Flexible Functional Split Design for Downlink C-RAN With Capacity-Constrained Fronthaul

Yong Zhou, *Member, IEEE*, Jie Li, Yuanming Shi, *Member, IEEE*, and Vincent W. S. Wong, *Fellow, IEEE*

***Abstract*—In cloud radio access networks, functional split refers to a division of signal processing functionalities between the baseband unit (BBU) pool and remote radio heads (RRHs). The functionality of baseband signal precoding can either be performed by the BBU pool or RRHs, which corresponds to different functional splits. The compression-after-precoding (CAP) and data-sharing (DS) strategies are the realizations of these two functional splits. In this paper, we propose a flexible functional split design to enable the dynamic functional configuration of each active RRH to use either CAP or DS strategy. Our goal is to minimize the aggregate power consumption, while taking into account limited fronthaul capacity, fronthaul power consumption, and quality-of-service requirement. We formulate a joint RRH mode (i.e., CAP, DS, and sleep) selection, precoding design, and fronthaul compression problem. The formulated problem is a non-convex quadratically constrained combinatorial optimization problem. Through sequential convex programming and ℓ_1-norm convex relaxation, the problem is transformed into a sequence of semidefinite programming problems. An efficient algorithm based on the majorization–minimization scheme is developed to solve the problem. Simulations demonstrate the importance of considering the limited fronthaul capacity and the performance improvement of the proposed algorithm compared with the pure CAP and DS strategies.**

***Index Terms*—Cloud radio access network, flexible functional split, capacity-constrained fronthaul, energy efficiency, semidefinite relaxation.**

I. Introduction

By IMPROVING the spatial frequency reuse and reducing the distance between the user equipments (UEs) and the access points, the ultra dense deployment of small cells is recognized as an efficient and effective method to boost the network capacity of the fifth generation (5G) wireless networks [1]. However, with the densification of small cells, every new cell adds to co-channel interference, which is a key performance-limiting factor in radio access networks (RANs).

With the virtualization of baseband signal processing functionalities, cloud RAN (C-RAN) has been proposed as a promising network architecture for 5G wireless networks [2]. In C-RAN, the baseband unit (BBU) pool, composed of multiple BBUs, performs centralized baseband signal processing and coordinates the transmissions of low-cost remote radio heads (RRHs). The digitized baseband inphase and quadrature samples of radio signals between the BBU pool and the RRHs are transmitted through low-latency optical fronthaul links. C-RAN can enhance the spectrum and energy efficiency by suppressing co-channel interference via cooperative transmission/reception [3], [4]. It can also reduce the network capital expenditure and operating expenditure by adapting to spatial and temporal traffic variations via statistical multiplexing.

The aforementioned benefits of C-RAN are achieved at the cost of imposing a significant burden on fronthaul links. However, the fronthaul links are usually capacity-constrained in practice [5]–[7], which may become the bottleneck of the centralized signal processing and affect the resource allocation processes across RRHs. The *compression-after-precoding* (CAP) and *data-sharing* (DS) strategies are two fundamental cooperative strategies in C-RAN. In the CAP strategy, the BBU pool performs centralized precoding and compresses the precoded baseband signals before delivering them to the corresponding RRHs through fronthaul links. On the other hand, in the DS strategy, the BBU pool transmits the precoding coefficients along with the original signals to the RRHs, which perform local precoding. Based on these two strategies, resource allocation [8], fronthaul compression [9], RRH clustering [10]–[12], and device-to-device (D2D) communications [13] are studied to alleviate the fronthaul capacity constraint. Specifically, Zhao *et al.* in [8] propose a joint transmit beamforming design and user data allocation scheme to minimize the requirement on fronthaul capacity. Given the finite capacity of fronthaul links, the weighted sum-rate of the CAP strategy can be enhanced by jointly compressing the precoded signals for different RRHs [9]. By balancing the tradeoff between the cooperation gain and fronthaul capacity constraint, a dynamic user-centric clustering scheme is investigated in [10] to maximize the weighted sum-rate. Under the fronthaul capacity constraint, we propose a multi-timescale resource allocation mechanism to guarantee efficient resource sharing among multiple service providers as well as to address the user mobility issue [11]. Moreover, the authors in [12] propose an approximate stochastic cutting plane algorithm to address the short-term precoding and long-term user-centric clustering problems for sum-rate maximization. Taking

Manuscript received October 25, 2018; revised January 22, 2019 and March 12, 2019; accepted April 10, 2019. Date of publication April 18, 2019; date of current version June 18, 2019. This work was supported in part by the ShanghaiTech University Start-Up Funds, in part by the National Nature Science Foundation of China under Grant 61601290, in part by the Shanghai Sailing Program under Grant 16YF1407700, and in part by the Natural Sciences and Engineering Research Council of Canada. The review of this paper was coordinated by Dr. S. Misra. *(Corresponding author: Yong Zhou.)*

Y. Zhou, J. Li, and Y. Shi are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: zhouyong@shanghaitech.edu.cn; lijie3@shanghaitech.edu.cn; shiym@shanghaitech.edu.cn).

V. W. S. Wong is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: vincentw@ece.ubc.ca).

Digital Object Identifier 10.1109/TVT.2019.2911934

into account dynamic traffic arrival, the authors in [13] formulate a stochastic optimization problem to maximize the overall throughput of C-RAN with D2D communications, which allow direct communication between two adjacent UEs without going through fronthaul links. However, the aforementioned studies focus on maximizing the spectrum efficiency without considering the power consumption issue in C-RAN.

With an increasing number of RRHs, minimizing the power consumption becomes an important design objective of C-RAN due to the economic concern of network operators [14]–[16]. By exploiting spatial and temporal traffic fluctuations, power consumption can be significantly reduced by switching off idle RRHs to provide on-demand services for UEs [17]. The authors in [18] and [19] propose dynamic RRHs and virtual base stations clustering and resource provisioning schemes to adapt to the fluctuations of UEs' capacity demand, which can enhance the energy efficiency and data rate. In C-RAN, the power consumption introduced by fronthaul links is comparable to that of RRHs and thus cannot be neglected. By taking into account the fronthaul power consumption, a joint RRH selection and precoding design problem is formulated in [20] to minimize the aggregate power consumption. To efficiently solve this problem, a low complexity algorithm based on group sparse precoding is proposed. Such an optimization framework is extended to account for both downlink and uplink transmissions in [21], and to address the generalized sparse and low-rank optimization in [22]. By modeling the fronthaul power consumption as a function of the fronthaul data rate, the energy efficiency of C-RAN is investigated in [23]. The authors in [24] exploit the benefit of non-orthogonal multiple access (NOMA) in C-RAN to enhance the energy efficiency. Moreover, to address the channel uncertainty, a robust beamforming problem is formulated in [25], where an alternating direction method of multipliers (ADMM)-based algorithm is proposed to solve the problem. However, the impact of the fronthaul capacity constraint on power consumption is not studied in the aforementioned studies.

To minimize the aggregate power consumption, it is necessary to take into account the limited fronthaul capacity as it affects the number of RRHs required to be active, which in turn determines the circuit and fronthaul power consumption. Hence, the effect of the limited fronthaul capacity on the aggregate power consumption of C-RAN should be investigated. The concept and benefit of flexible functional splits between the BBU pool and RRHs in the physical (PHY) and medium access control (MAC) layers are discussed in [26] and [27]. The authors in [28] formulate an integer linear programming problem to minimize the inter-cell interference by dynamically adjusting the functional split in PHY and MAC layers. However, the radio transmission of data streams between the RRHs and the UEs, as an indispensable component of C-RAN, is not taken into account. Differently, the CAP and DS strategies correspond to different divisions of signal processing functionalities between the BBU pool and RRHs. Specifically, the baseband signal precoding functionality in the CAP and DS strategies is performed centrally by the BBU pool and locally by the RRHs, respectively. The fronthaul capacity required by the CAP and DS strategies depends on different parameters. In particular, the fronthaul data rate of the CAP strategy depends on the precoding gain, quantization noise, and the number of antennas on the RRH, while the fronthaul data rate of the DS strategy is determined by the number of UEs served by the RRH. The maximizations of energy efficiency for downlink C-RAN using DS and CAP strategies are separately studied in [29]. Flexible functional split enables each RRH to support either the CAP or DS strategy, so as to fully utilize the fronthaul capacity based on the quality of service (QoS) requirement of UEs and channel conditions. However, utilizing flexible functional split design to reduce power consumption has not been studied. Moreover, as most existing works (e.g., [30], [31]) use the CAP strategy to maximize the spectrum efficiency, the impact of fronthaul compression on the tradeoff between the aggregate power consumption and the fronthaul capacity requirement has not been investigated.

Different from the aforementioned studies, in this paper we propose a flexible functional split design to minimize the aggregate power consumption of downlink C-RAN, while taking into account the fronthaul capacity constraint and the quality of service requirement. The power consumption under consideration includes the RRH transmit power, RRH circuit power, and fronthaul power consumption. Each RRH can be switched off to save power, which corresponds to the sleep mode. Each active RRH can flexibly be configured to support either the CAP or DS strategy to further reduce the power consumption, leading to a mixture of RRHs using the CAP and DS strategies in the network. Such a flexible functional split design takes the advantages of both the CAP and DS strategies, and enables the full utilization of the fronthaul capacity for given quality of service requirement. The main contributions of this paper are summarized as follows:

1) We formulate a joint RRH mode (i.e., CAP, DS, sleep) selection, precoding design, and fronthaul compression problem to minimize the aggregate power consumption, while taking into account the limited fronthaul capacity, per-RRH power constraint, and QoS requirement.
2) To tackle the non-convex quadratical constraints, we transform the formulated problem into a sequence of rank-constrained semidefinite programming (SDP) problems through sequential convex programming (SCP) and ℓ_1-norm convex relaxation. We handle the combinatorial RRH mode selection by using the group sparse precoding approach and develop an efficient algorithm based on the majorize minimization (MM) scheme to solve the problem.
3) Simulations demonstrate the convergence of the proposed algorithm and show that the fronthaul capacity constraint has a significant impact on the aggregate power consumption. In addition, the CAP strategy performs better than the DS strategy in the high target data rate and/or low fronthaul capacity regimes. By taking advantages of both the CAP and DS strategies, the proposed algorithm outperforms both baseline strategies in terms of the energy efficiency.

The remainder of this paper is organized as follows. Section II presents the network topology, the CAP and DS strategies, the signal reception model, and the power consumption

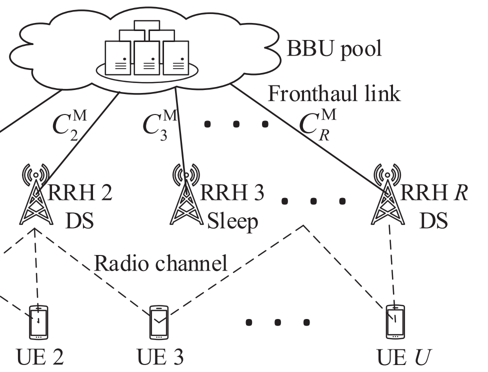


Fig. 1. An illustration of the architecture of a C-RAN, which consists of the BBU pool, the optical fronthaul links of finite capacity, the multi-antenna RRHs, the radio channels, and the single-antenna UEs. An RRH can either be in the active or sleep mode. Each active RRH can flexibly be configured to support either the CAP or DS strategy.

model. In Section III, we formulate a non-convex quadratically constrained optimization problem to minimize the aggregate power consumption and transform it into a sequence of rank-constrained SDP problems. The proposed algorithm is presented in Section IV. The performance of the proposed algorithm is evaluated in Section V. Finally, Section VI concludes this paper.

Notation: $\mathbb{R}$ and $\mathbb{C}$ denote the real and complex domains, respectively. The absolute value of a scalar is denoted as $|\cdot|$. The conjugate transpose and ℓ_p-norm of a vector are denoted as $(\cdot)^{\mathrm{H}}$ and $\|\cdot\|_p$, respectively. The inverse, trace, determinant, and rank of a matrix are denoted as $(\cdot)^{-1}$, $\mathrm{Tr}(\cdot)$, $\det(\cdot)$, and $\mathrm{rank}(\cdot)$, respectively. Denote $\mathbf{1}_x$ and $\mathbf{I}_x$ as the unit vector of length x and the identity matrix of order x, respectively. Indicator function $\mathbb{1}_{\{x\}}$ equals to 0 if $x = 0$, and 1 otherwise. $\mathbf{X} \succeq 0$ and $\mathbf{X} \succ 0$ indicate that matrix $\mathbf{X}$ is positive semidefinite and definite, respectively.

II. System Model

Consider the downlink transmission of a C-RAN, which consists of one BBU pool, R RRHs, and U UEs, as shown in Fig. 1. We denote $\mathcal{R} = \{1, 2, \ldots, R\}$ and $\mathcal{U} = \{1, 2, \ldots, U\}$ as the sets of the RRHs and UEs, respectively. The r-th RRH is equipped with N_r omni-directional antennas. Each UE has a single omni-directional antenna and it receives a single independent data stream from the BBU pool, which performs centralized baseband signal processing, cooperative strategy selection, and coordinated resource allocation. The data streams for all UEs are assumed to be available at the BBU pool. The BBU pool connects to each RRH via an optical fronthaul link of finite capacity. In addition to the radio frequency (RF) functionality (e.g., power amplification), each RRH has baseband signal processing capabilities such as precoding. After receiving the data streams from the BBU pool, the RRHs forward the data streams to the corresponding UEs over quasi-static radio channels. The global channel state information (CSI) is assumed to be available at the BBU pool, as in [8]–[10].

