Cloud-Edge Coordinated Processing: Low-Latency Multicasting Transmission

Shiwen He, *Member, IEEE*, Ju Ren, *Member, IEEE*, Jiaheng Wang, *Senior Member, IEEE*, Yongming Huang, *Senior Member, IEEE*, Yaoxue Zhang, *Senior Member, IEEE*, Weihua Zhuang, *Fellow, IEEE*, and Sherman (Xuemin) Shen, *Fellow, IEEE*

Abstract

Recently, edge caching and multicasting arise as two promising technologies to support high-data-rate and low-latency delivery in wireless communication networks. In this paper, we design three transmission schemes aiming to minimize the delivery latency for cache-enabled multigroup multicasting networks. In particular, full caching bulk transmission scheme is first designed as a performance benchmark for the ideal situation where the caching capability of each enhanced remote radio head (eRRH) is sufficient large to cache all files. For the practical situation where the caching capability of each eRRH is limited, we further design two transmission schemes, namely partial caching bulk transmission (PCBT) and partial caching pipelined transmission (PCPT) schemes. In the PCBT scheme, eRRHs first fetch the uncached requested files from the baseband unit (BBU) and then all requested files are simultaneously transmitted to the users. In the PCPT scheme, eRRHs first transmit the cached requested files while fetching the uncached requested files from the BBU. Then, the remaining cached requested files and fetched uncached requested files are simultaneously transmitted to the users. The design goal of the three transmission schemes is to minimize the delivery latency, subject to some practical constraints. Efficient algorithms are developed for the low-latency cloud-edge coordinated transmission strategies. Numerical results are provided to evaluate the performance of the proposed

S. He, J. Ren, and Y. Zhang are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China. They are also with the School of information Technology, Jiangxi University Of Finance and Economics, West Yuping Road, Nanchang 330032, China. (email: {shiwen.he.hn, renju, zyx}@csu.edu.cn).

J. Wang and Y. Huang are with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China. (email: {jhwang, huangym}@seu.edu.cn).

S. Shen and W. Zhuang are with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada N2L 3G1. (email: {sshen, wzhuang}@uwaterloo.ca).

transmission schemes and show that the PCPT scheme outperforms the PCBT scheme in terms of the delivery latency criterion.

Index Terms

Cache-enabled radio access networks, delivery latency, multigroup multicasting, non-convex optimization.

I. Introduction

Driven by the visions of ultra-high-definition video, intelligent driving, and Internet of Things, high-data-rate and low-latency delivery become two key performance indicators of future wireless communication networks [1]. The vast resources available in the cloud radio access server can be leveraged to deliver elastic computing power and storage to support resource-constrained end-user devices [2]. However, it is not suitable for a large set of cloud-based applications such as the delay-sensitive ones, since end devices in general are far away from the central cloud server, i.e., data center [3], [4]. To overcome these drawbacks, caching popular content at the network edge during the off-peak period is proved to be a powerful technique to realize low-latency delivery for some specific applications, such as real-time online gaming, virtual reality, and ultra-high-definition video streaming, in next-generation communication systems [5]–[7]. Consequently, an evolved network architecture, labeled as cache-enabled radio access networks (RANs), has emerged to satisfy the demands of ultra-low-latency delivery by migrating the computing and caching functions from the cloud to the network edge [7]–[9]. In the cache-enabled RANs, the cache-enabled radio access nodes named as enhanced remote radio heads (eRRHs) have the ability to enable processing at the network edge and to cache files at its local cache [9]–[11].

A. Related Works

The main motivation of caching frequently requested content at the network edge is to reduce the burden on the fronthaul links and the delivery latency. Existing studies on the cache-enabled RANs are mainly on two-fold, i.e., the pre-fetching phase and the delivery phase. The pre-fetching phase studies focus on caching strategies, while accounting for the caching capacity of eRRHs, popularity of contents and user distribution [12]–[18]. The delivery phase deals with the requested data transmission for different performance criteria [19]–[25].

1) Optimization of content placement: Content placement with a finite cache size is the key issue in caching design, since unplanned caching at the network edge will result in more inter-cell interference or delivery latency. Therefore, how to effectively cache the popular content at the network edge has attracted extensive attention from both academia and industry. The cache placement problem in cache-enabled RANs is investigated while accounting for the flexible physical-layer transmission and diverse content preferences of different users [12]. Hybrid caching together with relay clustering is studied to strike a balance between the signal cooperation gain achieved by caching the most popular contents and the largest content diversity gain by caching different contents [13]. In [14]–[18], an edge caching strategy is investigated for cache-enabled coordinated heterogeneous networks. Probabilistic content placement is designed to maximize the performance of content delivery for cache-enabled multi-antenna dense small cell network [14]. Dynamic content caching is studied via stochastic optimization for a hierarchical caching system consisting of original content servers, central cloud units and base stations (BSs) [15]. The successful transmission probabilities are analyzed for a two-tier large-scale cache-enabled wireless multicasting network [16], [17]. Proactive caching strategies are proposed to reduce the backhaul transmission for large-scale mobile edge networks [18]. Though efficient caching at the network edge can effectively reduce the burden on the fronthaul links, how to effectively design a transmission strategy is key problem, especially for content-centric ultra-dense massive-access networks.

