Cross-Layer Optimization for Industrial Internet of Things in NOMA-based C-RANs

Jie Tang, *Senior Member, IEEE,* Yanfei Zhao, Wanmei Feng, Xiaolan Zhao,
Xiu Yin Zhang, *Senior Member, IEEE,* Minqian Liu, *Member, IEEE,* and Kai-Kit Wong, *Fellow, IEEE*

Abstract—**This paper investigates non-orthogonal multiple access (NOMA)-based cloud radio access networks (C-RANs), where edge caching is adopted to cut down the crowdedness of the fronthaul links. We aim to maximize the energy efficency (EE) by jointly optimizing the power allocation, analog and digital precoding, which turns out to be an intractable non-convex optimization problem. To tackle this problem, we first select cluster heads using the selecting cluster-head (SCH) algorithm, where the analog precoding matrix can be resolved by means of maximizing the array gains. Then, the device grouping algorithm is proposed to group devices according to the equivalent channel correlations, and thus the NOMA devices in the same beam are capable of sharing the same digital precoding vector. Finally, joint digital precoding design and power allocation algorithm is proposed to decompose the resultant optimization problem into two subproblems and solve them iteratively by applying Taylor expansion operation and the minimum mean square error (MMSE) detection. Simulation results validate that the proposed NOMA-based C-RANs with hybrid precoding (HP) scheme can achieve higher SE and EE than traditional orthogonal multiple access (OMA)-based approach and two-stage HP scheme.**

Index Terms—**Cloud radio access network (C-RAN), non-orthogonal multiple access (NOMA), massive multiple-input multiple-output (mMIMO), hybrid precoding, power allocation.**

I. Introduction

INDUSTRIAL Internet of Things (IIoT), as a communication paradigm which integrates the operational technology and informational technology, connects the sensors, computers and machines, etc, and makes industrial operations more efficient and intelligent [1]. Nevertheless, traditional IIoT networks face the challenges in the explosive growth of data traffic caused by multiple IIoT devices. The combination of massive multiple-input multiple-output (mMIMO) and millimeter wave (mmWave) has been considered as a potential technique for IIoT since it can achieve higher spectral efficiency (SE) and broader bandwidth [2] [3]. In traditional MIMO systems, the fully digital precoding requires each antenna equipped with a dedicated radio-frequency (RF) chain, which makes it difficult to afford the hardware cost and energy consumption [4] [5]. On the other hand, hybrid precoding (HP) is considered as an energy-efficient architecture which balances the tradeoff between the system performance and hardware complexity [6]. The HP comprises two architecture: the fully-connected HP where each RF chain is connected to all antenna elements, and the sub-connected HP where each RF chain is connected to the equal number of antennas. In general, the fully-connected HP has higher SE while the sub-connected HP has higher energy efficiency (EE) [5].

Non-orthogonal multiple access (NOMA) is also one of the promising techniques for significantly enhancing the SE in mmWave mMIMO systems [7]–[9]. Different from orthogonal multiple access (OMA)-based systems, the NOMA-based systems can serve multiple devices in one beam at the same frequency-time resource block with the aid of intra-beam superposition coding (SC) at base station (BS) and successive interference cancellation (SIC) at the receivers [10], [11]. Moreover, the NOMA-based systems with HP architecture can significantly reduce the number of phase shifters (PS) and RF chains, which can mitigate the energy consumption while achieving good system performance.

Due to the large hardware and operating costs of mmWave mMIMO technique, the system performance is limited. The cloud radio access network (C-RAN) has been regarded a potential technique to realize centralized data processing and dynamic resource scheduling, which can reduce the hardware overhead. However, there is a long distance between the data center of the conventional C-RAN and the devices, which leads to a higher end-to-end delivery latency. To tackle this problem, a cache-enabled C-RAN is presented, where the enhanced remote radio heads (eRRHs) are deployed to execute baseband signal processing as well as cache the popular contents [12] [13]. In particular, the network edges can pre-fetch the most popular files and store them into their own local cache during the off-peak periods. In this case, the cache procedure is beneficial to decrease the end-to-end latency of devices and thus enhance the SE [14], [15]. The authors in [14] investigated the cache placement strategy for minimizing the average requested download delay subject to the limited cache and fronthaul

This work has been supported in part by National Key Research and Development Project under Grant 2019YFB1804100, in part by Key Research and Development Project of Guangdong Province under Grant 2019B010156003, in part by the National Natural Science Foundation of China under Grant 61971194, 62071364, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011607, and in part by the Research Fund Program of Guangdong Key Laboratory of Aerospace Communication and Networking Technology under Grant 2018B030322004. (*Corresponding author: Xiaolan Zhao.*)

J. Tang, Y. Zhao, W. Feng, X. Zhao and X. Zhang are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (email: eejtang@scut.edu.cn; zyf18843103189@outlook.com; eewmfeng@mail.scut.edu.cn; gzxll@scut.edu.cn; zhangxiuyin@scut.edu.cn).

M. Liu is with the School of Communication Engineering, Xidian University, Xian 710126, China (e-mail: mqliu@mail.xidian.edu.cn).

K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: kai-kit.wong@ucl.ac.uk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

capacity. A novel cache scheme was studied to maximize the average date rate whilst satisfying the constraint of finite service latency [15]. The authors in [16] studied a cooperative caching strategy with content request prediction, which can significantly improve the cache hit ratio and reduce content acquisition delay effectively. Besides, due to the variability of the C-RAN network environments, the increase of IIoT devices and the dense deployment of eRRHs, there is an urgent need for efficient resource management solutions to reduce the energy consumption. To this end, maximizing the EE whilst satisfying a certain SE requirement becomes a new optimization criterion in wireless communication systems. Besides, EE is also the key performance indicators (KPIs) for the fifth generation (5G) and beyond wireless communication networks [7].

A. Contributions

Previous works mainly focus on EE maximization using the NOMA technique without considering the effect of C-RANs. On the other hand, the works in [14], [15] aim to study the cache placement strategies for C-RANs, where NOMA is not considered to further enhance the system performance. To the best of our knowledge, the resource allocation for maximizing EE in NOMA-based C-RANs has not been well studied. Motivated by these observations, we consider hybrid precoding design and power allocation for EE maximization in NOMA-based C-RANs, including cluster-head selection, analog precoding, device grouping, digital precoding and power allocation. Specifically, our contributions can be summarized as follows.

- We propose a theoretical model for EE maximization in NOMA-based C-RANs, where the hybrid precoding and power allocation are jointly optimized. To tackle the mixed combinatorial non-convex optimization problem, we first design the analog precoding and device clustering, and then jointly optimize the digital precoding and power allocation.
- For the design of analog precoding and device clustering, we first propose the selecting cluster-head (SCH) algorithm to select cluster heads, and then resolve the analog precoding matrix by maximizing the array gains. After that, the device grouping algorithm is proposed to group devices according to the equivalent channel correlations.
- In the case of power allocation and digital precoding, the resultant EE maximization problem is still non-convex owing to the inter-cluster and intra-cluster interferences. To tackle this problem, a joint digital precoding design and power allocation algorithm is proposed to decompose the original non-convex problem into two subproblems, and solve them iteratively by applying Taylor expansion operation and the minimum mean square error (MMSE) detection.
- Numerical results reveal that the EE performance can be significantly improved by our proposed algorithms compared to the fully ZF precoding scheme. Moreover, these results also demonstrate that the proposed NOMA-based C-RANs with hybrid precoding scheme can achieve higher SE and EE than OMA-based approach and two-stage HP scheme.

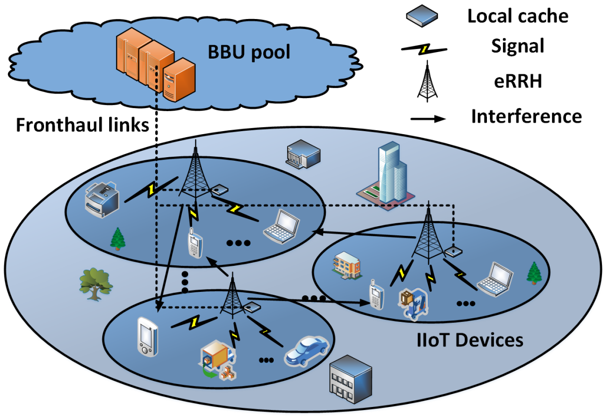

Fig. 1: Illustration of a NOMA-based C-RANs architecture.

B. Organization and Notation

The remainder of the paper is structured as follows. In Section II, the system model of the NOMA-based C-RANs is presented. In Section III, the SCH algorithm, analog precoding and device grouping are discussed. The joint digital precoding design and power allocation algorithm is presented in Section IV. Numerical results and the conclusion are discussed in Section V and Section VI, respectively.

Notation: We use bold upper case and lower case letters to denote matrices and column vectors, respectively. The $(\cdot)^{-1}$, $(\cdot)^T$ and $(\cdot)^H$ represent matrix inversion, transpose and conjugate transpose, respectively. $||\cdot||_p$ and tr$(\cdot)$ denote l_p norm operation and trace, respectively. The diagonal elements of diagonal matrix diag$(\mathbf{a})$ are the elements of vector $\mathbf{a}$. E$\{\cdot\}$ represents the expectation. The number of elements in set Γ is denoted as $|\Gamma|$ and $|\cdot|$ indicates the absolute value. $\lfloor x \rfloor$ denotes the integer closest to but not greater than x. The complex Gaussian distribution with mean $\mathbf{n}$ and covariance $\mathbf{R}$ is denoted as $\mathcal{CN}(\mathbf{n}, \mathbf{R})$. The Kronecker product is denoted as $\otimes$. $\mathbf{I}_N$ denotes the identity matrix with the dimension of N × N. $\bar{c}$ is the complement $1 - c$ of a binary variable $c \in \{0, 1\}$. The empty set is denoted as $\emptyset$. $\mathbb{C}$ indicates the complex field.

II. System Model And Problem Formulation

A. System Model

In this paper, the considered NOMA-based C-RANs architecture is shown in Fig. 1. There are K_R eRRHs establishing wireless communication links with K_U single-antenna devices, where $K = K_U/K_R$ denotes the number of devices connected with each eRRH. Let $\mathcal{K}_U = \{1, \cdots, K_U\}$ and $\mathcal{K}_R = \{1, \cdots, K_R\}$ be the set of devices and eRRHs, respectively. The BBU connects eRRH i through an error-free fronthaul link with capacity C_i bps/Hz, $i \in \mathcal{K}_R$.