To fully utilize the available resources (e.g., radio spectrum, transmit power, and fronthaul capacity) to meet the UEs' QoS requirement, we propose a flexible functional split design for C-RAN. In particular, each active RRH can flexibly be configured to support either the CAP or DS strategy, as shown in Fig. 2(a). The block diagrams of the CAP and DS strategies are illustrated in Fig. 2(b) and (c), respectively. In the CAP strategy, based on the CSI and UEs' QoS requirement, the BBU pool performs centralized precoding and delivers the compressed signals to the RRHs, as shown in Fig. 2(b). On the other hand, in the DS strategy, the BBU pool delivers both the signals and precoding vectors to the corresponding RRHs, which perform local precoding, as shown in Fig. 2(c). The CAP strategy and DS strategy correspond to the PHY-RF split and MAC-PHY split proposed for 5G RAN in [32], [33], respectively. The fronthaul interfaces supporting the CAP and DS strategies follow the common public radio interface (CPRI) and Fx interface [32], respectively. Hence, the fronthaul interface should change accordingly with the cooperative strategy (i.e., CAP or DS) selected by its connected RRH. We assume that the RRHs can switch among the sleep, CAP, and DS modes with negligible delay. Due to the differences in baseband signal processing and data sharing, the CAP and DS strategies are different in terms of the fronthaul data rate and the RRH transmit power, as discussed in detail as follows.

A. CAP Strategy

We denote s_u as the signal intended for UE $u \in \mathcal{U}$. Without loss of generality, the signals are assumed to be independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and unit variance. In the BBU pool, the precoded baseband signal for the r-th RRH supporting the CAP strategy, denoted as $\widehat{\mathbf{x}}_r \in \mathbb{C}^{N_r \times 1}$, is given by

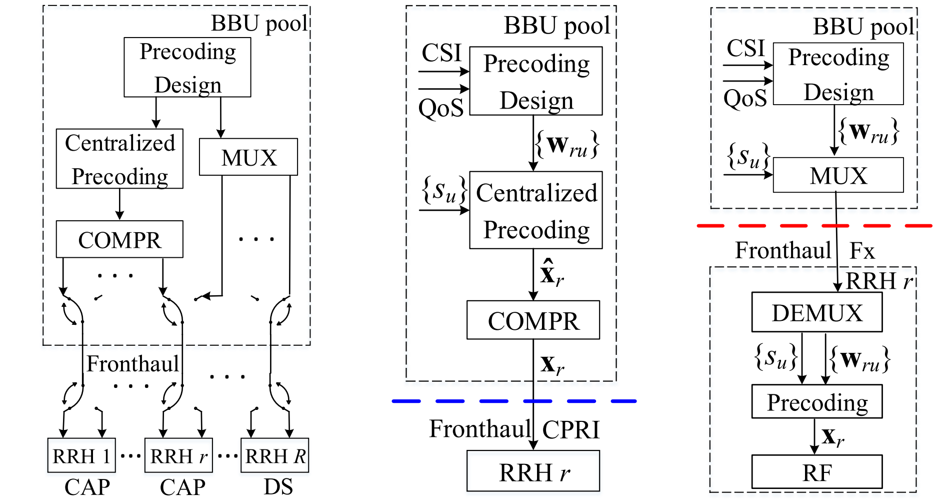
$$\widehat{\mathbf{x}}_r = \sum_{u \in \mathcal{U}} \mathbf{w}_{ru} s_u, \qquad \forall\, r \in \mathcal{R}^{\mathrm{C}}, \tag{1}$$

where $\mathbf{w}_{ru} \in \mathbb{C}^{N_r \times 1}$ denotes the precoding vector at RRH r for UE u, and $\mathcal{R}^{\mathrm{C}} \subseteq \mathcal{R}$ denotes the set of active RRHs using the CAP strategy. Note that the coefficients of the precoding vector $\mathbf{w}_{ru}$ should be set to 0 if RRH r is not serving UE u.

In order to reduce the amount of information delivered over the fronthaul links, the BBU pool compresses and quantizes the precoded baseband signals before transmitting them to the RRHs. Each $\widehat{\mathbf{x}}_r$ is independently compressed and quantized across the RRHs. Note that it is possible to leverage joint signal compression to further alleviate the fronthaul capacity constraint as in [9], which is out of the scope of this paper. The compressed signal for the r-th RRH using the CAP strategy can be expressed as

$$\mathbf{x}_r = \widehat{\mathbf{x}}_r + \mathbf{q}_r, \qquad \forall\, r \in \mathcal{R}^{\mathrm{C}}, \tag{2}$$

where $\mathbf{q}_r \in \mathbb{C}^{N_r \times 1}$ denotes the quantization noise vector, which is independent of $\widehat{\mathbf{x}}_r$ and is assumed to be Gaussian distributed with zero mean and variance $\sigma^2_{\mathrm{q},r} \mathbf{1}_{N_r}$. According to the rate-distortion theory [34], the achievable compression rate equals to the mutual information between the compressed signal $\mathbf{x}_r$ and the precoded baseband signal $\widehat{\mathbf{x}}_r$. As a result, for the CAP strategy, the data rate of the r-th fronthaul, $\forall\, r \in \mathcal{R}^{\mathrm{C}}$, can be



(a) Flexible functional split design (b) CAP strategy (c) DS strategy

Fig. 2. An illustration of the block diagram of the flexible functional split between the BBU pool and RRHs, where COMPR, MUX, and DEMUX represent the compression, multiplexer, and demultiplexer, respectively.

calculated by

$$B \log_2 \det\left(\sum_{u\in\mathcal{U}} \mathbf{w}_{ru}\mathbf{w}_{ru}^{\mathrm{H}} + \sigma_{\mathrm{q},r}^2 \mathrm{I}_{N_r}\right) - N_r B \log_2\left(\sigma_{\mathrm{q},r}^2\right), \quad (3)$$

where B denotes the channel bandwidth. According to (3), the fronthaul data rate of the CAP strategy depends on the values of the precoding coefficients $\mathbf{w}_{ru}$ and the quantization noise $\sigma_{\mathrm{q},r}^2$ as well as on the number of the antennas of RRH r. In particular, a higher precoding gain and smaller quantization noise lead to smaller signal distortion, but also a higher fronthaul data rate. In this paper, such a tradeoff is balanced by jointly optimizing the precoding coefficients and quantization noise.

B. DS Strategy

In the DS strategy, the BBU pool delivers both signal s_u and its corresponding precoding vectors $\{\mathbf{w}_{ru}\}$ to a cluster of RRHs serving UE u through fronthaul links. Similar to the CAP strategy, all coefficients of the precoding vector $\mathbf{w}_{ru}$ should be set to 0 if RRH r is not within the serving cluster of UE u. After receiving the signals and the corresponding precoding vectors, each RRH performs local precoding. As a result, the signal transmitted by the r-th RRH can be written as

$$\mathbf{x}_r = \sum_{u\in\mathcal{U}} \mathbf{w}_{ru} s_u, \qquad \forall\, r \in \mathcal{R}^{\mathrm{D}}, \quad (4)$$

where $\mathcal{R}^{\mathrm{D}} \subseteq \mathcal{R}$ denotes the set of active RRHs using the DS strategy. As each active RRH can be configured to support either the CAP or DS strategy, we have $\mathcal{R}^{\mathrm{C}} \cap \mathcal{R}^{\mathrm{D}} = \emptyset$.

According to (4), as the signals and corresponding precoding vectors are required to perform local precoding at each RRH, the fronthaul data rate is the summation of the data rates required by its serving UEs. For simplicity, the overhead introduced by CSI estimation and precoding vector delivery is ignored due to its negligible size compared with the data stream. As a result, for the DS strategy, the data rate of the r-th fronthaul can be expressed as

$$\sum_{u\in\mathcal{U}} \mathbb{1}_{\{\|\mathbf{w}_{ru}\|_2^2\}} B \log_2(1+\gamma_u), \qquad \forall\, r \in \mathcal{R}^{\mathrm{D}}, \quad (5)$$

where γ_u denotes the target signal-to-interference-plus-noise ratio (SINR) of UE u. According to (5), the fronthaul data rate of the DS strategy is determined by the number of UEs served by the RRH and the target SINR of all serving UEs. In particular, having more serving UEs at each RRH leads to a higher cooperation gain, but also a higher fronthaul data rate. Comparing with (3), different parameters influence the fronthaul data rates of the CAP and DS strategies.

C. Signal Reception Model

With full spatial frequency reuse, each UE can simultaneously receive its own signal transmitted from both the RRHs in $\mathcal{R}^{\mathrm{C}}$ and the RRHs in $\mathcal{R}^{\mathrm{D}}$ over radio channels. The signal received at UE u is given by

$$y_u = \sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}} \mathbf{x}_r + n_u, \qquad \forall\, u \in \mathcal{U}, \quad (6)$$

where $\mathbf{h}_{ru} \in \mathbb{C}^{N_r \times 1}$ denotes the channel fading vector between RRH r and UE u and incorporates the effects of both path loss and small-scale fading, and n_u denotes the additive white Gaussian noise (AWGN) at UE u with zero mean and variance $\sigma_{\mathrm{n},u}^2$.

By substituting (2) and (4) into (6), we have

$$y_u = \sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}} \mathbf{w}_{ru} s_u + \sum_{k\in\mathcal{U}\setminus\{u\}} \sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}} \mathbf{w}_{rk} s_k + \sum_{r\in\mathcal{R}^{\mathrm{C}}} \mathbf{h}_{ru}^{\mathrm{H}} \mathbf{q}_r + n_u, \qquad \forall\, u \in \mathcal{U}, \quad (7)$$

where the second term of the right hand side is the co-channel interference.

By using single user detection (i.e., treating the co-channel interference as noise), according to (7), the received SINR at

UE $u \in \mathcal{U}$ can be written as

$$\mathrm{SINR}_u = \frac{\left|\sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}}\mathbf{w}_{ru}\right|^2}{I_u + \sigma_{\mathrm{n},u}^2}, \tag{8}$$

where I_u denotes the summation of the co-channel interference and quantization noise power, given by

$$I_u = \sum_{k\in\mathcal{U}\backslash\{u\}} \left|\sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}}\mathbf{w}_{rk}\right|^2 + \left|\sum_{r\in\mathcal{R}^{\mathrm{C}}} \mathbf{h}_{ru}^{\mathrm{H}}\sigma_{\mathrm{q},r}\mathbf{1}_{N_r}\right|^2. \tag{9}$$

D. Power Consumption Model

The aggregate power consumption consists of RRH transmit power, RRH circuit power, and fronthaul power consumption. According to (2), the transmit power of the r-th RRH using the CAP strategy is given by

$$P_r^{\mathrm{tx}} = \sum_{u\in\mathcal{U}} \|\mathbf{w}_{ru}\|_2^2 + N_r\sigma_{\mathrm{q},r}^2, \qquad \forall\, r \in \mathcal{R}^{\mathrm{C}}. \tag{10}$$

Similarly, according to (4), the transmit power of the r-th RRH using the DS strategy can be expressed as

$$P_r^{\mathrm{tx}} = \sum_{u\in\mathcal{U}} \|\mathbf{w}_{ru}\|_2^2, \qquad \forall\, r \in \mathcal{R}^{\mathrm{D}}. \tag{11}$$

Based on (10) and (11), the RRH transmit power of the CAP strategy involves the quantization noise power, which is different from that of the DS strategy.

The RRH circuit power consists of the RF circuit and basic baseband processing power consumption. They depend on the transmission mode of the RRH (i.e., being in either active or sleep mode). In particular, the circuit power of RRH $r \in \mathcal{R}^\nu, \nu \in \{\mathrm{C}, \mathrm{D}\}$ is modeled by a piecewise function,

$$P_r^{\mathrm{cc},\nu} = \begin{cases} P_{\mathrm{a},r}^{\mathrm{cc},\nu}, & \text{if } P_r^{\mathrm{tx}} > 0, \\ P_{\mathrm{s},r}^{\mathrm{cc}}, & \text{if } P_r^{\mathrm{tx}} = 0, \end{cases} \tag{12}$$

where $P_{\mathrm{a},r}^{\mathrm{cc},\nu}$ and $P_{\mathrm{s},r}^{\mathrm{cc}}$ denote the circuit power of the r-th RRH in $\mathcal{R}^\nu$ in the active and sleep modes, respectively, and $P_{\mathrm{a},r}^{\mathrm{cc},\nu} > P_{\mathrm{s},r}^{\mathrm{cc}}$.

