumed to measure its channel vectors to its nearest Y RRHs, i.e., $\left|\widetilde{\mathcal{I}}_k\right| = Y, \forall k$. Unless stated otherwise, the system parameters are set as follows: $M = 2$, $K = 16$, $I = 20$, $N = 3$, system bandwidth $B = 10$ MHz, error tolerance $\delta = 10^{-3}$, noise power spectral density is -174 dBm/Hz, $P_i^{\text{active}} = 6.8$ Watt, $P_i^{\text{sleep}} = 4.3$ Watt, $\eta_i = 4, \rho_i = 0.5, P_{i,\max} = 2$ Watt, $\forall i$, $\theta = 10^{-5}$, $R_{\min} = 15$ bit/s/Hz, $\tilde{C}_{\max} = 3$, $X = 3$, $Y = 6$.

B. Numerical Results

1) Performance of the UE selection algorithm: Fig. 4 shows the average number of admitted UEs for three different algorithms. Specifically, 'Joint-BUES-alg.' denotes the joint beam direction and power allocation optimization algorithm in Algorithm 2, while 'MF-BUES-alg.' represents that the beam directions are fixed to be the channel direction, and the power allocation problem is solved by using Algorithm 2. Note that beam direction is not optimized in 'MF-BUES-alg.'. This scheme has lower complexity than 'Joint-BUES-alg.', but incurs inferior performance as seen in the following examples. For comparison, the optimal performance obtained by ex-

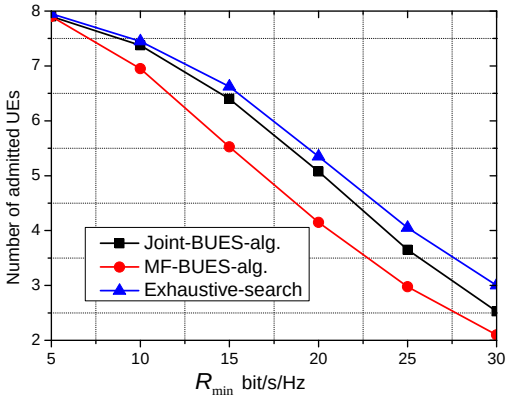

Fig. 4. The average number of admitted UEs for different algorithms with $K = 8$, $I = 12$, $P_{i,\max} = 2$ Watt, $\forall i$, $X = 3$, and $Y = 6$.

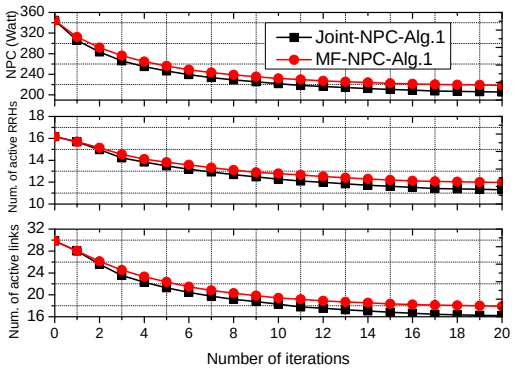

Fig. 5. Convergence behaviour for the NPC minimization algorithm.

haustive search [9] (denoted as 'Exhaustive-search') is also shown, which evaluates every possible subset of UEs and chooses the feasible subset with the maximum number of UEs. Due to the exponential complexity associated with the exhaustive search, we only simulate a small network with $K = 8$ and $I = 12$. As expected, the number of admitted UEs for all algorithms decrease with the increase in the rate requirements. The optimal exhaustive search performs better than the 'Joint-BUES-alg.', which comes at the cost of high computational complexity. However, the performance gap between these two algorithms is negligible when $R_{\min}$ is small (e.g., $R_{\min} < 20$ bit/s/Hz). By jointly optimizing the beam direction and power allocation, the 'Joint-BUES-alg.' outperforms the 'MF-BUES-alg.'. However, the performance gain decreases when $R_{\min}$ is large. The reason can be explained as follows. With the increase of $R_{\min}$, the number of admitted UEs decreases and these UEs are separated far away from each other. Hence, interference is not so significant and the channel matching beam direction approaches

[9]There are many existing MINLP solvers to solve the MINLP problems, such as the generalized Benders decomposition method in [36], [41] and the branch-and-cut (BnC) method in [33]. The main idea of these two methods is to decompose the original problem into several more tractable subproblems and iteratively solve the subproblems until convergence. The condition for convergence is that the globally optimal solution can be obtained. Unfortunately, we consider the multichannel case, wherein each subproblem is non-convex and globally optimal solution cannot be obtained as explained in [42]. Hence, the above two methods are not applicable. Instead, we adopt the exhaustive search method as the performance benchmark that is only simulated in a small network.

the optimal direction.