To reduce the energy consumption, the sub-connected HP architecture is adopted in the NOMA-based C-RANs. In particular, the number of RF chains in HP architectures is less than the number of antennas, which can be realized by a high-dimensional analog precoder and a low-dimensional digital

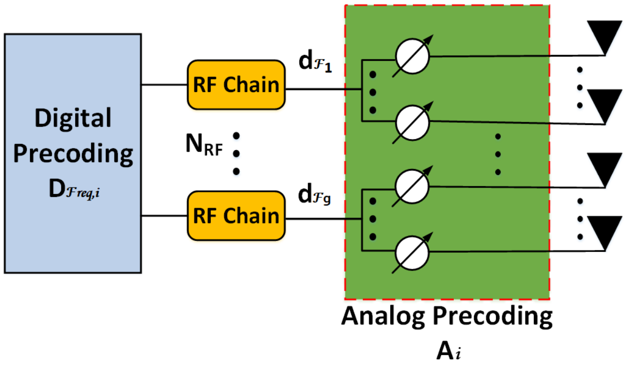
Digital
Precoding
$\mathbf{D}_{\mathcal{F}req,i}$
RF Chain
RF Chain
$\mathbf{d}_{\mathcal{F}_1}$
$\mathbf{d}_{\mathcal{F}g}$
$\mathbf{N}_{RF}$
N
Analog Precoding
$\mathbf{A}_i$

Fig. 2: The sub-connected hybrid precoding architecture.

precoder. The total antenna array elements at the eRRH is separated into multiple sub-arrays, wherein each RF chain is connected to a sub-array. The beam direction radiated from each sub-array is controlled by the value of the phase shifters. N transmit antennas and N_{RF} RF chains are allocated to every eRRH, where each RF chain connects to N/N_{RF} transmit antennas through phase shifters. Since the number of beams is restricted to the number of RF chains, we assume that the number of RF chains N_{RF} is equal to the number of beams G, i.e., $N_{RF} = G$.

1) Cache Model: Assume that the requested files can be divided into different sublibraries. There are G sublibraries which are transmitted by G beams. The requested files $|S_g|$ in the gth sublibrary $\mathcal{F}_{g,i}$ for eRRH i are transmitted through the gth beam, where $\mathcal{F}_{g,i} = \{f^i_{g,1}, f^i_{g,2}, \cdots, f^i_{g,|S_g|}\}$. Next, we can model the cache status of sublibrary $\mathcal{F}g$ at eRRH i by introducing binary variables $c_{\mathcal{F}g,i}$, $\mathcal{F}g \subseteq \mathcal{F}$, $i \in \mathcal{K}_R$, as

$$c_{\mathcal{F}g,i} = \begin{cases} 1, & \text{if eRRH } i \text{ caches subfile } \mathcal{F}g, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where the binary variable $\bar{c}_{\mathcal{F}g,i} = 1 - c_{\mathcal{F}g,i}$.

2) Channel Model: Owing to the limited scattering in mmWave channels, we adopt the geometric Saleh-Valenzuela model to embody the spatial correlation characteristics of mmWave communications [17]. Specifically, the mmWave channel model between the eRRH i and the device k with one line-of-sight (LoS) path and $L_{k,i} - 1$ non-LoS (NLoS) paths is adopted [18]. The $N \times 1$ channel vector $\mathbf{h}^i_k$ between the eRRH i and device k can be given by [19]

$$\mathbf{h}^i_k = \sqrt{\frac{N}{L_k}} \sum_{l=1}^{L_{k,i}} \alpha^{(l)}_{k,i} \mathbf{a}(\varphi^{(l)}_{k,i}, \theta^{(l)}_{k,i}), \tag{2}$$

where $\alpha^{(l)}_{k,i}$ is the complex gain of the lth path, $\theta^{(l)}_{k,i}$ and $\varphi^{(l)}_{k,i}$ represent the elevation and azimuth angle of departure (AoD) of the lth path, $\mathbf{a}(\varphi^{(l)}_{k,i}, \theta^{(l)}_{k,i}) \in \mathbb{C}^{N \times 1}$ indicates the array steering vector of a $N_1 \times N_2$ uniform linear array (ULA), which is represented by

$$\mathbf{a}(\varphi, \theta) = \mathbf{a}_a(\varphi) \otimes \mathbf{a}_e(\theta), \tag{3}$$

where $\mathbf{a}_a(\varphi) = \frac{1}{\sqrt{N_1}}[e^{j2\pi i(d_1/\lambda)\sin\varphi}]_{i\in\{0,1,\cdots,N_1-1\}}$, $\mathbf{a}_e(\theta) = \frac{1}{\sqrt{N_2}}[e^{j2\pi j(d_2/\lambda)\sin\theta}]_{j\in\{0,1,\cdots,N_2-1\}}$, the signal wavelength is denoted as λ, d_1 and d_2 indicate the horizontal and vertical antenna spacing, respectively. Similar to [17], we assume that $d_1 = d_2 = \lambda/2$ in mmWave communications. The channel state information (CSI) is assumed to be perfectly known at all the eRRHs [7] [20]. It is assumed that the parameters of the channel between the eRRH and IoT devices (i.e., the direction of arrival (DOA)) are estimated by the rank reduction method [21]. In particular, IoT devices first send training sequences to the eRRH in the uplink. Then, the covariance of the received signals is used to perform eigenvalue decomposition, and thereby the determination matrix is generated. Finally, when the DOA is consistent with the signal direction, it is determined that the eigenvector of the matrix corresponds to the minimum eigenvalue.

3) Signal Model: According to the cache model subsection, $c_{f,i}$ denotes the caching state of file f in eRRH i. If $c_{f,i} = 1$, the devices of eRRH i will retrieve the contents of files from the local cache. Otherwise, devices will retrieve the contents of files from the library located at the BBU. Generally, as for the uncached files, soft- and hard- fronthaul information transmission are two common kinds of transmission methods to fetch files from the BBU. For the soft fronthaul information transmission, a quantized version of the precoded signals is delivered through fronthaul links. In addition, the hard information of the uncached files is transferred through fronthaul links. When the devices request the uncached files in the eRRHs, the soft fronthaul information transmission is adopted to transfer the uncached files from the BBU to the eRRHs. As a result, the signal $x \in \mathbb{C}^{N \times 1}$ which contains the information of the uncached and cached files delivered by eRRH is written as

$$\mathbf{x}_i = \mathbf{A}_i \left(\sum_{f \in \mathcal{F}_{req}} \mathbf{d}_{f,i} s_f + \mathbf{z}_i \right), \tag{4}$$

where $\mathbf{d}_{f,i} = c_{f,i}\tilde{\mathbf{d}}_{f,i} + \bar{c}_{f,i}\bar{\mathbf{d}}_{f,i} \in \mathbb{C}^{N_{RF} \times 1}$, $\tilde{\mathbf{d}}_{f,i}$ is the digital precoding vector of cached baseband signal s_f, and $\bar{\mathbf{d}}_{f,i}$ is the digital precoding vector of uncached signal for eRRH i, respectively, the scalar quantity s_f is the data symbol of requested file f with $E\{|s_f|^2\} = 1$, the requested files set is denoted as $\mathcal{F}_{req}$. Supposing that the quantization noise $\mathbf{z}_i \in \mathbb{C}^{N_{RF} \times 1}$ is independent of the information of the requested files and distributed as $\mathbf{z}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$. The quantization covariance matrix $\mathbf{\Omega}_i$ is set as $\upsilon \mathbf{I}_{N_{RF}}$, where υ is a constant with $0 < \upsilon < 0.02$ [20]. Note that if eRRHs have cached all requested files, i.e., files $\mathcal{F}_{req}$ are prestored at the local cache, the quantization noise $\mathbf{z}_i$ is a zero vector. In addition, we suppose that the quantization noise $\mathbf{z}_i$ is unrelated to the eRRHs. $\mathbf{A}_i$ is the analog precoding matrix of size $N \times N_{RF}$. The sub-connected HP analog precoding matrix is written as [22]

$$\mathbf{A}_i = \begin{bmatrix} \bar{\mathbf{a}}_{i,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{a}}_{i,2} & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{a}}_{i,N_{RF}} \end{bmatrix}, \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

where $M = N/N_{RF}$ is presumed to be an integer. Thus, each RF chain connects M antennas through M phase shifters. The elements of $\bar{\mathbf{a}}_{i,n} \in \mathbb{C}^{M\times 1}$ have the same amplitude $1/\sqrt{M}$ but different phases for $n = 1, 2, \cdots, N_{RF}$.

Consequently, the signal y_k is given by

$$\begin{aligned} y_k &= \mathbf{h}_k^H \mathbf{A} \sum_{f\in\mathcal{F}_{req}} \mathbf{d}_f s_f + \mathbf{h}_k^H \mathbf{A}\mathbf{z} + v_k \\ &= \underbrace{\mathbf{h}_k^H \mathbf{A}\mathbf{d}_{f_k} s_{f_k}}_{\text{desired file}} + \underbrace{\sum_{f\in\mathcal{F}_{req}\setminus\{f_k\}} \mathbf{h}_k^H \mathbf{A}\mathbf{d}_f s_f}_{\text{interference files}} + \underbrace{\mathbf{h}_k^H \mathbf{A}\mathbf{z} + v_k}_{\text{noise}}, \end{aligned} \tag{6}$$

where $\mathbf{h}_k^H \triangleq \left[\mathbf{h}_{k,1}^H, \mathbf{h}_{k,2}^H, \cdots, \mathbf{h}_{k,K_R}^H\right] \in \mathbb{C}^{1\times NK_R}$, $\mathbf{h}_{k,i}^H \in \mathbb{C}^{1\times N}$ is the channel matrix between device k and eRRH i, $\mathbf{d}_f \triangleq \left[\mathbf{d}_{f,1}^H, \mathbf{d}_{f,2}^H, \cdots, \mathbf{d}_{f,K_R}^H\right]^H \in \mathbb{C}^{N_{RF}K_R\times 1}$ and $\mathbf{z} \triangleq \left[\mathbf{z}_1^H, \mathbf{z}_2^H, \cdots, \mathbf{z}_{K_R}^H\right]^H \in \mathbb{C}^{N_{RF}K_R\times 1}$, v_k is the noise following the distribution $\mathcal{CN}(0, \sigma_v^2)$. The analog precoding super-matrix $\mathbf{A}$ can be given by

$$\begin{aligned} \mathbf{A} &= \left[\left(\mathbf{A}_1\mathbf{P}_1^H\right)^H, \cdots, \left(\mathbf{A}_{K_R}\mathbf{P}_{K_R}^H\right)^H\right]^H \\ &= diag\Big(\mathbf{A}_1, \cdots, \mathbf{A}_{K_R}\Big), \end{aligned} \tag{7}$$

where the permutation matrix $\mathbf{P}_i$ is defined as

$$\mathbf{P}_i = \Big[\mathbf{0}_{N_{RF}\times(i-1)N_{RF}}, \mathbf{I}_{N_{RF}\times N_{RF}}, \mathbf{0}_{N_{RF}\times(K_R-i)N_{RF}}\Big]^T. \tag{8}$$

Then, the signal-to-interference-and-noise ratio (SINR) at the device k can be written as

$$\gamma_k = \frac{\mathbf{h}_k^H \mathbf{A}\mathbf{d}_{f_k}\mathbf{d}_{f_k}^H \mathbf{A}^H \mathbf{h}_k}{\xi_k}, \tag{9}$$

where

$$\xi_k = \sum_{f\in\mathcal{F}_{req}\setminus\{f_k\}} \mathbf{h}_k^H \mathbf{A}\mathbf{d}_f\mathbf{d}_f^H \mathbf{A}^H \mathbf{h}_k + \mathbf{h}_k^H \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^H\mathbf{h}_k + \sigma_v^2, \tag{10}$$

$$\boldsymbol{\Omega} = diag(\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_{K_R}).$$

To this end, the achievable data rate of device k is given by

$$R_k = \log_2(1 + \gamma_k). \tag{11}$$

Since the interference between eRRHs and devices in adjacent cells is very low, we assume that each device is associated with the nearest eRRH and can only be licensed to one eRRH [23]. The eRRHs are allocated to the orthogonal time-frequency resources which avoids the co-tier interference between the eRRHs [24] [25]. As for the eRRH i, the set of devices served by the gth beam is expressed as S_g^i, where $|S_g^i| \geq 1$ for $g = 1, 2, \cdots, G$ and $S_m^i \cap S_n^i = \emptyset$ for $m \neq n$. The total number of devices supported by all beams under each eRRH is K, i.e., $\sum_{g=1}^G |S_g^i| = K, i \in \mathcal{K}_R$.