2) Optimization of transmission strategy: How to timely transmit the cached and uncached requested files to users is another key problem for cache-enabled coordinated RANs [19]–[25]. Joint optimization of cloud-edge coordinated precoding using different pre-fetching strategies is investigated respectively for cache-enabled sub-6 GHz and millimeter-wave multi-antennas multiuser RANs in [19], [20]. He *et al.* propose a two-phase transmission scheme to reduce both burden on the fronthaul links and delivery latency with a fixed delay caused by fronthaul links for cache-enabled RANs [21]. Tao *et al.* investigate the joint design of multicast beamforming and eRRH clustering for the delivery phase with fixed pre-fetching to minimize the compound cost including the transmit power and fronthaul cost, subject to predefined delivery rate requirements [22]. Research on the energy efficiency of cache-enabled RANs shows that caching at the BSs can improve the network energy efficiency when power efficient cache hardware is used [23], [24]. Studies on cache-enabled physical layer security have shown that caching

can reduce the burden on fronthaul links, introduce additional secure degrees of freedom, and enable power-efficient communication [25]. However, maximizing the (minimum) delivery rate or minimizing the compound cost function may not necessarily minimize the delivery latency, especially when partial requested contents are not cached at the network edge.

B. Contributions and Organization

In general, the delivery latency is incurred at least by the propagation of fronthaul links, signal processing at the BBU, and signal transmission for wireless communication systems. Furthermore, the limited capacity of fronthaul links is a key factor in determining the delivery latency and is the key motivation for mitigating the cache and baseband signal processing to the network edge. However, to the best of the authors' knowledge, how to effectively exploit the cache and baseband signal processing at the network edge to minimize the delivery latency is an open problem for cache-enabled multi-antennas multigroup multicasting RANs with limited capacities of fronthaul links. In this paper, we study the minimization of delivery latency of three different transmission schemes, accounting for the delay caused by fetching the uncached requested files from the BBU and the signal processing at the BBU for cache-enabled multi-antennas multigroup multicasting RANs. The main contributions of this paper can be summarized as follows:

- When the caching capability of each eRRH is sufficient large to cache all files, a full caching bulk transmission (FCBT) scheme is proposed as a performance benchmark for minimizing the delivery latency;
- In practice, only a part of files is cached at network edge due to the limited caching capability of each eRRH. For this case, we first present a partial caching bulk transmission (PCBT) scheme that transmits simultaneously all requested files to the users and then a novel partial caching pipelined transmission (PCPT) scheme to further reduce the delivery latency;
- Three optimization problems are formulated to minimize the delivery latency that includes the delay caused by fetching the uncached files from the BBU, signal processing at the BBU, and transmitting all requested files to the users, subject to constraints on the fronthaul capacity, per-eRRH transmit power constraint, and file size;
- An efficient algorithm that is proved to converge to a Karush-Kuhn-Tucker (KKT) solution is designed to solve each of optimization problems, respectively;

- Numerical results are provided to validate the effectiveness of the proposed methods. Compared to the other non-FCBT transmission schemes, the PCPT scheme achieves obvious performance improvement in terms of delivery latency.

The remainder of this paper is organized as follows. The system model is described in Section II. In Section III, three transmission schemes are formulated. Design of optimization algorithms are investigated in Section IV. Numerical results are presented in Section V and conclusions are drawn in Section VI.

Notations: Bold lower case and upper case letters represent column vectors and matrices, respectively; $\mathrm{diag}\left(\mathbf{a}\right)$ is a diagonal matrix whose diagonal elements are the elements of vector $\mathbf{a}$; $\mathbf{0}_{N\times N}$ and $\mathbf{I}_{N\times N}$ denote the $N\times N$ zero and identity matrices, respectively; $\mathrm{tr}\left(\cdot\right)$, $\|\cdot\|_2$, and $\|\cdot\|_{\mathrm{F}}$ denote the trace, the Euclidean norm, and the Frobenius norm, respectively. The circularly symmetric complex Gaussian distribution with mean $\mathbf{u}$ and covariance matrix $\mathbf{R}$ is denotes by $\mathcal{CN}\left(\mathbf{u},\mathbf{R}\right)$; $\mathbf{A}\succeq\mathbf{0}$ is a positive semidefinite matrix; $\left[\mathbf{A}\right]_{(m,n)}$ represents the element in row m and column n of matrix $\mathbf{A}$, and $\mathrm{vec}\left(\mathbf{A}\right)$ denotes the column vector obtained by stacking the columns of matrix $\mathbf{A}$ on top of one another. Superscripts $\left(\cdot\right)^T$, $\left(\cdot\right)^*$, and $\left(\cdot\right)^H$ represent transpose, conjugate, and conjugate transpose operators, respectively. For set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of the set, while for complex number x, $|x|$ denotes the magnitude value of x; $\mathbb{R}$ and $\mathbb{C}$ are the fields of real and complex numbers, respectively. Function $\lfloor x\rfloor$ rounds x to the nearest integer not larger than x; $\overline{a}$ denotes the complement $1-a$ of a binary variable $a\in\{0,1\}$; and $\ln\left(\cdot\right)$ is the logarithm with base e. The circularly symmetric complex Gaussian distribution with mean μ and covariance matrix $\mathbf{R}$ is denoted by $\mathcal{CN}\left(\mu,\mathbf{R}\right)$. The symbols used frequently in this paper are summarized in Table I.