2) Convergence behaviour of the NPC minimization algorithm: Figure 5 shows the convergence behaviour for the NPC minimization algorithm (i.e., Algorithm 1), where 'Joint-NPC-Alg.1' denotes the joint beam direction and power allocation optimization performed by Algorithm 1, while 'MF-NPC-Alg.1' denotes that the beam direction is fixed to be channel direction and the power optimization is carried out by Algorithm 1. The top subplot shows the NPC trend, the middle and bottom subplots show the numbers of active RRHs and active links remained in each iteration, respectively. It is seen from this figure that all these values decrease rapidly and converge within twenty iterations. Both the convergence speed and the NPC performance for the considered two algorithms are very similar in this scenario. Moreover, for 'Joint-NPC-Alg.1', the NPC, the number of active RRHs and active links decrease about 65%, 45% and 94%, respectively, which confirm the effectiveness of the proposed algorithm in terms of power savings.

In the following, we will evaluate the effects of different system parameters on both the NPC minimization algorithm (i.e.,Algorithm 1) and UE selection algorithm (i.e., Algorithm 2). To compare the performance of the NPC minimization algorithm, the performance of the conventional transmit power minimization is also considered, where all the RRHs in each UE's candidate set are assumed to be active. 'Joint-Conven' and 'MF-Conven' denote the conventional method when beam direction and power allocation are jointly optimized and beam direction is fixed at channel direction, respectively.

3) Effects of the candidate size: Figs. 6 and 7 show the numbers of admitted UEs and NPC versus the candidate size X, respectively. The set of UEs that are admitted by the 'MF-BUES-alg.' are set as the initialization point for the NPC minimization algorithms, which is the same in the following simulations. As expected, the larger candidate size leads to more admitted UEs due to the increasing network degrees of freedom. However, the number of admitted UEs achieved by 'Joint-BUES-alg.' and 'MF-BUES-alg.' become flat in the large candidate size regime, which is consistent with the conclusion in [28]. This means that it is not necessary to consider the far away RRHs for each UE since they contribute less to their performance and the candidate size should not be larger than 4 to obtain a tradeoff between performance and implementation complexity. The similar trend holds for the 'Joint-NPC-Alg.1' and 'MF-NPC-Alg.1' in Figure 7. However, in Figure 7, the conventional transmit power minimization consumes much higher

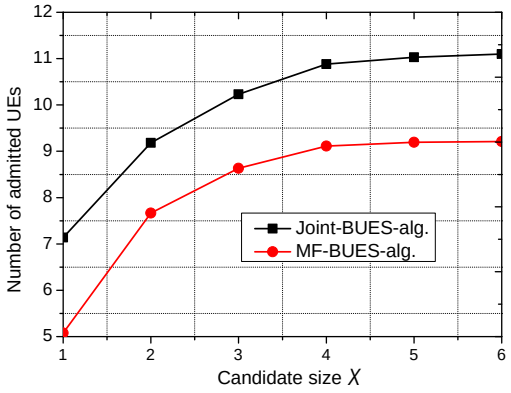

Fig. 6. Number of admitted UEs versus the candidate size X with $Y = 6$.

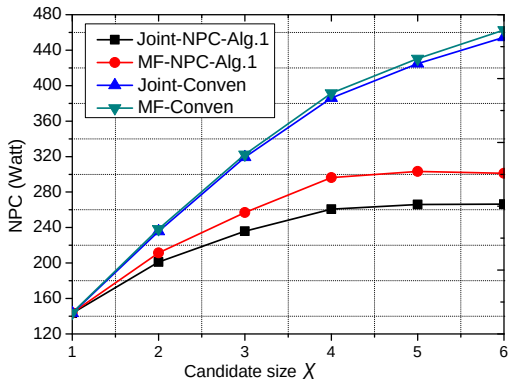

Fig. 7. NPC versus the candidate size X for different algorithms with $Y = 6$.

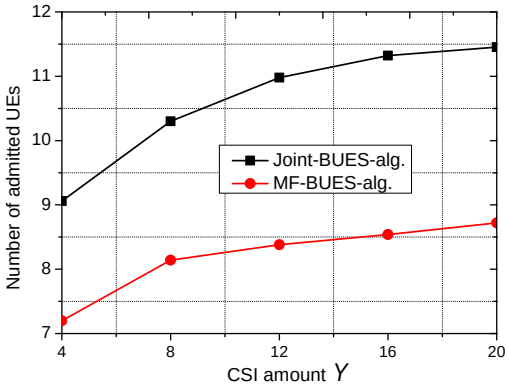

Fig. 8. Number of admitted UEs versus the CSI amount Y with $X = 3$.