According to the above cache model, the signal $\mathbf{x}_i \in \mathbb{C}^{N\times 1}$ contains the information of the uncached and cached files delivered by eRRH i, which is written as

$$\mathbf{x}_i = \mathbf{A}_i\Bigg(\mathbf{D}_{\mathcal{F}_{req},i}\mathbf{s}_{\mathcal{F}_{req},i} + \mathbf{z}_i\Bigg), \tag{12}$$

where $\mathbf{A}_i \in \mathbb{C}^{N\times N_{RF}}$ is the analog precoding matrix, $\mathbf{D}_{\mathcal{F}_{req},i} = [\mathbf{d}_{\mathcal{F}1,i}, \mathbf{d}_{\mathcal{F}2,i}, \cdots, \mathbf{d}_{\mathcal{F}G,i}] \in \mathbb{C}^{N_{RF}\times G}$ is the digital precoding matrix of eRRH i, and $\mathbf{d}_{\mathcal{F}g,i} = c_{\mathcal{F}g,i}\tilde{\mathbf{d}}_{\mathcal{F}g,i} + \bar{c}_{\mathcal{F}g,i}\bar{\mathbf{d}}_{\mathcal{F}g,i} \in \mathbb{C}^{N_{RF}\times 1}$, $\tilde{\mathbf{d}}_{\mathcal{F}g,i}$ denotes the digital precoding vector for caching the gth sublibrary $\mathcal{F}g$, and $\bar{\mathbf{d}}_{\mathcal{F}g,i}$ represents the digital precoding vector for unavailable sublibrary $\mathcal{F}g$ in eRRH i, respectively. The signal $\mathbf{s}_{\mathcal{F}_{req},i} \in \mathbb{C}^{G\times 1}$ represents the data symbol of requested file $\mathcal{F}_{req}$. $\mathbf{z}_i \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{i})$ is the quantization noise of size $N_{RF} \times 1$. Consequently, the signal received by the mth device of the gth beam in the ith eRRH can be given by

$$\begin{aligned} y_{g,m}^i &= (\mathbf{h}_{g,m}^i)^H \mathbf{A}_i \sum_{j=1}^{G}\sum_{k=1}^{|S_j|} \mathbf{d}_{\mathcal{F}j,i}\sqrt{p_{j,k}^i}s_{fj,k}^i + (\mathbf{h}_{g,m}^i)^H \mathbf{A}_i\mathbf{z}_i + v_{g,m}^i \\ &= \underbrace{(\mathbf{h}_{g,m}^i)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\sqrt{p_{g,m}^i}s_{fg,m}^i}_{\text{desired files}} + \underbrace{I_{intra} + I_{inter}}_{\text{interference}} \\ &\quad + \underbrace{\left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\mathbf{z}_i + v_{g,m}^i}_{\text{noise}}, \end{aligned} \tag{13}$$

where the intra-beam interference and inter-beam interference are given by

$$I_{intra} = \left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\Big(\sum_{k=1}^{m-1}\sqrt{p_{g,k}^i}s_{fg,k}^i + \sum_{k=m+1}^{|S_g|}\sqrt{p_{g,k}^i}s_{fg,k}^i\Big),$$

$$I_{inter} = \left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\sum_{j\neq g}\sum_{k=1}^{|S_j|}\mathbf{d}_{\mathcal{F}j,i}\sqrt{p_{j,k}^i}s_{fj,k}^i,$$

where $s_{fg,m}^i$ is the transmit signal with $E\{\left|s_{fg,m}^i\right|^2\} = 1$, $p_{g,k}^i$ denotes the transmit power for the kth device of gth cluster, $v_{g,m}^i \sim \mathcal{CN}(0, \sigma_{i,v}^2)$ denotes the additive white Gaussian noise.

To mitigate the intra-cluster interference, NOMA schemes are exploited in this paper. In particular, intra-beam SC is performed at the ith eRRH and the receivers perform SIC to remove the interference of the mth device from the kth device (for all $k > m$) in the gth beam. In general, we assume that $\left\|\left(\mathbf{h}_{g,1}^i\right)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2 \geq \left\|\left(\mathbf{h}_{g,2}^i\right)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2 \geq \cdots \geq \left\|\left(\mathbf{h}_{g,|S_g|}^i\right)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2$, $g = 1, 2, \cdots, G$. Then, the remaining received signal of the mth device in the gth beam can be rewritten as

$$\begin{aligned} y_{g,m}^i = &\underbrace{(\mathbf{h}_{g,m}^i)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\sqrt{p_{g,m}^i}s_{fg,m}^i}_{\text{desired files}} \\ &+ \underbrace{\left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\sum_{k=1}^{m-1}\sqrt{p_{g,k}^i}s_{fg,k}^i}_{\text{intra-beam interference}} \\ &+ \underbrace{\left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\sum_{j\neq g}\sum_{k=1}^{|S_j|}\mathbf{d}_{\mathcal{F}j,i}\sqrt{p_{j,k}^i}s_{fj,k}^i}_{\text{inter-beam interference}} \\ &+ \underbrace{\left(\mathbf{h}_{g,m}^i\right)^H \mathbf{A}_i\mathbf{z}_i + v_{g,m}^i}_{\text{noise}}, \end{aligned} \tag{14}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

and the SINR of the mth device in the gth beam is written as

$$\gamma_{g,m}^{i}=\left\|\left(\mathbf{h}_{g,m}^{i}\right)^{H}\mathbf{A}_{i}\mathbf{d}_{\mathcal{F}g,i}\right\|_{2}^{2}p_{g,m}^{i}/\xi_{g,m}^{i}, \tag{15}$$

where $\xi_{g,m}^{i}$ is given by (16) at the top of the next page. Consequently, the data rate of the mth device in the gth beam can be given by

$$R_{g,m}^{i}=\log_{2}(1+\gamma_{g,m}^{i}). \tag{17}$$

The sum data rate in the ith eRRH can be written as

$$R_{i}=\sum_{g=1}^{G}\sum_{m=1}^{|S_{g}|}R_{g,m}^{i}. \tag{18}$$

4) Power Consumption Model: For the considered NOMA-based C-RANs, the total energy consumption is composed of the circuit power and the transmit power. Thus, similar to [20] [26], we can express the total power consumption of eRRH i as

$$P_{total}^{i}=P_{c}^{i}+\eta\bigg(\sum_{g=1}^{G}\sum_{m=1}^{|S_{g}|}p_{g,m}^{i}\bigg), \tag{19}$$

where the constant $\eta\geq 1$ represents the inefficiency of power amplifier [27]. The circuit power consumption P_{c}^{i} of eRRH i is given by

$$P_{c}^{i}=N_{RF}P_{R}+N_{P}P_{P}+N_{A}P_{A}+P_{B}, \tag{20}$$

where P_{R} and P_{P} denote the power consumed by a RF chain and phase shifter, respectively. N_{P} and N_{A} are the number of phase shifters and power amplifies. P_{A} and P_{B} represent the power consumed by power amplifies and baseband signal processing, respectively.

B. Problem Formulation

Our aim is to maximize the EE of NOMA-based C-RANs under the constraints of the fronthaul transmission rate, the transmit power of eRRH, and precoding vectors. Accordingly, the resulting optimization problem is formulated as

$$\textbf{(P1):}\ \max_{\{p_{g,m}^{i}\},\{\mathbf{A}_{i}\},\{\mathbf{d}_{\mathcal{F}g,i}\}}\sum_{i=1}^{K_{R}}\frac{\sum_{g=1}^{G}\sum_{m=1}^{|S_{g}|}R_{g,m}^{i}}{P_{c}^{i}+\eta\left(\sum_{g=1}^{G}\sum_{m=1}^{|S_{g}|}p_{g,m}^{i}\right)} \tag{21a}$$

$$\text{s.t.}\quad g(\mathbf{d}_{\mathcal{F}g,i})\leq C_{i},\qquad \forall i\in\mathcal{K}_{\mathcal{R}}, \tag{21b}$$

$$\sum_{g=1}^{G}\sum_{m=1}^{|S_{g}|}p_{g,m}^{i}\leq P_{i},\quad \forall i\in\mathcal{K}_{\mathcal{R}}, \tag{21c}$$

$$\sum_{g=1}^{G}\|\mathbf{A}_{i}\mathbf{d}_{\mathcal{F}g,i}\|^{2}\leq 1,\ \forall i\in\mathcal{K}_{R}, \tag{21d}$$

where P_{i} is the maximum transmit power of eRRH i. Constraint (21b) implies that the data rate on each fronthaul link cannot exceed the limited capacity C_{i}. We thus can express the fronthaul rate $g(\mathbf{d}_{\mathcal{F}g,i})$ as

$$\begin{aligned}g(\mathbf{d}_{\mathcal{F}g,i})\triangleq&\log_{2}\det\left(\sum_{\mathcal{F}_{g}\subseteq\mathcal{F}_{req}}\bar{c}_{\mathcal{F}_{g},i}\bar{\mathbf{d}}_{\mathcal{F}_{g},i}\bar{\mathbf{d}}_{\mathcal{F}_{g},i}^{H}+\boldsymbol{\Omega}_{i}\right)\\&-\log_{2}\det(\boldsymbol{\Omega}_{i}).\end{aligned} \tag{22}$$

Constraint (21c) corresponds to the transmit power constraint of eRRH i. Constraint (21d) denotes the normalization limitation of precoding vectors. The considered EE optimization problem, which jointly optimizes the analog precoding, digital precoding and power allocation, is mixed combinatorial non-convex and cannot be solved directly. Therefore, we concentrate on solving these variables in an alternative manner. Since the beam pattern design needs to obtain beam scanning angles, the cluster heads should be selected first in order to compute the beam scanning angles. Then, the analog precoding and device clustering can be designed based on the selected cluster heads. Finally, the considered problem is decoupled into two subproblems, where the power allocation and digital precoding are optimized iteratively by using Taylor expansion operation and the MMSE detection.