Similarly, the power consumption of the r-th fronthaul in the active and sleep modes are denoted as $P_{\mathrm{a},r}^{\mathrm{fh}}$ and $P_{\mathrm{s},r}^{\mathrm{fh}}$, respectively, and $P_{\mathrm{a},r}^{\mathrm{fh}} > P_{\mathrm{s},r}^{\mathrm{fh}}$. By denoting $\mathcal{R}^{\mathrm{S}} \subseteq \mathcal{R}$ as the set of the RRHs in the sleep mode, the total power consumption of RRH circuits and fronthaul links is given by

$$\begin{aligned} P^{\mathrm{cf}} &= \sum_{r\in\mathcal{R}^{\mathrm{C}}} \left(P_{\mathrm{a},r}^{\mathrm{cc,C}} + P_{\mathrm{a},r}^{\mathrm{fh}}\right) + \sum_{r\in\mathcal{R}^{\mathrm{D}}} \left(P_{\mathrm{a},r}^{\mathrm{cc,D}} + P_{\mathrm{a},r}^{\mathrm{fh}}\right) \\ &\quad + \sum_{r\in\mathcal{R}^{\mathrm{S}}} \left(P_{\mathrm{s},r}^{\mathrm{cc}} + P_{\mathrm{s},r}^{\mathrm{fh}}\right) \\ &= \sum_{r\in\mathcal{R}^{\mathrm{C}}} \left(P_{\mathrm{a},r}^{\mathrm{cc,C}} + P_{\mathrm{a},r}^{\mathrm{fh}} - P_{\mathrm{s},r}^{\mathrm{cc}} - P_{\mathrm{s},r}^{\mathrm{fh}}\right) \\ &\quad + \sum_{r\in\mathcal{R}^{\mathrm{D}}} \left(P_{\mathrm{a},r}^{\mathrm{cc,D}} + P_{\mathrm{a},r}^{\mathrm{fh}} - P_{\mathrm{s},r}^{\mathrm{cc}} - P_{\mathrm{s},r}^{\mathrm{fh}}\right) \\ &\quad + \sum_{r\in\mathcal{R}} \left(P_{\mathrm{s},r}^{\mathrm{cc}} + P_{\mathrm{s},r}^{\mathrm{fh}}\right). \end{aligned}$$

By denoting $P_{r,\mathrm{C}}^{\mathrm{dif}} = P_{\mathrm{a},r}^{\mathrm{cc,C}} + P_{\mathrm{a},r}^{\mathrm{fh}} - P_{\mathrm{s},r}^{\mathrm{cc}} - P_{\mathrm{s},r}^{\mathrm{fh}} > 0$, $P_{r,\mathrm{D}}^{\mathrm{dif}} = P_{\mathrm{a},r}^{\mathrm{cc,D}} + P_{\mathrm{a},r}^{\mathrm{fh}} - P_{\mathrm{s},r}^{\mathrm{cc}} - P_{\mathrm{s},r}^{\mathrm{fh}} > 0$, and omitting the constant term $\sum_{r\in\mathcal{R}} \left(P_{\mathrm{s},r}^{\mathrm{cc}} + P_{\mathrm{s},r}^{\mathrm{fh}}\right)$, minimizing the aggregate power consumption is equivalent to minimizing P_{agg}, which is given by

$$\begin{aligned} P_{\mathrm{agg}} = &\sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \sum_{u\in\mathcal{U}} \frac{1}{\eta_r} \|\mathbf{w}_{ru}\|_2^2 + \sum_{r\in\mathcal{R}^{\mathrm{C}}} \left(\frac{1}{\eta_r} N_r\sigma_{\mathrm{q},r}^2 + P_{r,\mathrm{C}}^{\mathrm{dif}}\right) \\ &+ \sum_{r\in\mathcal{R}^{\mathrm{D}}} P_{r,\mathrm{D}}^{\mathrm{dif}}, \end{aligned} \tag{13}$$

where $\eta_r > 0$ denotes the drain efficiency [35] of the RF power amplifier of the r-th RRH.

Discussions: There exist performance tradeoff between the DS and CAP strategies in terms of the cooperation gain, the fronthaul data rate, and the power consumption. The main advantage of the DS strategy is that each RRH receives the original signals of its serving UEs without distortion. However, the cooperation gain that can be achieved by the DS strategy is determined by the RRH cluster size (i.e., the number of RRHs cooperatively transmitting the same signal). In particular, a larger cluster size contributes to a higher cooperation gain, but also leads to a larger fronthaul data rate, as each cooperating RRH is required to receive a copy of the original signal. Note that a larger cluster size for each UE corresponds to more UEs served by each RRH. Therefore, the fronthaul capacity constraint limits the cluster size and the cooperation gain. In the high traffic load regime (e.g., high target data rate and large number of UEs), the DS strategy may require more RRHs to be active than the CAP strategy, so as to achieve a large enough cooperation gain to meet the target data rate requirement of UEs, leading to higher circuit power consumption. For the CAP strategy, the RRHs can receive the precoded baseband signal by utilizing all UEs' signals, and hence, achieving full cooperation. The fronthaul data rate of the CAP strategy can be adjusted by changing the quantization noise, which determines the compression resolution. Compared to the DS strategy, the main disadvantage of the CAP strategy is the signal distortion due to quantization noise, which leads to larger transmit power consumption. In the low traffic load regime (e.g., low target data rate and small number of UEs), the DS strategy is able to achieve full cooperation, and hence can consume less power than the CAP strategy, which suffers from quantization noise.

III. Problem Formulation and Transformation

To minimize the aggregate power consumption, we need to reduce both the RRH transmit power and the number of active RRHs and corresponding fronthaul links. However, there exists a tradeoff between these two aspects. Specifically, to reduce the RRH transmit power, more RRHs are required to be active to meet UEs' QoS requirement. On the other hand, having less active RRHs leads to lower RRH circuit and fronthaul power consumption, but also higher RRH transmit power. Such a tradeoff is further affected by other factors, including the capacity constraints of fronthaul links, maximum transmit power constraints of the RRHs, and UEs' QoS constraints. Hence, the RRH mode (i.e., CAP, DS, sleep) selection, precoding design, and fronthaul compression should jointly be optimized to minimize the aggregate power consumption. Note that the precoding coefficients and the quantization noise not only affect the RRH modes for the reduction of RRH circuit power and fronthaul power consumption, but also play an important role in further reducing the transmit power consumption when the RRH modes are fixed. Based on the above discussions, the aggregate power consumption minimization problem is formulated as

$$\underset{\substack{\mathcal{R}^{\mathrm{C}},\mathcal{R}^{\mathrm{D}},\mathcal{R}^{\mathrm{S}}\\ \{\mathbf{w}_{ru}\},\{\sigma^2_{\mathrm{q},r}\}}}{\text{minimize}} \; P_{\mathrm{agg}} \tag{14a}$$

$$\text{subject to} \sum_{u\in\mathcal{U}} \|\mathbf{w}_{ru}\|_2^2 + N_r\sigma^2_{\mathrm{q},r} \le P_r^{\mathrm{M}}, \quad \forall\, r\in\mathcal{R}^{\mathrm{C}}, \tag{14b}$$

$$\sum_{u\in\mathcal{U}} \|\mathbf{w}_{ru}\|_2^2 \le P_r^{\mathrm{M}}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}, \tag{14c}$$

$$\sum_{u\in\mathcal{U}} \|\mathbf{w}_{ru}\|_2^2 = 0, \quad \forall\, r\in\mathcal{R}^{\mathrm{S}}, \tag{14d}$$

$$B\log_2\det\left(\sum_{u\in\mathcal{U}} \mathbf{w}_{ru}\mathbf{w}_{ru}^{\mathrm{H}} + \sigma^2_{\mathrm{q},r}\mathrm{I}_{N_r}\right) - N_r B\log_2\left(\sigma^2_{\mathrm{q},r}\right) \le C_r^{\mathrm{M}}, \forall\, r\in\mathcal{R}^{\mathrm{C}}, \tag{14e}$$

$$\sum_{u\in\mathcal{U}} \mathbb{1}_{\left\{\|\mathbf{w}_{ru}\|_2^2\right\}} B\log_2(1+\gamma_u) \le C_r^{\mathrm{M}}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}, \tag{14f}$$

$$\frac{\left|\sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}} \mathbf{h}_{ru}^{\mathrm{H}}\mathbf{w}_{ru}\right|^2}{I_u + \sigma^2_{\mathrm{n},u}} \ge \gamma_u, \quad \forall\, u\in\mathcal{U}, \tag{14g}$$

$$\mathcal{R}^{\mathrm{C}} \cap \mathcal{R}^{\mathrm{D}} = \emptyset, \tag{14h}$$

$$\left(\mathcal{R}^{\mathrm{C}} \cup \mathcal{R}^{\mathrm{D}}\right) \cap \mathcal{R}^{\mathrm{S}} = \emptyset, \tag{14i}$$

$$\mathcal{R}^{\mathrm{C}} \cup \mathcal{R}^{\mathrm{D}} \cup \mathcal{R}^{\mathrm{S}} = \mathcal{R}, \tag{14j}$$

where P_r^{M} and C_r^{M} denote the maximum transmit power of the r-th RRH and the capacity of the r-th fronthaul link, respectively. Constraints (14b)–(14d) represent the maximum transmit power constraints of the RRHs in $\mathcal{R}^{\mathrm{C}}$, $\mathcal{R}^{\mathrm{D}}$, and $\mathcal{R}^{\mathrm{S}}$, respectively. Constraints (14h)–(14j) ensure that each RRH can be configured to support one of the CAP, DS, and sleep modes. The aggregate power consumption minimization problem in (14) is a non-convex quadratically constrained combinatorial optimization problem, which is generally difficult to solve and imposes the following challenges: First, the objective function (14a) is a combinatorial function due to both the RRH selection (i.e., either being in the active or sleep mode) and the cooperative strategy selection (i.e., supporting either the CAP or DS strategy). Second, the capacity constraints of fronthaul links supporting the CAP and DS strategies (i.e., (14e) and (14f)), and the QoS constraints of the UEs in terms of the SINR (i.e., (14g)) are non-convex quadratically constrained.

To address the aforementioned challenges, we transform problem (14) into a sequence of rank-constrained SDP problems. We define precoding matrix $\mathbf{W}_u = \mathbf{w}_u\mathbf{w}_u^{\mathrm{H}} \in \mathbb{C}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$ as a new optimization variable for UE $u\in\mathcal{U}$, where $N_{\mathrm{T}} = \sum_{r=1}^{R} N_r$ and $\mathbf{w}_u = [\mathbf{w}_{1u}^{\mathrm{H}}, \mathbf{w}_{2u}^{\mathrm{H}}, \ldots, \mathbf{w}_{Ru}^{\mathrm{H}}]^{\mathrm{H}} \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$. We have constraints $\mathbf{W}_u \succeq \mathbf{0}$ and $\mathrm{rank}(\mathbf{W}_u) = 1$ for UE $u\in\mathcal{U}$. Precoding vector $\mathbf{w}_u$ is given by the eigenvector of $\mathbf{W}_u$. The maximum transmit power constraints of the RRHs using the CAP strategy (i.e., (14b)) can equivalently be expressed as

$$\sum_{u\in\mathcal{U}} \mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) + N_r\sigma^2_{\mathrm{q},r} \le P_r^{\mathrm{M}}, \qquad \forall\, r\in\mathcal{R}^{\mathrm{C}}, \tag{15}$$

where $\mathbf{B}_r \in \mathbb{R}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$ denotes a block diagonal matrix with identity matrix $\mathbf{I}_{N_r}$ as the r-th main diagonal block matrix and zeros elsewhere.

Similarly, the maximum transmit power constraints of the RRHs using the DS strategy and being in the sleep mode (i.e., (14c) and (14d)) are, respectively, given by

$$\sum_{u\in\mathcal{U}} \mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) \le P_r^{\mathrm{M}}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}, \tag{16}$$

$$\sum_{u\in\mathcal{U}} \mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) = 0, \qquad \forall\, r\in\mathcal{R}^{\mathrm{S}}. \tag{17}$$

By defining $\mathbf{B}_{\mathrm{c},r} \in \mathbb{R}^{N_{\mathrm{T}}\times N_r}$ as the matrix composed of the columns from $\sum_{k=1}^{r-1} N_k + 1$ to $\sum_{k=1}^{r} N_k$ of matrix $\mathbf{B}_r$, we have $\mathbf{w}_{ru}\mathbf{w}_{ru}^{\mathrm{H}} = \mathbf{B}_{\mathrm{c},r}^{\mathrm{H}}\mathbf{W}_u\mathbf{B}_{\mathrm{c},r}$. The capacity constraints of the fronthaul links with the CAP strategy (i.e., (14e)) can be expressed as

$$B\log_2\det\left(\sum_{u\in\mathcal{U}} \mathbf{B}_{\mathrm{c},r}^{\mathrm{H}}\mathbf{W}_u\mathbf{B}_{\mathrm{c},r} + \left(\sigma^2_{\mathrm{q},r} + \epsilon\right)\mathbf{I}_{N_r}\right) - N_r B\log_2\left(\sigma^2_{\mathrm{q},r} + \epsilon\right) \le C_r^{\mathrm{M}}, \quad \forall\, r\in\mathcal{R}^{\mathrm{C}}, \tag{18}$$

where $\epsilon > 0$ is a small fixed regularization parameter.