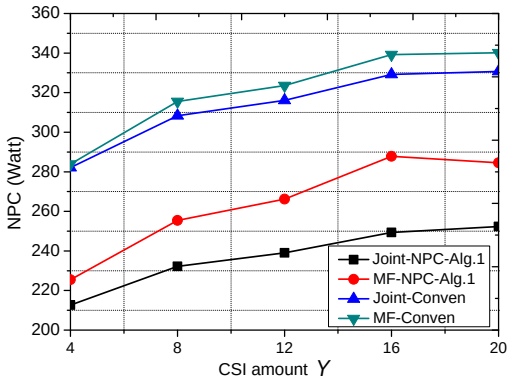

Fig. 9. NPC versus the CSI amount Y for different algorithms with $X = 3$.

power than the proposed 'Joint-NPC-Alg.1' and 'MF-NPC-Alg.1', and the gap increases with the increase of candidate size. The reason is that with the increase of candidate size, more RRHs will be in the active mode, which requires large amount of circuit power consumption. It is surprising to see from Figure 7 that for the conventional method, 'MF-Conven' requires slightly higher power consumption than 'Joint-Conven'. On the other hand, 'MF-NPC-Alg.1' requires much higher power than 'Joint-NPC-Alg.1'. These two facts confirm that joint beam direction and power allocation optimization is more important for RRH and link selection.

4) Effects of the amount of CSI: Now, we investigate the effects of limited CSI on the performance of the proposed algorithms. Figs. 8 and 9 illustrate the number of admitted UEs

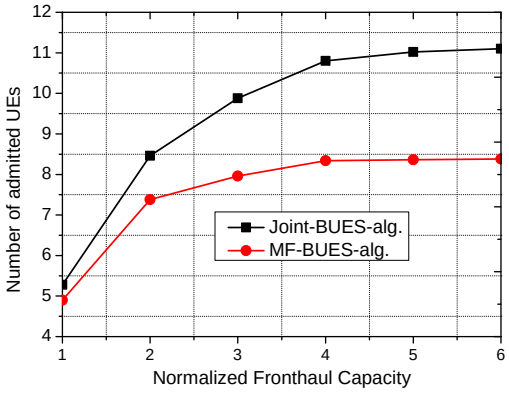

Fig. 10. Number of admitted UEs versus the normalized fronthaul capacity $\tilde{C}_{\max}$.

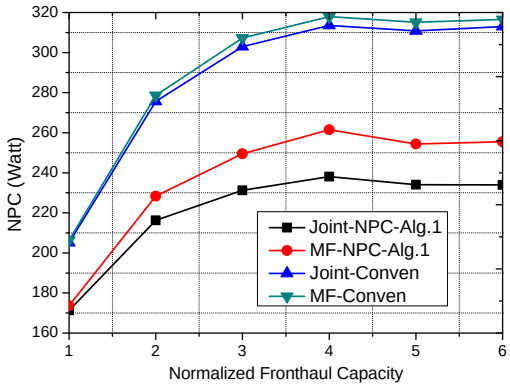

Fig. 11. NPC versus the normalized fronthaul capacity $\tilde{C}_{\max}$ for different algorithms.

and NPC versus the amount of CSI Y, respectively. Note that Y denotes the number of (nearest) RRHs from which CSI is measured. As expected, the number of admitted UEs increase as the amount of CSI increases, since multi-user interference can be more accurately suppressed. From Figure 8, it is seen that the number of admitted UEs increases quickly when $Y < 12$ and increases slowly in the high amount CSI regime. This result indicates that only a moderate amount of CSI is sufficient for the proposed algorithms to achieve good performance, which can significantly reduce the channel estimation overhead. The corresponding NPC increases sightly with the amount of CSI due to more UEs are admitted. The proposed algorithms are again observed to perform much better than the conventional transmit power minimization method, highlighting the importance of joint optimization of transmit power, RRH and link selection.