III. Analog Precoding Design And Device Clustering

For the reason that the number of devices K is much larger than that of RF chains N_{RF} in each cell, we divide devices into N_{RF} clusters, where the number of beams is equal to the number of RF chains, i.e. $G=N_{RF}$. In particular, the correlation-based SCH Algorithm is proposed to select the cluster head for each beam. Then, the analog precoding matrix is solved through maximizing the array gain. In addition, devices grouping is determined by the equivalent channel correlation between the remaining devices and cluster-heads.

A. The Proposed Correlation-Based SCH Algorithm

To enhance the system performance, we select G cluster heads in terms of the channel conditions of devices. Firstly, we define $\varrho_{i,j}=|\mathbf{h}_{i}^{H}\mathbf{h}_{j}|/\|\mathbf{h}_{i}\|_{2}\|\mathbf{h}_{j}\|_{2}$ as the channel correlation coefficient between the device j and device i, where $\varrho_{i,j}$ indicates the lower correlation. In this case, inter-beam interference is suppressed for low channel correlation of devices in different beams.

In the proposed correlation-based SCH algorithm, we measure and calculate the correlation of the cluster-heads through an adaptive threshold σ, which is transformed into $\sigma=\sigma+0.1*(1-\sigma)$ in the next iteration. Specifically, as for the first beam, the device whose channel gain is the highest is assigned to the first cluster head, and then the devices are chosen as the cluster-head candidates for other clusters when their channel correlations are less than the threshold σ. Particularly, we pick out the device with the highest channel gain in the cluster head candidates as the second cluster head. Based on this principle, G cluster heads can be chosen. The specific steps and details of the algorithm are summarized in Table I, where Γ is the set of selected G cluster heads.

In particular, the highest complexity from step 8 to step 13 in Table I reaches $(2+2(K-1))(K-1)$, while that is $2(K-1)$ from step 14 to step 17. Thus, the computational complexity of the proposed correlation-based SCH algorithm reaches $\mathcal{O}(GK^{2})$ [28].

B. Analog Precoding With Finite Phase Shifters

A typical two-stage HP scheme is considered, where the HP structure is divided into the analog and digital precoding [29]. Particularly, for analog precoding, it is difficult to seek

$$\xi^i_{g,m} = \sum_{k=1}^{m-1} \left\| \left(\mathbf{h}^i_{g,m}\right)^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i} \right\|_2^2 p^i_{g,k} + \sum_{j\neq g}\sum_{k=1}^{|S_j|} \left\| \left(\mathbf{h}^i_{g,m}\right)^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}j,i} \right\|_2^2 p^i_{j,k} + \left(\mathbf{h}^i_{g,m}\right)^H \mathbf{A}_i \Omega_i \mathbf{A}_i^H \mathbf{h}^i_{g,m} + \sigma^2_{i,v}. \tag{16}$$

TABLE I: **THE PROPOSED SCH ALGORITHM**

```
INPUT:
    The number of devices and beams: $K$ and $G$;
    The channel correlation threshold: $\sigma$;
    The channel vectors: $\mathbf{h}_k$ for $k = 1, 2, \cdots, K$.
OUTPUT:
    The cluster head set $\Upsilon$.
$\boldsymbol{\mathcal{H}} = [\|\mathbf{h}_1\|_2, \|\mathbf{h}_2\|_2, \cdots, \|\mathbf{h}_K\|_2]$;
$[\sim, O] = sort(\boldsymbol{\mathcal{H}},' descend')$;
$\Upsilon$ = O(1);
$\Upsilon^c$ = O/$\Upsilon$;
$\Theta = \Upsilon^c$;
$g = 2$;
while $g \leq G$ do
  if $\Theta == \Phi$ then
    while $\Theta == \Phi$ do
      $\sigma = \sigma + 0.1 * (1 - \sigma)$;
      $\Theta = \{i \in \Upsilon^c | \varrho_{i,j} < \sigma, \forall j \in \Upsilon\}$.
    end while
  end if
  $\Theta = \{i \in \Theta | \varrho_{i,j} < \sigma, \forall j \in \Upsilon\}$;
  $\Upsilon = \Upsilon \cup \Theta(1)$;
  $\Upsilon^c$ = O/$\Upsilon$;
  $g = g + 1$.
end while
return $\Upsilon$
```

precise control of phase shifters. To tackle this problem, we exploit finite-resolution phase shifters. It is assumed that the quantized phase shifters is B bits, and we can extract the non-zero elements of the analog precoding matrix A as:

$$\mathbf{A}(i,j) \in \mathcal{F}_A = \frac{1}{\sqrt{M}} \left\{ e^{j\frac{2\pi l}{2^B}} : l = 0, 1, \cdots, 2^B - 1 \right\}. \tag{23}$$

For the cluster-head set Υ acquired in the proposed SCH algorithm, the analog precoding can be obtained via the channel vectors of cluster-heads in Υ. Specifically, we first obtain the analog precoding vectors by maximizing the array gain $|\mathbf{h}^H_{\Upsilon(g)} \bar{\mathbf{a}}_g|^2$, where $g = 1, 2, \cdots, G$. Subsequently, we align the phases of the channel vector in the cluster head and the precoding vector, and then quantize the phase of the precoding vector to the nearest phase element in the feasible set $\mathcal{F}_A$ according to the Euclidean distance. To minimize the phase difference of $\mathbf{h}_{\Upsilon(g)}$ and $\bar{\mathbf{a}}_g$, we obtain the ith element of analog precoding vector in the gth cluster $\bar{\mathbf{a}}_g$ via the following formula:

$$\bar{\mathbf{a}}_g(i) = \frac{1}{\sqrt{M}} e^{j\frac{2\pi \hat{l}}{2^B}}, \tag{24}$$

where $i = (g-1)M + 1, (g-1)M + 2, \cdots, gM$, and

$$\hat{l} = \underset{l \in \{0,1,\cdots,2^B-1\}}{argmin} \left| angle\Big(\mathbf{h}_{\Upsilon(g)}(i)\Big) - \frac{2\pi l}{2^B} \right|. \tag{25}$$

C. Equivalent Correlations-Based Device Clustering

Based on the obtained analog precoding matrix, the equivalent channel vectors can be expressed as:

$$\bar{\mathbf{h}}_k^H = \mathbf{h}_k^H \mathbf{A}, \forall k = 1, 2, \cdots, K. \tag{26}$$

The device will be chosen to the cluster when the cluster head and the selected device have the highest correlation of equivalent channels. In particular, we classify the device m ($m \notin \Upsilon$) into the $\hat{g}$th beam based on the following condition

$$\hat{g} = \underset{g \in \{0,1,\cdots,G\}}{argmax} \frac{|\bar{\mathbf{h}}_m^H \bar{\mathbf{h}}_{\Upsilon(g)}|}{\|\bar{\mathbf{h}}_m\|_2 \|\bar{\mathbf{h}}_{\Upsilon(g)}\|_2}. \tag{27}$$

Since the devices in different clusters have low correlations, the interference between beams can be diminished, thereby improving the multiplexing gain.

IV. JOINT DIGITAL PRECODING DESIGN AND POWER ALLOCATION

Based on the acquired analog precoding matrix, we focus on jointly optimizing power allocation and digital precoding for eRRH i. Then, (P1) is transformed as

$$\textbf{(P2):} \quad \max_{\{p^i_{g,m}\},\{\mathbf{d}_{\mathcal{F}g,i}\}} \frac{\sum_{g=1}^{G}\sum_{m=1}^{|S_g|} R^i_{g,m}}{P^i_c + \eta\big(\sum_{g=1}^{G}\sum_{m=1}^{|S_g|} p^i_{g,m}\big)} \tag{28a}$$

$$\text{s.t.} \quad g(\mathbf{d}_{\mathcal{F}g,i}) \leq C_i, \tag{28b}$$

$$\sum_{g=1}^{G}\sum_{m=1}^{|S_g|} p^i_{g,m} \leq P_i, \tag{28c}$$

$$\sum_{g=1}^{G} \|\mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}\|^2 \leq 1. \tag{28d}$$

Since (P2) is non-convex due to the non-convexity of the objective function (28a) as well as constraints (28b), it is difficult to solve directly. Thus, we decouple (P2) into two subproblems, i.e., designing the power allocation with the fixed digital precoding and optimizing the digital precoding with the fixed power allocation. These two non-convex subproblems will be converted to convex optimization problems via the Taylor expansion operation and the minimum mean square error (MMSE) detection, and we solve them iteratively.

A. Power Allocation Optimization

With a fixed digital precoding matrix, the original problem is converted into a power allocation optimization problem. Since the objective function is non-linear fractional, it is difficult to solve directly. To address this problem, the Dinkelbach method [30] is exploited to transform the non-linear fractional problem to a subtractive form, which can facilitate the solutions. Based on [30], we have the following proposition.

TABLE II: **THE PROPOSED EE BASED POWER ALLOCATION ALGORITHM**

```
INPUT:
    The index of iteration: $r = 0$;
    The $r$th iteration's achievable EE: $q_i^{(r)} = 0$;
    The stopping criterion: $\epsilon > 0$.
OUTPUT:
    The final solution of power allocation: $\{p_{g,m}^{i*}\}$;
    The maximum achievable EE: $q_i^*$.
1: REPEAT
2:   For a given $q_i^{(r)}$, Solve problem (34a) - (34b)
     to obtain the power allocation $\{p_{g,m}^{i(r)}\}$;
3:   IF $U_R(\{p_{g,m}^{i(r)}\}) - q_i^{(r)} U_T(\{p_{g,m}^{i(r)}\}) \leq \epsilon$
4:     Convergence = TURE;
5:     RETURN $\{p_{g,m}^{i*}\} = \{p_{g,m}^{i(r)}\}$, $q_i^* = q_i^{(r)}$;
6:   ELSE
7:     Convergence = FALSE;
8:     Set $r = r + 1$ and $q_i^{(r)} = \frac{U_R(\{p_{g,m}^{i(r-1)}\})}{U_T(\{p_{g,m}^{i(r-1)}\})}$;
9:   END IF
10:UNTIL Convergence = TURE.
```

Proposition 1: The maximum achievable EE q_i^* can be obtained as follows

$$\begin{aligned}&\max_{\{p^i_{g,m}\}} \; U_R(\{p^i_{g,m}\}) - q_i^* U_T(\{p^i_{g,m}\}) \\ &= U_R(\{p^{i*}_{g,m}\}) - q_i^* U_T(\{p^{i*}_{g,m}\}) = 0\end{aligned} \tag{29}$$

for $U_R(\{p^i_{g,m}\}) \geq 0$ and $U_T(\{p^i_{g,m}\}) \geq 0$, and

$$U_R(\{p^i_{g,m}\}) = \sum_{g=1}^{G} \sum_{m=1}^{|S_g^i|} \log_2 \left(1 + \gamma^i_{g,m}\right), \tag{30}$$

$$U_T(\{p^i_{g,m}\}) = P_c^i + \eta \biggl(\sum_{g=1}^{G} \sum_{m=1}^{|S_g^i|} p^i_{g,m} \biggr), \tag{31}$$

$$q_i^* = \frac{U_R(\{p^{i*}_{g,m}\})}{U_T(\{p^{i*}_{g,m}\})}, \tag{32}$$

where $\{p^i_{g,m}\}$ is the set of the power allocation for device m in the gth beam, $\forall g = [1, 2, \cdots, G], m = [1, 2, \cdots, |S_g^i|]$.

