Group Sparse Beamforming for Green Cloud-RAN

Yuanming Shi, *Student Member, IEEE,* Jun Zhang, *Member, IEEE,* and Khaled B. Letaief, *Fellow, IEEE*

***Abstract*—A cloud radio access network (Cloud-RAN) is a network architecture that holds the promise of meeting the explosive growth of mobile data traffic. In this architecture, all the baseband signal processing is shifted to a single baseband unit (BBU) pool, which enables efficient resource allocation and interference management. Meanwhile, conventional powerful base stations can be replaced by low-cost low-power remote radio heads (RRHs), producing a green and low-cost infrastructure. However, as all the RRHs need to be connected to the BBU pool through optical transport links, the transport network power consumption becomes significant. In this paper, we propose a new framework to design a green Cloud-RAN, which is formulated as a joint RRH selection and power minimization beamforming problem. To efficiently solve this problem, we first propose a greedy selection algorithm, which is shown to provide near-optimal performance. To further reduce the complexity, a novel group sparse beamforming method is proposed by inducing the group-sparsity of beamformers using the weighted ℓ_1/ℓ_2-norm minimization, where the group sparsity pattern indicates those RRHs that can be switched off. Simulation results will show that the proposed algorithms significantly reduce the network power consumption and demonstrate the importance of considering the transport link power consumption.**

***Index Terms*—Cloud-RAN, green communications, coordinated beamforming, greedy selection, group-sparsity.**

I. Introduction

MOBILE data traffic has been growing enormously in recent years, and it is expected that cellular networks will have to offer a 1000x increase in capacity in the following decade to meet this demand [1]. Massive MIMO [2] and heterogeneous and small cell networks (HetSNets) [1] are regarded as two most promising approaches to achieve this goal. By deploying a large number of antennas at each base station (BS), massive MIMO can exploit spatial multiplexing gain in a large scale and also improve energy efficiency. However, the performance of massive MIMO is limited by correlated scattering with the antenna spacing constraints, which also brings high deployment cost to maintain the minimum spacing [1]. HetSNets exploit the spatial reuse by deploying more and more access points (APs). Meanwhile, as stated in [3], placing APs based on the traffic demand is an effective way for compensating path-loss, resulting in energy efficient cellular networks. However, efficient interference management is challenging for dense small-cell networks. Moreover, deploying more and more small-cells will cause significant cost and operating challenges for operators.

Cloud radio access network (Cloud-RAN) has recently been proposed as a promising network architecture to unify the above two technologies in order to jointly manage the interference (via coordinated multiple-point process (CoMP)), increase network capacity and energy efficiency (via network densification), and reduce both the network capital expenditure (CAPEX) and operating expense (OPEX) (by moving baseband processing to the baseband unit (BBU) pool) [4], [5]. A large-scale distributed cooperative MIMO system will thus be formed. Cloud-RAN can therefore be regarded as the ultimate solution to the "spectrum crunch" problem of cellular networks.

There are three key components in a Cloud-RAN: (i) a pool of BBUs in a datacenter *cloud*, supported by the real-time virtualization and high performance processors, where all the baseband processing is performed; (ii) a high-bandwidth low-latency optical transport network connecting the BBU pool and the remote radio heads (RRHs); and (iii) distributed transmission/reception points (i.e., RRHs). The key feature of Cloud-RAN is that RRHs and BBUs are separated, resulting a centralized BBU pool, which enables efficient cooperation of the transmission/reception among different RRHs. As a result, significant performance improvements through joint scheduling and joint signal processing such as coordinated beamforming or multi-cell processing[6] can be achieved. With efficient interference suppression, a network of RRHs with a very high density can be deployed. This will also reduce the communication distance to the mobile terminals and can thus significantly reduce the transmission power. Moreover, as baseband signal processing is shifted to the BBU pool, RRHs only need to support basic transmission/reception functionality, which further reduces their energy consumption and deployment cost.

The new architecture of Cloud-RAN also indicates a paradigm shift in the network design, which causes some technical challenges for implementation. For instance, as the data transmitted between the RRHs and the BBU pool is typically oversampled real-time I/Q digital data streams in the order of Gbps, high-bandwidth optical transport links with low latency will be needed. To support CoMP and enable computing resource sharing among BBUs, new virtualization technologies need to be developed to distribute or group the BBUs into a centralized entity [4]. Another important aspect is the energy efficiency consideration, due to the increased power consumption of a large number of RRHs and also of the transport links.

Conventionally, the transport network (i.e., backhaul links

Manuscript received September 26, 2013; revised January 9, 2014; accepted January 31, 2014. The associate editor coordinating the review of this paper and approving it for publication was K. Huang.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (e-mail: {yshiac, eejzhang, eekhaled}@ust.hk).

This work was supported by the Hong Kong Research Grant Council under Grant No. 610212. The work of J. Zhang was supported by the Hong Kong RGC Direct Allocation Grant DAG11EG03.

Part of this work was presented at the IEEE Global Communications Conference (GLOBECOM), Atlanta, GA, Dec. 2013.

Digital Object Identifier 10.1109/TWC.2014.040214.131770

between the core network and base stations (BSs)) power consumption can be ignored as it is negligible compared to the power consumption of macro BSs. Therefore, all the previous works investigating the energy efficiency of cellular networks only consider the BS power consumption [7], [8]. Recently, the impact of the backhaul power consumption in cellular networks was investigated in [9], where it was shown through simulations that the backhaul power consumption will affect the energy efficiency of different cellular network deployment scenarios. Subsequently, Rao *et al.* in [10] investigated the spectral efficiency and energy efficiency tradeoff in homogeneous cellular networks when taking the backhaul power consumption into consideration.

In Cloud-RAN, the transport network power consumption will have a more significant impact on the network energy efficiency. Hence, allowing the transport links and the corresponding RRHs to support the sleep mode will be essential to reduce the network power consumption for the Cloud-RAN. Moreover, with the spatial and temporal variation of the mobile traffic, it would be feasible to switch off some RRHs while still maintaining the quality of service (QoS) requirements. It will be also practical to implement such an idea in the Cloud-RAN with the help of centralized signal processing at the BBU pool. As energy efficiency is one of the major objectives for future cellular networks [5], in this paper we will focus on the design of green Cloud-RAN by jointly considering the power consumption of the transport network and RRHs.

A. Contributions

The main objective of this paper is to minimize the network power consumption of Cloud-RAN, including the transport network and radio access network power consumption, with a QoS constraint at each user. Specifically, we formulate the design problem as a joint RRH selection and power minimization beamforming problem, where the transport network power consumption is determined by the set of active RRHs, while the transmit power consumption of the active RRHs is minimized through coordinated beamforming. This is a mixed-integer non-linear programming (MINLP) problem, which is NP-hard. We will focus on designing low-complexity algorithms for practical implementation. The major contributions of the paper are summarized as follows:

1) We formulate the network power consumption minimization problem for the Cloud-RAN by enabling both the transport links and RRHs to support the sleep mode. In particular, we provide a group sparse beamforming (GSBF) formulation of the design problem, which assists the problem analysis and algorithm design.
2) We first propose a greedy selection (GS) algorithm, which selects one RRH to switch off at each step. It turns out that the RRH selection rule is critical, and we propose to switch off the RRH that *maximizes the reduction in the network power consumption* at each step. From the simulations, the proposed GS algorithm often yields optimal or near-optimal solutions, but its complexity may still be prohibitive for a large-sized network.
3) To further reduce the complexity, we propose a three-stage group sparse beamforming (GSBF) framework, by adopting the weighted mixed ℓ_1/ℓ_p-norm to induce the group sparsity for the beamformers. In contrast to all the previous works applying the mixed ℓ_1/ℓ_p-norm to induce group sparsity, we exploit the additional prior information (i.e., transport link power consumption, power amplifier efficiency, and instantaneous effective channel gains) to design the weights for different beamformer coefficient groups, resulting in a significant performance gain. Two GSBF algorithms with different complexities are proposed: namely, a bi-section GSBF algorithm and an iterative GSBF algorithm.
4) We shall show that the GS algorithm always provides near-optimal performance. Hence, it would be a good option if the number of RRHs is relatively small, such as in clustered deployment. With a very low computational complexity, the bi-section GSBF algorithm is an attractive option for a large-scale Cloud-RAN. The iterative GSBF algorithm provides a good tradeoff between the complexity and performance, which makes it a good candidate for a medium-size network.

B. Related Works

A main design tool applied in this paper is optimization with the group sparsity induced norm. With the recent theoretical breakthrough in compressed sensing [11], [12], the sparsity patterns in different applications in signal processing and communications have been exploited for more efficient system design, e.g., for pilot aided sparse channel estimation [13]. The sparsity inducing norms have been widely applied in high-dimensional statistics, signal processing, and machine learning in the last decade [14]. The ℓ_1-norm regularization has been successfully applied in compressed sensing [11], [12]. More recently, mixed ℓ_1/ℓ_p-norms are widely investigated in the case where some variables forming a group will be selected or removed simultaneously, where the mixed ℓ_1/ℓ_2-norm [15] and mixed ℓ_1/ℓ_∞-norm [16] are two commonly used ones to induce group sparsity for their computational and analytical convenience.

In Cloud-RAN, one RRH will be switched off only when all the coefficients in its beamformer are set to zeros. In other words, all the coefficients in the beamformer at one RRH should be selected or ignored simultaneously, which requires group sparsity rather than individual sparsity for the coefficients as commonly used in compressed sensing. In this paper, we will adopt the mixed ℓ_1/ℓ_p-norm to promote group sparsity for the beamformers instead of ℓ_1-norm, which only promotes individual sparsity. Recently, there are some works [17]–[19] adopting the mixed ℓ_1/ℓ_p-norm to induce group-sparsity in a large-scale cooperative wireless cellular network. Specifically, Hong *et al.* [17] adopted the mixed ℓ_1/ℓ_2-norm and Zhao *et al.* [18] used the ℓ_2-norm to induce the group sparsity of the beamformers, which reduce the amount of the shared user data among different BSs. The squared mixed ℓ_1/ℓ_∞-norm was investigated in [19] for antenna selection.

All of the above works simply adopted the un-weighted mixed ℓ_1/ℓ_p-norms to induce group-sparsity, in which, no

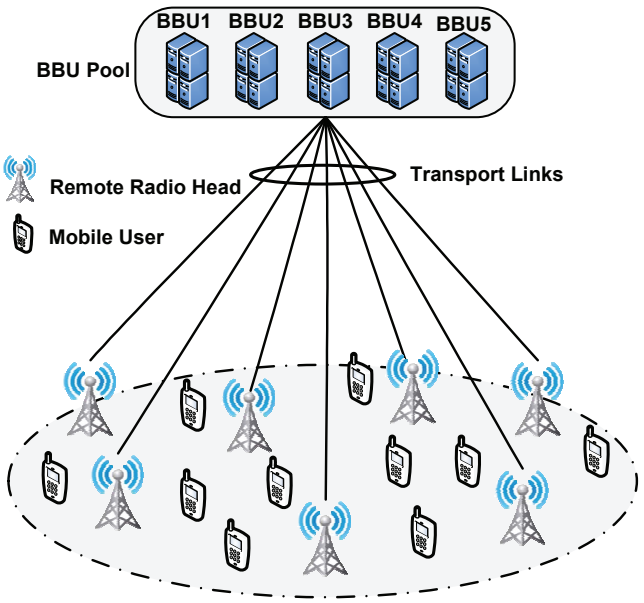


Fig. 1. The architecture of Cloud-RAN, in which, all the RRHs are connected to a BBU pool through transport links.

prior information of the unknown signal is assumed other than the fact that it is sufficiently sparse. By exploiting the prior information in terms of system parameters, the weights for different beamformer coefficient groups can be more rigorously designed and performance can be enhanced. We demonstrate through simulations that the proposed three-stage GSBF framework, which is based on the weighted mixed ℓ_1/ℓ_p-norm minimization, outperforms the conventional unweighted mixed ℓ_1/ℓ_p-norm minimization based algorithms substantially.