II. System Model

Consider the downlink transmission of a cache-enabled multigroup multicasting RAN, as illustrated in Fig. 1, comprising one baseband unit (BBU), K_{R} eRRHs, and K_{U} single-antenna users. In the system, eRRH $i\in\mathcal{K}_{\mathrm{R}}=\{1,\cdots,K_{\mathrm{R}}\}$ is equipped with a cache that can store B_i nats, where B_i is the normalized cache size; eRRH i is equipped with N_{t} antennas and connected to the BBU through an error-free fronthaul link with normalized capacity C_i nats/Hz/s. The user set $\mathcal{K}_{\mathrm{U}}=\{1,\cdots,K_{\mathrm{U}}\}$ is partitioned into G multicast groups, denoted by $\mathcal{G}_1$, $\cdots$, $\mathcal{G}_G$. Each

TABLE I. List of important mathematical symbols.

Symbol	Meaning	Symbol	Meaning
K_{U}, K_{R}	Numbers of users and eRRHs	$\mathcal{K}_{\mathrm{U}}$, $\mathcal{K}_{\mathrm{R}}$	Sets of users and eRRHs
N_{t}	Numbers of antennas equipped at each eRRH	P_i	Maximum transmit power of eRRH i
C_i	Capacity of fronthaul link to eRRH i	B_i	Normalized cache size of eRRH i
G	Number of multicast groups	G_g	g-th group
F	Number of files in the library	$\mathcal{F}$, $\mathcal{F}_{\mathrm{req}}$	Sets of all files and all requested files
S	Normalized size of files	f_g	Index of file requested by the g-th group
$c_{f,i}$	Binary caching variable of file f at eRRH i	$\overline{c}_{f,i}$	Complement of $\overline{c}_{f,i}$
$\mathbf{y}_k$	Signal received by user k	$\mathbf{h}_{k,i}$	Channel matrix from eRRH i to user k
σ_k^2	Noise variance at user k	$\mathcal{V}$	Set of beamforming vector $\mathbf{v}_{g,i}$
$\mathbf{\Omega}_i$	Quantization noise covariance matrix of eRRH i	$\mathcal{O}$	Set of quantization noise covariance matrices
$\gamma_{k,i}$	SINR of user k for full caching case or partial caching case		
$r_{k,i}$	Achievable rate of user k for full caching case or partial caching case, $i = 1, 2$		
$\mathbf{w}_{g,i}$	Beamforming vector for basedband signal $\mathbf{s}_{f_g}$ at eRRH i for full caching cse		
$\mathbf{u}_{g,i}$	Beamforming vector for cached basedband signal $\mathbf{s}_{f_g}$ at eRRH i for partial caching case		
$\mathbf{v}_{g,i}$	Beamforming vector for uncached basedband signal $\mathbf{s}_{f_g}$ at eRRH i for partial caching case		
$\mathbf{h}_k$, $\mathbf{w}_g$	$\mathbf{h}_k = \left[\mathbf{h}_{k,1}^H, \cdots, \mathbf{h}_{k,K_{\mathrm{R}}}^H\right]^H$, $\mathbf{w}_g = \left[c_{f_g,1}\mathbf{w}_{g,1}^H, \cdots, c_{f_g,K_{\mathrm{R}}}\mathbf{w}_{g,K_{\mathrm{R}}}^H\right]^H$		
$\overline{\mathbf{w}}_k$, $\mathbf{u}_g$, $\mathbf{u}_g$	$\overline{\mathbf{w}}_g = \mathbf{u}_g + \mathbf{v}_g$, $\mathbf{u}_g = \left[c_{f_g,1}\mathbf{u}_{g,1}^H, \cdots, c_{f_g,K_{\mathrm{R}}}\mathbf{u}_{g,K_{\mathrm{R}}}^H\right]^H$, $\mathbf{v}_g = \left[\overline{c}_{f_g,1}\mathbf{v}_{g,1}^H, \cdots, \overline{c}_{f_g,K_{\mathrm{R}}}\mathbf{v}_{g,K_{\mathrm{R}}}^H\right]^H$		