5) Effects of fronthaul capacity constraints: Figs. 10 and 11 show the number of admitted UEs and NPC versus the normalized fronthaul capacity $C_{\max}$, respectively. It is seen from Figure 10 that the numbers of admitted UEs for both the 'Joint-BUES-alg.' and 'MF-BUES-alg.' increase with $C_{\max}$ initially due to the fact that more UEs can be supported by each fronthaul link for large $C_{\max}$. However, the number of admitted UEs will be saturated in the large $C_{\max}$ regime. It is shown that $C_{\max} = 4$ is enough to achieve a large portion of the optimal performance, which indicates that the fronthaul link capacity is not necessary to be very large and the wireless fronthaul link such as mmWave communication technologies may be applicable in dense C-RAN network. Figure 10 also shows that 'Joint-BUES-alg.' outperforms 'MF-BUES-alg.' in terms of

the number of admitted UEs and the performance gain increases with $C_{\max}$. From Figure 11, it can be seen that the NPC performances of 'Joint-NPC-Alg.1' and 'MF-NPC-Alg.1' increase with $C_{\max}$ when $C_{\max} <= 4$, since more UEs are admitted in this regime as seen from Figure 10. However, when $C_{\max} >= 4$, the NPC value experiences a slight decrease though the numbers of UEs are almost the same as seen in Figure 10. This is due to the fact that more flexibility in the fronthaul link can be exploited to reduce the numbers of active links and RRHs.

VII. Conclusions

This paper provided a complete framework to handle the challenges arising in the dense C-RAN. More specifically, the downlink beam-vectors, RRH selection and UE-RRH associations were jointly optimized to minimize the total NPC for dense C-RAN with incomplete CSI subject to fronthaul capacity constraints, UEs' QoS targets and per-RRH power constraints. We formulated this problem as an MINLP problem, which is NP-hard. In addition, the incomplete CSI makes the QoS constraints difficult to handle. We first replaced the exact expression of data rate with its lower bound. Then, we developed a low-complexity single-layer iterative algorithm to solve the NPC minimization problem based on the successive convex approximation technique and the equivalent relationship between data rate and MSE. Also, a low-complexity UE selection algorithm was proposed to guarantee the feasibility of the NPC problem. Simulation results showed that the proposed UE selection can achieve near-optimal performance compared to the optimal exhaustive UE search method. Moreover, the proposed single-layer iterative algorithm can achieve significant power savings in various setups. Simulation results also showed that only nearest four RRHs are sufficient to be the candidate set of each UE and limited CSI can contribute large portion of performance gain from the full CSI case.

The future work lies in the joint optimization of cluster sizes and beam-vectors when taking into account the cost of computational complexity and channel training overhead. Also, it is worth studying how to extend the work to the scenario where each UE is equipped with multiple antennas.

APPENDIX A
ACCURATE CLOSED-FORM EXPRESSION OF DATA RATE FOR SPECIAL CASE

For the simplicity of notations, the SC index n is omitted in the following derivations. The SINR for UE k in (14) can be rewritten as

$$\gamma_k = \frac{|X_k|^2}{\sum_{l\neq k, l\in\mathcal{U}} |Y_{l,k}|^2 + \sigma_k^2}, \tag{A.1}$$

where $X_k = \bar{\mathbf{h}}_{k,k}\bar{\mathbf{w}}_k$ and $Y_{l,k} = \bar{\mathbf{h}}_{l,k}\bar{\mathbf{w}}_l$. Note that $\bar{\mathbf{h}}_{k,k}$ is perfectly known and $\bar{\mathbf{w}}_l, \forall l$ are deterministic, X_k is a deterministic value and only $\{Y_{l,k}, \forall l \in \mathcal{U}, l \neq k\}$ are random variables. According to the first two assumptions, all elements in $\bar{\mathbf{h}}_{l,k}$ are unknown and follow the circular symmetric complex Gaussian distribution. Specifically, the distribution of $\bar{\mathbf{h}}_{l,k}$ is given by $\mathcal{CN}(\mathbf{0}, \mathbf{R}_{l,k})$, where $\mathbf{R}_{l,k}$ is a diagonal matrix. To obtain the expression of $\mathbf{R}_{l,k}$, we define the indices of $\mathcal{I}_l$ as $\mathcal{I}_l = \{s_1^l, \cdots, s_{|\mathcal{I}_l|}^l\}$. Then, based on the third assumption, $\mathbf{R}_{l,k}$ can be easily calculated as $\mathbf{R}_{l,k} = \text{blkdiag}\left(\left|\alpha_{s_1^l,k}\right|^2 \mathbf{I}_{M\times M}, \cdots, \left|\alpha_{s_{|\mathcal{I}_l|}^l,k}\right|^2 \mathbf{I}_{M\times M}\right)$. Then, given beam-vector $\bar{\mathbf{w}}_l$, $Y_{l,k}$ is a Gaussian random variable with zero mean and variance given by $\varpi_{l,k} = \bar{\mathbf{w}}_l^H \mathbf{R}_{l,k}\bar{\mathbf{w}}_l$, i.e., $Y_{l,k} \sim \mathcal{CN}(\mathbf{0}, \varpi_{l,k})$. For convenience, denote $Z_k = \sum_{l\neq k, l\in\mathcal{U}} |Y_l|^2$. Then, Z_k follows a generalized chi-squared distribution, given by [43]