C. Organization

The remainder of the paper is organized as follows. Section II presents the system and power model. In Section III, the network power consumption minimization problem is formulated, followed by some analysis. Section IV presents the GS algorithm, which yields near-optimal solutions. The three-stage GSBF framework is presented in Section V. Simulation results will be presented in Section VI. Finally, conclusions and discussions are presented in Section VII.

Notations: $\|\cdot\|_{\ell_p}$ is the ℓ_p-norm. Boldface lower case and upper case letters represent vectors and matrices, respectively. $(\cdot)^T$, $(\cdot)^\dagger$, $(\cdot)^{\mathsf{H}}$ and $\mathrm{Tr}(\cdot)$ denote the transpose, conjugate, Hermitian and trace operators, respectively. $\Re(\cdot)$ denotes the real part.

II. System and Power Model

A. System Model

We consider a Cloud-RAN with L remote radio heads (RRHs), where the l-th RRH is equipped with N_l antennas, and K single-antenna mobile users (MUs), as shown in Fig. 1. In this network architecture, all the base band units (BBUs) are moved into a single BBU pool, creating a set of shared processing resources, and enabling efficient interference management and mobility management. With the baseband signal processing functionality migrated to the BBU pool, the RRHs can be deployed in a large scale with low-cost. The BBU pool is connected to the RRHs using the common public radio interface (CPRI) transport technology via a high-bandwidth, low-latency optical transport network [4]. In order to enable full cooperation among RRHs, it is assumed that all the user data are routed to the BBU pool from the core network through the backhaul links [4], i.e., all users can access all the RRHs. The digitized baseband complex inphase (I) and quadrature (Q) samples of the radio signals are transported over the transport links between the BBUs and RRHs. The key technical and economic issue of the Cloud-RAN is that this architecture requires significant transport network resources. As the focus of this paper is on network power consumption, we will assume all the transport links have sufficiently high capacity and negligible latency[1].

Due to the high density of RRHs and the joint transmission among them, the energy used for signal transmission will be reduced significantly. However, the power consumption of the transport network becomes enormous and cannot be ignored. Therefore, it is highly desirable to switch off some transport links and the corresponding RRHs to reduce the network power consumption based on the data traffic requirements, which forms the main theme of this work.

Let $\mathcal{L} = \{1, ..., L\}$ denote the set of RRH indices, $\mathcal{A} \subseteq \mathcal{L}$ denote the active RRH set, $\mathcal{Z}$ denote the inactive RRH set with $\mathcal{A} \cup \mathcal{Z} = \mathcal{L}$, and $\mathcal{S} = \{1, ..., K\}$ denote the index set of scheduled users. In a beamforming design framework, the baseband transmit signals are of the form:

$$\mathbf{x}_l = \sum_{k=1}^{K} \mathbf{w}_{lk} s_k, \forall l \in \mathcal{A}, \tag{1}$$

where s_k is a complex scalar denoting the data symbol for user k and $\mathbf{w}_{lk} \in \mathbb{C}^{N_l}$ is the beamforming vector at RRH l for user k. Without loss of generality, we assume that $E[|s_k|^2] = 1$ and s_k's are independent with each other. The baseband signals $\mathbf{x}_l$'s will be transmitted to the corresponding RRHs, but not the data information s_k's [4], [21]. The baseband received signal at user k is given by

$$y_k = \sum_{l \in \mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}} \mathbf{w}_{lk} s_k + \sum_{i \neq k} \sum_{l \in \mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}} \mathbf{w}_{li} s_i + z_k, k \in \mathcal{S}, \tag{2}$$

where $\mathbf{h}_{kl} \in \mathbb{C}^{N_l}$ is the channel vector from RRH l to user k, and $z_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the additive Gaussian noise.

We assume that all the users are employing single user detection (i.e., treating interference as noise), so that they can use the receivers with a low-complexity and energy-efficient structure. Moreover, in the low interference region, treating interference as noise can be optimal [22]. The corresponding signal-to-interference-plus-noise ratio (SINR) for user k is hence given by

$$\mathrm{SINR}_k = \frac{|\sum_{l \in \mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}} \mathbf{w}_{lk}|^2}{\sum_{i \neq k} |\sum_{l \in \mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}} \mathbf{w}_{li}|^2 + \sigma_k^2}, \forall k \in \mathcal{S}. \tag{3}$$

[1]The impact of limited-capacity transport links on compression in Cloud-RAN was recently investigated in [20], [21], and its impact in our setting is left to future work.

Each RRH has its own transmit power constraint

$$\sum_{k=1}^{K} \|\mathbf{w}_{lk}\|_{\ell_2}^2 \leq P_l, \forall l \in \mathcal{A}. \quad (4)$$

B. Power Model

The network power model is critical for the investigation of the energy efficiency of Cloud-RAN, which is described as follows.

1) RRH Power Consumption Model: We will adopt the following empirical linear model [23] for the power consumption of an RRH:

$$P_l^{\text{rrh}} = \begin{cases} P_{a,l}^{\text{rrh}} + \frac{1}{\eta_l} P_l^{\text{out}}, & \text{if } P_l^{\text{out}} > 0, \\ P_{s,l}^{\text{rrh}}, & \text{if } P_l^{\text{out}} = 0. \end{cases} \quad (5)$$

where $P_{a,l}^{\text{rrh}}$ is the active power consumption, which depends on the number of antennas N_l, $P_{s,l}^{\text{rrh}}$ is the power consumption in the sleep mode, P_l^{out} is the transmit power, and η_l is the drain efficiency of the radio frequency (RF) power amplifier. For the Pico-BS, the typical values are $P_{a,l}^{\text{rrh}} = 6.8W$, $P_{s,l}^{\text{rrh}} = 4.3W$, and $\eta_l = 1/4$ [23]. Based on this power consumption model, we conclude that it is essential to put the RRHs into sleep whenever possible.

2) Transport Network Power Consumption Model: Although there is no superior solution to meet the low-cost, high-bandwidth, low-latency requirement of transport networks for the Cloud-RAN, the future passive optical network (PON) can provide cost-effective connections between the RRHs and the BBU pool [24]. PON comprises an optical line terminal (OLT) that connects a set of associated optical network units (ONUs) through a single fiber. Implementing a sleep mode in the optical network unit (ONU) has been considered as the most cost-effective and promising power-saving method [25] for the PON, but the OLT cannot go into the sleep mode and its power consumption is fixed[25]. Hence, the total power consumption of the transport network is given by [25]

$$P^{\text{tn}} = P_{\text{olt}} + \sum_{l=1}^{L} P_l^{\text{tl}}, \quad (6)$$

where P_{olt} is the OLT power consumption, $P_l^{\text{tl}} = P_{a,l}^{\text{tl}}$ and $P_l^{\text{tl}} = P_{s,l}^{\text{tl}}$ denote the power consumed by the ONU l (or the transport link l) in the active mode and sleep mode, respectively. The typical values are $P_{\text{olt}} = 20W$, $P_{a,l}^{\text{tl}} = 3.85W$ and $P_{s,l}^{\text{tl}} = 0.75W$ [25]. Thus, we conclude that putting some transport links into the sleep mode is a promising way to reduce the power consumption of Cloud-RAN.

3) Network Power Consumption: Based on the above discussion, we define $P_l^a \triangleq P_{a,l}^{\text{rrh}} + P_{a,l}^{\text{tl}}$ ($P_l^s \triangleq P_{s,l}^{\text{rrh}} + P_{s,l}^{\text{tl}}$) as the active (sleep) power consumption when both the RRH and the corresponding transport link are switched on (off). Therefore, the network power consumption of the Cloud-RAN is given by

$$\begin{aligned} \hat{p}(\mathcal{A}) &= \sum_{l\in\mathcal{A}} \frac{1}{\eta_l} P_l^{\text{out}} + \sum_{l\in\mathcal{A}} P_l^a + \sum_{l\in\mathcal{Z}} P_l^s + P_{\text{olt}} \\ &= \sum_{l\in\mathcal{A}} \frac{1}{\eta_l} P_l^{\text{out}} + \sum_{l\in\mathcal{A}} (P_l^a - P_l^s) + \sum_{l\in\mathcal{L}} P_l^s + P_{\text{olt}} \\ &= \sum_{l\in\mathcal{A}} \sum_{k=1}^{K} \frac{1}{\eta_l} \|\mathbf{w}_{lk}\|_{\ell_2}^2 + \sum_{l\in\mathcal{A}} P_l^c + \sum_{l\in\mathcal{L}} P_l^s + P_{\text{olt}}, \quad (7) \end{aligned}$$

where $P_l^{\text{out}} = \sum_{k=1}^{K} \|\mathbf{w}_{lk}\|_{\ell_2}^2$ and $P_l^c = P_l^a - P_l^s$, and the second equality in (7) is based on the fact $\sum_{l\in\mathcal{Z}} P_l^s = \sum_{l\in\mathcal{L}} P_l^s - \sum_{l\in\mathcal{A}} P_l^s$. Given a Cloud-RAN with the RRH set $\mathcal{L}$, the term $(\sum_{l\in\mathcal{L}} P_l^s + P_{\text{olt}})$ in (7) is a constant. Therefore, minimizing the total network power consumption $\hat{p}(\mathcal{A})$ (7) is equivalent to minimizing the following *re-defined* network power consumption by omitting the constant term $(\sum_{l\in\mathcal{L}} P_l^s + P_{\text{olt}})$:

$$p(\mathcal{A}, \mathbf{w}) = \sum_{l\in\mathcal{A}} \sum_{k=1}^{K} \frac{1}{\eta_l} \|\mathbf{w}_{lk}\|_{\ell_2}^2 + \sum_{l\in\mathcal{A}} P_l^c, \quad (8)$$

where $\mathbf{w} = [\mathbf{w}_{11}^T, \ldots, \mathbf{w}_{1K}^T, \ldots, \mathbf{w}_{L1}^T, \ldots, \mathbf{w}_{LK}^T]^T$. The advantage of introducing the term P_l^c is that we can rewrite the network power consumption model (7) in a more compact form as in (8) and extract the relevant parameters for our system design. In the following discussion, we refer to P_l^c as the *relative transport link power consumption* for simplification. Therefore, the first part of (8) is the total transmit power consumption and the second part is the total relative transport link power consumption.

Note 1: The re-defined network power consumption model (8) reveals two key design parameters: the transmit power consumption ($\frac{1}{\eta_l} \sum_{k=1}^{K} \|\mathbf{w}_{lk}\|_{\ell_2}^2$) and the relative transport link power consumption P_l^c. With the typical values provided in Section II-B1 and Section II-B2, the maximum transmit power consumption, i.e., $\frac{1}{\eta_l} P_l^{\text{out}} = 4W$, is comparable with the relative transport link power consumption, i.e., $P_l^c = P_l^a - P_l^s = (P_{a,l}^{\text{rrh}} + P_{a,l}^{\text{tl}}) - (P_{s,l}^{\text{rrh}} + P_{s,l}^{\text{tl}}) = 5.6W$. This implies that a joint RRH selection (and the corresponding transport link selection) and power minimization beamforming is required to minimize the network power consumption.

III. Problem Formulation and Analysis

Based on the power consumption model, we will formulate the network power consumption minimization problem in this section.

A. Power Saving Strategies and Problem Formulation

The network power consumption model (8) indicates the following two strategies to reduce the network power consumption:

- Reduce the transmission power consumption;
- Reduce the number of active RRHs and the corresponding transport links.

However, the two strategies conflict with each other. Specifically, in order to reduce the transmission power consumption, more RRHs are required to be active to exploit a

higher beamforming gain. On the other hand, allowing more RRHs to be active will increase the power consumption of transport links. As a result, the network power consumption minimization problem requires a joint design of RRH (and the corresponding transport link) selection and coordinated transmit beamforming.