Proof : Please refer to [30] for a proof of *Proposition 1*. ■

Proposition 1 supplies a compulsory and abundant condition in order to obtain the optimal digital precoding and power allocation. Generally speaking, an equivalent subtractive form-objective function of the optimization problem, i.e., $U_R(\{p^i_{g,m}\}) - q_i^* U_T(\{p^i_{g,m}\})$, can be exploited to replace the original objective function, where two optimization problems have the same solution. Moreover, [30] also implies that the optimal solutions can be acheived based on the equality condition in (29). Therefore, the original optimization problem (P2) is transformed to the equivalent optimization problem according to *Proposition 1*. A power allocation algorithm is proposed to handle the equivalent optimization problem, which is summarized in TABLE II.

In order to demonstrate the feasibility of the EE based power allocation algorithm, it is substantial to prove its convergence. First of all, the EE (i.e., q_i) increases in each iteration. Next, when the number of iterations is large enough, it is evident that q_i converges to the optimal value q_i^*. Notice that q_i^* reaches the optimal conditions in *Proposition 1*, i.e., $U_R(\{p^{i*}_{g,m}\}) - q_i^* U_T(\{p^{i*}_{g,m}\}) = F(q_i^*) = 0$. Assuming that $\{p^{i(r)}_{g,m}\}$ is the optimal power allocation set in the rth iteration, $q_i^{(r)} \neq q_i^*$ and $q_i^{(r+1)} \neq q_i^*$ indicate the EE of the NOMA-based C-RAN system in the rth and the $(r+1)$th iteration respectively. It follows that $F(q_i^{(r)}) > 0$ and $F(q_i^{(r+1)}) > 0$, where $q_i^{(r+1)} = \frac{U_R(\{p^{i(r)}_{g,m}\})}{U_T(\{p^{i(r)}_{g,m}\})}$ [30]. As a consequence, we have the following expression

$$\begin{aligned} F(q_i^{(r)}) &= U_R\left(\{p^{i(r)}_{g,m}\}) - q_i^{(r)} U_T(\{p^{i(r)}_{g,m}\}\right) \\ &= q_i^{(r+1)} U_T(\{p^{i(r)}_{g,m}\}) - q_i^{(r)} U_T(\{p^{i(r)}_{g,m}\}) \\ &= \left(q_i^{(r+1)} - q_i^{(r)}\right) U_T(\{p^{i(r)}_{g,m}\}). \end{aligned} \tag{33}$$

Since $U_T(\{p^{i(r)}_{g,m}\}) = P_c^i + \eta \Bigl(\sum_{g=1}^{G} \sum_{m=1}^{|S_g^i|} p^{i(r)}_{g,m} \Bigr) > 0$ and $F(q_i^{(r)}) > 0$, we can obtain $q_i^{(r+1)} > q_i^{(r)}$. As the number of iterations r is large enough, $q_i^{(r+1)}$ and $F(q_i^{(r+1)})$ are capable of reaching the optimality condition in *Proposition 1*, i.e., $q_i^{(r+1)} \rightarrow q_i^{(r)}$ and $F(q_i^{(r+1)}) \rightarrow 0$ hold [31].

As described in the TABLE II, (P2) is converted to the equivalent optimization problem (P3) according to the fixed q_i (step 2 in TABLE II)

$$\textbf{(P3):} \max_{\{p^i_{g,m}\}} \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} R^i_{g,m} - q_i \left(P_c^i + \eta \Bigl(\sum_{g=1}^{G} \sum_{m=1}^{|S_g|} p^i_{g,m} \Bigr) \right) \tag{34a}$$

$$\text{s.t.} \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} p^i_{g,m} \leq P_i. \tag{34b}$$

For convenience, the channel power gain is defined as

$$G^i_{g,m,j}(\mathbf{h}^i_{g,m}, \mathbf{d}_{\mathcal{F}j,i}) = \left\| (\mathbf{h}^i_{g,m})^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}j,i} \right\|_2^2. \tag{35}$$

As a result, we formulate the data rate $R^i_{g,m}$ as

$$R^i_{g,m} = \widehat{R}^1_{g,m} - \widehat{R}^2_{g,m}, \tag{36}$$

where

$$\begin{aligned} \widehat{R}^1_{g,m} = \log_2 \Biggl(&\sum_{k=1}^{m} G^i_{g,m,g} p^i_{g,k} + \sum_{j \neq g} \sum_{k=1}^{|S_j|} G^i_{g,m,j} p^i_{j,k} \\ &+ (\mathbf{h}^i_{g,m})^H \mathbf{A}_i \Omega_i \mathbf{A}_i^H \mathbf{h}^i_{g,m} + \sigma^2_{i,v} \Biggr), \end{aligned}$$

$$\begin{aligned} \widehat{R}^2_{g,m} = \log_2 \Biggl(&\sum_{k=1}^{m-1} G^i_{g,m,g} p^i_{g,k} + \sum_{j \neq g} \sum_{k=1}^{|S_j|} G^i_{g,m,j} p^i_{j,k} \\ &+ (\mathbf{h}^i_{g,m})^H \mathbf{A}_i \Omega_i \mathbf{A}_i^H \mathbf{h}^i_{g,m} + \sigma^2_{i,v} \Biggr). \end{aligned}$$

Noted that $\widehat{R}^1_{g,m}$ and $\widehat{R}^2_{g,m}$ are both concave with respect to the transmit power $\{p^i_{g,m}\}$. However, $R^i_{g,m}$ is not concave, and hence it is difficult to deal with the non-convex problem (P3). Based on the following *Proposition 2*, $R^i_{g,m}$ can be converted to a convex function by applying the successive convex optimization algorithm.

Proposition 2: The non-convex function $\widehat{R}^2_{g,m}$ can be converted to a convex one as

$$\widehat{R}^2_{g,m} \le \widehat{R}^{2ub}_{g,m} = \log_2(S^{i,r}_{g,m}) + \frac{\log_2(e)}{S^r_{g,m}} \sum_{k=1}^{m-1} G^i_{g,m,g}(p^i_{g,k} - p^{i,r}_{g,k}) + \frac{\log_2(e)}{S^r_{g,m}} \sum_{j\neq g} \sum_{k=1}^{|S_j|} G^i_{g,m,j}(p^i_{j,k} - p^{i,r}_{g,k}),$$

where

$$S^{i,r}_{g,m} = \sum_{k=1}^{m-1} G^i_{g,m,g} p^{i,r}_{g,k} + \sum_{j\neq g} \sum_{k=1}^{|S_j|} G^i_{g,m,j} p^{i,r}_{j,k} + (\mathbf{h}^i_{g,m})^H \mathbf{A}_i \Omega_i \mathbf{A}_i^H \mathbf{h}^i_{g,m} + \sigma^2_{i,v}. \tag{37}$$

Proof: Since $\widehat{R}^2_{g,m}$ is a concave function, its first-order Taylor series expansion at any specific point is its global upper bound [32]. For a given points $\{p^r_{i,j}, \forall i, j\}$, the first-order Taylor expansion operation of $\widehat{R}^2_{g,m}$ can be easily obtained. ■

Based on *Proposition 2*, we can convert (P3) into (P3.1) with the given $\{p^{i,r}_{j,k}, \forall j, k\}$ and q_i

$$\textbf{(P3.1):} \max_{\{p^i_{g,m}\}} \widehat{R}_1 - \widehat{R}^{ub}_2 - q_i^{(r)} \left(P^i_c + \eta (\sum_{g=1}^{G} \sum_{m=1}^{|S_g|} p^i_{g,m}) \right) \tag{38a}$$

$$\text{s.t.} \quad \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} p^i_{g,m} \le P_i, \tag{38b}$$

where

$$\widehat{R}_1 = \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} \widehat{R}^1_{g,m}, \widehat{R}^{ub}_2 = \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} \widehat{R}^{2ub}_{g,m}.$$

Obviously, (P3.1) is convex optimization problem, and we can solve it using the standard convex optimization methods [33]. The proposed power allocation algorithm includes two layers. In particular, the outer iteration repeats until the subtractive form-objective function is less than the stopping criterion ϵ. For the inner iteration, we solve problem (P3) iteratively. Let I_i and I_o represent the maximum number of inner and outer iteration, respectively. Then, the computitional complexity of the proposed dual-layer power allocation algorithm is $I_i I_o V^2$ with the number of dual variables V [31].

B. Digital Precoding Design

After tackling the power allocation subproblem, the goal of this subsection is to optimize the set of digital precoding vector. It is obvious that the denominator of EE is unrelated to digital precoding. Thus, we omit the denominator and problem (P2) is rewritten as

$$\textbf{(P4):} \max_{\{\mathbf{d}_{\mathcal{F}g,i}\}} \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} R^i_{g,m} \tag{39a}$$

$$\text{s.t.} \quad g(\mathbf{d}_{\mathcal{F}g,i}) \le C_i, \tag{39b}$$

$$\sum_{g=1}^{G} \|\mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}\|^2 \le 1. \tag{39c}$$

The objective function is expressed as

$$\max_{\{\mathbf{d}_{\mathcal{F}g}\}} \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} R^i_{g,m} = \sum_{g=1}^{G} \sum_{m=1}^{|S_g|} \log_2(1 + \gamma^i_{g,m}). \tag{40}$$

Nevertheless, (40) is still non-convex. Therefore, an iterative optimization algorithm is developed to tackle the non-convex objective function. In particular, the Sherman-Morrison-Woodbury formula can be presented as follows [34]

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + \mathbf{CDA}^{-1}\mathbf{B})^{-1}\mathbf{CDA}^{-1}. \tag{41}$$

Substituting $(1 + \gamma^i_{g,m})$ into (41), we can obtain $(1 + \gamma^i_{g,m})^{-1}$, which is given by (42) at the top of the next page. In order to solve $s^i_{fg,m}$ in (13), we adopt the minimum mean square error (MMSE) detection [28], which can be denoted as

$$a^{i*}_{g,m} = arg \min_{a^i_{g,m}} e^i_{g,m}, \tag{43}$$

where the mean square error (MSE) is written as

$$e^i_{g,m} = E\left\{ |s^i_{fg,m} - a^i_{g,m} y^i_{g,m}|^2 \right\}, \tag{44}$$

and the channel equalization coefficient is expressed as $a^i_{g,m}$. Then, substituting (13) into (44), we have the following equality

$$e^i_{g,m} = 1 - 2Re\left(a^i_{g,m} (\mathbf{h}^i_{g,m})^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i} \sqrt{p^i_{g,m}} \right) + |a^i_{g,m}|^2 \left(\|(\mathbf{h}^i_{g,m})^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}\|^2_2 p^i_{g,m} + \xi^i_{g,m} \right). \tag{45}$$

Next, substituting (45) into (43), we are capable of receiving the optimal equalization coefficient $a^{i*}_{g,m}$, which is given by (46) at the top of the next page.