$$f(z_k) = \sum\nolimits_{l\neq k, l\in\mathcal{U}} T_{l,k} e^{-z_k/\varpi_{l,k}}, \tag{A.2}$$

where $T_{l,k}$ is given by

$$T_{l,k} = \frac{1}{\varpi_{l,k}\prod_{j\in\mathcal{U}, j\neq l,k}\left(1 - \frac{\varpi_{j,k}}{\varpi_{l,k}}\right)}.$$

Then, the data rate is derived as

$$\begin{aligned}
&\int_0^\infty \log_2\left(1 + \frac{|X_k|^2}{z_k + \sigma_k^2}\right) f(z_k) dz_k \\
&= \sum_{l\neq k, l\in\mathcal{U}} T_{l,k} \int_0^\infty \log_2\left(1 + \frac{|X_k|^2}{z_k + \sigma_k^2}\right) e^{-\frac{z_k}{\varpi_{l,k}}} dz_k \\
&= \sum_{l\neq k, l\in\mathcal{U}} -\frac{T_{l,k}\varpi_{l,k}}{\ln 2} \int_0^\infty \left[\ln\left(z_k + \sigma_k^2 + |X_k|^2\right) - \ln\left(z_k + \sigma_k^2\right)\right] de^{-\frac{z_k}{\varpi_{l,k}}} \\
&= \sum_{l\neq k, l\in\mathcal{U}} \frac{T_{l,k}\varpi_{l,k}}{\ln 2}\left[\ln\left(1 + \frac{|X_k|^2}{\sigma_k^2}\right) + \int_0^\infty \frac{e^{-\frac{z_k}{\varpi_{l,k}}}}{z_k + \sigma_k^2 + |X_k|^2} dz_k - \int_0^\infty \frac{e^{-\frac{z_k}{\varpi_{l,k}}}}{z_k + \sigma_k^2} dz_k\right] \qquad \text{(A.3)} \\
&= \sum_{l\neq k, l\in\mathcal{U}} \frac{T_{l,k}\varpi_{l,k}}{\ln 2}\left[\ln\left(1 + \frac{|X_k|^2}{\sigma_k^2}\right) - e^{\frac{\sigma_k^2 + |X_k|^2}{\varpi_{l,k}}} \text{Ei}\left(-\frac{\sigma_k^2 + |X_k|^2}{\varpi_{l,k}}\right) + e^{\frac{\sigma_k^2}{\varpi_{l,k}}} \text{Ei}\left(-\frac{\sigma_k^2}{\varpi_{l,k}}\right)\right] \qquad \text{(A.4)}
\end{aligned}$$

where $\mathrm{Ei}(x) = -\int_{-x}^{\infty} (e^{-t}/t)\, dt$ is an exponential integral function, (A.3) is obtained by using integration by parts, and (A.4) is achieved by invoking [Eq. (3.352.4), [44]].

APPENDIX B
PROOF OF LEMMA 1

We prove that $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$ is a lower bound of $\tilde{r}_k^{(n)}(\mathbf{w})$ by showing that given $\mathbf{w}$, the maximum of $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$ is equal to $\tilde{r}_k^{(n)}(\mathbf{w})$.

Obviously, function $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$ is respectively concave over $u_k^{(n)}$, $q_k^{(n)}$ and $\mathbf{w}$ when the other two are fixed. As a result, the optimal $u_k^{(n)}$ and $q_k^{(n)}$ to achieve the maximum value of $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$ are obtained by setting the first order of $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$ to zero, which are given in (28) and (29), respectively.

By inserting the expression of $u_k^{(n)\star}$ in (28) into (26), the expression of $\epsilon_k^{(n)}(\mathbf{u}^\star, \mathbf{w})$ can be obtained by (30). By substituting the optimal $u_k^{(n)\star}$ in (28) and $q_k^{(n)\star}$ in (29) into function $\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)}, u_k^{(n)}\right)$, we have

$$\begin{aligned}
\Psi_k^{(n)}\left(\mathbf{w}, q_k^{(n)\star}, u_k^{(n)\star}\right) &= \log_2 e \ln\left(1 - \frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2 + \sum\limits_{l\in\mathcal{U}, l\neq k} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_k^{(n)} + \sigma_k^2}\right)^{-1} && \text{(B.1)}\\
&= \log_2\left(1 + \frac{\left|\bar{\mathbf{h}}_{k,k}^{(n)}\bar{\mathbf{w}}_k^{(n)}\right|^2}{\sum\limits_{l\in\mathcal{U}, l\neq k} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}\mathbf{A}_{l,k}^{(n)}\bar{\mathbf{w}}_k^{(n)} + \sigma_k^2}\right) && \text{(B.2)}\\
&= \tilde{r}_k^{(n)}(\mathbf{w}). && \text{(B.3)}
\end{aligned}$$

Hence, the proof is complete.