In this work, we assume perfect channel state information (CSI) available at the BBU pool. With target SINRs $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$, the network power consumption minimization problem can be formulated as

$$\begin{aligned} \mathscr{P}: \underset{\{\mathbf{w}_{lk}\},\mathcal{A}}{\text{minimize}} \quad & p(\mathcal{A}, \mathbf{w}) \\ \text{subject to} \quad & \frac{|\sum_{l\in\mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}}\mathbf{w}_{lk}|^2}{\sum_{i\neq k}|\sum_{l\in\mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}}\mathbf{w}_{li}|^2+\sigma_k^2} \geq \gamma_k, \\ & \sum_{k=1}^{K} \|\mathbf{w}_{lk}\|_{\ell_2}^2 \leq P_l, l \in \mathcal{A}. \end{aligned} \tag{9}$$

Problem $\mathscr{P}$ is a joint RRH set selection and transmit beamforming problem, which is difficult to solve in general. In the following, we will analyze and reformulate it.

B. Problem Analysis

We first consider the case with a given active RRH set $\mathcal{A}$ for problem $\mathscr{P}$, resulting a network power minimization problem $\mathscr{P}(\mathcal{A})$. Let $\mathbf{w}_k = [\mathbf{w}_{lk}^T]^T \in \mathbb{C}^{\sum_{l\in\mathcal{A}} N_l}$ indexed by $l \in \mathcal{A}$, and $\mathbf{h}_k = [\mathbf{h}_{lk}^T]^T \in \mathbb{C}^{\sum_{l\in\mathcal{A}} N_l}$ indexed by $l \in \mathcal{A}$, such that $\mathbf{h}_k^{\mathsf{H}}\mathbf{w}_k = \sum_{l\in\mathcal{A}} \mathbf{h}_{kl}^{\mathsf{H}}\mathbf{w}_{lk}$. Since the phases of $\mathbf{w}_k$ will not change the objective function and constraints of $\mathscr{P}(\mathcal{A})$ [26], the SINR constraints are equivalent to the following second order cone (SOC) constraints:

$$\mathcal{C}_1(\mathcal{A}): \sqrt{\sum_{i\neq k} |\mathbf{h}_k^{\mathsf{H}}\mathbf{w}_i|^2 + \sigma_k^2} \leq \frac{1}{\sqrt{\gamma_k}} \Re(\mathbf{h}_k^{\mathsf{H}}\mathbf{w}_k), k \in \mathcal{S}. \tag{10}$$

The per-RRH power constraints (4) can be rewritten as

$$\mathcal{C}_2(\mathcal{A}): \sqrt{\sum_{k=1}^{K} \|\mathbf{A}_{lk}\mathbf{w}_k\|_{\ell_2}^2} \leq \sqrt{P_l}, l \in \mathcal{A}, \tag{11}$$

where $\mathbf{A}_{lk} \in \mathbb{C}^{\sum_{l\in\mathcal{A}} N_l \times \sum_{l\in\mathcal{A}} N_l}$ is a block diagonal matrix with the identity matrix $\mathbf{I}_{N_l}$ as the l-th main diagonal block square matrix and zeros elsewhere. Therefore, given the active RRH set $\mathcal{A}$, the network power minimization problem is given by

$$\begin{aligned} \mathscr{P}(\mathcal{A}): \underset{\mathbf{w}_1,\ldots,\mathbf{w}_K}{\text{minimize}} \quad & \sum_{l\in\mathcal{A}} \left(\sum_{k=1}^{K} \frac{1}{\eta_l} \|\mathbf{A}_{lk}\mathbf{w}_k\|_{\ell_2}^2 + P_l^c \right) \\ \text{subject to} \quad & \mathcal{C}_1(\mathcal{A}), \mathcal{C}_2(\mathcal{A}), \end{aligned} \tag{12}$$

with the optimal value denoted as $p^\star(\mathcal{A})$. This is a second-order cone programming (SOCP) problem, and can be solved efficiently, e.g., via interior point methods [27].

Based on the solution of $\mathscr{P}(\mathcal{A})$, the network power minimization problem $\mathscr{P}$ can be solved by searching over all the possible RRH sets, i.e.,

$$p^\star = \underset{Q\in\{J,\ldots,L\}}{\text{minimize}} \; p^\star(Q), \tag{13}$$

where $J \geq 1$ is the minimum number of RRHs that makes the network support the QoS requirements, and $p^\star(Q)$ is determined by

$$p^\star(Q) = \underset{\mathcal{A}\subseteq\mathcal{L}, |\mathcal{A}|=Q}{\text{minimize}} \; p^\star(\mathcal{A}), \tag{14}$$

where $p^\star(\mathcal{A})$ is the optimal value of the problem $\mathscr{P}(\mathcal{A})$ in (12) and $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. The number of subsets $\mathcal{A}$ of size m is $\binom{L}{m}$, which can be very large. Thus, in general, the overall procedure will be exponential in the number of RRHs L and thus cannot be applied in practice. Therefore, we will reformulate this problem to develop more efficient algorithms to solve it.

C. Group Sparse Beamforming Formulation

One way to solve problem $\mathscr{P}$ is to reformulate it as a MINLP problem [28], and the generic algorithms for solving MINLP can be applied. Unfortunately, due to the high complexity, such an approach can only provide a performance benchmark for a simple network setting. In the following, we will pursue a different approach, and try to exploit the problem structure.

We will exploit the group sparsity of the optimal aggregative beamforming vector $\mathbf{w}$, which can be written as a partition:

$$\mathbf{w} = [\underbrace{\mathbf{w}_{11}^T, \ldots, \mathbf{w}_{1K}^T}_{\tilde{\mathbf{w}}_1^T}, \ldots, \underbrace{\mathbf{w}_{L1}^T, \ldots, \mathbf{w}_{LK}^T}_{\tilde{\mathbf{w}}_L^T}]^T, \tag{15}$$

where all the coefficients in a given vector $\tilde{\mathbf{w}}_l = [\mathbf{w}_{l1}^T, \ldots, \mathbf{w}_{lK}^T]^T \in \mathbb{C}^{KN_l}$ form a *group*. When the RRH l is switched off, the corresponding coefficients in the vector $\tilde{\mathbf{w}}_l$ will be set to zeros simultaneously. Overall there may be multiple RRHs being switched off and the corresponding beamforming vectors will be set to zeros. That is, $\mathbf{w}$ has a group sparsity structure, with the priori knowledge that the blocks of variables in $\tilde{\mathbf{w}}_l$'s should be selected (the corresponding RRH will be switched on) or ignored (the corresponding RRH will be switched off) simultaneously.

Define $N = K\sum_{l=1}^{L} N_l$ and an index set $\mathcal{V} = \{1, 2, \ldots, N\}$ with its power set as $2^{\mathcal{V}} = \{\mathcal{I}, \mathcal{I} \subseteq \mathcal{V}\}$. Furthermore, define the sets $\mathcal{G}_l = \{K\sum_{i=1}^{l-1} N_i + 1, \ldots, K\sum_{i=1}^{l} N_i\}, l = 1, \ldots, L$, as a partition of $\mathcal{V}$, such that $\tilde{\mathbf{w}}_l = [w_i]$ is indexed by $i \in \mathcal{G}_l$. Define the support of beamformer $\mathbf{w}$ as

$$\mathcal{T}(\mathbf{w}) = \{i | w_i \neq 0\}, \tag{16}$$

where $\mathbf{w} = [w_i]$ is indexed by $i \in \mathcal{V}$. Hence, the total relative transport link power consumption can be written as

$$F(\mathcal{T}(\mathbf{w})) = \sum_{l=1}^{L} P_l^c I(\mathcal{T}(\mathbf{w}) \cap \mathcal{G}_l \neq \emptyset), \tag{17}$$

where $I(\mathcal{T} \cap \mathcal{G}_l \neq \emptyset)$ is an indicator function that takes value 1 if $\mathcal{T} \cap \mathcal{G}_l \neq \emptyset$ and 0 otherwise. Therefore, the network power minimization problem $\mathscr{P}$ is equivalent to the following group sparse beamforming (GSBF) formulation

$$\begin{aligned} \mathscr{P}_{\text{sparse}}: \underset{\mathbf{w}}{\text{minimize}} \quad & T(\mathbf{w}) + F(\mathcal{T}(\mathbf{w})) \\ \text{subject to} \quad & \mathcal{C}_1(\mathcal{L}), \mathcal{C}_2(\mathcal{L}), \end{aligned} \tag{18}$$

where $T(\mathbf{w}) = \sum_{l=1}^{L}\sum_{k=1}^{K} \frac{1}{\eta_l}\|\mathbf{w}_{lk}\|_{\ell_2}^2$ represents the total transmit power consumption. The equivalence means that if $\mathbf{w}^\star$ is a solution to $\mathscr{P}_{\text{sparse}}$, then $(\{\mathbf{w}_{lk}^\star\}, \mathcal{A}^\star)$ with $\mathcal{A}^\star = \{l : \mathcal{T}(\mathbf{w}^\star) \cap \mathcal{G}_l \neq \emptyset\}$ is a solution to $\mathscr{P}$, and vice versa.

Note that the group sparsity of $\mathbf{w}$ is fundamentally different from the conventional sparsity measured by the ℓ_0-norm of

$\mathbf{w}$, which is often used in compressed sensing [11], [12]. The reason is that although the ℓ_0-norm of $\mathbf{w}$ will result in a sparse solution for $\mathbf{w}$, the zero entries of $\mathbf{w}$ will not necessarily align to a same group $\tilde{\mathbf{w}}_l$ to lead to switch off one RRH. As a result, the conventional ℓ_1-norm relaxation [11], [12] to the ℓ_0-norm will not work for our problem. Therefore, we will adopt the mixed ℓ_1/ℓ_p-norm [14] to induce group sparsity for $\mathbf{w}$. The details will be presented in Section V. Note that the "group" in this work refers to the collection of beamforming coefficients associated with each RRH, but not a subset of RRHs.

Since obtaining the global optimization solutions to problem $\mathscr{P}$ is computationally difficult, in the following sections, we will propose two low-complexity algorithms to solve it. We will first propose a greedy algorithm in Section IV, which can be viewed as an approximation to the iteration procedure of (13). In order to further reduce the complexity, based on the GSBF formulation $\mathscr{P}_{\text{sparse}}$, a three-stage GSBF framework will then be developed based on the group-sparsity inducing norm minimization in Section V.

IV. Greedy Selection Algorithm

In this section, we develop a heuristic algorithm to solve $\mathscr{P}$ based on the backward greedy selection, which was successfully applied in spare filter design [29] and has been shown to often yield optimal or near-optimal solutions. The backward greedy selection algorithm iteratively selects one RRH to switch off at each step, while re-optimizing the coordinated transmit beamforming for the remaining active RRH set. The key design element for this algorithm is the selection rule of the RRHs to determine which one should be switched off at each step.

A. Greedy Selection Procedure

Denote the iteration number as $i = 0, 1, 2, \ldots$. At the ith iteration, $\mathcal{A}^{[i]} \subseteq \mathcal{L}$ shall denote the set of active RRHs, and $\mathcal{Z}^{[i]}$ denotes the inactive RRH set with $\mathcal{Z}^{[i]} \cup \mathcal{A}^{[i]} = \mathcal{L}$. At iteration i, an additional RRH $r^{[i]} \in \mathcal{A}^{[i]}$ will be added to $\mathcal{Z}^{[i]}$, resulting in a new set $\mathcal{Z}^{[i+1]} = \mathcal{Z}^{[i]} \cup \{r^{[i]}\}$ after this iteration. We initialize by setting $\mathcal{Z}^{[0]} = \emptyset$. In our algorithm, once an RRH is added to the set $\mathcal{Z}$, it cannot be removed. This procedure is a simplification of the exact search method described in Section III-B. At iteration i, we need to solve the network power minimization problem $\mathscr{P}(\mathcal{A}^{[i]})$ in (12) with the given active RRH set $\mathcal{A}^{[i]}$.