Substituting (46) into (45), the closed-form expression of MMSE can be obtained by

$$e^{i*}_{g,m} = 1 - \|(\mathbf{h}^i_{g,m})^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}\|^2_2 p^i_{g,m} \left(\|(\mathbf{h}^i_{g,m})^H \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}\|^2_2 p^i_{g,m} + \xi^i_{g,m} \right)^{-1}, \tag{47}$$

which equals to $(1 + \gamma^i_{g,m})^{-1}$ in (42), and we have

$$(1 + \gamma^i_{g,m})^{-1} = \min_{a^i_{g,m}} e^i_{g,m}. \tag{48}$$

Subsequently, the data rate of the mth device in the gth beam is rewritten as

$$R^i_{g,m} = \log_2(1 + \gamma^i_{g,m}) = \max_{a^i_{g,m}} (-\log_2 e^i_{g,m}). \tag{49}$$

Apparently, the latter expression in (49) has removed the polynomial division $\gamma^i_{g,m}$, which simplifies the objective funtion. However, the log function still exists, which is difficult to solve. Then we have the following *Proposition 3* [7], [9].

Proposition 3: Let $f(b) = -\frac{bc}{\ln 2} + \log_2 b + \frac{1}{\ln 2}$ and b be a positive real number, we have

$$\max_{b>0} f(b) = -\log_2 c, \tag{50}$$

where the optimum value of b is $b^* = \frac{1}{c}$.

By using *Proposition 3*, we can rewrite (49) as

$$R^i_{g,m} = \max_{a^i_{g,m}} \max_{b^i_{g,m}>0} \left(-\frac{b^i_{g,m} e^i_{g,m}}{\ln 2} + \log_2 b^i_{g,m} + \frac{1}{\ln 2} \right). \tag{51}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

$$(1+\gamma_{g,m}^{i})^{-1} = 1 - \|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\|_2^2 p_{g,m}^{i}\left(\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\|_2^2 p_{g,m}^{i} + \xi_{g,m}^{i}\right)^{-1}, \quad (42)$$

$$a_{g,m}^{i*} = \left((\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\sqrt{p_{g,m}^{i}}\right)^{*}\left(\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\|_2^2 p_{g,m}^{i} + \xi_{g,m}^{i}\right)^{-1}. \quad (46)$$

Consequently, we reformulate the optimization problem (P4) as

$$\textbf{(P4.1):} \max_{\{\mathbf{d}_{\mathcal{F}g,i}\}} \sum_{g=1}^{G}\sum_{m=1}^{|S_g|} \max_{a_{g,m}^{i}} \max_{b_{g,m}^{i}>0} \left(-\frac{b_{g,m}^{i}e_{g,m}^{i}}{\ln 2} + \log_2 b_{g,m}^{i}\right) \quad (52a)$$

$$\text{s.t.} \quad g(\mathbf{d}_{\mathcal{F}g,i}) \le C_i, \quad (52b)$$

$$\sum_{g=1}^{G}\|\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\|^2 \le 1. \quad (52c)$$

To tackle problem (P4.1), we employ the iterative optimization method to optimize $\{a_{g,m}^{i}\}, \{b_{g,m}^{i}\}$ and $\{\mathbf{d}_{\mathcal{F}g,i}\}$ separately. In particular, based on the optimal digital precoding vector $\{\mathbf{d}_{\mathcal{F}g,i}^{(l-1)}\}$ in the $(l-1)$th iteration, the optimum solution of $\{a_{g,m}^{i(l)}\}$ in the lth iteration is achieved based on (53) at the top of the next page, where

$$\begin{aligned}\xi_{g,m}^{i(l-1)} = &\sum_{k=1}^{m-1}\left\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}^{(l-1)}\right\|_2^2 p_{g,k}^{i} \\ &+\sum_{j\neq g}\sum_{k=1}^{|S_j|}\left\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}j,i}^{(l-1)}\right\|_2^2 p_{j,k}^{i} \\ &+ (\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\Omega_i\mathbf{A}_i^{H}\mathbf{h}_{g,m}^{i} + \sigma_{i,v}^2.\end{aligned} \quad (54)$$

At the same time, the $b_{g,m}^{i(l)}$ is given by (55) at the top of the next page according to (47), (49) and (50). Then, the objective function can be transformed into a convex function with respect to the digital precoding $\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}$ as follows

$$\min_{\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}} \sum_{g=1}^{G}\sum_{m=1}^{|S_g|} b_{g,m}^{i(l)}e_{g,m}^{i(l)}, \quad (56)$$

where

$$\begin{aligned}e_{g,m}^{i(l)} &= 1 - 2Re(a_{g,m}^{i(l)}(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}^{(l)}\sqrt{p_{g,m}^{i}}) \\ &\quad + |a_{g,m}^{i(l)}|^2\left(\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}^{(l)}\|_2^2 p_{g,m}^{i} + \xi_{g,m}^{i(l)}\right) \\ &= 1 - 2Re(a_{g,m}^{i(l)}(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}^{(l)}\sqrt{p_{g,m}^{i}}) \\ &\quad + |a_{g,m}^{i(l)}|^2\Bigg(\sum_{k=1}^{m}\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}^{(l)}\|_2^2 p_{g,k}^{i} \\ &\quad + \sum_{j\neq g}\sum_{k=1}^{|S_j|}\left\|(\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}j,i}\right\|_2^2 p_{j,k}^{i} \\ &\quad + (\mathbf{h}_{g,m}^{i})^{H}\mathbf{A}_i\Omega_i\mathbf{A}_i^{H}\mathbf{h}_{g,m}^{i} + \sigma_{i,v}^2\Bigg).\end{aligned} \quad (57)$$

For the non-convexity of $g(\mathbf{d}_{\mathcal{F}g,i})$, we exploit the concavity of $\log_2\det(\cdot)$, and obtain $g(\mathbf{d}_{\mathcal{F}g,i}) \le \bar{g}(\mathbf{d}_{\mathcal{F}g,i}, \mathbf{L}_i)$. Here, $\bar{g}(\mathbf{d}_{\mathcal{F}g,i}, \mathbf{L}_i)$ is calculated by

$$\begin{aligned}\bar{g}(\mathbf{d}_{\mathcal{F}g,i},\mathbf{L}_i) \triangleq &\log_2\det(\mathbf{L}_i) - \log_2\det(\mathbf{\Omega}_i) - N_{RF} \\ &+ Tr\left(\mathbf{L}_i^{-1}\left(\sum_{\mathcal{F}_g\subseteq\mathcal{F}_{req}}\bar{c}_{\mathcal{F}_g,i}\bar{\mathbf{d}}_{\mathcal{F}_g,i}\bar{\mathbf{d}}_{\mathcal{F}_g,i}^{H} + \mathbf{\Omega}_i\right)\right).\end{aligned} \quad (58)$$

Constraint (52b) is equal to $\bar{g}(\mathbf{d}_{\mathcal{F}g,i}, \mathbf{L}_i) \le C_i$ when

$$\mathbf{L}_i = \sum_{\mathcal{F}_g\subseteq\mathcal{F}_{req}}\bar{c}_{\mathcal{F}_g,i}\bar{\mathbf{d}}_{\mathcal{F}_g,i}\bar{\mathbf{d}}_{\mathcal{F}_g,i}^{H} + \mathbf{\Omega}_i. \quad (59)$$

Accordingly, constraint (52b) is replaced by $\bar{g}(\mathbf{d}_{\mathcal{F}g,i}, \mathbf{L}_i) \le C_i$. As a result, (P4.1) is reformulated as

$$\textbf{(P4.2):} \min_{\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}} \sum_{g=1}^{G}\sum_{m=1}^{|S_g|} b_{g,m}^{i(l)}e_{g,m}^{i(l)} \quad (60a)$$

$$\text{s.t.} \quad \bar{g}(\mathbf{d}_{\mathcal{F}g,i}, \mathbf{L}_i) \le C_i, \quad (60b)$$

$$\sum_{g=1}^{G}\|\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\|^2 \le 1. \quad (60c)$$

Problem (P4.2) is convex with the fixed $b_{g,m}^{i}$ and $\mathbf{L}_i$. Therefore, we tackle the problem iteratively to obtain a suboptimal solution. In each step, we first update $b_{g,m}^{i}$ and $\mathbf{L}_i$ by using (55) and (59). Then, classical convex optimization methods are utilized to solve the optimal value $\{\mathbf{d}_{\mathcal{F}g,i}\}$ [33], while keeping ($b_{g,m}^{i}$,$\mathbf{L}_i$)'s fixed. In particular, since $b_{g,m}^{i(l)}$, $\mathbf{L}_i^{(l)}$ and $\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}$ are the optimal solution of the lth iteration, updating these variables will increase the value of the objective function in (P4.1). As a consequence, the proposed digital precoding scheme can converge to a local optimum solution. The details of the proposed scheme are discribed in TABLE III. Afterwards, the devices of each beam in each eRRH will be reordered such that $\left\|(\mathbf{h}_{g,1}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2 \ge \left\|(\mathbf{h}_{g,2}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2 \ge \cdots \ge \left\|(\mathbf{h}_{g,|S_g|}^{i})^{H}\mathbf{A}_i\mathbf{d}_{\mathcal{F}g,i}\right\|_2$, $g = 1, 2, \cdots, G$, $i \in \mathcal{K}_R$, which is assumed in Section II for SIC.