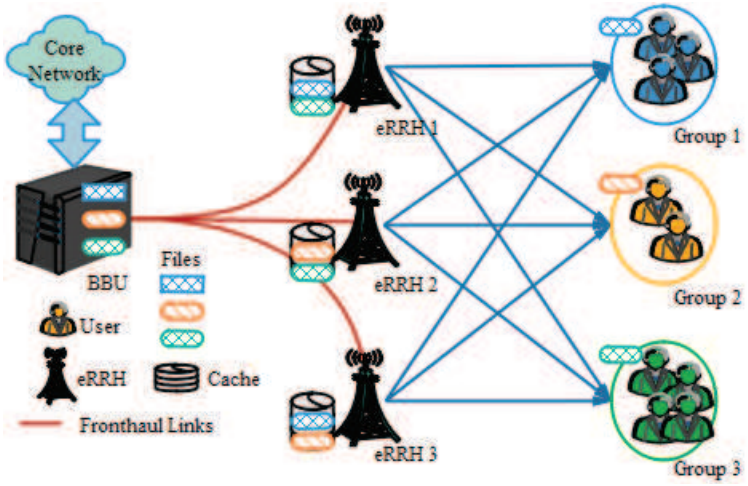

Fig. 1. Illustration of downlink transmission of a cache-enabled multigroup multicasting RAN.

user, $k \in \mathcal{K}_{\mathrm{U}}$, independently requests only a single file in a given transmission interval, i.e., each user belongs to at most one multicast group.

Without loss of generality, we assume that all files in library $\mathcal{F} = \{1, \cdots, F\}$ at the BBU

have the same size of S nats/Hz, where S is the normalized file size. The assumption of equal file sizes is standard and reasonable in that the most frequently requested and cached files by users are chunks of videos, e.g., fragments of a given duration, which are often segmented with the same length [6]. In general, eRRH $i \in \mathcal{K}_{\mathrm{R}}$ selectively pre-fetches some popular files from library $\mathcal{F}$ to its local cache during the off-peak period, according to the content popularity and predefined caching strategies [19], [22]. The cache status of file $f \in \mathcal{F}$ can be modeled as a binary variable $c_{f,i}$, $f \in \mathcal{F}$, $i \in \mathcal{K}_{\mathrm{R}}$, given by

$$c_{f,i} = \begin{cases} 1, & \text{if file } f \text{ is cached by eRRH } i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The complement of $c_{f,i}$ is denoted as $\overline{c}_{f,i} = 1 - c_{f,i}$. In this work, we focus on transmission strategies given cache state information, i.e., $c_{f,i}$, $f \in \mathcal{F}$, $i \in \mathcal{K}_{\mathrm{R}}$. Let $\mathcal{F}_{\mathrm{req}} = \cup_{g \in \mathcal{G}} \{f_g\} \subseteq \mathcal{F}$ be the index set of requested files of all user groups, where f_g is the index of the requested file by the users in group $\mathcal{G}_g$. We denote the respective group index of user k by a positive integer g_k.

III. Design of Transmission Schemes

In this section, we consider three transmission schemes according to the caching capabilities of eRRHs and then formulate the corresponding optimization problems for cache-enabled multi-group multicasting RANs. In particular, the design objective is to minimize the delivery latency subject to the constraints on the fronthaul links, eRRH transmit power, and file size.

A. Full Caching Bulk Transmission (FCBT) Scheme

In this subsection, we assume that all files are cached at the local cache, i.e., $c_{f,i} = 1, \forall i \in \mathcal{K}_{\mathrm{R}}$, $\forall f \in \mathcal{F}$. Consequently, all requested files can be directly retrieved from the local caches of eRRHs and transmitted to the users[1]. We refer to this transmission scheme as full caching bulk transmission (FCBT) scheme, which is similar with the coordinated multiple points (CoMP) mechanism in wireless communication systems [26].

[1]The reason of considering full caching is to provide a benchmark for performance comparison. In practical communication networks, eRRH cannot cache all requested files even if it has sufficient large caching capacity due to the diversity of files, massive users, and mobility of users.

In the FCBT scheme, the users first send the requirement of files to the eRRHs and then the eRRHs coordinately transmit the requested files to the users. Consequently, the received baseband signal at user k can be expressed as

$$y_{k,1} = \sum_{g \in \mathcal{G}} \mathbf{h}_k^H \mathbf{w}_g s_{f_g} + n_k \tag{2}$$

where $\mathbf{w}_g = \left[c_{f_g,1}\mathbf{w}_{g,1}^H, \cdots, c_{f_g,K_\mathrm{R}}\mathbf{w}_{g,K_\mathrm{R}}^H\right]^H$, with $\mathbf{w}_{g,i} \in \mathbb{C}^{N_\mathrm{t}\times 1}$ being the beamforming vector for the users in $\mathcal{G}_g$ at eRRH i, $\mathbf{h}_k = \left[\mathbf{h}_{k,1}^H, \cdots, \mathbf{h}_{k,K_\mathrm{R}}^H\right]^H$ with $\mathbf{h}_{k,i} \in \mathbb{C}^{N_\mathrm{t}\times 1}$ denoting the channel coefficient between user k and eRRH i, s_{f_g} represents the baseband signal of requested file f_g for the users in $\mathcal{G}_g$, and n_k denotes the additive white Gaussian noise with $\mathcal{CN}\left(0, \sigma_k^2\right)$. Thus, the signal-to-interference-plus-noise ratio (SINR) of user k is calculated as