By defining $\mathbf{\Omega}_r = \sum_{u\in\mathcal{U}} \mathbf{B}_{\mathrm{c},r}^{\mathrm{H}}\mathbf{W}_u\mathbf{B}_{\mathrm{c},r} + \left(\sigma^2_{\mathrm{q},r} + \epsilon\right)\mathbf{I}_{N_r}$, the non-convex term in constraint (18) can be linearized by using SCP [36]. Hence, the non-convex fronthaul capacity constraint can be tackled in an iterative manner. In the $(m+1)$-th iteration ($m = 0, 1, 2, \ldots$), constraint (18) can be rewritten as

$$\log_2\det\left(\mathbf{\Omega}_r^{(m+1)}\right) + \frac{1}{\ln 2}\mathrm{Tr}\left(\left(\mathbf{\Omega}_r^{(m+1)}\right)^{-1}\left(\mathbf{\Omega}_r - \mathbf{\Omega}_r^{(m+1)}\right)\right) - N_r\log_2\left(\sigma^2_{\mathrm{q},r} + \epsilon\right) \le \frac{C_r^{\mathrm{M}}}{B}, \quad \forall\, r\in\mathcal{R}^{\mathrm{C}}, \tag{19}$$

where

$$\mathbf{\Omega}_r^{(m+1)} = \sum_{u\in\mathcal{U}} \mathbf{B}_{\mathrm{c},r}^{\mathrm{H}}\mathbf{W}_u^{(m)}\mathbf{B}_{\mathrm{c},r} + \left(\sigma^{2(m)}_{\mathrm{q},r} + \epsilon\right)\mathbf{I}_{N_r}, \tag{20}$$

and $\mathbf{W}_u^{(m)}$ and $\sigma^{2(m)}_{\mathrm{q},r}$ are obtained from the m-th iteration.

Since $\|\mathbf{w}_{ru}\|_2^2 = \mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right)$, the capacity constraints of the fronthaul links supporting the DS strategy (i.e., (14f)) can be written as

$$\sum_{u\in\mathcal{U}} \mathbb{1}_{\{\mathrm{Tr}(\mathbf{B}_r\mathbf{W}_u)\}} \log_2(1+\gamma_u) \le \frac{C_r^{\mathrm{M}}}{B}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}. \tag{21}$$

The indicator function in constraint (21) can equivalently be expressed as an ℓ_0-norm of a scalar, which indicates whether or not this scalar is equal to zero. Thereby, constraint (21) can be written as

$$\sum_{u\in\mathcal{U}} \left\|\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right)\right\|_0 \log_2(1+\gamma_u) \le \frac{C_r^{\mathrm{M}}}{B}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}. \tag{22}$$

Such a non-convex ℓ_0-norm can be approximated by a convex reweighted ℓ_1-norm, which is widely used in compressive sensing [37]. Similar to (19), in the $(m+1)$-th iteration, constraint (22) can be rewritten as

$$\sum_{u\in\mathcal{U}} \beta_{ru}^{(m+1)}\, \mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right)\log(1+\gamma_u) \le \frac{C_r^{\mathrm{M}}}{B}, \quad \forall\, r\in\mathcal{R}^{\mathrm{D}}, \tag{23}$$

where $\beta_{ru}^{(m+1)}$ can be iteratively updated according to

$$\beta_{ru}^{(m+1)} = \frac{1}{\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u^{(m)}\right) + c_1} \tag{24}$$

and $c_1 > 0$ is a constant regularization factor.

To achieve the target SINR, the QoS constraint of UE u can be rewritten as

$$\frac{\mathbf{h}_u^{\mathrm{H}}\mathbf{W}_u\mathbf{h}_u}{\mathbf{h}_u^{\mathrm{H}}\left(\sum_{k\in\mathcal{U}\setminus\{u\}}\mathbf{W}_k\right)\mathbf{h}_u + \mathbf{h}_u^{\mathrm{H}}\mathbf{\Lambda}_{\mathrm{q}}\mathbf{h}_u + \sigma_{\mathrm{n},u}^2} \ge \gamma_u, \quad \forall\, u\in\mathcal{U}, \tag{25}$$

where $\mathbf{h}_u = [\mathbf{h}_{1u}^{\mathrm{H}},\ldots,\mathbf{h}_{Ru}^{\mathrm{H}}]^{\mathrm{H}} \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$, and $\mathbf{\Lambda}_{\mathrm{q}} \in \mathbb{R}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$ is a block diagonal matrix with identity matrix $\sigma_{\mathrm{q},r}^2\mathbf{I}_{N_r}$ as the r-th main diagonal block square matrix. Note that $\sigma_{\mathrm{q},r}^2 = 0$ for RRH $r\notin\mathcal{R}^{\mathrm{C}}$.

Based on the above transformation, problem (14) can be tackled by iteratively solving the following problem,

$$\begin{aligned}
\mathcal{P}^{(m+1)}: \underset{\substack{\mathcal{R}^{\mathrm{C}},\mathcal{R}^{\mathrm{D}},\mathcal{R}^{\mathrm{S}}\\ \{\mathbf{W}_u\},\{\sigma_{\mathrm{q},r}^2\}}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{D}}}\sum_{u\in\mathcal{U}}\frac{1}{\eta_r}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) \\
& + \sum_{r\in\mathcal{R}^{\mathrm{C}}}\left(\frac{1}{\eta_r}N_r\sigma_{\mathrm{q},r}^2 + P_{r,\mathrm{C}}^{\mathrm{dif}}\right) \\
& + \sum_{r\in\mathcal{R}^{\mathrm{D}}} P_{r,\mathrm{D}}^{\mathrm{dif}} \qquad (26\mathrm{a}) \\
\text{subject to} \quad & \text{constraints (14h)–(14j), (15)–(17),} \\
& \text{(19), (23), (25),} \\
& \mathrm{rank}\left(\mathbf{W}_u\right) = 1, \quad \forall\, u\in\mathcal{U}, \quad (26\mathrm{b}) \\
& \mathbf{W}_u \succeq 0, \quad \forall\, u\in\mathcal{U}. \qquad (26\mathrm{c})
\end{aligned}$$

Problem $\mathcal{P}^{(m+1)}$ still cannot directly be solved due to the combinatorial objective function (26a) and the non-convex rank-one constraint (26b). Given RRH sets $\mathcal{R}^{\mathrm{C}}$, $\mathcal{R}^{\mathrm{D}}$, and $\mathcal{R}^{\mathrm{S}}$, problem $\mathcal{P}^{(m+1)}$ is a rank-constrained SDP problem. By dropping the rank-one constraint [38], the convex relaxation problem can be efficiently solved by using the interior-point method [39]. Finally, the aggregate power minimization problem in (14) can be solved by developing an MM algorithm to iteratively update parameters $\{\Omega_r^{(m)}\}$ and $\{\beta_{ru}^{(m)}\}$ according to (20) and (24) by solving (26).

IV. Group Sparse Precoding Algorithm

In this section, we develop an efficient algorithm to tackle the combinatorial challenge based on the group sparse precoding approach and mitigate the non-convex rank-one constraint. The proposed algorithm is composed of two stages, as discussed in the following two sub-sections.

A. Stage One: Identify Active RRHs

In the first stage, we identify the RRHs that are required to be active to meet UEs' QoS requirement. Suppose all active RRHs are initially configured to support the CAP strategy (i.e., $\mathcal{R}^{\mathrm{D}} = \emptyset$), problem $\mathcal{P}^{(m+1)}$ can be simplified as

$$\begin{aligned}
\underset{\substack{\mathcal{R}^{\mathrm{C}},\mathcal{R}^{\mathrm{S}}\\ \{\sigma_{\mathrm{q},r}^2\},\{\mathbf{W}_u\}}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}^{\mathrm{C}}}\frac{1}{\eta_r}\left(\sum_{u\in\mathcal{U}}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) + N_r\sigma_{\mathrm{q},r}^2\right) \\
& + \sum_{r\in\mathcal{R}^{\mathrm{C}}} P_{r,\mathrm{C}}^{\mathrm{dif}} \\
\text{subject to} \quad & \text{constraints (15), (17), (19), (25),} \\
& \text{(26b), (26c),} \\
& \mathcal{R}^{\mathrm{C}}\cap\mathcal{R}^{\mathrm{S}} = \emptyset, \\
& \mathcal{R}^{\mathrm{C}}\cup\mathcal{R}^{\mathrm{S}} = \mathcal{R}.
\end{aligned} \tag{27}$$

When RRH r is switched off, all coefficients of precoding vector $\widetilde{\mathbf{w}}_r = [\mathbf{w}_{r1}^{\mathrm{H}},\ldots,\mathbf{w}_{rU}^{\mathrm{H}}]^{\mathrm{H}}$ should be set to 0, yielding $\|\widetilde{\mathbf{w}}_r\|_2^2 = \sum_{u\in\mathcal{U}}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) = 0$ and a group-sparsity structure of precoding vector $\mathbf{w} = [\widetilde{\mathbf{w}}_1^{\mathrm{H}},\ldots,\widetilde{\mathbf{w}}_R^{\mathrm{H}}]^{\mathrm{H}}$. As a result, problem (27) can be expressed as

$$\begin{aligned}
\underset{\{\mathbf{W}_u\},\{\sigma_{\mathrm{q},r}^2\}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}}\frac{1}{\eta_r}\left(\sum_{u\in\mathcal{U}}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) + N_r\sigma_{\mathrm{q},r}^2\right) \\
& + \sum_{r\in\mathcal{R}}\mathbb{1}_{\left\{\sum_{u\in\mathcal{U}}\mathrm{Tr}(\mathbf{B}_r\mathbf{W}_u)+N_r\sigma_{\mathrm{q},r}^2\right\}} P_{r,\mathrm{C}}^{\mathrm{dif}} \qquad (28) \\
\text{subject to} \quad & \text{constraints (15), (19), (25), (26b), (26c),}
\end{aligned}$$

where $\forall\, r\in\mathcal{R}^{\mathrm{C}}$ in constraints (15) and (19) is replaced by $\forall\, r\in\mathcal{R}$. Problem (28) is non-convex due to the indicator function in the objective function. An indicator function is equivalent to the ℓ_0-norm of a scalar, which can further be approximated by a convex reweighted ℓ_1-norm. Thus, we have

$$\begin{aligned}
& \mathbb{1}_{\left\{\sum\limits_{u\in\mathcal{U}}\mathrm{Tr}(\mathbf{B}_r\mathbf{W}_u)+N_r\sigma_{\mathrm{q},r}^2\right\}} \\
& \approx \mu_r^{(m+1)}\left(\sum_{u\in\mathcal{U}}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u\right) + N_r\sigma_{\mathrm{q},r}^2\right),
\end{aligned}$$

where $\mu_r^{(m+1)}$ can be iteratively updated according to

$$\mu_r^{(m+1)} = \frac{1}{\sum_{u\in\mathcal{U}}\mathrm{Tr}\left(\mathbf{B}_r\mathbf{W}_u^{(m)}\right) + N_r\sigma_{\mathrm{q},r}^{2(m)} + c_2}, \tag{29}$$

and $c_2 > 0$ is a constant regularization factor.

Algorithm 1: An Algorithm for Identifying Active RRHs.

```
Initialize variables $\{\mathbf{W}_u^{(0)}\}$ and $\{\sigma_{\mathbf{q},r}^{2(0)}\}$ satisfying
    constraints (15) for all $r \in \mathcal{R}$ and (25), and calculate
    $P_{\text{Alg1}}^{(0)}$.
$m := 0$ and $\Delta_1^{(m)} := 10^{10}$.
while $\Delta_1^{(m)} > \delta_1$ and $m < \phi_1$ do
    Update parameters $\{\mathbf{\Omega}_r^{(m+1)}\}$ and $\{\mu_r^{(m+1)}\}$
        according to (20) and (29), respectively.
    Solve problem (30) with parameters $\{\mathbf{\Omega}_r^{(m+1)}\}$ and
        $\{\mu_r^{(m+1)}\}$, and obtain $\{\mathbf{W}_u^{(m+1)}\}$, $\{\sigma_{\text{q},r}^{2(m+1)}\}$,
        and $P_{\text{Alg1}}^{(m+1)}$.
    $\Delta_1^{(m+1)} := \left|P_{\text{Alg1}}^{(m+1)} - P_{\text{Alg1}}^{(m)}\right|$.
    $m := m + 1$.
$\widetilde{\mathcal{R}}^{\text{S}} := \{r \mid \left(\sum_{u\in\mathcal{U}} \text{Tr}\left(\mathbf{B}_r \mathbf{W}_u^{(m)}\right) + N_r \sigma_{\text{q},r}^{2(m)}\right) < \varphi\}$.
$\widetilde{\mathcal{R}}^{\text{C}} := \mathcal{R} \setminus \widetilde{\mathcal{R}}^{\text{S}}$.
```

Through convexifying the indicator function in the objective function, we need to solve the following optimization problem,

$$\begin{aligned} \underset{\{\mathbf{W}_u\},\{\sigma_{\text{q},r}^2\}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}} \left(\frac{1}{\eta_r} + \mu_r^{(m+1)} P_{r,\text{C}}^{\text{dif}}\right) \\ & \times \left(\sum_{u\in\mathcal{U}} \text{Tr}\left(\mathbf{B}_r \mathbf{W}_u\right) + N_r \sigma_{\text{q},r}^2\right), \qquad (30) \\ \text{subject to} \quad & \text{constraints (15), (19), (25), (26b), (26c),} \end{aligned}$$

where $\forall\, r \in \mathcal{R}^{\text{C}}$ in constraints (15) and (19) is replaced by $\forall\, r \in \mathcal{R}$. After dropping rank-one constraint (26b), problem (30) is an SDP problem, which can be efficiently solved by convex programming solver (e.g., CVX [40]). We show the tightness of the rank-one constraint relaxation as follows.