1) RRH Selection Rule: How to select $r^{[i]}$ at the ith iteration is critical for the performance of the greedy selection algorithm. Based on our objective, we propose to select $r^{[i]}$ to maximize the decrease in the network power consumption. Specifically, at iteration i, we obtain the network power consumption $p^\star(\mathcal{A}_m^{[i]})$ with $\mathcal{A}_m^{[i]} \cup \{m\} = \mathcal{A}^{[i]}$ by removing any $m \in \mathcal{A}^{[i]}$ from the active RRH set $\mathcal{A}^{[i]}$. Thereafter, $r^{[i]}$ is chosen to yield the smallest network power consumption after switching off the corresponding RRH, i.e.,

$$r^{[i]} = \arg \min_{m \in \mathcal{A}^{[i]}} p^\star(\mathcal{A}_m^{[i]}). \tag{19}$$

We assume that $p^\star(\mathcal{A}_m^{[i]}) = +\infty$ if problem $\mathscr{P}(\mathcal{A}_m^{[i]})$ is infeasible. The impact of switching off one RRH is reducing the transport network power consumption while increasing the total transmit power consumption. Thus, the proposed selection rule actually aims at minimizing the impact of turning off one RRH at each iteration.

Denote $\mathcal{J}$ as the set of candidate RRHs that can be turned off, the greedy selection algorithm is described as follows:

Algorithm 1: The Greedy Selection Algorithm

Step 0: Initialize $\mathcal{Z}^{[0]} = \emptyset$, $\mathcal{A}^{[0]} = \{1, \ldots, L\}$ and $i = 0$;
Step 1: Solve the optimization problem $\mathscr{P}(\mathcal{A}^{[i]})$ (12);

1) **If** (12) is feasible, obtain $p^\star(\mathcal{A}^{[i]})$;
   - **If** $\forall m \in \mathcal{A}^{[i]}$, problem $\mathscr{P}(\mathcal{A}_m^{[i]})$ is infeasible, obtain $\mathcal{J} = \{0, \ldots, i\}$, **go to Step 2**;
   - **If** $\exists m \in \mathcal{A}^{[i]}$ makes problem $\mathscr{P}(\mathcal{A}_m^{[i]})$ feasible, find the $r^{[i]}$ according to (19) and update the set $\mathcal{Z}^{[i+1]} = \mathcal{Z}^{[i]} \cup \{r^{[i]}\}$ and the iteration number $i \leftarrow i + 1$, **go to Step 1**;
2) **If** (12) is infeasible, when $i = 0$, $p^\star = \infty$, **go to End**; when $i > 0$, obtain $\mathcal{J} = \{0, 1, \ldots, i-1\}$, **go to Step 2**;

Step 2: Obtain the optimal active RRH set $\mathcal{A}^{[j^\star]}$ with $j^\star = \arg\min_{j \in \mathcal{J}} p^\star(\mathcal{A}^{[j]})$ and the transmit beamformers minimizing $\mathscr{P}(\mathcal{A}^{[j^\star]})$;
End

B. Complexity Analysis

At the i-th iteration, we need to solve $|\mathcal{A}^{[i]}|$ SCOP problems $\mathscr{P}(\mathcal{A}_m^{[i]})$ by removing the RRH m from the set $\mathcal{A}^{[i]}$ to determine which RRH should be selected. For each of the SOCP problem $\mathscr{P}(\mathcal{A})$, using the interior-point method, the computational complexity is $\mathcal{O}((K\sum_{l\in\mathcal{A}} N_l)^{3.5})$ [27]. The total number of iterations is bounded by L. As a result, the total number of SOCP problems required to be solved grows *quadratically* with L. Although this reduces the computational complexity significantly compared with the mixed-integer conic programming based algorithms in [30] and [31], the complexity is still prohibitive for large-scale networks. Therefore, in the next section we will propose a group sparse beamforming framework to further reduce the complexity.

V. Group Sparse Beamforming Framework

In this section, we will develop two low-complexity algorithms based on the GSBF formulation $\mathscr{P}_{\text{sparse}}$, namely a bi-section GSBF algorithm and an iterative GSBF algorithm, for which, the overall number of SOCP problems to solve grows *logarithmically* and *linearly* with L, respectively. The main motivation is to induce group sparsity in the beamformer, which corresponds to switching off RRHs.

In the bi-section GSBF algorithm, we will minimize the *weighted* mixed ℓ_1/ℓ_2-norm to induce group-sparsity for the beamformer. By exploiting the additional prior information (i.e., power amplifier efficiency, relative transport link power consumption, and channel power gain) available in our setting, the proposed bi-section GSBF algorithm will be demonstrated through rigorous analysis and simulations to

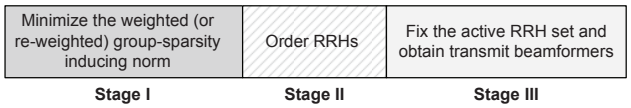


Fig. 2. A three-stage GSBF framework.

outperform the conventional *unweighted* mixed ℓ_1/ℓ_p-norm minimization substantially[17]–[19]. By minimizing the *re-weighted* mixed ℓ_1/ℓ_2-norm iteratively to enhance the group sparsity for the beamformer, the proposed iterative GSBF algorithm will further improve the performance.

The proposed GSBF framework is a three-stage approach, as shown in Fig. 2. Specifically, in the first stage, we minimize a weighted (or re-weighted) group-sparsity inducing norm to induce the group-sparsity in the beamformer. In the second stage, we propose an ordering rule to determine the priority for the RRHs that should be switched off, based on not only the (approximately) sparse beamformer obtained in the first stage, but also some key system parameters. Following the ordering rule, a selection procedure is performed to determine the optimal active RRH set, followed by the coordinated beamforming. The details will be presented in the following subsections.

A. Preliminaries on Group-Sparsity Inducing Norms

The mixed ℓ_1/ℓ_p-norm has recently received lots of attention and is shown to be effective to induce group sparsity [14], which is defined as follows:

Definition 1: Consider the vector $\mathbf{w} = [\mathbf{w}_{lk}]$ indexed by $l \in \mathcal{L}$ and $k \in \mathcal{S}$ as define in (15). Its mixed ℓ_1/ℓ_p-norm is defined as follows:

$$\mathcal{R}(\mathbf{w}) = \sum_{l=1}^{L} \beta_l \|\tilde{\mathbf{w}}_l\|_{\ell_p},\ p > 1, \tag{20}$$

where $\beta_1, \beta_2, \ldots, \beta_L$ are positive weights.

Define the vector $\mathbf{r} = [\|\tilde{\mathbf{w}}_1\|_{\ell_p}, \ldots, \|\tilde{\mathbf{w}}_L\|_{\ell_p}]^T$, then the mixed ℓ_1/ℓ_p-norm behaves as the ℓ_1-norm on the vector $\mathbf{r}$, and therefore, inducing group sparsity (i.e., each vector $\tilde{\mathbf{w}}_l$ is encouraged to be set to zero) for $\mathbf{w}$. Note that, within the group $\tilde{\mathbf{w}}_l$, the ℓ_p-norm does not promote sparsity as $p > 1$. By setting $p = 1$, the mixed ℓ_1/ℓ_p-norm becomes a weighted ℓ_1-norm, which will not promote group sparsity. The mixed ℓ_1/ℓ_2-norm and ℓ_1/ℓ_∞-norm are two commonly used norms for inducing group sparsity. For instance, the mixed ℓ_1/ℓ_2-norm is used with the name *group least-absolute selection* and *shrinkage operator* (or *Group-Lasso*) in machine learning [15]. In high dimensional statistics, the mixed ℓ_1/ℓ_∞-norm is adopted as a regularizer in the linear regression problems with sparsity constraints for its computational convenience [16].

B. Bi-Section GSBF Algorithm

In this section, we propose a binary search based GSBF algorithm, in which, the overall number of SOCP problems required to be solved grows logarithmically with L, instead of quadratically for the GS algorithm.

1) Group-Sparsity Inducing Norm Minimization: With the combinatorial function $F(\cdot)$ in the objective function $p(\mathbf{w}) = T(\mathbf{w}) + F(\mathcal{T}(\mathbf{w}))$, the problem $\mathscr{P}_{\text{sparse}}$ becomes computationally intractable. Therefore, we first construct an appropriate convex relaxation for the objective function $p(\mathbf{w})$ as a surrogate objective function, resulting a weighted mixed ℓ_1/ℓ_2-norm minimization problem to induce group sparsity for the beamformer. Specifically, we first derive its tightest positively homogeneous lower bound $p_h(\mathbf{w})$, which has the property $p_h(\lambda\mathbf{w}) = \lambda p_h(\mathbf{w}), 0 < \lambda < \infty$. Since $p_h(\mathbf{w})$ is still not convex, we further calculate its Fenchel-Legendre biconjugate $p_h^{**}(\mathbf{w})$ to provide a tightest convex lower bound for $p_h(\mathbf{w})$. We call $p_h^{**}(\mathbf{w})$ as the *convex positively homogeneous lower bound* (the details can be found in [32]) of function $p(\mathbf{w})$, which is provided in the following proposition:

Proposition 1: The tightest convex positively homogeneous lower bound of the objective function in $\mathscr{P}_{\text{sparse}}$, denoted as $p(\mathbf{w})$, is given by

$$\Omega(\mathbf{w}) = 2\sum_{l=1}^{L} \sqrt{\frac{P_l^c}{\eta_l}} \|\tilde{\mathbf{w}}_l\|_{\ell_2}. \tag{21}$$

Proof: Please refer to Appendix A. ∎

This proposition indicates that the group-sparsity inducing norm (i.e., the weighted mixed ℓ_1/ℓ_2-norm) can provide a convex relaxation for the objective function $p(\mathbf{w})$. Furthermore, it encapsulates the additional prior information in terms of system parameters into the weights for the groups. Intuitively, the weights indicate that the RRHs with a higher transport link power consumption and lower power amplifier efficiency will have a higher chance being forced to be switched off. Using the weighted mixed ℓ_1/ℓ_2-norm as a surrogate for the objective function, we minimize the weighted mixed ℓ_1/ℓ_2-norm $\Omega(\mathbf{w})$ to induce the group-sparsity for the beamformer $\mathbf{w}$:

$$\begin{aligned} \mathscr{P}_{\text{GSBF}}: \underset{\mathbf{w}}{\text{minimize}} \quad & \Omega(\mathbf{w}) \\ \text{subject to} \quad & \mathcal{C}_1(\mathcal{L}), \mathcal{C}_2(\mathcal{L}), \end{aligned} \tag{22}$$

which is an SOCP problem and can be solved efficiently.

2) RRH Ordering: After obtaining the (approximately) sparse beamformer $\hat{\mathbf{w}}$ via solving the weighted group-sparsity inducing norm minimization problem $\mathscr{P}_{\text{GSBF}}$, the next question is how to determine the active RRH set. We will first give priorities to different RRHs, so that an RRH with a higher priority should be switched off before the one with a lower priority. Most previous works [17]–[19] applying the idea of group-sparsity inducing norm minimization directly to map the sparsity to their application, e.g., in [19], the transmit antennas corresponding to the smaller coefficients in the group (measured by the ℓ_∞-norm) will have a higher priority to be switched off. In our setting, one might be tempted to give a higher priority for an RRH l with a smaller coefficient $r_l = (\sum_{k=1}^{K} \|\hat{\mathbf{w}}_{lk}\|_{\ell_2}^2)^{1/2}$, as it may provide a lower beamforming gain and should be encouraged to be turned off. It turns out that such an ordering rule is not a good option and will bring performance degradation.

To get a better performance, the priority of the RRHs should be determined by not only the beamforming gain but also other key system parameters that indicate the impact of the RRHs

on the network performance. In particular, the channel power gain $\kappa_l = \sum_{k=1}^{K} \|\mathbf{h}_{kl}\|_{\ell_2}^2$ should be taken into consideration. Specifically, by the broadcast channel (BC)-multiple-access channel (MAC) duality [33], we have the sum capacity of the Cloud-RAN as:

$$C_{\text{sum}} = \log\det(\mathbf{I}_N + \mathsf{snr}\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}), \tag{23}$$

where we assume equal power allocation to simplify the analysis, i.e., $\mathsf{snr} = P/\sigma^2, \forall k = 1, \ldots, K$. One way to upper-bound C_{sum} is through upper-bounding the capacity by the total receive SNR, i.e., using the following relation

$$\begin{aligned}\log\det(\mathbf{I}_N + \mathsf{snr}\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}) &\leq \text{Tr}(\mathsf{snr}\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}) \\ &= \mathsf{snr}\sum_{l=1}^{L}\kappa_l, \end{aligned}\tag{24}$$

which relies on the inequality $\log(1+x) \leq x$. Therefore, from the capacity perspective, the RRH with a higher channel power gain κ_l contributes more to the sum capacity, i.e., it provides a higher power gain and should not be encouraged to be switched off.