$$\gamma_{k,1} = \frac{\left|\mathbf{h}_k^H \mathbf{w}_{g_k}\right|^2}{\sum\limits_{g \in \mathcal{G}\setminus\{g_k\}} \left|\mathbf{h}_k^H \mathbf{w}_g\right|^2 + \sigma_k^2}. \tag{3}$$

Based on Shannon capacity, the corresponding achievable rate on unit bandwidth in one second of user k is given by $R_{k,1} = \ln\left(1 + \gamma_{k,1}\right)$ in unit of nats/Hz/s. Let $r_{g,1}$ in unit of nats/Hz/s be the delivery rate of group g for the cached files transmission. We consider minimizing the delivery latency and meanwhile providing fairness for all multicast groups. To achieve this two-fold goal, the design problem is formulated as

$$\min \max_{g \in \mathcal{G}} \frac{S}{\mathrm{r}_{g,1}} \tag{4a}$$

$$\text{s.t. } \mathrm{r}_{g,1} \le \mathrm{R}_{k,1}, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{4b}$$

$$\sum_{g \in \mathcal{G}} \left\|c_{f_g,i}\mathbf{w}_{g,i}\right\|^2 \le P_i, \forall i \in \mathcal{K}_\mathrm{R} \tag{4c}$$

where the optimization variables are $\mathbf{w}_g$ and $\mathrm{r}_{g,1}$, $\forall g \in \mathcal{G}$. Constraints (4b) means that the delivery rate of the file requested by user k is constrained by the achievable rate. Constraints (4c) is the power constraint per eRRH. The goal of problem (4) is to minimize the maximum group delivery latency, as the accomplishment of the transmission of requested files is determined by the worst group for the whole data transmission.

B. Partial Caching Bulk Transmission Scheme

In practice, each eRRH can only cache a fractional of files at its local cache due to the limited caching capabilities of eRRHs. As a result, some requested files may not be cached at the cache

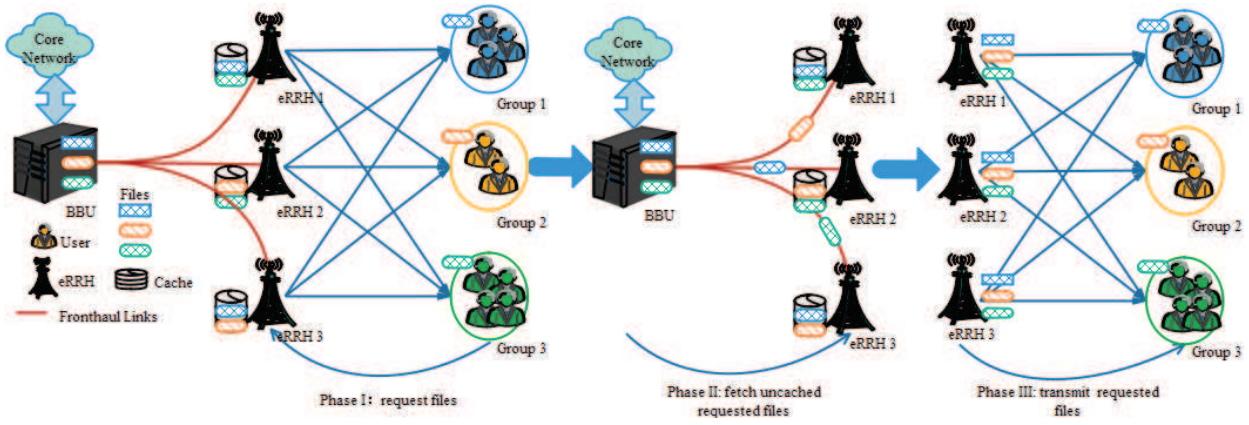

Fig. 2. Flowchart of partial caching buck transmission scheme.

of the network edge. In this subsection, we explore the problem of minimizing the delivery latency for the case that only a fractional of requested files are cached at the local cache of eRRHs. In this case, a common transmission scheme contains three phases [19], as illustrated in Fig. 2, which is termed as partial caching bulk transmission (PCBT) scheme. In particular, in phase I, the users first send the requirement of files to the eRRHs. Because only partial required files are cached at the local cache, the eRRHs fetch the uncached requested files from the BBU in phase II. After obtaining the uncached requested files, in phase III, the eRRHs coordinately transmit the cached and uncached requested files to the users.