APPENDIX C
PROOF OF THEOREM 1

Before proving the theorem, we first construct the following auxiliary problem

$$\begin{aligned}
\mathcal{P}_X: \ \min_{\mathbf{w}} \quad & \hat{P}_{\mathrm{tot},\theta}(\mathbf{w}) && \text{(C.1a)}\\
\text{s.t.} \quad & \mathrm{C3}, \mathrm{C4}, \mathrm{C8}. && \text{(C.1b)}
\end{aligned}$$

Note that Problem $\mathcal{P}_X$ has the same objective function as Problem $\mathcal{P}_7$. In the following, we show that Algorithm 1 actually solves Problem $\mathcal{P}_X$. In addition, the only difference between the original Problem $\mathcal{P}_4$ and Problem $\mathcal{P}_X$ is that the indicator function is replaced by the concave smooth function.

Since $\{\mathbf{w}(0), \forall k\}$ is initialized by using the output from the UE selection algorithm, $\mathbf{w}(0)$ is a feasible solution of Problem $\mathcal{P}_4$. By using the fact that $f_\theta(x) < 1, \forall x$, we conclude that $\mathbf{w}(0)$ is also feasible for $\mathcal{P}_X$. Note that $\mathbf{u}(0)$ and $\mathbf{q}(0)$ are calculated by using (28) and (29) with $\mathbf{w}(0)$. Then by using Lemma 1, $\mathbf{w}(0)$ is a feasible solution of Problem $\mathcal{P}_7$ with fixed $\mathbf{u}(0)$ and $\mathbf{q}(0)$. It is easy to check that $\mathbf{w}(0)$ is also a feasible solution of Problem $\mathcal{P}_8$ with $\{\beta_i(0), \chi_{i,k}(0), \mathbf{G}_k(0), \tau_{i,k}(0), \tilde{C}_{i,\max}(0), \forall i, k\}$. Now, we consider step 2 of the first iteration (i.e., $t = 1$) of Algorithm 1. Since $\mathbf{w}(1)$ is the optimal solution of Problem $\mathcal{P}_8$, we have

$$\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}(1)\mathbf{G}_k(0)\bar{\mathbf{w}}_k^{(n)}(1) \leq \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{U}} \bar{\mathbf{w}}_k^{(n)\mathrm{H}}(0)\mathbf{G}_k(0)\bar{\mathbf{w}}_k^{(n)}(0). \tag{C.2}$$

For the simplicity of representation, denote $\psi_i(t) = P_i^{\mathrm{tr}}(\mathbf{w}(t))$ and $\xi_{i,k}(t) = P_{i,k}^{\mathrm{tr}}(\mathbf{w}(t))$. Then, we have

$$\begin{aligned}
&\mathrm{Obj}(\mathbf{w}(1))\\
&= \sum_{i\in\mathcal{I}} \left(\eta_i\psi_i(1) + f_\theta\left(\psi_i(1)\right) P_i^{\mathrm{c}} + \rho_i \sum_{k\in\mathcal{U}_i} f_\theta\left(\xi_{i,k}(1)\right) R_{k,\min} \right)\\
&\overset{(a)}{\leq} \sum_{i\in\mathcal{I}} \left(\eta_i\psi_i(1) + f_\theta\left(\psi_i(0)\right) P_i^{\mathrm{c}} + \beta_i(0) P_i^{\mathrm{c}}\left(\psi_i(1) - \psi_i(0)\right)\right) +\\
&\quad \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{U}_i} \rho_i R_{k,\min}\left(f_\theta\left(\xi_{i,k}(0)\right) + \chi_{i,k}(0)\left(\xi_{i,k}(1) - \xi_{i,k}(0)\right)\right)\\
&\overset{(b)}{\leq} \sum_{i\in\mathcal{I}} \left(\eta_i\psi_i(0) + f_\theta\left(\psi_i(0)\right) P_i^{\mathrm{c}} + \beta_i(0) P_i^{\mathrm{c}}\left(\psi_i(0) - \psi_i(0)\right)\right) +\\
&\quad \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{U}_i} \rho_i R_{k,\min}\left(f_\theta\left(\xi_{i,k}(0)\right) + \chi_{i,k}(0)\left(\xi_{i,k}(0) - \xi_{i,k}(0)\right)\right)\\
&= \sum_{i\in\mathcal{I}} \left(\eta_i\psi_i(0) + f_\theta\left(\psi_i(0)\right) P_i^{\mathrm{c}} + \rho_i \sum_{k\in\mathcal{U}_i} f_\theta\left(\xi_{i,k}(0)\right) R_{k,\min} \right)\\
&= \mathrm{Obj}(\mathbf{w}(0))
\end{aligned}$$