Therefore, different from the previous democratic assumptions (e.g., [17]–[19]) on the mapping between the sparsity and their applications directly, we exploit the prior information in terms of system parameters to refine the mapping on the group-sparsity. Specifically, considering the key system parameters, we propose the following ordering criterion to determine which RRHs should be switched off, i.e.,

$$\theta_l := \sqrt{\frac{\kappa_l\eta_l}{P_l^c}}\left(\sum_{k=1}^{K}\|\hat{\mathbf{w}}_{lk}\|_{\ell_2}\right)^{1/2}, \forall l = 1, \ldots, L, \tag{25}$$

where the RRH with a smaller θ_l will have a higher priority to be switched off. This ordering rule indicates that the RRH with a lower beamforming gain, lower channel power gain, lower power amplifier efficiency, and higher relative transport link power consumption should have a higher priority to be switched off. The proposed ordering rule will be demonstrated to significantly improve the performance of the GSBF algorithm through simulations.

3) Binary Search Procedure: Based on the ordering rule (25), we sort the coefficients in the ascending order: $\theta_{\pi_1} \leq \theta_{\pi_2} \leq \cdots \leq \theta_{\pi_L}$ to fix the final active RRH set. We set the first J smallest coefficients to zero, as a result, the corresponding RRHs will be turned. Denote J_0 as the maximum number of RRHs that can be turned off, i.e., the problem $\mathscr{P}(\mathcal{A}^{[i]})$ is infeasible if $i > J_0$, where $\mathcal{A}^{[i]} \cup \mathcal{Z}^{[i]} = \mathcal{L}$ with $\mathcal{Z}^{[i]} = \{\pi_0, \pi_1, \ldots, \pi_i\}$ and $\pi_0 = \emptyset$. A binary search procedure can be adopted to determine J_0, which only needs to solve no more than $(1 + \lceil\log(1+L)\rceil)$ SOCP problems. In this algorithm, we regard $\mathcal{A}^{[J_0]}$ as the final active RRH set and the solution of $\mathscr{P}(\mathcal{A}^{[J_0]})$ is the final transmit beamformer.

Therefore, the bi-section GSBF algorithm is presented as follows:

Algorithm 2: The Bi-Section GSBF Algorithm

Step 0: Solve the weighted group-sparsity inducing norm minimization problem $\mathscr{P}_{\text{GSBF}}$;
1) **If** it is infeasible, set $p^\star = \infty$, **go to End**;
2) **If** it is feasible, obtain the solution $\hat{\mathbf{w}}$, calculate ordering criterion (25), and sort them in the ascending order: $\theta_{\pi_1} \leq \cdots \leq \theta_{\pi_L}$, **go to Step 1**;

Step 1: Initialize $J_{\text{low}} = 0$, $J_{\text{up}} = L$, $i = 0$;

Step 2: Repeat
1) Set $i \leftarrow \lfloor\frac{J_{\text{low}}+J_{\text{up}}}{2}\rfloor$;
2) Solve the optimization problem $\mathscr{P}(\mathcal{A}^{[i]})$ (12): if it is infeasible, set $J_{\text{low}} = i$; otherwise, set $J_{\text{up}} = i$;

Step 3: Until $J_{\text{up}} - J_{\text{low}} = 1$, obtain $J_0 = J_{\text{low}}$ and obtain the optimal active RRH set $\mathcal{A}^\star$ with $\mathcal{A}^\star \cup \mathcal{J} = \mathcal{L}$ and $\mathcal{J} = \{\pi_1, \ldots, \pi_{J_0}\}$;

Step 4: Solve the problem $\mathscr{P}(\mathcal{A}^\star)$, obtain the minimum network power consumption and the corresponding transmit beamformers;

End

C. Iterative GSBF Algorithm

Under the GSBF framework, the main task of the first two stages is to order the RRHs according to the criterion (25), which depends on the sparse solution to $\mathscr{P}_{\text{GSBF}}$, i.e., $\{\hat{\mathbf{w}}_{lk}\}$. However, when the minimum of $r_l = (\sum_{k=1}^{K}\|\hat{\mathbf{w}}_{lk}\|_{\ell_2}^2)^{1/2} > 0$ is not close to zero, it will introduce large bias in estimating which RRHs can be switched off. To resolve this issue, we will apply the idea from the majorization-minimization (MM) algorithm [34] (please refer to appendix B for details on this algorithm), to enhance group-sparsity for the beamformer to better estimate which RRHs can be switched off.

The MM algorithms have been successfully applied in the re-weighted ℓ_1-norm (or mixed ℓ_1/ℓ_2-norm) minimization problem to enhance sparsity [18], [19], [35]. However, these algorithms failed to exploit the additional system prior information to improve the performance. Specifically, they used the un-weighted ℓ_1-norm (or mixed ℓ_1/ℓ_p-norm) minimization as the start point of the iterative algorithms and re-weighted the ℓ_1-norm (or mixed ℓ_1/ℓ_p-norm) only using the estimate of the coefficients obtained in the last minimization step. Different from the above conventional re-weighted algorithms, we exploit the additional system prior information at each step (including the start step) to improve the estimation on the group sparsity of the beamformer.

1) Re-weighted Group-Sparsity Inducing Norm Minimization: One way to enhance the group-sparsity compared with using the weighted mixed ℓ_1/ℓ_2 norm $\Omega(\mathbf{w})$ in (21) is to minimize the following combinatorial function directly:

$$\mathcal{R}(\mathbf{w}) = 2\sum_{l=1}^{L}\sqrt{\frac{P_l^c}{\eta_l}}I(\|\tilde{\mathbf{w}}_l\|_{\ell_2} > 0), \tag{26}$$

for which the convex function $\Omega(\mathbf{w})$ in (21) can be regarded as an ℓ_1-norm relaxation. Unfortunately, minimizing $\mathcal{R}(\mathbf{w})$ will lead to a non-convex optimization problem. In this subsection, we will provide a sub-optimal algorithm to solve (25) by adopting the idea from the MM algorithm to enhance sparsity.

Based on the following fact in [36]

$$\lim_{\epsilon\to 0}\frac{\log(1+x\epsilon^{-1})}{\log(1+\epsilon^{-1})}=\begin{cases}0 & \text{if } x=0,\\ 1 & \text{if } x>0,\end{cases} \tag{27}$$

we rewrite the indicator function in (26) as

$$I(\|\tilde{\mathbf{w}}_l\|_{\ell_2}>0)=\lim_{\epsilon\to 0}\frac{\log(1+\|\tilde{\mathbf{w}}_l\|_{\ell_2}\epsilon^{-1})}{\log(1+\epsilon^{-1})},\forall l\in\mathcal{L}. \tag{28}$$

The surrogate objective function $\mathcal{R}(\mathbf{w})$ can then be approximated as

$$f(\mathbf{w})=\lambda_\epsilon\sum_{l=1}^{L}\sqrt{\frac{P_l^c}{\eta_l}}\log(1+\|\tilde{\mathbf{w}}_l\|_{\ell_2}\epsilon^{-1}), \tag{29}$$

by neglecting the limit in (28) and choosing an appropriate $\epsilon>0$, where $\lambda_\epsilon=\frac{2}{\log(1+\epsilon^{-1})}$. Compared with $\Omega(\mathbf{w})$ in (21), the log-sum penalty function $f(\mathbf{w})$ has the potential to be much more sparsity-encouraging. The detailed explanations can be found in [35].

Since $\log(1+x), x\geq 0$, is a concave function, we can construct a majorization function for f by the first-order approximation of $\log(1+\|\tilde{\mathbf{w}}_l\|_{\ell_2}\epsilon^{-1})$, i.e.,

$$f(\mathbf{w})\leq\lambda_\epsilon\sum_{l=1}^{L}\sqrt{\frac{P_l^c}{\eta_l}}\left(\underbrace{\frac{\|\tilde{\mathbf{w}}_l\|_{\ell_2}}{\|\tilde{\mathbf{w}}_l^{[m]}\|_{\ell_2}+\epsilon}+c(\mathbf{w}^{[m]})}_{g(\mathbf{w}|\mathbf{w}^{[m]})}\right), \tag{30}$$

where $\mathbf{w}^{[m]}$ is the minimizer at the $(m-1)$-th iteration, and $c(\mathbf{w}^{[m]})=\log(1+\|\tilde{\mathbf{w}}_l^{[m]}\|_{\ell_2})-\|\tilde{\mathbf{w}}_l^{[m]}\|_{\ell_2}/(\|\tilde{\mathbf{w}}_l^{[m]}\|_{\ell_2}+\epsilon)$ is a constant provided that $\mathbf{w}^{[m]}$ is already known at the current m-th iteration.

By omitting the constant part of $g(\mathbf{w}|\mathbf{w}^{[m]})$ at the m-th iteration, which will not affect the solution, we propose a re-weighted GSBF framework to enhance the group-sparsity:

$$\begin{aligned}\mathscr{P}_{\text{iGSBF}}^{[m]}:\{\tilde{\mathbf{w}}_l^{[m+1]}\}_{l=1}^{L}=&\arg\min\sum_{l=1}^{L}\beta_l^{[m]}\|\tilde{\mathbf{w}}_l\|_{\ell_2}\\ &\text{subject to } \mathcal{C}_1(\mathcal{L}),\mathcal{C}_2(\mathcal{L}),\end{aligned} \tag{31}$$

where

$$\beta_l^{[m]}=\sqrt{\frac{P_l^c}{\eta_l}}\frac{1}{(\|\tilde{\mathbf{w}}_l^{[m]}\|_{\ell_2}+\epsilon)},\forall l=1,\ldots,L, \tag{32}$$

are the weights for the groups at the m-th iteration. At each step, the mixed ℓ_1/ℓ_2-norm optimization is re-weighted using the estimate of the beamformer obtained in the last minimization step.

As this iterative algorithm cannot guarantee the global minimum, it is important to choose a suitable starting point to obtain a good local optimum. As suggested in [18], [19], [35], this algorithm can be initiated with the solution of the unweighted ℓ_1-norm minimization, i.e., $\beta_l^{[0]}=1,\forall l=1,\ldots,L$. In our setting, however, the prior information on the system parameters can help us generate a high quality stating point for the iterative GSBF framework. Specifically, with the available channel state information, we choose the ℓ_2-norm of the initial beamformer at the l-th RRH $\|\tilde{\mathbf{w}}_l^{[0]}\|_{\ell_2}$ to be proportional to its corresponding channel power gain κ_l, arguing that the RRH with a low channel power gain should be encouraged to be switched off as justified in Section V-B. Therefore, from (32), we set the following weights as the initiation weights for $\mathscr{P}_{\text{iGSBF}}^{[0]}$:

$$\beta_l^{[0]}=\sqrt{\frac{P_l^c}{\eta_l\kappa_l}},\forall l=1,\ldots,L. \tag{33}$$

The weights indicate that the RRHs with a higher relative transport link consumption, lower power amplifier efficiency and lower channel power gain should be penalized more heavily.

As observed in the simulations, this algorithm converges very fast (typically within 20 iterations). We set the maximum number of iterations as $m_{\max}=L$ in our simulations.

2) Iterative Search Procedure: After obtaining the (approximately) sparse beamformers using the above re-weighted GSBF framework, we still adopt the same ordering criterion (25) to fix the final active RRH set.

Different from the aggressive strategy in the bi-section GSBF algorithm, which assumes that the RRH should be switched off as many as possible and thus results a minimum transport network power consumption, we adopt a conservative strategy to determine the final active RRH set by realizing that the minimum network power consumption may not be attained when the transport network power consumption is minimized.

Specifically, denote J_0 as the maximum number of RRHs that can be switched off, the corresponding inactive RRH set is $\mathcal{J}=\{\pi_0,\pi_1,\ldots,\pi_{J_0}\}$. The minimum network power consumption should be searched over all the values of $\mathscr{P}^*(\mathcal{A}^{[i]})$, where $\mathcal{A}^{[i]}=\mathcal{L}\backslash\{\pi_0,\pi_1,\ldots,\pi_i\}$ and $0\leq i\leq J_0$. This can be accomplished using an iterative search procedure that requires to solve no more than L SOCP problems.