The uncached files in the BBU are quantized and precoded and then delivered to the eRRHs via the fronthaul links. Let $\widetilde{\mathbf{x}}_i$ denote the precoded signal of the requested files that are not stored at eRRH i, which is given by

$$\widetilde{\mathbf{x}}_i = \sum_{g \in \mathcal{G}} \overline{c}_{f_g,i} \mathbf{v}_{g,i} \mathbf{s}_{f_g} \tag{5}$$

where, $\mathbf{v}_{g,i} \in \mathbb{C}^{N_\mathrm{t} \times 1}$ is the beamforming vector for the uncached requested file f_g at eRRH i. Let $\widehat{\mathbf{x}}_i = \widetilde{\mathbf{x}}_i + \mathbf{z}_i$ be the quantized version of precoded signal $\widetilde{\mathbf{x}}_i$ at the BBU, where $\mathbf{z}_i \in \mathbb{C}^{N_\mathrm{t} \times 1}$ is the quantization noise independent of $\widetilde{\mathbf{x}}_i$ with distribution $\mathbf{z}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$. We assume that the quantization noise $\mathbf{z}_i$ is independent across the eRRHs, i.e., the signals intended for different eRRHs are quantized independently [27]. Let $\mathbf{\Omega}$ be the covariance matrix of quantization noise, i.e., $\mathbf{\Omega} = \mathrm{Diag}(\mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_{K_\mathrm{R}})$.

The signal $\mathbf{x}_i$ transmitted by eRRH i is a superposition of two signals, where one is the

locally precoded signal of the cached requested files and the other is the precoded and quantized signal of the uncached requested files stored at the BBU, which is delivered to eRRH i via the error-free fronthaul link. Therefore, we have

$$\mathbf{x}_i = \sum_{g\in\mathcal{G}} c_{f_g,i}\mathbf{u}_{g,i}\mathbf{s}_{f_g} + \widehat{\mathbf{x}}_i = \sum_{g\in\mathcal{G}} c_{f_g,i}\mathbf{u}_{g,i}\mathbf{s}_{f_g} + \sum_{g\in\mathcal{G}} \overline{c}_{f_g,i}\mathbf{v}_{g,i}\mathbf{s}_{f_g} + \mathrm{z}_i \tag{6}$$

where $\mathbf{u}_{g,i} \in \mathbb{C}^{N_\mathrm{t}\times 1}$ denotes the beamforming vectors for the cached requested file f_g at eRRH i. The received baseband signal at user k is expressed as

$$y_{k,2} = \sum_{g\in\mathcal{G}} \mathbf{h}_k^H \overline{\mathbf{w}}_g s_{f_g} + \mathbf{h}_k^H \mathbf{z} + n_k \tag{7}$$

where $\overline{\mathbf{w}}_g = \mathbf{u}_g + \mathbf{v}_g$, $\mathbf{u}_g = \left[c_{f_g,1}\mathbf{u}_{g,1}^H, \cdots, c_{f_g,K_\mathrm{R}}\mathbf{u}_{g,K_\mathrm{R}}^H\right]^H$, $\mathbf{v}_g = \left[\overline{c}_{f_g,1}\mathbf{v}_{g,1}^H, \cdots, \overline{c}_{f_g,K_\mathrm{R}}\mathbf{v}_{g,K_\mathrm{R}}^H\right]^H$, and $\mathbf{z} = \left[\left(\mathbf{z}_1\right)^H, \cdots, \left(\mathbf{z}_{K_\mathrm{R}}\right)^H\right]^H$. It is easy to see that $\overline{\mathbf{w}}_{g,i} = c_{f_g,i}\mathbf{u}_{g,i} + \overline{c}_{f_g,i}\mathbf{v}_{g,i}$. Furthermore, only one of $\overline{\mathbf{w}}_{g,i} = \mathbf{u}_{g,i}$ and $\overline{\mathbf{w}}_{g,i} = \mathbf{v}_{g,i}$ holds. Thus, the SINR at user k is given by

$$\gamma_{k,2} = \frac{\left|\mathbf{h}_k^H \overline{\mathbf{w}}_{g_k}\right|^2}{\sum\limits_{g\in\mathcal{G}\setminus\{g_k\}} \left|\mathbf{h}_k^H \overline{\mathbf{w}}_g\right|^2 + \mathbf{h}_k^H \mathbf{\Omega} \mathbf{h}_k + \sigma_k^2}. \tag{8}$$

The corresponding achievable rate on unit bandwidth in one second of user k is calculated as $R_{k,2} = \ln\left(1+\gamma_{k,2}\right)$ in unit of nats/Hz/s.