where $\mathrm{Obj}(\mathbf{w}(t))$ denotes the objective value of Problem $\mathcal{P}_7$ or $\mathcal{P}_X$, (a) follows by using (36) and (37), (b) follows due to (C.2).

Next, we show that $\mathbf{w}(1)$ is also a feasible solution of Problem $\mathcal{P}_X$. Obviously, $\mathbf{w}(1)$ satisfies C3, i.e., the power constraints. We only need to prove that $\mathbf{w}(1)$ satisfies C4 and C8.

Since $\mathbf{w}(1)$ is the optimal solution of Problem $\mathcal{P}_8$, we have

$$\sum\nolimits_{n\in\mathcal{N}} \Psi_k^{(n)}\left(\mathbf{w}(1), q_k^{(n)}(0), u_k^{(n)}(0)\right) \geq R_{k,\min}. \tag{C.3}$$

In step 4 of the first iteration of Algorithm 1, $\mathbf{u}(1)$ and $\mathbf{q}(1)$ are updated by using (28) and (29) with $\mathbf{w}(1)$. Then according to Lemma 1, we have

$$\sum_{n\in\mathcal{N}} \tilde{r}_k^{(n)}(\mathbf{w}(1)) = \sum_{n\in\mathcal{N}} \Psi_k^{(n)}\left(\mathbf{w}(1), q_k^{(n)}(1), u_k^{(n)}(1)\right) \tag{C.4}$$

$$\geq \sum_{n\in\mathcal{N}} \Psi_k^{(n)}\left(\mathbf{w}(1), q_k^{(n)}(0), u_k^{(n)}(0)\right) \tag{C.5}$$

$$\geq R_{k,\min}. \tag{C.6}$$

Hence, C4 is satisfied.

In addition, we have

$$C_{i,\max} \geq \sum_{k\in\mathcal{U}_i} \left(f_\theta\left(\xi_{i,k}(0)\right) + \chi_{i,k}(0)\left(\xi_{i,k}(1) - \xi_{i,k}(0)\right)\right) R_{k,\min} \tag{C.7}$$

$$\geq \sum_{k\in\mathcal{U}_i} f_\theta\left(\xi_{i,k}(1)\right) R_{k,\min}, \tag{C.8}$$

where (C.7) follows since $\mathbf{w}(1)$ is the solution of Problem $\mathcal{P}_8$ given $\mathbf{u}(0)$ and $\mathbf{q}(0)$, and (C.8) follows by using (37). Hence, C8 is satisfied.

As a result, we can conclude that $\mathbf{w}(1)$ is also feasible for Problem $\mathcal{P}_X$. By using the similar method, we can obtain

$$\mathrm{Obj}(\mathbf{w}(0)) \geq \mathrm{Obj}(\mathbf{w}(1)) \geq \mathrm{Obj}(\mathbf{w}(2)) \geq \cdots. \tag{C.9}$$

Obviously, the objective value of Problem $\mathcal{P}_7$ (also $\mathcal{P}_X$) is lower bounded by zero. Hence, Algorithm 1 is guaranteed to converge in objective values.

Next, we prove the second part of the theorem: Given the feasible input $\mathbf{w}(0)$, the solution obtained from Algorithm 1 will converge to a unique point. Obviously, when $\mathbf{w}$ is given, $\mathbf{u}$ and $\mathbf{q}$ can be uniquely determined by using (28) and (29), respectively. Since $\{\mathbf{G}_k, \forall k\}$ are positive definite matrices, the objective function in Problem $\mathcal{P}_8$ is a strictly convex function of $\mathbf{w}$. Furthermore, it can be easily proved that the constraints in Problem $\mathcal{P}_8$ are convex. As a

result, Problem $\mathcal{P}_8$ is a strictly convex optimization problem. According to [Page 137 in [35]], the globally optimal solution is unique. Then, by iteratively updating step 2 to step 4 in Algorithm 1, the algorithm will converge to a unique solution. We emphasize that since Problem $\mathcal{P}_X$ is a non-convex optimization problem, it may have multiple locally optimal solutions and its converged unique solution depends on the initial point. However, once the initial point is given, Algorithm 1 will converge to a unique solution.