Theorem 1: Let $\mathbf{W}_u^\star$ denote the precoding matrix of UE $u \in \mathcal{U}$ as the solution of problem (30) without rank-one constraint (26b), then $\text{rank}\left(\mathbf{W}_u^\star\right) = 1$ always holds.

Proof: Please refer to Appendix A. ■

The MM algorithm [41] can be used to solve a sequence of convex optimization problems (i.e., problem (30) without the rank-one constraint) in an iterative manner. We denote $P_{\text{Alg1}}^{(m+1)}$ as the value of the objective function of problem (30) in the $(m+1)$-th iteration. The convergence threshold and the maximum number of iterations are denoted as δ_1 and ϕ_1, respectively. The proposed algorithm based on the MM scheme to identify the active RRHs is summarized in Algorithm 1. It is shown in [41] that the MM algorithm always converges to a stationary point of the original problem. After solving problem (27) by using Algorithm 1, we can obtain the set of RRHs in sleep mode as $\widetilde{\mathcal{R}}^{\text{S}} = \{r \mid \left(\sum_{u\in\mathcal{U}} \text{Tr}\left(\mathbf{B}_r \mathbf{W}_u\right) + N_r \sigma_{\text{q},r}^2\right) < \varphi\}$, and set of RRHs using the CAP strategy as $\widetilde{\mathcal{R}}^{\text{C}} = \mathcal{R} \setminus \widetilde{\mathcal{R}}^{\text{S}}$, where φ is a predefined small constant. Besides, we obtain the converged objective value of problem (27) denoted by $P_{\text{agg}}^{\text{C}}$ and quantization noises for active RRHs given by $\{\widetilde{\sigma}_{\text{q},r}^2, r \in \widetilde{\mathcal{R}}^{\text{C}}\}$.

B. Stage Two: Identify Cooperative Strategies and Optimize Precoding Matrices and Quantization Noise

In the second stage, we determine the set of active RRHs switching to support the DS strategy, and optimize the precoding matrices and quantization noise, to further reduce the power consumption. We utilize the following ordering criterion to determine the priorities of RRHs using the CAP strategy to be switched to support the DS strategy,

$$\theta_r = \frac{1}{\eta_r} N_r \widetilde{\sigma}_{\text{q},r}^2, \quad \forall\, r \in \widetilde{\mathcal{R}}^{\text{C}}. \qquad (31)$$

The RRH with a larger θ_r has a higher priority to support the DS strategy. In particular, the RRHs with more transmit antennas, smaller drain efficiency, and larger quantization noise are likely to consume more power and generate higher interference according to (8) and (10). We denote the number of active RRHs (i.e., cardinality of $\widetilde{\mathcal{R}}^{\text{C}}$) as α. Based on the ordering criterion (31), we order the RRHs in a descending order, i.e., $\theta_{\pi_1} \geq \theta_{\pi_2} \geq \cdots \geq \theta_{\pi_\alpha}$, to determine the set of active RRHs using the DS strategy. For simplicity, we iteratively select the active RRHs to support the DS strategy. Thus, we introduce another iteration which is outside the iterations used to update $\mathbf{\Omega}_r^{(m+1)}$ and $\beta_{ru}^{(m+1)}$. The RRH sets supporting the DS and CAP strategies in the τ-th outer iteration are denoted as $\widetilde{\mathcal{R}}^{\text{D}(\tau)} = \{\pi_1, \pi_2, \ldots, \pi_\tau\}$ and $\widetilde{\mathcal{R}}^{\text{C}(\tau)} = \{\pi_{\tau+1}, \pi_{\tau+2}, \ldots, \pi_\alpha\}$, respectively. Based on the above definitions, we have $\widetilde{\mathcal{R}}^{\text{D}(\tau)} \cup \widetilde{\mathcal{R}}^{\text{C}(\tau)} = \widetilde{\mathcal{R}}^{\text{C}}$. Given RRH sets $\widetilde{\mathcal{R}}^{\text{C}(\tau)}$, $\widetilde{\mathcal{R}}^{\text{D}(\tau)}$, and $\widetilde{\mathcal{R}}^{\text{S}}$, the RRH circuit and fronthaul power consumption is fixed. Hence, $\sum_{r\in\widetilde{\mathcal{R}}^{\text{C}(\tau)}} P_{r,\text{C}}^{\text{dif}} + \sum_{r\in\widetilde{\mathcal{R}}^{\text{D}(\tau)}} P_{r,\text{D}}^{\text{dif}}$ is a constant and can be omitted in the objective function. As a result, we can solve the following problem in the $(m+1)$-th inner iteration to reduce the aggregate power consumption,

$$\begin{aligned} \underset{\{\mathbf{W}_u\},\{\sigma_{\text{q},r}^2\}}{\text{minimize}} \quad & \sum_{r\in\widetilde{\mathcal{R}}^{\text{C}(\tau)}} \frac{1}{\eta_r} \left(\sum_{u\in\mathcal{U}} \text{Tr}\left(\mathbf{B}_r \mathbf{W}_u\right) + N_r \sigma_{\text{q},r}^2\right) \\ & + \sum_{r\in\widetilde{\mathcal{R}}^{\text{D}(\tau)}} \frac{1}{\eta_r} \sum_{u\in\mathcal{U}} \text{Tr}\left(\mathbf{B}_r \mathbf{W}_u\right) \\ \text{subject to} \quad & \text{constraints (15)-(17), (19), (23),} \\ & \text{(25), (26b), (26c),} \qquad (32) \end{aligned}$$

where $\forall\, r \in \mathcal{R}^{\text{C}}$ and $\forall\, r \in \mathcal{R}^{\text{D}}$ in all constraints are replaced by $\forall\, r \in \widetilde{\mathcal{R}}^{\text{C}(\tau)}$ and $\forall\, r \in \widetilde{\mathcal{R}}^{\text{D}(\tau)}$, respectively. Similarly, problem (32) without rank-one constraint (26b) is an SDP problem and can efficiently be solved. The tightness of the rank-one constraint relaxation is shown in the following theorem.

Theorem 2: Let $\mathbf{W}_u^\star$ denote the precoding matrix of UE $u \in \mathcal{U}$ as the solution of problem (32) without rank-one constraint (26b), then $\text{rank}\left(\mathbf{W}_u^\star\right) = 1$ always holds.

Proof: Please refer to Appendix B ■

We denote $P_{\text{Alg2}}^{(m+1)}$ as the value of the objective function of problem (32) in the $(m+1)$-th inner iteration. The convergence threshold and the maximum number of iterations are denoted as δ_2 and ϕ_2, respectively. The proposed algorithm to solve problem (32) is summarized in Algorithm 2. We denote $P_{\text{agg}}^{(\tau)}$ as the converged objective value of problem (32) for the τ-th

Algorithm 2: An Algorithm for Optimizing Precoding Matrices and Quantization Noise.

```
Initialize variables $\{\mathbf{W}_u^{(0)}\}$ and $\{\sigma_{\mathrm{q},r}^{2(0)}\}$ satisfying
  constraints (15) for all $r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}$, (16) for all
  $r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}$, and (25), and calculate $P_{\mathrm{Alg2}}^{(0)}$. $m := 0$,
  $\Delta_2^{(m)} := 10^{10}$.
while $\Delta_2^{(m)} > \delta_2$ and $m < \phi_2$ do
    Update parameters $\{\mathbf{\Omega}_r^{(m+1)}\}$ and $\{\beta_{ru}^{(m+1)}\}$
      according to (20) and (24), respectively.
    Solve problem (32) with parameters $\{\mathbf{\Omega}_r^{(m+1)}\}$ and
      $\{\beta_{ru}^{(m+1)}\}$, and obtain $\{\mathbf{W}_u^{(m+1)}\}$, $\{\sigma_{\mathrm{q},r}^{2(m+1)}\}$,
      and $P_{\mathrm{Alg2}}^{(m+1)}$.
    $\Delta_2^{(m+1)} := \left|P_{\mathrm{Alg2}}^{(m+1)} - P_{\mathrm{Alg2}}^{(m)}\right|$.
    $m := m + 1$.
```

outer iteration. Finally, combining the above two stages, the algorithm for solving the aggregate power consumption minimization problem (14) is given in Algorithm 3. The set of RRHs using the DS strategy, $\widetilde{\mathcal{R}}^{\mathrm{D}(0)}$, and the iteration index, τ, are initialized in Step 1. By using Algorithm 1, we solve problem (27) to check the feasibility and identify the set of RRHs required to be active (Step 2). If problem (27) is feasible, then we determine the set of RRHs using the CAP strategy, $\widetilde{\mathcal{R}}^{\mathrm{C}}$, the set of RRHs in the sleep mode, $\widetilde{\mathcal{R}}^{\mathrm{S}}$, the quantization noise, $\{\widetilde{\sigma}_{\mathrm{q},r}^2\}$, as well as the aggregate power consumption, $P_{\mathrm{agg}}^{\mathrm{C}}$, and then sort the ordering criterion (31) in a descending order (Steps 3–5). Otherwise, the algorithm terminates (Steps 6 and 7). We initialize $\widetilde{\mathcal{R}}^{\mathrm{C}(0)}$ in Step 8. In the τ-th iteration, we move one active RRH from set $\widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}$ to set $\widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}$ based on the ordering of active RRHs (Steps 10 and 11), and solve problem (32) using Algorithm 2 to obtain $P_{\mathrm{agg}}^{(\tau)}$ (Step 12). If the aggregate power consumption in the τ-th iteration is smaller than $P_{\mathrm{agg}}^{\mathrm{C}}$, then we update the values of $P_{\mathrm{agg}}^{\mathrm{C}}$ and τ (Steps 13 – 15). Otherwise, we break the loop (Steps 16 and 17). The loop stops when either $\tau > \alpha$ or $P_{\mathrm{agg}}^{(\tau)} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}} P_{r,\mathrm{C}}^{\mathrm{dif}} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}} P_{r,\mathrm{D}}^{\mathrm{dif}} \geq P_{\mathrm{agg}}^{\mathrm{C}}$. In Step 18, we determine sets $\widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}$ and $\widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}$, and recover precoding vectors $\{\widehat{\mathbf{w}}_u\}$ and quantization noise $\{\widehat{\sigma}_{\mathrm{q},r}^2\}$ to serve all UEs. Note that the final precoding vector $\widehat{\mathbf{w}}_u$ is the eigenvector of $\widehat{\mathbf{W}}_u, \forall\, u \in \mathcal{U}$. By using the iteratively reweighted method and the MM-based algorithm, the solution of the proposed algorithm is always a stationary point of the original problem [42].

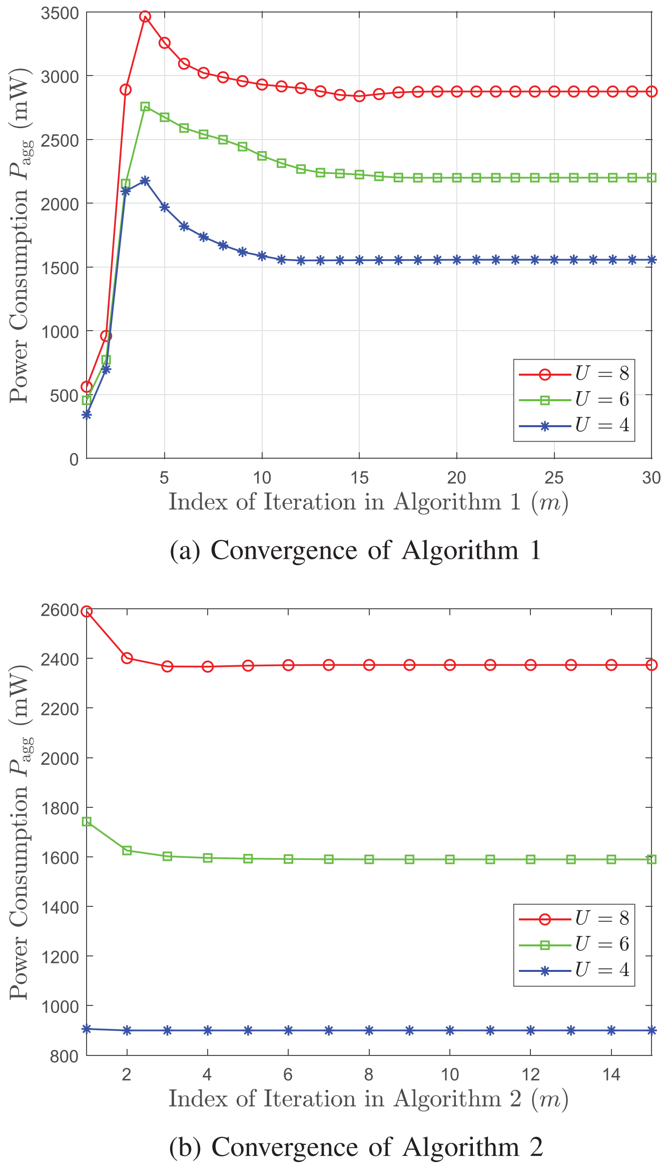
The overall algorithm (i.e., Algorithm 3) runs Algorithm 1 once and Algorithm 2 at most R times. Algorithms 1 and 2 solve a sequence of SDP problems, i.e., problems (30) and (32) without rank-one constraint, respectively. To solve the SDP problem with U matrix optimization variables of size $N_{\mathrm{T}} \times N_{\mathrm{T}}$, the interior-point method takes $\mathcal{O}(\sqrt{UN_{\mathrm{T}}}\log(1/\varepsilon))$ iterations and $\mathcal{O}(UN_{\mathrm{T}}^6)$ floating point operations to achieve an optimal solution with accuracy $\varepsilon > 0$. Note that the maximum number of SDP problems required to be solved for Algorithms 1 and 2 are ϕ_1 and ϕ_2, respectively. Hence, the overall computational complexity of the proposed algorithm is given by $\mathcal{O}((\phi_1 + R\phi_2)U^{1.5}N_{\mathrm{T}}^{6.5}\log(1/\varepsilon))$.