At the same time, the proposed digital precoding scheme with the fixed power allocation has a polynomial complexity. In particular, the computational complexity of acquiring $a_{g,m}$ in (53) and $b_{g,m}$ in (55) for each iteration is linear in the number of devices, i.e., $\mathcal{O}(K)$. Similarly, we can prove that the computational complexity of obtaining $\mathbf{L}$ in (59) is $\mathcal{O}(N_{RF}^2)$. In addition, the worst-case complexity of solving (P4.2) is $\mathcal{O}(T_{max}N_{RF}^{4.5}\log_2(1/\varepsilon))$ with a given solution accuracy $\varepsilon > 0$ [35]. In consequence, the computational complexity of the raised digital scheme is $\mathcal{O}(T_{max}N_{RF}^{4.5}\log_2(1/\varepsilon))$.

$$a_{g,m}^{i(l)} = \left((\mathbf{h}_{g,m}^{i})^{H} \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}^{(l-1)} \sqrt{p_{g,m}^{i}} \right)^{*} \left(\|(\mathbf{h}_{g,m}^{i})^{H} \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}^{(l-1)}\|_2^2 p_{g,m}^{i} + \xi_{g,m}^{i(l-1)} \right)^{-1}, \tag{53}$$

$$b_{g,m}^{i(l)} = 1 \Big/ e_{g,m}^{i*} = 1 \Big/ \left(1 - \left\| (\mathbf{h}_{g,m}^{i})^{H} \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}^{(l-1)} \right\|_2^2 p_{g,m}^{i} \left(\left\| (\mathbf{h}_{g,m}^{i})^{H} \mathbf{A}_i \mathbf{d}_{\mathcal{F}g,i}^{(l-1)} \right\|_2^2 p_{g,m}^{i} + \xi_{g,m}^{i(l-1)} \right)^{-1} \right), \tag{55}$$

TABLE III: **THE PROPOSED DIGITAL PRECODING SCHEME FOR (P4)**

```
INPUT:
    The maximum iteration times: $T_{max}$;
    The initialized digital precoding: $\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}$ for $l = 0$;
    The channel vectors: $\mathbf{h}_{g,m}^{i}$ for $g = 1, 2, \cdots, G$,
    $m = 1, 2, \cdots, |S_g^i|$.
OUTPUT:
    The final solution of digital precoding: $\{\mathbf{d}_{\mathcal{F}g,i}^{*}\}$.
REPEAT
    $l = l + 1$;
    Compute auxiliary variables $\mathbf{L}_i^{(l)}$ via (59);
    Compute $a_{g,m}^{i(l)}$ via (53);
    Compute $b_{g,m}^{i(l)}$ via (55);
    Solve the problem (60a)-(60c) to obtain $\{\mathbf{d}_{\mathcal{F}g,i}^{(l)}\}$;
UNTIL Convergence=$T_{max}$
```

C. Joint Digital Precoding Design and Power Allocation

The joint digital precoding design and power allocation strategy for solving the original problem (P2) can be summarized as follows

$$\begin{aligned} &\underbrace{\{p_{g,m}^{i(0)}\},\{\mathbf{d}_{\mathcal{F}g,i}^{(0)}\} \rightarrow q_i^{(0)}}_{\text{Initialization}} \rightarrow \cdots \rightarrow \underbrace{\{p_{g,m}^{i(t)}\}, q_i^{(t)} \rightarrow \{\mathbf{d}_{\mathcal{F}g,i}^{(t)}\}}_{\text{Iteration t}} \\ &\rightarrow \underbrace{\{p_{g,m}^{i(t+1)}\}, q_i^{(t+1)} \rightarrow \{\mathbf{d}_{\mathcal{F}g}^{(t+1)}, i\}}_{\text{Iteration t+1}} \rightarrow \cdots, \end{aligned} \tag{61}$$

where $\{p_{g,m}^{i(t)}\}$, $q_i^{(t)}$ and $\{\mathbf{d}_{\mathcal{F}g,i}^{(t)}\}$ are the set of power allocation, EE and digital precoding in the tth iteration, respectively. In general, the overall procedure of the algorithm for solving (P2) is summarized in TABLE IV. Obviously, the EE q_i increases in each iteration and finally converges to the optimal value. Accordingly, the computational complexity of the joint digital precoding design and power allocation scheme is $\mathcal{O}(I_i I_o T_{max} V^2 N_{RF}^{4.5} \log_2(1/\varepsilon))$.

V. SIMULATION RESULTS

In this section, we provide numerical results to verify the performance of the proposed schemes in NOMA-based C-RANs. The eRRHs are all equipped with an ULA with $N = 64$ antennas. In addition, $N_{RF} = 4$ RF chains are employed to serve $K = 6$ devices simultaneously. All the devices are grouped into $G = N_{RF} = 4$ clusters. As for the channel vector between the kth device and the eRRH i, we set $L_{k,i} = 3$, which includes one LoS path and two NLoS path, and $\alpha_{k,i}^{(1)} \sim \mathcal{CN}(0, 1)$ while $\alpha_{k,i}^{(l)} \sim \mathcal{CN}(0, 10^{-1})$ for $2 \leq l \leq L_{k,i}$. Besides, $\theta_{k,i}^{(l)}$ and $\varphi_{k,i}^{(l)}$ satisfy uniform distribution $\mathcal{U}(-\pi, \pi)$, $1 \leq l \leq L_{k,i}$. The quantized phase shifters are set to 4 bits. The maximum transmit power for all eRRHs is set to $P_i = 10$mW, $\forall i = 1, \cdots, K_R$. The fronthaoul capacity is set to $C_i = 10R_{zf}$ where R_{zf} is the achievable sum-rate through the fully digital zero-forcing (ZF) precoding among all K devices. The probability of the file cached by eRRH i is set to 0.5, and the requested files $\mathcal{F}_{req}$ are randomly generated. In addition, the quantization covariance matrix $\mathbf{\Omega}_i$ is designed to $\upsilon \mathbf{I}_{N_{RF}}$ with $\upsilon = 0.01$. The power consumption of the baseband, per RF chain and power amplifies are set to $P_B = 200$ mW, $P_R = 300$ mW and $P_A = 20$ mW, respectively. The power consumption of phase shifter with 4 bits is $P_P = 40$ mW [36]. The inefficiency of power amplifier is assumed to $\eta = 1/0.38$.

TABLE IV: **THE PROPOSED JOINT DIGITAL PRECODING DESIGN AND POWER ALLOCATION ALGORITHM**

```
INPUT:
    The channel vectors: $\mathbf{h}_{g,m}^{i}$ for $g = 1, 2, \cdots, G$,
    $m = 1, 2, \cdots, |S_g^i|$;
    The stopping criterion: $\delta$;
    The initialized digital precoding: $\{p_{g,m}^{i(t)}\}$ for $t = 0$;
    The initialized power allocation: $\{\mathbf{d}_{\mathcal{F}g,i}^{(t)}\}$ for $t = 0$;
    The calculated $q_i^{(0)}$ based on $\{p_{g,m}^{i(0)}\}$ and $\{\mathbf{d}_{\mathcal{F}g,i}^{(0)}\}$.
OUTPUT:
    The final solution of digital precoding: $\{\mathbf{d}_{\mathcal{F}g,i}^{*}\}$;
    The final solution of power allocation: $\{p_{g,m}^{i*}\}$.
REPEAT:
    $t = t + 1$;
    Solve the P3 under the fixed digital precoding
        $\{\mathbf{d}_{\mathcal{F}g,i}^{(t-1)}\}$ according to TABLE II;
        obtain the set of power allocation $\{p_{g,m}^{i(t)}\}$ and $q_i^{(t)}$;
    Solve the P5 under the fixed power allocation
        $\{p_{g,m}^{i(t)}\}$ in terms of TABLE III;
        obtain the set of digital precoding $\{\mathbf{d}_{\mathcal{F}g,i}^{(t)}\}$;
UNTIL converge, i.e., $|q_i^{(t)} - q_i^{(t-1)}|^2 \leq \delta$.
```

We first study the convergence behavior of the proposed algorithm under different multiple access techniques with sub-connected HP. For comparison, we consider the "Sub-Connected HP-NOMA" scheme, where the sub-onnected HP architecture is adopted in the NOMA-based C-RAN system, and the "Sub-Connected HP-OMA" scheme where the OMA technique is utilized for devices in each beam. As shown in Fig. 3 and Fig. 4, it can be seen that both the "Sub-Connected HP-NOMA" and "Sub-Connected HP-OMA" can converge to

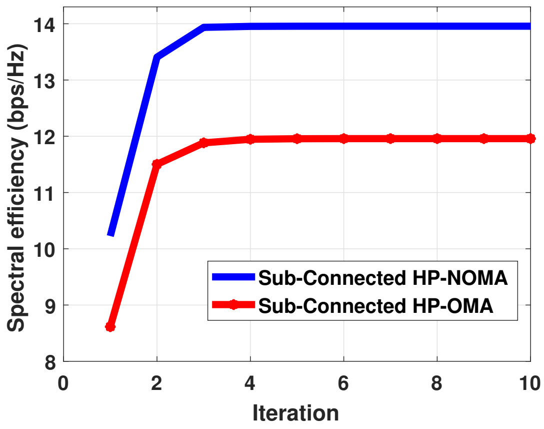

Fig. 3: Convergence in terms of the SE under different multiple access schemes.

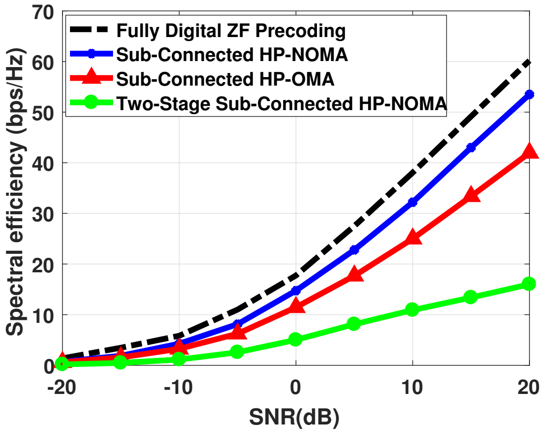

Fig. 5: The SE comparision versus SNR under different precoding schemes.

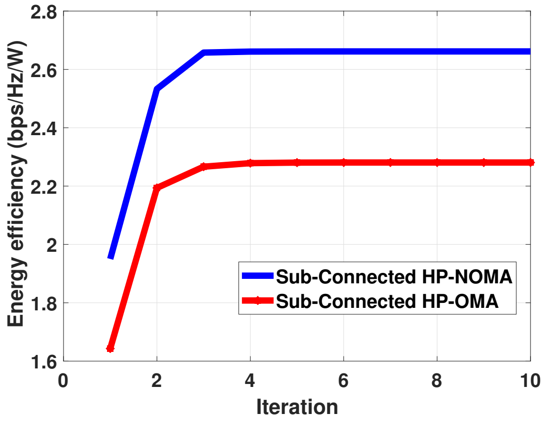

Fig. 4: Convergence in terms of the EE under different multiple access schemes.