Therefore, the overall iterative GSBF algorithm is presented as Algorithm 3.

Algorithm 3: The Iterative GSBF Algorithm

Step 0: Initialize the weights $\beta_l^{[0]}, l=1,\ldots,L$ as in (33) and the iteration counter as $m=0$;
Step 1: Solve the weighted GSBF problem $\mathscr{P}_{\text{iGSBF}}^{[m]}$ (31): **if** it is infeasible, set $p^\star=\infty$ and **go to End**; otherwise, set $m=m+1$, **go to Step 2**;
Step 2: Update the weights using (32);
Step 3: If converge or $m=m_{\max}$, obtain the solution $\hat{\mathbf{w}}$ and calculate the selection criterion (25), and sort them in the ascending order: $\theta_{\pi_1}\leq\cdots\leq\theta_{\pi_L}$, **go to Step 4**; otherwise, **go to Step 1**;
Step 4: Initialize $\mathcal{Z}^{[0]}=\emptyset$, $\mathcal{A}^{[0]}=\{1,\ldots,L\}$, and $i=0$;
Step 5: Solve the optimization problem $\mathscr{P}(\mathcal{A}^{[i]})$ (12);

1) **If** (12) is feasible, obtain $p^*(\mathcal{A}^{[i]})$, update the set $\mathcal{Z}^{[i+1]}=\mathcal{Z}^{[i]}\cup\{\pi_{i+1}\}$ and $i=i+1$, **go to Step 5**;
2) **If** (12) is infeasible, obtain $\mathcal{J}=\{0,1,\ldots,i-1\}$, **go to Step 6**;

Step 6: Obtain optimal RRH set $\mathcal{A}^{[j^\star]}$ and beamformers minimizing $\mathscr{P}(\mathcal{A}^{[j^\star]})$ with $j^\star=\arg\min_{j\in\mathcal{J}}p^*(\mathcal{A}^{[j]})$;
End

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Path-loss at distance d_{kl} (km)	148.1+37.6 $\log_2(d_{kl})$ dB
Standard deviation of log-norm shadowing σ_s	8 dB
Small-scale fading distribution $\mathbf{g}_{kl}$	$\mathcal{CN}(\mathbf{0}, \mathbf{I})$
Noise power σ_k^2 [1] (10 MHz bandwidth)	-102 dBm
Maximum transmit power of RRH P_l [1]	1 W
Power amplifier efficiency η_l [23]	25%
Transmit antenna power gain	9 dBi

D. Complexity Analysis and Optimality Discussion

We have demonstrated that the maximum number of iterations is linear and logarithmical to L for the "Iterative GSBF Algorithm" and the "Bi-Section GSBF Algorithm," respectively. Therefore, the convergence speed of the proposed GSBF algorithms scales well for large-scale Cloud-RAN (e.g., with $L = 100$). However, the main computational complexity of the proposed algorithms is related to solving an SOCP problem at each iteration. In particular, with a large number of RRHs, the computational complexity of solving an SOCP problem using the interior-point method is proportional to $\mathcal{O}(L^{3.5})$. Therefore, in order to solve a large-sized SOCP problem, other approaches need to be explored (e.g., the *alternating direction method of multipliers* (ADMM) method [37]). This is an on-going research topic, and we will leave it as our future research direction.

Furthermore, the proposed group sparse beamforming algorithm is a convex relaxation to the original combinatorial optimization problem using the group-sparsity inducing norm, i.e., the mixed ℓ_1/ℓ_2-norm. It is very challenging to quantify the performance gap due to the convex relaxation, for which normally specific prior information is needed, e.g., in compressive sensing, the sparse signal is assumed to obey a power law (see Eq.(1.8) in [12]). However, we do not have any prior information about the optimal solution. This is the fundamental difference between our problem and the existing ones in the field of sparse signal processing. The optimality analysis of the group sparse beamforming algorithms will be left to our future work.

VI. SIMULATION RESULTS

In this section, we simulate the performance of the proposed algorithms. We consider the following channel model

$$\mathbf{h}_{kl} = 10^{-L(d_{kl})/20}\sqrt{\varphi_{kl}s_{kl}}\mathbf{g}_{kl}, \tag{34}$$

where $L(d_{kl})$ is the path-loss at distance d_{kl}, , as given in Table I, s_{kl} is the shadowing coefficient, φ_{kl} is the antenna gain and $\mathbf{g}_{kl}$ is the small scale fading coefficient. We use the standard cellular network parameters as showed in Table I. Each point of the simulation results is averaged over 50 randomly generated network realizations. The network power consumption is given in (7). We set $P_{s,l}^{\text{rrh}} = 4.3W$ and $P_{s,l}^{\text{tl}} = 0.7W$, $\forall l$, and $P_{\text{olt}} = 20W$.

The proposed algorithms are compared to the following algorithms:

- **Coordinated beamforming (CB) algorithm**: In this algorithm, all the RRHs are active and only the total transmit power consumption is minimized [7].

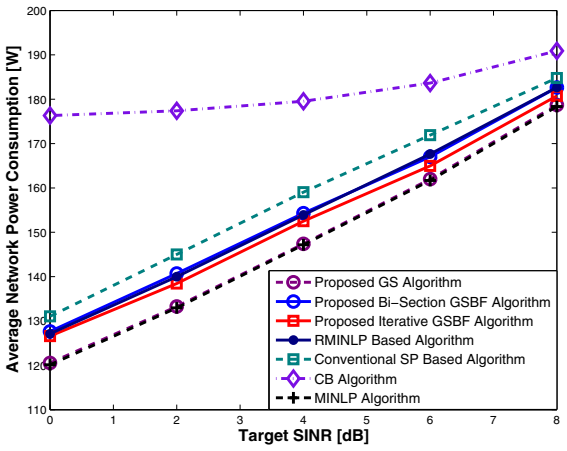


Fig. 3. Average network power consumption versus target SINR.

- **Mixed-integer nonlinear programming (MINLP) algorithm**: This algorithm [30], [31] can obtain the global optimum. Since the complexity of the algorithm grows exponentially with the number of RRHs L, we only run it in a small-size network.
- **Conventional sparsity pattern (SP) based algorithm**: In this algorithm, the unweighted mixed ℓ_1/ℓ_p-norm is adopted to induce group sparsity as in [17] and [19]. The ordering of RRHs is determined only by the group-sparsity of the beamformer, i.e., $\theta_l = (\sum_{k=1}^{K} \|\hat{\mathbf{w}}_{lk}\|_{\ell_2})^{1/2}, \forall l = 1, \ldots, L$, instead of (25). The complexity of the algorithm grows logarithmically with L.
- **Relaxed mixed-integer nonlinear programming (RMINLP) based algorithm**: In this algorithm, a deflation procedure is performed to switch off RRHs one-by-one based on the solutions obtained via solving the relaxed MINLP by relaxing the integers to the unit intervals [31]. The complexity of the algorithm grows linearly with L.

A. Network Power Consumption versus Target SINR

Consider a network with $L = 10$ 2-antenna RRHs and $K = 15$ single-antenna MUs uniformly and independently distributed in the square region $[-1000\ 1000] \times [-1000\ 1000]$ meters. We set all the relative transport link power consumption to be $P_l^c = (5 + l)W, l = 1, \ldots, L$, which is to indicate the inhomogeneous power consumption on different transport links and RRHs. Fig. 3 demonstrates the average network power consumption with different target SINRs.

This figure shows that the proposed GS algorithm can always achieve global optimum (i.e., the optimal value from the MINLP algorithm), which confirms the effectiveness of the proposed RRH selection rule for the greedy search procedure. With only logarithmic complexity, the proposed bi-section GSBF algorithm achieves almost the same performance as the RMINLP algorithm, which has a linear complexity. Moreover, with the same complexity, the gap between the conventional SP based algorithm and the proposed bi-section GSBF algorithm is large. Furthermore, the proposed iterative GSBF

algorithm always outperforms the RMINLP algorithm, while both of them have the same computational complexity. These confirm the effectiveness of the proposed GSBF framework to minimize the network power consumption. Overall, this figure shows that our proposed schemes have the potential to reduce the power consumption by 40% in the low QoS regime, and by 20% in the high QoS regime.

This figure also demonstrates that, when the target SINR increases[2], the performance gap between the CB algorithm and the other algorithms becomes smaller. In particular, when the target SINR is relatively high (e.g., 8 dB), all the other algorithms achieve almost the same network power consumption as the CB algorithm. This implies that almost all the RRHs need to be switched on when the QoS requirements are extremely high. In the extreme case with all the RRHs active, all the algorithms will yield the same network power consumption, as all of them will perform coordinated beamforming with all the RRHs active, resulting in the same total transmit power consumption.

1) Impact of Different Components of Network Power Consumption: Consider the same network setting as in Fig. 3. The corresponding average total transmit power consumption $p_1(\mathcal{A}) = \sum_{l\in\mathcal{A}} \frac{1}{\eta_l} \sum_{k=1}^{K} \|\mathbf{w}_{lk}\|_{\ell_2}^2$ is demonstrated in Fig. 4, and the corresponding average total relative transport link power consumption $p_2(\mathcal{A}) = \sum_{l\in\mathcal{A}} P_l^c$ is shown in Fig. 5. Table II shows the average numbers of RRHs that are switched off with different algorithms. From Fig. 4 and Fig. 5, we see that the CB algorithm, which intends to minimize the total transmit power consumption, achieves the lowest total transmit power consumption due to the highest beamforming gain with all the RRH active, but it has the highest total relative transport link power consumption. This implies that a joint RRH selection and power minimization beamforming is required to minimize the network power consumption.

From Table II, we see that the proposed GS algorithm can switch off almost the same number of RRHs as the MINLP algorithm. Furthermore, the proposed GSBF algorithms can switch off more RRHs than the RMINLP based algorithm and the conventional SP based algorithm on average. Overall, the proposed algorithms achieve a lower total relative transport link power consumption, as shown in Fig. 5. In particular, the proposed iterative GSBF algorithm can achieve a higher beamforming gain to minimize the total transmit power consumption, as shown in Fig. 4. Therefore, the results in Fig. 4, Fig. 5, and Table II demonstrate the effectiveness of our proposed RRH selection rule and RRH ordering rule for the GS algorithm and the GSBF algorithms, respectively. Furthermore, the results in Table II verify the group sparsity assumption in the GSBF algorithms.

[2]We will show, in Table II and Fig. 4, both the number of active RRHs and the total transmit power consumption will increase simultaneously to meet the QoS requirements.

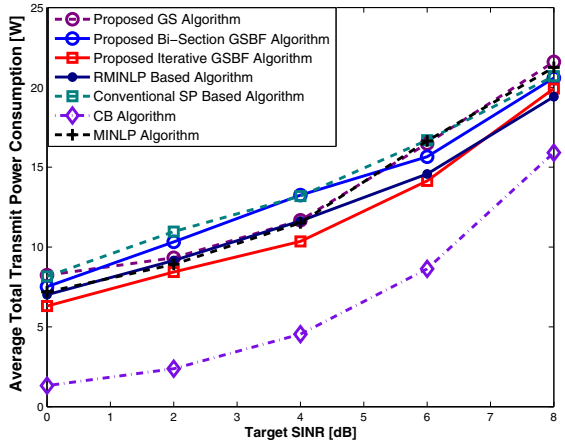


Fig. 4. Average total transmit power consumption versus target SINR.

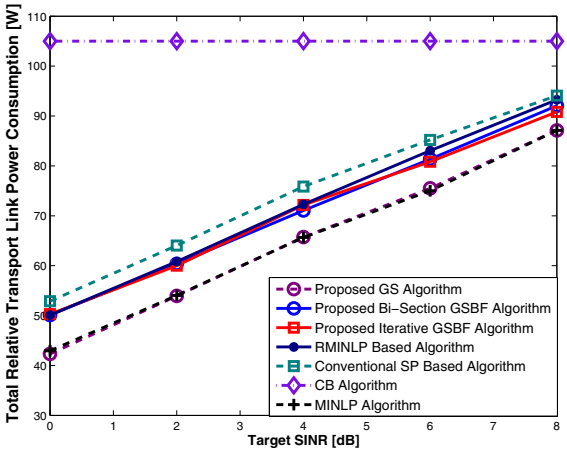


Fig. 5. Average total relative transport link power consumption versus target SINR.