Algorithm 3: Aggregate Power Minimization Algorithm.

```
1  Initialize RRH set $\widetilde{\mathcal{R}}^{\mathrm{D}(0)} := \emptyset$ and $\tau := 1$.
2  Solve problem (27) using Algorithm 1.
3  if problem (27) is feasible then
4      Obtain $\widetilde{\mathcal{R}}^{\mathrm{C}}$, $\widetilde{\mathcal{R}}^{\mathrm{S}}$, $\{\widetilde{\sigma}_{\mathrm{q},r}^2\}$, $\alpha$, and $P_{\mathrm{agg}}^{\mathrm{C}}$.
5      Calculate the ordering criterion (31) and sort them
         in a descending order: $\theta_{\pi_1} \geq \theta_{\pi_2} \geq \cdots \geq \theta_{\pi_\alpha}$.
6  else
7      Go to End.
8  $\widetilde{\mathcal{R}}^{\mathrm{C}(0)} := \widetilde{\mathcal{R}}^{\mathrm{C}}$.
9  while $\alpha - \tau \geq 0$ do
10     $\widetilde{\mathcal{R}}^{\mathrm{C}(\tau)} := \widetilde{\mathcal{R}}^{\mathrm{C}(\tau-1)} \setminus \{\pi_\tau\}$.
11     $\widetilde{\mathcal{R}}^{\mathrm{D}(\tau)} := \widetilde{\mathcal{R}}^{\mathrm{D}(\tau-1)} \cup \{\pi_\tau\}$.
12     Solve problem (32) using Algorithm 2 and obtain
         $P_{\mathrm{agg}}^{(\tau)}$.
13     if $P_{\mathrm{agg}}^{(\tau)} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}} P_{r,\mathrm{C}}^{\mathrm{dif}} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}} P_{r,\mathrm{D}}^{\mathrm{dif}} < P_{\mathrm{agg}}^{\mathrm{C}}$
         then
14         $P_{\mathrm{agg}}^{\mathrm{C}} :=$
             $P_{\mathrm{agg}}^{(\tau)} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}} P_{r,\mathrm{C}}^{\mathrm{dif}} + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}} P_{r,\mathrm{D}}^{\mathrm{dif}}$.
15         $\tau := \tau + 1$.
16     else
17         Break.
18 Use RRHs given in sets $\widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}$ and $\widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}$, precoding
     vectors $\{\widehat{\mathbf{w}}_u\}$, and quantization noise $\{\widehat{\sigma}_{\mathrm{q},r}^2\}$ to serve
     all UEs.
19 End
```

V. Performance Evaluation

In this section, we evaluate the energy efficiency of the proposed flexible functional split design for downlink C-RAN and compare the aggregate power consumption with that of the pure CAP and DS strategies. Specifically, in the CAP and DS strategies, all active RRHs work in the CAP and DS modes, respectively. In the simulations, the RRHs and UEs are randomly distributed in a circular network coverage area with radius 500 m. We consider quasi-static Rayleigh fading channels and set the path loss exponent to be 4. The channel bandwidth B and noise power $\sigma_{\mathrm{n},u}^2$ are set to be 10 MHz and -100 dBm, respectively. The number of RRHs (i.e., R) in the network coverage area is 10. The maximum transmit power of the r-th RRH (i.e., $P_r^{\mathrm{M}}, \forall\, r \in \mathcal{R}$) is 80 mW. Each RRH using the DS strategy only needs to superimpose the received signals weighted by the corresponding precoding coefficients, which is a simple operation and consumes less power than the quantization codebook based signal decompression operation performed by each RRH using the CAP strategy. Hence, the power differences between the active and sleep modes for the CAP and DS strategies (i.e., $P_{r,\mathrm{C}}^{\mathrm{dif}}$ and $P_{r,\mathrm{D}}^{\mathrm{dif}}, \forall\, r \in \mathcal{R}$) are set to be 500 mW and 400 mW, respectively. The drain efficiency of the RF power amplifier of the



(a) Convergence of Algorithm 1

(b) Convergence of Algorithm 2

Fig. 3. Convergence of Algorithms 1 and 2 for different number of UEs in the network when $C_r^{\text{M}} = 80$ Mbps and $\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$.

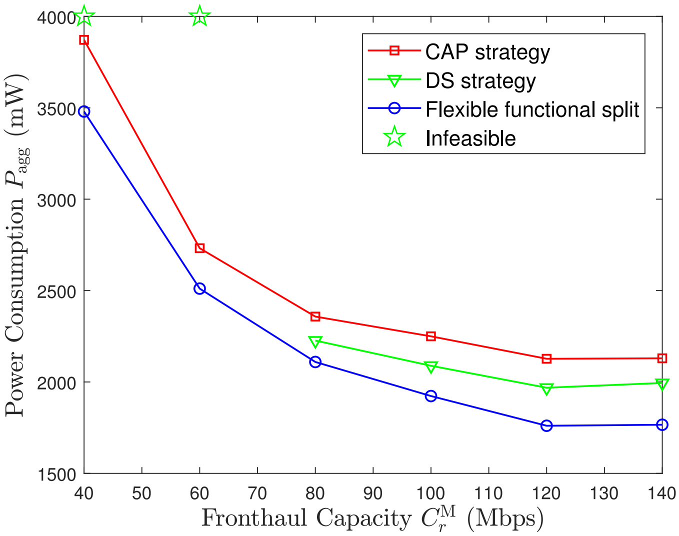


Fig. 4. Aggregate power consumption versus fronthaul capacity when $U = 8$ and $\kappa_u = 20$ Mbps, $\forall\, u \in \mathcal{U}$.

r-th RRH (i.e., $\eta_r, \forall\, r \in \mathcal{R}$) is 0.25. The constant regularization factors (i.e., ϵ, c_1, and c_2) are all set to be 10^{-5}. The convergence thresholds (i.e., δ_1 and δ_2) are set to 1, and the predefined small constant (i.e., φ) is set to 10^{-3}. The maximum number of iterations (i.e., ϕ_1 and ϕ_2) used in Algorithms 1 and 2 are set to be 30 and 15, respectively. Each RRH is equipped with two antennas and each UE is equipped with a single antenna. We denote the target data rate as $\kappa_u = B \log_2(1 + \gamma_u), \forall\, u \in \mathcal{U}$.

In Fig. 3, we first evaluate the convergence of the proposed algorithm for the flexible functional split design in downlink C-RAN with different number of UEs (i.e., U) when $C_r^{\text{M}} = 80$ Mbps and $\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$. According to Algorithm 3, the convergence of the proposed algorithm is guaranteed as long as Algorithms 1 and 2 converge. The maximum number of iterations for the loops in Algorithms 1 and 2 are set as 30 and 15, respectively. The objective values obtained by Algorithms 1 and 2 after each iteration are plotted in Fig. 3(a) and (b), respectively. As can be seen, Algorithm 1 converges after about 12 to 18 iterations, while Algorithm 2 converges after about 2 to 5 iterations. In particular, the larger the number of UEs in the network, the larger the number of iterations is required for the algorithm to converge. Overall, Algorithm 3 always converges after a small number of iterations.

In Fig. 4, we then investigate the impact of the limited fronthaul capacity on the aggregate power consumption when $U = 8$ and $\kappa_u = 20$ Mbps, $\forall\, u \in \mathcal{U}$. With the variation of the fronthaul capacity, the aggregate power consumption changes significantly, which demonstrates the importance of taking into account the limited fronthaul capacity. For the DS strategy, the fronthaul capacity constraint limits the number of cooperating RRHs for each UE, which in turn limits the achievable cooperation gain. Hence, in the low fronthaul capacity regime, the DS strategy is less likely to meet the QoS requirement of all UEs due to the limited cooperation gain. In particular, when $C_r^{\text{M}} = 40$ Mbps or 60 Mbps, the DS strategy is infeasible (i.e., the QoS requirement of all UEs cannot be simultaneously satisfied). Hence, the corresponding points are marked with stars, as shown in Fig. 4. On the other hand, the CAP strategy is feasible in the low fronthaul capacity regime. Hence, by transforming the advantage of generating low fronthaul data rates to the requirement of activating less RRHs, the CAP strategy outperforms the DS strategy when the fronthaul capacity is small. With the increase of C_r^{M} from 60 Mbps to 120 Mbps, the aggregate power consumption of all considered strategies decreases as less RRHs are required to be active to meet the QoS requirement of all UEs. When $C_r^{\text{M}} > 120$ Mbps, the aggregate power consumption cannot be further reduced by increasing the fronthaul capacity. When the fronthaul capacity is large enough to deliver multiple data streams, the DS strategy not only requires a similar number of active RRHs as that of the CAP strategy, but also consumes less transmit power (e.g., no quantization noise in the DS strategy) and processing power (e.g., no signal decompression is needed in the DS strategy). As a result, the DS strategy outperforms the CAP strategy when the fronthaul capacity is large. The proposed flexible functional split design

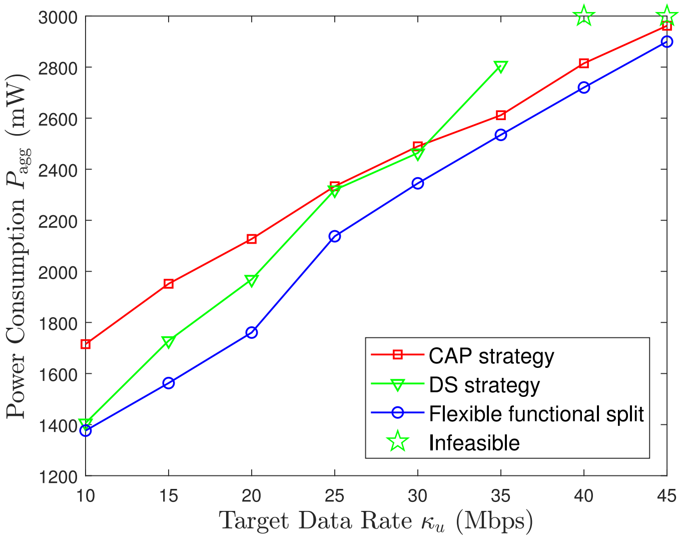


Fig. 5. Aggregate power consumption versus target data rate when $U = 8$ and $C_r^{\text{M}} = 120$ Mbps, $\forall\, r \in \mathcal{R}$.

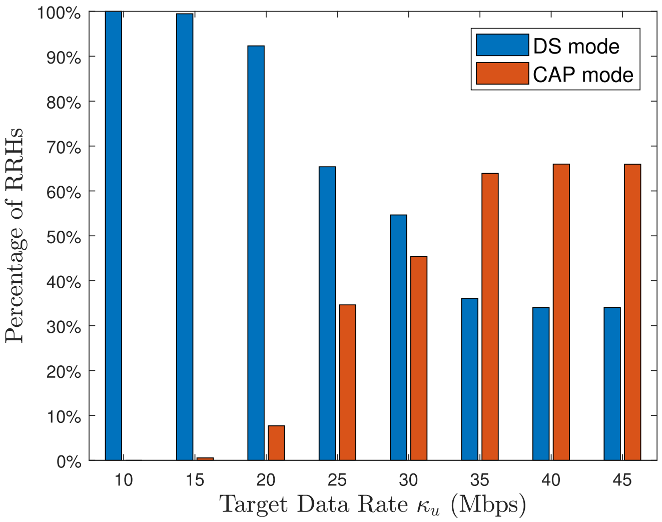


Fig. 6. Percentage of RRHs in the DS and CAP modes versus target data rate for the flexible functional split design when $U = 8$ and $C_r^{\text{M}} = 120$ Mbps, $\forall\, r \in \mathcal{R}$.

exploits the advantages of both the CAP and DS strategies, i.e., activating less RRHs and using a lower transmit power, respectively. Hence, the flexible functional split design always achieves a better performance than both the CAP and DS strategies for all values of the fronthaul capacity.