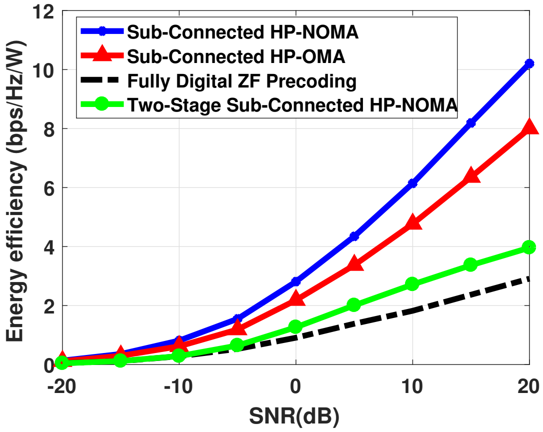

Fig. 6: The EE comparision versus SNR under different precoding schemes.

a fixed value, which verifies the convergence of the proposed hybrid precoding design and power allocation algorithm. In addition, the “Sub-Connected HP-NOMA” achieves better performance than the “Sub-Connected HP-OMA” in terms of both SE and EE. This is because NOMA technique can serve multiple devices in the same resource block, resulting in achieving multi-device diversity, and thereby enhances the SE and EE.

We then investigate the SE and EE of the NOMA-based C-RAN system under different SNR values. To show the SE and EE performance, we compare with the “Fully Digital ZF Precoding” scheme, where each antenna is linked to a RF chain and the ZF technique is applied. Moreover, the “Two-Stage Sub-Connected HP-NOMA” algorithm in [28] is also presented for comparison, which devides the HP design into two step, i.e., analog precoding and digital precoding, and then investigates the power allocation based on the solved HP. As shown in Fig. 5 and Fig. 6, the SE and EE achieved by all schemes are non-decreasing with the SNR value. In addition, our proposed NOMA-based resource allocation scheme achieves higher SE and EE than both the “Sub-Connected HP-OMA” and “Two-Stage Sub-Connected HP-NOMA” schemes. This is due to the fact that NOMA technique enables multiple devices receiving signal from eRRHs simultaneously, and the proposed algorithm optimizes the digital precoding and power allocation iteratively, which thereby converges to at least a local optimal solution. Furthermore, it is apparent that the “Fully Digital ZF Precoding” scheme achieves the highest SE as described in Fig. 5. This is because each antenna is equipped with a dedicated RF chain in order to take full advantage of space diversity. However, the “Fully Digital ZF Precoding” scheme has the lowest EE since the required RF chains consume much more energy. In contrast, the “Sub-Connected HP-NOMA” requires less RF chains than the “Fully Digital ZF Precoding” scheme and thus can achieve higher EE, which verifies the effectiveness of the proposed HP-NOMA scheme.

In the next simulation, we study the EE versus SNR under different quantization bits of phase shifter in the proposed NOMA-based C-RANs. From Fig. 7, it can be seen that the sub-connected HP-NOMA with different bit of phase shifter achieves good EE performance when SNR>-10dB. Besides, the HP-NOMA with lower quantization bits can achieve higher EE when SNR≤15dB, but the phase shifter of 3-bit has higher EE performance than that of 2-bit for the case when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

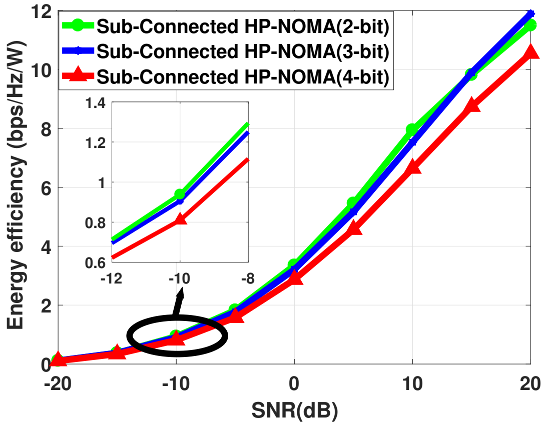

Fig. 7: The EE comparision of proposed HP-NOMA scheme versus SNR under different quantization bits of phase shifter.

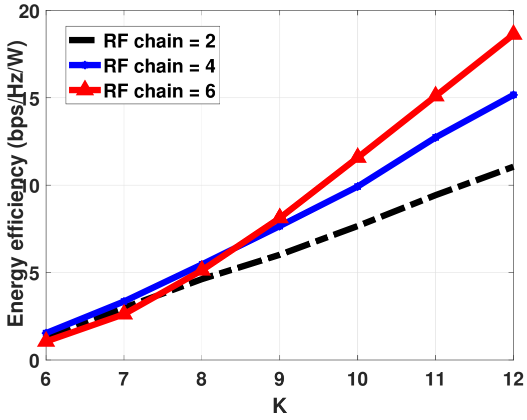

Fig. 8: The EE comparision of proposed HP-NOMA scheme versus K under different number of RF chains.

SNR>15dB. This is because the SE enhancement caused by the higher SNR can compensate for the energy consumption caused by 3-bit phase shifter, and hence a higher EE can be obtained.

Finally, the EE versus the number of devices under different number of RF chains for the NOMA-based C-RANs is evaluated. For sub-connected structures, since the number of antennas must be an integer multiple of the number of RF chains, we set $N = 72$. Fig. 8 shows that the EE is monotonically non-decreasing with respect to the number of devices. It can also be seen that increasing the number of RF chains can enhance EE since it can achieve higher SE. However, the sub-connected HP-NOMA scheme with 6 RF chains achieves the worst EE performance when K≤8. This is due to the fact that the improvement of SE cannot compensate for the extra energy consumption of additional RF chains. Therefore, we conclude that there exists a trade-off between EE performance and the number of RF chains, especially for the system with less devices.

VI. Conclusions

This paper investigated the EE performance for the NOMA-based C-RANs. To maximize the EE, we jointly optimized the power allocation, analog and digital precoding under the constraints of the fronthaul link capacity and the total transmit power. Since the formulated problem is non-convex, we proposed a hybrid precoding design and power allocation algorithm to decompose the original problem into three sub-problems, and optimize analog precoding, devices clustering, digital precoding and power allocation sequentially. Firstly, the analog precoding was optimized by maximizing the array gain. Next, we grouped devices based on the selected cluster heads. Then, a joint digital precoding design and power allocation algorithm was proposed to further improve the EE. Simulation results indicated that our proposed schemes can achieve faster convergence. In addition, the EE performance can be significantly improved by our proposed algorithms compared to the traditional OMA-based approach and two-stage HP scheme. In the future, we will consider more sophisticated hybrid precoding design for the NOMA-based C-RANs to further improve the performance. Besides, the analysis of the effect of imperfect CSI is an interesting topic for future work.

References

[1] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, “Industrial internet of things: Challenges, opportunities, and directions,” *IEEE Trans. Industr. Inform.*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018.

[2] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, “Millimeter-wave massive MIMO: The next wireless revolution?” *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[3] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. Abdul Latiff, and N. F. Za’bah, “On the spectral-efficiency of low-complexity and resolution hybrid precoding and combining transceivers for mmwave MIMO systems,” *IEEE Access*, vol. 7, pp. 109 259–109 277, Aug. 2019.

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

[5] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[6] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[7] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, “Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[8] W. Yuan, V. Kalokidou, S. M. D. Armour, A. Doufexi, and M. A. Beach, “Application of non-orthogonal multiplexing to mmwave multi-user systems,” in *2017 IEEE 85th VTC Spring*, Nov. 2017, pp. 1–6.

[9] Q. Zhang, Q. Li, and J. Qin, “Robust beamforming for nonorthogonal multiple-access systems in MISO channels,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 231–10 236, Dec. 2016.

[10] Z. Ding, M. Peng, and H. V. Poor, “Cooperative non-orthogonal multiple access in 5G systems,” *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[11] Z. Ding, F. Adachi, and H. V. Poor, “The application of MIMO to non-orthogonal multiple access,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[12] M. Peng, S. Yan, K. Zhang, and C. Wang, “Fog-computing-based radio access networks: Issues and challenges,” *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul. 2016.

[13] H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang, “Mobile edge cloud system: Architectures, challenges, and approaches,” *IEEE Syst. J.*, vol. 12, no. 3, pp. 2495–2508, Sep. 2018.

[14] X. Peng, J. Shen, J. Zhang, and K. B. Letaief, “Backhaul-aware caching placement for wireless networks,” *IEEE Global Commun. Conf.*, pp. 1–6, Dec. 2015.

[15] J. Kwak, Y. Kim, L. B. Le, and S. Chong, “Hybrid content caching in 5G wireless networks: Cloud versus edge caching,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3030–3045, May. 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3138461, IEEE Internet of Things Journal

[16] R. Wang, Z. Kan, Y. Cui, D. Wu, and Y. Zhen, "Cooperative caching strategy with content request prediction in internet of vehicles," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8964–8975, Jun. 2021.

[17] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[18] T. S. Rappaport, S. Shu, R. Mayzus, Z. Hang, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, no. 1, pp. 335–349, May 2013.

[19] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May. 2018.

[20] S. He, J. Wang, W. Huang, Y. Huang, M. Xiao, and Y. Zhang, "Energy-efficient transceiver design for cache-enabled millimeter-wave systems," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3876–3889, Jun. 2020.

[21] D. Fan, F. Gao, B. Ai, G. Wang, Z. Zhong, Y. Deng, and A. Nallanathan, "Channel estimation and self-positioning for UAV swarm," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7994–8007, Nov. 2019.

[22] W. Feng, J. Tang, N. Zhao, X. Zhang, X. Wang, K.-K. Wong, and J. Chambers, "Hybrid beamforming design and resource allocation for UAV-aided wireless-powered mobile edge computing networks with NOMA," *IEEE J. Select. Areas Commun.*, pp. 1–1, Jun. 2021.

[23] X. Wen, H. Zhang, H. Zhang, and F. Fang, "Interference pricing resource allocation and user-subchannel matching for NOMA hierarchy fog networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 467–479, Jun. 2019.

[24] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, "Energy-efficient resource allocation in NOMA heterogeneous networks," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, Apr. 2018.

[25] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1351–1364, Feb. 2019.

[26] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. M. Leung, "Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems," *IEEE J. Select. Areas Commun.*, vol. 38, no. 9, pp. 2074–2085, Sep. 2020.

[27] C. G. Tsinos, S. Maleki, S. Chatzinotas, and B. Ottersten, "On the energy-efficiency of hybrid analog–digital transceivers for single- and multi-carrier large antenna array systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1980–1995, Sep. 2017.

[28] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[29] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[30] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, pp. 492–498, Mar. 1967.

[31] J. Tang, J. Luo, M. Liu, D. K. C. So, E. Alsusa, G. Chen, K. Wong, and J. A. Chambers, "Energy efficiency optimization for NOMA with SWIPT," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 452–466, Jun. 2019.

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

[33] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.2," Jan. 2020. [Online]. Available: http://cvxr.com/cvx.

[34] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. New York, NY, USA: Wiley, 1988.

[35] 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.

[36] X. Gao, L. Dai, Y. Sun, S. Han, and C.-L. I, "Machine learning inspired energy-efficient hybrid precoding for mmwave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (IEEE ICC)*, May 2017, pp. 1–6.