B. Network Power Consumption versus Transport Links Power Consumption

Consider a network involving[3] $L = 20$ 2-antenna RRHs and $K = 15$ single-antenna MUs uniformly and independently distributed in the square region $[-2000\ 2000] \times [-2000\ 2000]$ meters. We set all the relative transport link power consumption to be the same, i.e., $P_c = P_l^c, \forall l = 1, \ldots, L$ and set the target SINR as 4 dB. Fig. 6 presents average network power consumption with different relative transport link power consumption.

This figure shows that both the GS algorithm and the iterative GSBF algorithm significantly outperform other algorithms, especially in the high transport link power consumption regime. Moreover, the proposed bi-section GSBF algorithm provides better performance than the conventional SP based algorithm and is close to the RMINLP based algorithm, while with a lower complexity. This result clearly indicates the importance of considering the key system parameters when

[3]In [4, Section 6.1], some field trials were demonstrated to verify the feasibility of Cloud-RAN, in which, a BBU pool can typically support 18 RRHs.

TABLE II
THE AVERAGE NUMBER OF INACTIVE RRHS WITH DIFFERENT ALGORITHMS

Target SINR [dB]	0	2	4	6	8
Proposed GS Algorithm	5.00	4.00	3.02	2.35	1.40
Proposed Bi-Section GSBF Algorithm	4.92	3.98	2.96	2.04	1.13
Proposed Iterative GSBF Algorithm	4.94	4.00	2.94	2.15	1.25
RMINLP Based Algorithm	4.88	3.90	2.79	1.85	1.00
Conventional SP Based Algorithm	4.88	3.90	2.81	1.94	1.10
CB Algorithm	0.00	0.00	0.00	0.00	0.00
MINLP Algorithm	5.00	4.00	3.08	2.42	1.44

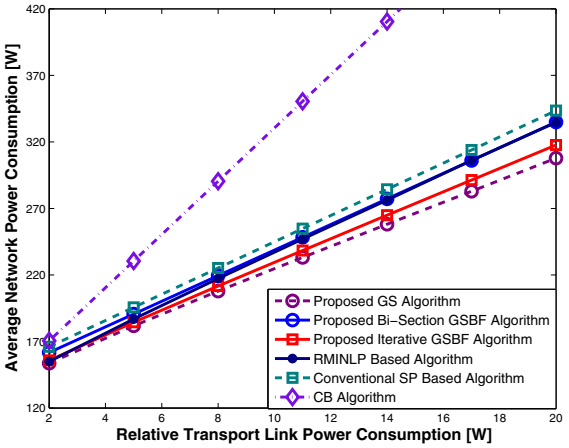


Fig. 6. Average network power consumption versus relative transport links power consumption.

applying the group sparsity beamforming framework.

Furthermore, this figure shows that all the algorithms achieve almost the same network power consumption when the relative transport link power consumption is relatively low (e.g., $2W$). This implies that almost all the RRHs need to be switched on to get a high beamforming gain to minimize the total transmit power consumption when the relative transport link power consumption can be ignored, compared to the RRH transmit power consumption.

C. Network Power Consumption versus the Number of Mobile Users

Consider a network with $L = 20$ 2-antenna RRHs uniformly and independently distributed in the square region $[-2000\ 2000] \times [-2000\ 2000]$ meters. We set all the relative transport link power consumption to be the same, i.e., $P_l^c = 20W, \forall l = 1, \ldots, L$ and set the target SINR as 4 dB. Fig. 7 presents the average network power consumption with different numbers of MUs, which are uniformly and independently distributed in the same region.

Overall, this figure further confirms the following conclusions:

1) With the $\mathcal{O}(L^2)$ computational complexity, the proposed GS algorithm has the best performance among all the low-complexity algorithms.
2) With the $\mathcal{O}(L)$ computational complexity, the proposed iterative GSBF algorithm outperforms the RMINLP algorithm, which has the same complexity.

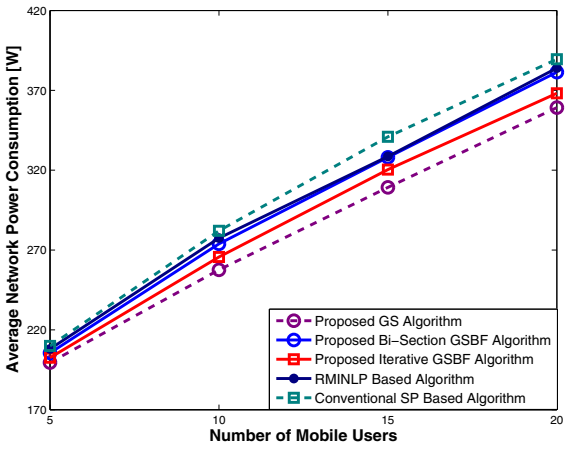


Fig. 7. Average network power consumption versus the number of mobile users.

3) With $\mathcal{O}(\log(L))$ computational complexity, the proposed bi-section GSBF algorithm has almost the same performance with the RMINLP algorithm and outperforms the conventional SP based algorithm, which has the same complexity. Therefore, the bi-section GSBF algorithm is very attractive for practical implementation in large-scale Cloud-RAN.

VII. CONCLUSIONS AND DISCUSSIONS

In this paper, we proposed a new framework to improve the energy efficiency of cellular networks with the new architecture of Cloud-RAN. It was shown that the transport network power consumption can not be ignored when designing green Cloud-RAN. By jointly selecting the active RRHs and minimizing the transmit power consumption through coordinated beamforming, the overall network power consumption can be significantly reduced, especially in the low QoS regime. The proposed group sparse formulation $\mathscr{P}_{\text{sparse}}$ serves as a powerful design tool for developing low complexity GSBF algorithms. Through rigorous analysis and careful simulations, the proposed GSBF framework was demonstrated to be very effective to provide near-optimal solutions. Especially, for the large-scale Cloud-RAN, the proposed bi-section GSBF algorithm will be a prior option due to its low complexity, while the iterative GSBF algorithm can be applied to provide better performance in a medium-size network. Simulation also showed that the proposed GS algorithm can always achieve nearly optimal performance, which makes it very attractive in the small-size clustered deployment of Cloud-RAN.

This initial investigation demonstrated the advantage of Cloud-RAN in terms of the network energy efficiency. More works will be needed to exploit the full benefits and overcome the main challenges of Cloud-RAN. Future research directions include theoretical analysis of the optimality of the proposed group sparse beamforming algorithms, more efficient beamforming algorithms for very large-scale Cloud-RAN deployment, joint beamforming and compression when considering the limited-capacity transport links, joint user scheduling, and effective CSI acquisition methods.

APPENDIX A
PROOF OF PROPOSITION 1

We begin by deriving the tightest positively homogeneous lower bound of $p(\mathbf{w})$, which is given by [32], [38]

$$p_h(\mathbf{w}) = \inf_{\lambda>0} \frac{p(\lambda\mathbf{w})}{\lambda} = \inf_{\lambda>0} \lambda T(\mathbf{w}) + \frac{1}{\lambda} F(\mathcal{T}(\mathbf{w})). \quad (35)$$

Setting the gradient of the objective function to zero, the minimum is obtained at $\lambda = \sqrt{F(\mathcal{T}(\mathbf{w}))/T(\mathbf{w})}$. Thus, the positively homogeneous lower bound of the objective function becomes

$$p_h(\mathbf{w}) = 2\sqrt{T(\mathbf{w})F(\mathcal{T}(\mathbf{w}))}, \quad (36)$$

which combines two terms multiplicatively.

Define diagonal matrices $\mathbf{U} \in \mathbb{R}^{N\times N}$, $\mathbf{V} \in \mathbb{R}^{N\times N}$ with $N = K\sum_{l=1}^{L} N_l$, for which the l-th block elements are $\eta_l \mathbf{I}_{KN_l}$ and $\frac{1}{\eta_l}\mathbf{I}_{KN_l}$, respectively. Next, we calculate the convex envelope of $p_h(\mathbf{w})$ via computing its conjugate:

$$\begin{aligned} p_h^*(\mathbf{y}) &= \sup_{\mathbf{w}\in\mathbb{C}^N} \left(\mathbf{y}^T\mathbf{U}^T\mathbf{V}\mathbf{w} - 2\sqrt{T(\mathbf{w})F(\mathcal{T}(\mathbf{w}))}\right), \\ &= \sup_{\mathcal{I}\subseteq\mathcal{V}} \sup_{\mathbf{w}_\mathcal{I}\in\mathbb{C}^{|\mathcal{I}|}} \left(\mathbf{y}_\mathcal{I}^T\mathbf{U}_{\mathcal{I}\mathcal{I}}^T\mathbf{V}_{\mathcal{I}\mathcal{I}}\mathbf{w}_\mathcal{I} - 2\sqrt{T(\mathbf{w}_\mathcal{I})F(\mathcal{I})}\right) \\ &= \begin{cases} 0 & \text{if } \Omega^*(\mathbf{y}) \le 1 \\ \infty, & \text{otherwise.} \end{cases} \end{aligned} \quad (37)$$

where $\mathbf{y}_\mathcal{I}$ is the $|\mathcal{I}|$-dimensional vector formed with the entries of $\mathbf{y}$ indexed by $\mathcal{I}$ (similarly for $\mathbf{w}$), and $\mathbf{U}_{\mathcal{I}\mathcal{I}}$ is the $|\mathcal{I}|\times|\mathcal{I}|$ matrix formed with the rows and columns of $\mathbf{U}$ indexed by $\mathcal{I}$ (similarly for $\mathbf{V}$), and $\Omega^*(\mathbf{y})$ defines a dual norm of $\Omega(\mathbf{w})$:

$$\Omega^*(\mathbf{y}) = \sup_{\mathcal{I}\subseteq\mathcal{V},\mathcal{I}\neq\emptyset} \frac{\|\mathbf{y}_\mathcal{I}\mathbf{U}_\mathcal{I}\|_{\ell_2}}{2\sqrt{F(\mathcal{I})}} = \frac{1}{2}\max_{l=1,\ldots,L} \sqrt{\frac{\eta_l}{P_l^c}}\|\mathbf{y}_{\mathcal{G}_l}\|_{\ell_2}. \quad (38)$$

The first equality in (38) can be obtained by the Cauchy-Schwarz inequality:

$$\begin{aligned} \mathbf{y}_\mathcal{I}^T\mathbf{U}_{\mathcal{I}\mathcal{I}}^T\mathbf{V}_{\mathcal{I}\mathcal{I}}\mathbf{w}_\mathcal{I} &\le \|\mathbf{y}_\mathcal{I}\mathbf{U}_\mathcal{I}\|_{\ell_2} \cdot \|\mathbf{V}_{\mathcal{I}\mathcal{I}}\mathbf{w}_\mathcal{I}\|_{\ell_2} \\ &= \|\mathbf{y}_\mathcal{I}\mathbf{U}_\mathcal{I}\|_{\ell_2} \cdot \sqrt{T(\mathbf{w}_\mathcal{I})}. \end{aligned} \quad (39)$$

The second equality in (38) can be justified by

$$\begin{aligned} \Omega^*(\mathbf{y}) &\ge \sup_{\mathcal{I}\subseteq\mathcal{V},\mathcal{I}\neq\emptyset} \left(\frac{1}{2\sqrt{F(\mathcal{I})}} \max_{l=1,\ldots,L} \|\mathbf{y}_{\mathcal{I}\cap\mathcal{G}_l}\mathbf{U}_{\mathcal{I}\cap\mathcal{G}_l}\|_{\ell_2}\right) \\ &= \frac{1}{2}\max_{l=1,\ldots,L} \sqrt{\frac{\eta_l}{P_l^c}}\|\mathbf{y}_{\mathcal{G}_l}\|_{\ell_2}, \end{aligned} \quad (40)$$

and

$$\begin{aligned} \Omega^*(\mathbf{y}) &\le \sup_{\mathcal{I}\subseteq\mathcal{V},\mathcal{I}\neq\emptyset} \left(\frac{\|\mathbf{y}_\mathcal{I}\mathbf{U}_\mathcal{I}\|_{\ell_2}}{2\min_{l=1,\ldots,L}\sqrt{F(\mathcal{I}\cap\mathcal{G}_l)}}\right) \\ &= \frac{1}{2}\max_{l=1,\ldots,L} \sqrt{\frac{\eta_l}{P_l^c}}\|\mathbf{y}_{\mathcal{G}_l}\|_{\ell_2}. \end{aligned} \quad (41)$$