Fig. 5 shows the impact of the target data rate of UEs on the aggregate power consumption when $U = 8$ and $C_r^{\text{M}} = 120$ Mbps, $\forall\, r \in \mathcal{R}$. For a given number of UEs, the target data rates of UEs reflect the traffic load in the network. As we can see, the aggregate power consumption of all strategies under consideration increases with the target data rates of UEs. This is because supporting higher data rates requires higher fronthaul data rates, which in turn requires more active RRHs as each RRH is connected to a fronthaul link with limited capacity. Hence, in the low traffic load regime, maximizing the number of RRHs in the sleep mode is crucial in minimizing the aggregate power consumption. As can be seen, neither the DS nor CAP strategy dominates the other across the entire target data rate regime. For example, when the target data rate is less than 25 Mbps, the DS strategy achieves a better performance than the CAP strategy. On the other hand, when the target data rate is larger than 30 Mbps, the CAP strategy achieves a better performance than the DS strategy. This is because the fronthaul data rate of the DS strategy directly depends on the target data rate and the number of serving UEs, while the fronthaul data rate of the CAP strategy depends on the logarithm of the SINR and increases slowly with the target data rate. In the high traffic load regime (e.g., the target data rate is 40 Mbps or above), the sets of RRHs using the CAP and DS strategies are critical optimization variables. As shown in Fig. 5, the DS strategy becomes infeasible in this regime and the corresponding points are plotted with stars. By appropriately setting the transmission mode for each RRH, the flexible functional split design outperforms both the CAP and DS strategies in terms of the energy efficiency.

In Fig. 6, we compare the percentage of RRHs in the DS and CAP modes for the flexible functional split design with different target data rates of UEs when $U = 8$ and $C_r^{\text{M}} = 120$ Mbps, $\forall\, r \in \mathcal{R}$. As we can see, when the target data rate is low (i.e., less than 15 Mbps), almost all active RRHs are switched to the DS mode, as the fronthaul capacity is not the dominant performance-limiting factor in the low data rate regime. With the increase of the target data rate from 15 Mbps to 35 Mbps, the percentage of RRHs in the DS mode decreases, while the percentage of RRHs in the CAP mode increases, i.e., less active RRHs are switched to the DS mode. This is because the fronthaul link is not able to support the transmission of multiple data streams without compression. By further increasing the target data rate, the percentage of RRHs in the DS mode almost remains at about 34%. As we can see, in the moderate and high target data rate regimes, the proposed flexible functional split design adjusts the transmission modes of all RRHs according to their channel conditions so as to fully exploit the advantages of both DS and CAP modes.

Fig. 7 illustrates the impact of the number of UEs on the aggregate power consumption of all strategies under consideration when $C_r^{\text{M}} = 120$ Mbps and $\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$. With the increase of the number of UEs, the traffic load in the network increases, which imposes a higher requirement on the fronthaul capacity. As a result, more RRHs are required to be active to support the QoS requirement of all UEs in the network, leading to higher power consumption. When the number of UEs is small, the required fronthaul data rate of the DS strategy is smaller than the fronthaul capacity, and hence, the DS strategy outperforms the CAP strategy in terms of the energy efficiency. When the number of UEs is large, the DS strategy becomes infeasible, while the CAP strategy becomes more favourable by activating less RRHs. Overall, the proposed flexible functional split design adapts to the network traffic load and outperforms both the CAP and DS strategies for all values of the number of UEs.

Fig. 8 shows the impact of the number of UEs on the fraction of active RRHs in downlink C-RAN when $C_r^{\text{M}} = 120$ Mbps and

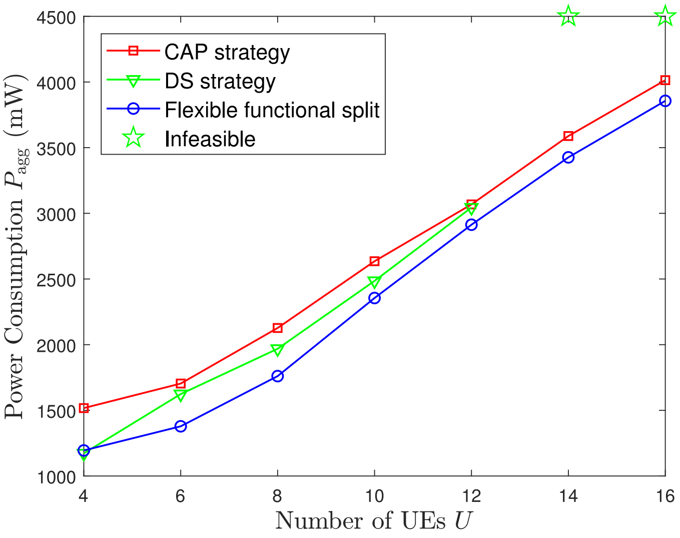


Fig. 7. Aggregate power consumption versus number of UEs when $C_r^{\mathrm{M}} = 120$ Mbps and $\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$.

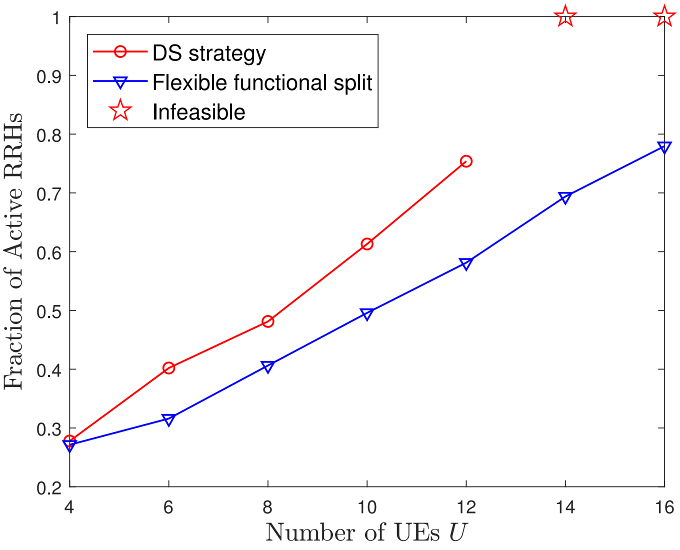


Fig. 8. Fraction of active RRHs versus number of UEs when $C_r^{\mathrm{M}} = 120$ Mbps and $\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$.

$\kappa_u = 20$ Mbps, $\forall\, r \in \mathcal{R}, u \in \mathcal{U}$. Similar to the trends observed in Fig. 7, the fraction of the active RRHs increases with the number of UEs. As we can see, the fraction of the active RRHs of the proposed flexible functional split design is always smaller than that of the DS strategy due to its better utilization of the fronthaul capacity. When the number of UEs reaches 14 or more, the DS strategy becomes infeasible, while the proposed flexible functional split design can still guarantee the QoS requirement of all UEs. Moreover, the gap in terms of the active RRHs between the DS strategy and the flexible functional split design also increases with the number of UEs.

VI. Conclusion

In this paper, we proposed a flexible functional split between the BBU pool and the RRHs in downlink C-RAN with limited fronthaul capacity. We formulated a joint RRH mode (i.e., CAP, DS, sleep) selection, precoding design, and fronthaul compression problem to minimize the aggregate power consumption. We took into account both the fronthaul capacity constraint and fronthaul power consumption, and tackled the non-convex fronthaul capacity constraints by using the SCP and ℓ_1-norm convex relaxation techniques. We transformed the non-convex optimization problem into a sequence of rank-constrained SDP problems. An iterative algorithm based on group sparse precoding approach and MM scheme was proposed to solve the problem. Simulation results showed that the fronthaul capacity constraint has a significant impact on aggregate power consumption and the proposed flexible functional split design outperforms both the pure CAP and DS strategies in terms of aggregate power consumption. For future work, we will consider the uncertainty of radio channels and the millimeter wave-based fronthaul, and investigate their impact on the energy efficiency of C-RAN.

Appendix

A. *Proof of Theorem 1*

For notational simplicity, we denote $g_r = \mu_r^{(m+1)} P_r^{\mathrm{dir}} + \frac{1}{\eta_r}, \forall\, r \in \mathcal{R}$. In addition, we denote $\zeta_r, \lambda_r, \nu_u \geq 0$, and Hermitian matrix $\mathbf{X}_u \succeq \mathbf{0}$ as the Lagrangian multipliers of constraints (15), (19) for all $r \in \mathcal{R}$, (25), and (26c), respectively. Hence, the Lagrangian of problem (30) is given by

$$\begin{aligned}&\mathcal{L}_1(\{\mathbf{W}_u\}, \{\sigma_{\mathrm{q},r}^2\}, \{\zeta_r\}, \{\lambda_r\}, \{\nu_u\}, \{\mathbf{X}_u\})\\ &= \sum_{u \in \mathcal{U}} \mathrm{Tr}\left(\mathbf{W}_u \left(\sum_{r \in \mathcal{R}} \left(g_r \mathbf{B}_r + \zeta_r \mathbf{B}_r + \lambda_r \mathbf{\Xi}_r\right)\right.\right.\\ &\quad \left.\left. + \sum_{k \in \mathcal{U} \setminus \{u\}} \nu_k \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} - \nu_u \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}} - \mathbf{X}_u\right)\right) + \Gamma_1,\end{aligned}$$

where $\mathbf{\Xi}_r = \frac{1}{\ln 2} \mathbf{B}_{\mathrm{c},r} (\mathbf{\Omega}_r^{(m+1)})^{-1} \mathbf{B}_{\mathrm{c},r}^{\mathrm{H}}$, Γ_1 depends on $\{\sigma_{\mathrm{q},r}^2\}$, $\{\zeta_r\}$, $\{\lambda_r\}$, $\{\nu_u\}$, and other constant parameters in problem (30). The dual problem of problem (30) is given by

$$\underset{\{\zeta_r\},\{\lambda_r\},\{\nu_u\},\{\mathbf{X}_u\}}{\text{maximize}} \quad \inf_{\{\mathbf{W}_u\},\{\sigma_{\mathrm{q},r}^2\}} \quad \mathcal{L}_1. \tag{33}$$

We denote $\mathbf{\Phi}^\star = \left(\{\mathbf{W}_u^\star\}, \{\sigma_{\mathrm{q},r}^{2\star}\}\right)$ and $\mathbf{\Psi}^\star = (\{\zeta_r^\star\}, \{\lambda_r^\star\}, \{\nu_u^\star\}, \{\mathbf{X}_u^\star\})$ as the solutions of primal and dual problems, respectively. Hence, the Karush-Kuhn-Tucker (KKT) conditions can be written as

$$\nabla_{\mathbf{W}_u} \mathcal{L}_1 \big|_{\mathbf{\Phi}^\star, \mathbf{\Psi}^\star} = \mathbf{0}, \qquad \forall\, u \in \mathcal{U}, \tag{34a}$$

$$\mathbf{X}_u^\star \mathbf{W}_u^\star = \mathbf{0}, \qquad \forall\, u \in \mathcal{U}, \tag{34b}$$

$$\zeta_r^\star \geq 0, \lambda_r^\star \geq 0, \nu_r^\star \geq 0, \qquad \forall\, r \in \mathcal{R}, \tag{34c}$$

where $\nabla_{\mathbf{W}_u} \mathcal{L}_1 \big|_{\mathbf{\Phi}^\star, \mathbf{\Psi}^\star}$ denotes the gradient of the Lagrangian in (33) with respect to $\mathbf{W}_u$ at $\mathbf{\Phi}^\star$ and $\mathbf{\Psi}^\star$. According to (34a), for UE $u \in \mathcal{U}$, we have

$$\begin{aligned}&\sum_{r \in \mathcal{R}} \left(g_r \mathbf{B}_r + \zeta_r^\star \mathbf{B}_r + \lambda_r^\star \mathbf{\Xi}_r\right)\\ &+ \sum_{k \in \mathcal{U} \setminus \{u\}} \nu_k^\star \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} - \nu_u^\star \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}} - \mathbf{X}_u^\star = \mathbf{0}, \quad \forall\, u \in \mathcal{U}.\end{aligned} \tag{35}$$

Based on (34b) and (35), for UE $u \in \mathcal{U}$, we have

$$\mathbf{W}_u^\star \left(\sum_{r \in \mathcal{R}} \left(g_r \mathbf{B}_r + \zeta_r^\star \mathbf{B}_r + \lambda_r^\star \mathbf{\Xi}_r \right) + \sum_{k \in \mathcal{U} \backslash \{u\}} \nu_k^\star \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} - \nu_u^\star \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}} \right) = \mathbf{W}_u^\star \mathbf{X}_u^\star = \mathbf{0}. \tag{36}$$

Hence, we have,

$$\mathbf{W}_u^\star \left(\mathbf{Y}_u^\star - \mathbf{Z}_u^\star \right) = \mathbf{0} \Leftrightarrow \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{Y}_u^\star \right) = \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{Z}_u^\star \right), \tag{37}$$

where $\mathbf{Y}_u^\star = \sum_{r \in \mathcal{R}} (g_r \mathbf{B}_r + \zeta_r^\star \mathbf{B}_r + \lambda_r^\star \mathbf{\Xi}_r) + \sum_{k \in \mathcal{U} \backslash \{u\}} (\nu_k^\star \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}})$ and $\mathbf{Z}_u^\star = \nu_u^\star \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}}$. By taking into account $g_r > 0$, constraint (34c), and the definition of $\mathbf{B}_r$, we have $\mathbf{Y}_u^\star \succ \mathbf{0}$, and thus $\operatorname{rank}(\mathbf{Y}_u^\star) = N_{\mathrm{T}}, \forall\, u \in \mathcal{U}$. As a result, we have

$$\begin{aligned} \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{Y}_u^\star \right) &= \operatorname{rank}\left(\mathbf{W}_u^\star \right) = \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{Z}_u^\star \right) \\ &\leq \operatorname{rank}\left(\mathbf{Z}_u^\star \right) = 1, \quad \forall\, u \in \mathcal{U}. \end{aligned}$$

As $\mathbf{W}_u = \mathbf{0}$ cannot be the solution of problem (30) due to UEs' QoS requirement, we conclude that solving problem (30) without the rank-one constraint always achieves $\operatorname{rank}(\mathbf{W}_u^\star) = 1, \forall\, u \in \mathcal{U}$. Hence, the proof of Theorem 1 is complete.