REFERENCES

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] H. Zhu and J. Wang, "Chunk-based resource allocation in ofdma systems - part i: chunk allocation," *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2734–2744, 2009.

[3] ——, "Chunk-based resource allocation in ofdma systems - part ii: Joint chunk, power and bit allocation," *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 499–509, 2012.

[4] H. Zhu, "Radio resource allocation for ofdma systems in high speed environments," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 748–759, 2012.

[5] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, 2016.

[6] 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 Communications*, vol. 21, no. 6, pp. 126–135, 2014.

[7] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[8] H. Zhu, "Performance comparison between distributed antenna and microcellular systems," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1151–1163, 2011.

[9] J. Wang, H. Zhu, and N. J. Gomes, "Distributed antenna systems for mobile communications in high speed trains," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 675–683, 2012.

[10] C. Mobile, "C-RAN: the road towards green RAN," *White Paper, ver*, vol. 2, 2011.

[11] G. Caire, S. A. Ramprashad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Information Theory and Applications Workshop (ITA), 2010*, 2010, pp. 1–10.

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

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

[14] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1037–1050, 2016.

[15] C. PAN, H. Zhu, N. Gomes, and J. Wang, "Joint precoding and RRH selection for user-centric green MIMO C-RAN," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1–1, 2017.

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

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

[18] D. W. K. Ng and R. Schober, "Secure and green SWIPT in distributed antenna networks with limited backhaul capacity," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5082–5097, 2015.

[19] L. Liu and R. Zhang, "Downlink SINR balancing in C-RAN under limited fronthaul capacity," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 3506–3510.

[20] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-RANs," in *2014 IEEE Global Communications Conference*, 2014, pp. 4054–4059.

[21] 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, 2016.

[22] Y. Shi, J. Zhang, and K. B. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 5154–5159.

[23] J. Kim, H. W. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1126–1138, 2014.

[24] T. R. Lakshmana, A. T?lli, R. Devassy, and T. Svensson, "Precoder design with incomplete feedback for joint transmission," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1923–1936, 2016.

[25] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2063–2077, 2016.

[26] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.

[27] 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, 2011.

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

[29] A. Papadogiannis, H. J. Bang, D. Gesbert, and E. Hardouin, "Downlink overhead reduction for multi-cell cooperative processing enabled wireless networks," in *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2008, pp. 1–5.

[30] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2017.

[31] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

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

[33] Y. Cheng, M. Pesavento, and A. Philipp, "Joint network optimization and downlink beamforming for CoMP transmissions using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, 2013.

[34] A. J. Fehske, P. Marsch, and G. P. Fettweis, "Bit per joule efficiency of cooperating base stations in cellular networks," in *GLOBECOM Workshops (GC Wkshps), 2010 IEEE*, 2010, pp. 1406–1411.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[36] R. Ramamonjison, A. Haghnegahdar, and V. K. Bhargava, "Joint optimization of clustering and cooperative beamforming in green cognitive wireless networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 982–997, 2014.

[37] Q. T. Dinh and M. Diehl, "Local convergence of sequential convex programming for nonconvex optimization," in *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 93–102.

[38] C. Pan, W. Xu, W. Zhang, J. Wang, H. Ren, and M. Chen, "Weighted sum energy efficiency maximization in ad hoc networks," *IEEE Wireless Communications Letters*, vol. 4, no. 3, pp. 233–236, 2015.

[39] 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, 2008.

[40] G. T. RAN and T. . v7.1.0, "Physical layer aspects for evolved UTRA," *3GPP Technical Specification TR*, Sep.

[41] D. W. K. Ng, Y. Wu, and R. Schober, "Power efficient resource allocation for full-duplex radio distributed antenna networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2896–2911, 2016.

[42] C. Pan, J. Wang, W. Zhang, B. Du, and M. Chen, "Power minimization in multi-band multi-antenna cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5056–5069, 2014.

[43] Z. Ye, C. Pan, H. Zhu, and J. Wang, "Tradeoff caching strategy of outage probability and fronthaul usage in cloud-RAN," *arXiv preprint arXiv:1611.02660*, 2016.

[44] A. Jeffrey and D. Zwillinger, *Table of integrals, series, and products*. Academic Press, 2007.