Therefore, the tightest convex positively homogeneous lower bound of the function $p(\mathbf{w})$ is

$$\begin{aligned} \Omega(\mathbf{w}) &= \sup_{\Omega^*(\mathbf{y})\le 1} \mathbf{w}^T\mathbf{y} \\ &\le \sup_{\Omega^*(\mathbf{y})\le 1} \sum_{l=1}^{L} \|\mathbf{w}_{\mathcal{G}_l}\|_{\ell_2}\|\mathbf{y}_{\mathcal{G}_l}\|_{\ell_2} \\ &\le \sup_{\Omega^*(\mathbf{y})\le 1} \left(\sum_{l=1}^{L} \sqrt{\frac{P_l^c}{\eta_l}}\|\mathbf{w}_{\mathcal{G}_l}\|_{\ell_2}\right)\left(\max_{l=1,\ldots,L}\sqrt{\frac{\eta_l}{P_l^c}}\|\mathbf{y}_{\mathcal{G}_l}\|_{\ell_2}\right) \\ &= 2\sum_{l=1}^{L}\sqrt{\frac{P_l^c}{\eta_l}}\|\mathbf{w}_{\mathcal{G}_l}\|_{\ell_2}. \end{aligned} \quad (42)$$

This upper bound actually holds with equality. Specifically, we let $\bar{\mathbf{y}}_{\mathcal{G}_l} = 2\sqrt{\frac{P_l^c}{\eta_l}}\frac{\mathbf{w}_{\mathcal{G}_l}^\dagger}{\|\mathbf{w}_{\mathcal{G}_l}^\dagger\|_{\ell_2}}$, such that $\Omega^*(\bar{\mathbf{y}}) = 1$. Therefore,

$$\begin{aligned} \Omega(\mathbf{w}) &= \sup_{\Omega^*(\mathbf{y})\le 1} \mathbf{w}^T\mathbf{y} \\ &\ge \sum_{l=1}^{L} \mathbf{w}_{\mathcal{G}_l}^T\bar{\mathbf{y}}_{\mathcal{G}_l} = 2\sum_{l=1}^{L}\sqrt{\frac{P_l^c}{\eta_l}}\|\mathbf{w}_{\mathcal{G}_l}\|_{\ell_2}. \end{aligned} \quad (43)$$

APPENDIX B
PRELIMINARIES ON MAJORIZATION-MINIMIZATION ALGORITHMS

The majorization-minimization (MM) algorithm, being a powerful tool to find a local optimum by minimizing a surrogate function that majorizes the objective function iteratively, has been widely used in statistics, machine learning, etc., [34]. We introduce the basic idea of MM algorithms, which allows us to derive our main results.

Consider the problem of minimizing $f(\mathbf{x})$ over $\mathcal{F}$. The idea of MM algorithms is as follows. First, we construct a majorization function $g(\mathbf{x}|\mathbf{x}^{[m]})$ for $f(\mathbf{x})$ such that

$$g(\mathbf{x}|\mathbf{x}^{[m]}) \ge f(\mathbf{x}), \forall\ \mathbf{x}\in\mathcal{F}, \quad (44)$$

and the equality is attained when $\mathbf{x} = \mathbf{x}^{[m]}$. In an MM algorithm, we will minimize the majorization function $g(\mathbf{x}|\mathbf{x}^{[m]})$ instead of the original function $f(\mathbf{x})$. Let $\mathbf{x}^{[m+1]}$ denote the minimizer of the function $g(\mathbf{x}|\mathbf{x}^{[m]})$ over $\mathcal{F}$ at the m-th iteration, i.e.,

$$\mathbf{x}^{[m+1]} = \arg\min_{\mathbf{x}\in\mathcal{F}}\ g(\mathbf{x}|\mathbf{x}^{[m]}), \quad (45)$$

then we can see that this iterative procedure will decrease the value of $f(\mathbf{x})$ monotonically after each iteration, i.e.,

$$f(\mathbf{x}^{[m+1]}) \le g(\mathbf{x}^{[m+1]}|\mathbf{x}^{[m]}) \le g(\mathbf{x}^{[m]}|\mathbf{x}^{[m]}) = f(\mathbf{x}^{[m]}), \quad (46)$$

which is a direct result from the definitions (44) and (45). The decreasing property makes an MM algorithm numerically stable. More details can be found in a tutorial on MM algorithms [34] and references therein.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers and the associate editor for their constructive comments.

REFERENCES

[1] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, pp. 20–27, June 2013.

[2] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[3] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, pp. 37–43, Mar. 2011.

[4] China Mobile, "C-RAN: the road towards green RAN," White Paper, ver. 2.5, Oct. 2011.

[5] J. Wu, "Green wireless communications: from concept to reality [industry perspectives]," *IEEE Wireless Commun.*, vol. 19, pp. 4–5, Aug. 2012.

[6] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: a new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, pp. 1380–1408, Sep. 2010.

[7] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1748–1759, Sep. 2010.

[8] C. Li, J. Zhang, and K. Letaief, "Energy efficiency analysis of small cell networks," in *Proc. 2013 IEEE Int. Conf. Commun.*, pp. 4404–4408, June 2013.

[9] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Proc. 2011 IEEE Global Commun. Conf.*, pp. 1–5.

[10] J. Rao and A. Fapojuwo, "On the tradeoff between spectral efficiency and energy efficiency of homogeneous cellular networks with outage constraint," *IEEE Trans. Veh. Technol.*, vol. 62, pp. 1801–1814, May 2013.

[11] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[12] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[13] C. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, pp. 164–174, Nov. 2010.

[14] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations Trends Mach. Learning*, vol. 4, pp. 1–106, Jan. 2012.

[15] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[16] S. Negahban and M. Wainwright, "Simultaneous support recovery in high dimensions: benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Trans. Inf. Theory*, vol. 57, pp. 3841–3863, June 2011.

[17] M. Hong, R. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, pp. 226–240, Feb. 2013.

[18] J. Zhao, T. Q. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 2762–2775, June 2013.

[19] O. Mehanna, N. Sidiropoulos, and G. Giannakis, "Joint multicast beamforming and antenna selection," *IEEE Trans. Signal Process.*, vol. 61, pp. 2660–2674, May 2013.

[20] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, pp. 692–703, Feb. 2013.

[21] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, pp. 5646–5658, Nov. 2013.

[22] V. Cadambe and S. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, pp. 3425–3441, Aug. 2008.

[23] 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.*, vol. 18, pp. 40–49, Oct. 2011.

[24] J. Kani and H. Nakamura, "Recent progress and continuing challenges in optical access network technologies," in *Proc. 2011 IEEE Int. Conf. Photonics*, pp. 66–70.

[25] A. Dhaini, P.-H. Ho, G. Shen, and B. Shihada, "Energy efficiency in TDMA-based next-generation passive optical access networks," *IEEE/ACM Trans. Netw.*, vol. PP, no. 99, p. 1, 2013.

[26] A. Wiesel, Y. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, pp. 161–176, Jan. 2006.

[27] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[28] Y. Shi, J. Zhang, and K. Letaief, "Group sparse beamforming for green cloud radio access networks," in *Proc. 2013 IEEE Global Commun. Conf.*, pp. 4635–4640.

[29] T. Baran, D. Wei, and A. Oppenheim, "Linear programming algorithms for sparse filter design," *IEEE Trans. Signal Process.*, vol. 58, pp. 1605–1617, Mar. 2010.

[30] S. Leyffer, *Mixed Integer Nonlinear Programming*, vol. 154. Springer, 2012.

[31] 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, pp. 3972–3987, Aug. 2013.

[32] G. Obozinski and F. Bach, "Convex relaxation for combinatorial penalties," arXiv preprint arXiv:1205.1240, 2012.

[33] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.

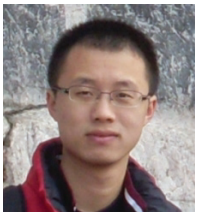
[34] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[35] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, Dec. 2008.

[36] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Mach. Learning*, vol. 85, pp. 3–39, Oct. 2011.

[37] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends Mach. Learning*, vol. 3, pp. 1–122, July 2011.

[38] R. T. Rockafellar, *Convex Analysis*, vol. 28. Princeton University Press, 1997.



Yuanming Shi (S'13) received his B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011. He is currently working towards the Ph.D. degree in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST). His research interests include 5G wireless communication networks, Cloud-RAN, optimization theory, and large-scale optimization and its applications.

Jun Zhang (S'06-M'10) received the B.Eng. degree in electronic engineering from the University of Science and Technology of China in 2004, the M.Phil. degree in information engineering from the Chinese University of Hong Kong in 2006, and the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2009. He is currently a Visiting Assistant Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST). Dr. Zhang is co-author of the book *Fundamentals of LTE* (Prentice-Hall, 2010). His research interests include MIMO communications, heterogeneous networks, cognitive radio, and green communications. He has served on TPCs of different international conferences including IEEE ICC, VTC, Globecom, WCNC, PIMRC, etc. He served as a MAC track co-chair for IEEE WCNC 2011.

Khaled B. Letaief (S'85-M'86-SM'97-F'03) received the B.S. degree *with distinction* in electrical engineering (1984) from Purdue University, USA. He also received the M.S. and Ph.D. degrees in electrical engineering from Purdue University in 1986 and 1990, respectively.

From January 1985 and as a Graduate Instructor at Purdue, he taught courses in communications and electronics. From 1990 to 1993, he was a faculty member at the University of Melbourne, Australia. Since 1993, he has been with the Hong Kong University of Science and Technology (HKUST) where he is currently Chair Professor and the Dean of Engineering, with expertise in wireless communications and networks. In these areas, he has over 470 journal and conference papers and has given invited keynote talks as well as courses all over the world. He has 13 patents including 11 US patents.

Dr. Letaief serves as a consultant for different organizations and is the founding Editor-in-Chief of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has served on the editorial board of other prestigious journals including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS-WIRELESS SERIES (as Editor-in-Chief). He has been involved in organizing a number of major international conferences. These include WCNC'07 in Hong Kong; ICC'08 in Beijing; ICC'10 in Cape Town; TTM'11 in Hong Kong; and ICCC'12 in Beijing.

Professor Letaief has been a dedicated teacher committed to excellence in teaching and scholarship. He received the *Mangoon Teaching Award* from Purdue University in 1990; the HKUST Engineering Teaching Excellence Award; and the Michael Gale Medal for Distinguished Teaching (*Highest University-wide Teaching Award* at HKUST). He is also the recipient of many other distinguished awards including the 2007 IEEE Communications Society Publications Exemplary Award; the 2009 IEEE Marconi Prize Award in Wireless Communications; the 2010 Purdue University Outstanding Electrical and Computer Engineer Award; the 2011 IEEE Communications Society Harold Sobol Award; the 2011 IEEE Wireless Communications Technical Committee Recognition Award; and 10 IEEE Best Paper Awards.

Dr. Letaief is a Fellow of IEEE and a Fellow of HKIE. He has served as an elected member of the IEEE Communications Society (ComSoc) Board of Governors, as an IEEE Distinguished lecturer, IEEE ComSoc Treasurer, and IEEE ComSoc Vice-President for Conferences.

He is currently serving as the IEEE ComSoc Vice-President for Technical Activities as a member of the IEEE Product Services and Publications Board, and is a member of the IEEE Fellow Committee. He is also recognized by Thomson Reuters as an *ISI Highly Cited Researcher*.