B. Proof of Theorem 2

The Lagrangian of problem (32) without constraint (26b) can be written as

$$\begin{aligned} &\mathcal{L}_2\big(\{\mathbf{W}_u\}, \{\sigma_{\mathrm{q},r}^2\}, \{\zeta_r^{\mathrm{C}}\}, \{\zeta_r^{\mathrm{D}}\}, \{\zeta_r^{\mathrm{S}}\}, \{\lambda_r^{\mathrm{C}}\}, \{\lambda_r^{\mathrm{D}}\}, \\ &\quad \{\nu_u\}, \{\mathbf{X}_u\} \big) \\ &= \sum_{u \in \mathcal{U}} \mathrm{Tr} \Bigg(\mathbf{W}_u \Bigg(\sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}} \left(\frac{\mathbf{B}_r}{\eta_k} + \zeta_r^{\mathrm{C}} \mathbf{B}_r + \lambda_r^{\mathrm{C}} \mathbf{\Xi}_r \right) \\ &\quad + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}} \left(\frac{\mathbf{B}_r}{\eta_k} + \zeta_r^{\mathrm{D}} \mathbf{B}_r + \lambda_r^{\mathrm{D}} \beta_{ru}^{(m+1)} \log_2 (1 + \gamma_u) \right) \\ &\quad + \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{S}}} \zeta_r^{\mathrm{S}} \mathbf{B}_r + \sum_{k \in \mathcal{U} \backslash \{u\}} \nu_k \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} - \nu_u \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}} - \mathbf{X}_u \Bigg) \Bigg) \\ &\quad + \Gamma_2, \end{aligned} \tag{38}$$

where $\zeta_r^{\mathrm{C}}, \lambda_r^{\mathrm{C}}, \zeta_r^{\mathrm{D}}, \lambda_r^{\mathrm{D}}, \zeta_r^{\mathrm{S}}, \nu_u \geq 0$, and $\mathbf{X}_u \succeq 0$ are the Lagrangian multipliers for constraints (15), (19) for all $r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}$, (16), (23) for all $r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}$, (17) for all $r \in \widetilde{\mathcal{R}}^{\mathrm{S}}$, (25), and (26c), respectively, and Γ_2 includes all other terms unrelated to $\mathbf{W}_u$ and $\mathbf{X}_u$. Following similar steps in the proof of Theorem 1, for UE u, we have

$$\mathbf{W}_u^\star \left(\mathbf{P}_u^\star - \mathbf{Q}_u^\star \right) = \mathbf{0} \Leftrightarrow \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{P}_u^\star \right) = \operatorname{rank}\left(\mathbf{W}_u^\star \mathbf{Q}_u^\star \right), \tag{39}$$

where

$$\begin{aligned} \mathbf{P}_u^\star = &\sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{C}(\tau)}} \left(\frac{\mathbf{B}_r}{\eta_k} + (\zeta_r^{\mathrm{C}})^\star \mathbf{B}_r + (\lambda_r^{\mathrm{C}})^\star \mathbf{\Xi}_r \right) \\ &+ \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{D}(\tau)}} \left(\frac{\mathbf{B}_r}{\eta_k} + (\zeta_r^{\mathrm{D}})^\star \mathbf{B}_r + (\lambda_r^{\mathrm{D}})^\star \beta_{ru}^{(m+1)} \log_2 (1 + \gamma_u) \right) \\ &+ \sum_{r \in \widetilde{\mathcal{R}}^{\mathrm{S}}} (\zeta_r^{\mathrm{S}})^\star \mathbf{B}_r + \sum_{k \in \mathcal{U} \backslash \{u\}} \nu_k^\star \gamma_k \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}}, \end{aligned} \tag{40}$$

and $\mathbf{Q}_u^\star = \nu_u^\star \mathbf{h}_u \mathbf{h}_u^{\mathrm{H}}$, and $\{(\zeta_r^{\mathrm{C}})^\star\}$, $\{(\lambda_r^{\mathrm{C}})^\star\}$, $\{(\zeta_r^{\mathrm{D}})^\star\}$, $\{(\lambda_r^{\mathrm{D}})^\star\}$, $\{(\zeta_r^{\mathrm{S}})^\star\}$, and $\{\nu_u^\star\}$ denote the solution of the dual problem. Thus, we have $\mathbf{P}_u^\star \succ \mathbf{0}$ and $\operatorname{rank}(\mathbf{W}_u^\star \mathbf{P}_u^\star) = \operatorname{rank}(\mathbf{W}_u^\star)$. As $\operatorname{rank}(\mathbf{W}_u^\star \mathbf{Q}_u^\star) \leq \operatorname{rank}(\mathbf{Q}_u^\star) = 1$, we obtain $\operatorname{rank}(\mathbf{W}_u^\star) \leq 1$. Due to QoS requirement of UEs, we have $\operatorname{rank}(\mathbf{W}_u^\star) = 1, \forall\, u \in \mathcal{U}$. Hence, the proof of Theorem 2 is complete.

REFERENCES

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] "C-RAN: The Road Towards Green RAN," version 3.0, China Mobile, Beijing, China, Dec. 2013.

[3] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[4] Z. Li, S. Cui, J. Gu, and K. G. Shin, "Coordinated multi-point transmissions based on interference alignment and neutralization," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Mar. 2016, pp. 1–9.

[5] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: Beyond baseband samples and point-to-point links," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 90–97, Oct. 2015.

[6] A. Pizzinat, P. Chanclou, F. Saliou, and T. Diallo, "Things you should know about fronthaul," *J. Lightw. Technol.*, vol. 33, no. 5, pp. 1077–1083, Mar. 2015.

[7] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 105–111, Oct. 2015.

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

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

[10] 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.

[11] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. S. Wong, "A dynamic resource sharing mechanism for cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8325–8338, Dec. 2016.

[12] A. Liu and V. K. Lau, "Two-timescale user-centric RRH clustering and precoding optimization for cloud RAN via local stochastic cutting plane," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 64–76, Jan. 2018.

[13] Y. Mo, M. Peng, H. Xiang, Y. Sun, and X. Ji, "Resource allocation in cloud radio access networks with device-to-device communications," *IEEE Access*, vol. 5, pp. 1250–1262, 2017.

[14] R. L. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter, and U. Türke, "Toward energy-efficient 5G wireless communications technologies: Tools for decoupling the scaling of networks from the growth of operating power," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 24–34, Nov. 2014.

[15] W. Wu, J. Wang, M. Li, K. Liu, and J. Luo, "Energy-efficient transmission with data sharing," in *Proc. IEEE Conf. Comput. Commun.*, Hong Kong, Apr. 2015, pp. 73–81.

[16] M. Feng, S. Mao, and T. Jiang, "Boost: Base station on-off switching strategy for energy efficient massive MIMO HetNets," in *Proc. 35th Annu. IEEE Conf. Comput. Commun.*, San Francisco, CA, USA, Mar. 2016, pp. 1–9.

[17] S. Han, C. Yang, and A. F. Molisch, “Spectrum and energy efficient cooperative base station doze,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 285–296, Feb. 2014.

[18] D. Pompili, A. Hajisami, and H. Viswanathan, “Dynamic provisioning and allocation in cloud radio access networks (C-RANs),” *Ad Hoc Netw.*, vol. 30, pp. 128–143, Mar. 2015.

[19] D. Pompili, A. Hajisami, and T. X. Tran, “Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN,” *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.

[20] 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, May 2014.

[21] 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, Jan. 2015.

[22] Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, “Generalized sparse and low-rank optimization for ultra-dense networks,” *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 42–48, Jun. 2018.

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

[24] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, “Energy-efficient NOMA enabled heterogeneous cloud radio access networks,” *IEEE Netw.*, vol. 32, no. 2, pp. 152–160, Apr. 2018.

[25] D. Yan, R. Wang, E. Liu, and Q. Hou, “ADMM-based robust beamforming design for downlink cloud radio access networks,” *IEEE Access*, vol. 6, pp. 27912–27922, 2018.

[26] A. Maeder *et al.*, “Towards a flexible functional split for cloud-RAN networks,” in *Proc. Eur. Conf. Netw. Commun.*, Bologna, Italy, Jun. 2014, pp. 1–5.

[27] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, “Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems,” *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[28] D. Harutyunyan and R. Riggio, “Flexible functional split in 5G networks,” in *Proc. 13th Int. Conf. Netw. Service Manage.*, Tokyo, Japan, Nov. 2017, pp. 1–9.

[29] T. T. Vu, D. T. Ngo, M. N. Dao, S. Durrani, D. H. Nguyen, and R. H. Middleton, “Energy efficiency maximization for downlink cloud radio access networks with data sharing and data compression,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 4955–4970, Aug. 2018.

[30] S.-H. Park, O. Simeone, O. Sahin, and S. S. Shitz, “Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory,” *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[31] P. Patil and W. Yu, “Hybrid compression and message-sharing strategy for the downlink cloud radio-access network,” in *Proc. Inf. Theory Appl. Workshop*, San Diego, CA, USA, Feb. 2014, pp. 1–6.

[32] *Transport Network Support of IMT-2020/5G*, ITU-T Standard Q11/15, Feb. 2018.

[33] *Study on New Radio Access Technology: Radio Access Architecture and Interfaces (Release 14)*, 3GPP TR 38.801 V14.0.0, Mar. 2017.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. 2nd ed. New York, NY, USA: Wiley, 2006.

[35] G. Auer *et al.*, “How much energy is needed to run a wireless network?” *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[36] S. Boyd, “Sequential convex programming,” Lecture Notes, Stanford Univ., Stanford, CA, USA, 2008.

[37] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Oct. 2008.

[38] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

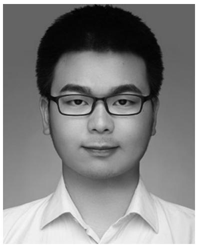
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[40] M. Grant and S. Boyd, “CVX: MATLAB software for disciplined convex programming, version 2.1,” Jun. 2015. [Online]. Available: http://cvxr.com/cvx

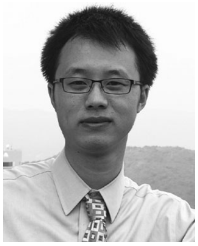
[41] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, “A majorization-minimization approach to the sparse generalized eigenvalue problem,” *Mach. Learn.*, vol. 85, no. 1, pp. 3–39, Dec. 2011.

[42] H. Wang, F. Zhang, Q. Wu, Y. Hu, and Y. Shi, “Nonconvex and nonsmooth sparse optimization via adaptively iterative reweighted methods,” 2018. [Online]. Available: https://arxiv.org/abs/1810.10167

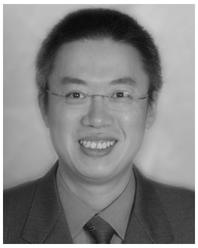
Yong Zhou (S’13–M’16) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From 2015 to 2017, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. He is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. He has served as a Technical Program Committee Member for several conferences. His research interests include performance analysis and resource allocation of 5G and Internet of Things networks.



Jie Li received the B.S. degree in communication engineering from Anhui University, Hefei, China, in 2014. He is currently working toward the M.S. degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include age of information and Internet of Things networks.



Yuanming Shi (S’13–M’15) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2015. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, where he is currently a tenured Associate Professor. He visited the University of California, Berkeley, CA, USA, from October 2016 to February 2017. His research interests include optimization, statistics, machine learning, signal processing, and their applications to wireless communications and quantitative finance. He is a recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, and the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society.



Vincent W. S. Wong (S’94–M’00–SM’07–F’16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada, in 2000. From 2000 to 2001, he was a Systems Engineer with PMC-Sierra Inc. (now Microchip Technology Inc.). In 2002, he joined the Department of Electrical and Computer Engineering, UBC, where he is currently a Professor. His research interests include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grids, mobile edge computing, and Internet of Things. He is an Executive Editorial Committee Member of the IEEE Transactions on Wireless Communications, an Area Editor for the IEEE Transactions on Communications, and an Associate Editor for the IEEE Transactions on Mobile Computing. He has served as a Guest Editor for the IEEE Journal on Selected Areas in Communications and the IEEE Wireless Communications. He has also served on the editorial boards of the IEEE Transactions on Vehicular Technology and the *Journal of Communications and Networks*. He was a Tutorial Co-Chair of the 2018 IEEE Global Communications Conference (Globecom), a Technical Program Co-Chair of the 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), as well as a Symposium Co-Chair of the 2018 IEEE International Conference on Communications, IEEE SmartGridComm 2013 and 2017, and IEEE Globecom 2013. He is the Chair of the IEEE Vancouver Joint Communications Chapter and has served as the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications. He received the 2014 UBC Killam Faculty Research Fellowship. He is an IEEE Communications Society Distinguished Lecturer for 2019–2020.