In general, fetching a requested file from the BBU via the fronthaul link incurs a certain delay since the propagation of fronthaul links and signal processing at the BBU. Such a delay is mainly determined by the worst transfer of the fronthaul links. We define the worst delay τ, in unit of second, as follows[2]

$$\tau = \tau_0 + \frac{S}{\min\limits_{i\in\mathcal{K}_\mathrm{R}} \left(\ln\left(|\mathbf{A}_i|\right) - \ln\left(|\mathbf{\Omega}_i|\right)\right)} \tag{9}$$

where τ_0 denotes the constant delay for constant route time and signal processing at the BBU and $\mathbf{A}_i = \sum\limits_{g\in\mathcal{G}} \overline{c}_{f_g,i}\mathbf{v}_{g,i}\mathbf{v}_{g,i}^\mathrm{H} + \mathbf{\Omega}_i$. The second term in (9) accounts for the worst transfer delay of the propagation of fronthaul links. Consequently, the delivery latency minimization problem is formulated as follows

$$\min \max_{g\in\mathcal{G}} \frac{S}{\mathrm{r}_{g,2}} + \tau \tag{10a}$$

[2]If eRRH $i \in \mathcal{K}_\mathrm{R}$ has cached all requested files at its local cache, i.e., $\sum\limits_{g\in\mathcal{G}} \overline{c}_{f_g,i} = 0$, the value of $\ln\left(|\mathbf{A}_i|\right) - \ln\left(|\mathbf{\Omega}_i|\right)$ in the denominator of the second item in (9) is set to be a very large constant value.

$$\text{s.t. } \mathrm{r}_{g,2} \leq \mathrm{R}_{k,2}, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{10b}$$

$$\sum_{g \in \mathcal{G}} \left\| \overline{\mathbf{w}}_{g,i} \right\|^2 + \mathrm{Tr}\left(\mathbf{\Omega}_i\right) \leq P_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{10c}$$

$$g_i\left(\mathcal{V}, \mathcal{O}\right) \leqslant C_i, \mathbf{\Omega}_i \succeq \mathbf{0}, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{10d}$$

where the optimization variables are $\mathbf{u}_g$, $\mathbf{v}_g$, $\mathbf{\Omega}$, and $\mathrm{r}_{g,2}$, $\forall g \in \mathcal{G}$, $r_{g,2}$ in unit of nats/Hz/s denotes the delivery rate of group g, and $g_i\left(\mathcal{V}, \mathcal{O}\right)$ denotes the rate on the fronthaul link of eRRH i and is given by

$$g_i\left(\mathcal{V}, \mathcal{O}\right) \triangleq I\left(\widetilde{\mathbf{x}}_i; \widehat{\mathbf{x}}_i\right) = \ln\left(\left|\mathbf{A}_i\right|\right) - \ln\left(\left|\mathbf{\Omega}_i\right|\right) \tag{11}$$

where $\mathcal{V} \triangleq \left\{\mathbf{v}_{g,i}\right\}_{g \in \mathcal{G}, i \in \mathcal{K}_{\mathrm{R}}}$ and $\mathcal{O} \triangleq \left\{\mathbf{\Omega}_i\right\}_{i \in \mathcal{K}_{\mathrm{R}}}$. Constraints (10b) means that the delivery rate of file requested by the users in group $\mathcal{G}_g$ is no larger than the achievable rate of user k that belongs to group $\mathcal{G}_g$. Constraints (10c) is the power constraint per eRRH. Constraint (10d) is the fronthaul capacity constraint ensuring signal $\widehat{\mathbf{x}}_i$ can be reliably recovered by eRRH i [28, Ch. 3]. Note that when no requested files are cached at the eRRHs, i.e., $c_{f,i} = 0$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, $\forall f \in \mathcal{F}_{\mathrm{req}}$, the delivery latency minimization problem can still be formulated by (10). When all requested files are stored at the eRRHs, i.e., $\sum_{g \in \mathcal{G}} c_{f_g,i} = \left|\mathcal{F}_{\mathrm{req}}\right|$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, the PCBT scheme reduces to the FCBT scheme, i.e., problem (10) is equivalent to problem (4).

C. Partial Caching Pipelined Transmission Scheme

In the PCBT scheme, the eRRHs have to wait for the arrival of the uncached requested files before transmitting all requested files to the users, that is, waiting for delay τ. In practice, an eRRH is able to receive data from its fronthaul link while sending wireless signals. Thus, we design a novel partial caching pipelined transmission (PCPT) scheme that also contains three phases, as shown in Fig. 3. Specifically, in phase I, the users first send the requirement of files to the eRRHs. After receiving the requirements of the users, in phase II, according to the caching status of the requested files, the eRRHs transmit the cached requested files to the users while fetching the uncached requested files from the BBU. In phase III, after the arrival of the uncached requested files, the eRRHs transmit the remaining cached requested files and uncached requested files to the users. Different from the PCBT scheme, the eRRHs do not have to wait for the uncached requested files to arrive before sending the cached requested files.

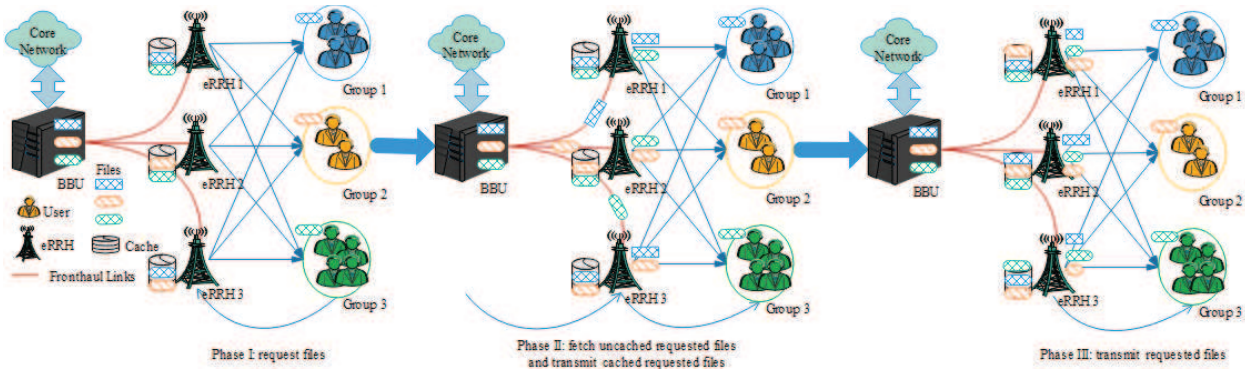

Fig. 3. Flowchart of partial caching pipelined transmission scheme.

The duration of phase II is delay τ given by (9), i.e., the time of fetching the uncached requested files from the BBU and the signal processing at the BBU, etc. In phase II, the eRRHs cooperatively transmit the cached requested files to the users. Thus, the received signal at each user is expressed as (2). In phase III, after the quantized precoded signals of the uncached requested files arrives at all eRRHs, the remaining cached requested files and uncached requested files are simultaneously transmitted to all users as in the PCBT scheme. Hence, the received signal at each user is expressed as (7). Consequently, the delivery latency minimization problem is formulated as[3]

$$\min \max_{g\in\mathcal{G}} \frac{S-\tau \mathrm{r}_{g,1}}{\mathrm{r}_{g,2}}+\tau \tag{12a}$$

$$\text{s.t. } (4b), (4c), (10b), (10c), (10d) \tag{12b}$$

$$\tau \mathrm{r}_{g,1} \leq S, \forall g \in \mathcal{G} \tag{12c}$$

where the optimization variables are $\mathbf{w}_g$, $\mathbf{u}_g$, $\mathbf{v}_g$, $\mathbf{\Omega}$, and $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P} = \{1,2\}$. In (12a), $S-\tau \mathrm{r}_{g,1}$ denotes the remaining cached requested files after the transmission of phase II. Constraints (12c) imposes that the amount of data transmission of each file be limited by file size S. Note that in problem (12), when all requested files are locally cached by the eRRHs, i.e., $\sum\limits_{g\in\mathcal{G}} c_{f_g,i} = |\mathcal{F}_{\mathrm{req}}|$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, the value of τ is zero and problem (12) is equivalent to problem (4).

[3]The existing work in [21] maximizes the minimum delivery rate with fixed delay τ incurred by the propagation of fronthaul links and signal processing at the BBU. However, this work aims to minimize the delivery latency and optimize the delay τ. As a consequence, the problem considered in this paper is more comprehensive as compared with that in [21].

When no requested files are stored at the cache of the network edge, i.e., $\sum_{g\in\mathcal{G}}\sum_{i\in\mathcal{K}_{\mathrm{R}}} c_{f_g,i}=0$, problem (12) reduces to problem (10). When requested file f_g is not cached at the eRRHs, i.e., $c_{f_g,i}=0$, $\forall i\in\mathcal{K}_{\mathrm{R}}$, constraints $\mathrm{r}_{g,1}\leq \mathrm{R}_{k,1}$, $\forall k\in\mathcal{G}_g$, and $\tau \mathrm{r}_{g,1}\leq S$ corresponding to group g in (4b) and (12c) are redundant.

IV. Design of Optimization Algorithms

In this section, we focus on designing an efficient optimization algorithm to solve the corresponding optimization problem of each transmission scheme. The common characteristics of these problems are the non-convex group rate constraints (4b) and (10b), due to the existence of non-convex achievable rate in the right side of constraints (4b) and (10b). Problems (10) and (12) are even more challenging due to the non-convex fractional objective function and fronthaul capacity constraints. Therefore, problems (4), (10) and (12) are non-convex and it is difficult to obtain the global optimum. In what follows, we design heuristic optimization algorithms to solve the problems locally via the penalty dual decomposition (PDD) method and the successive convex approximation (SCA) methods [30]–[33].

A. Solving Problem (4) *FCBT Scheme*

In this subsection, we focus on solving problem (4). The main barriers of solving problem (4) are the non-convexity of the objective function (4a) and the group rate constraints (4b). First, we need to convert the non-convex forms into convex ones. Introducing auxiliary variables $\overline{\gamma}_{k,1}$, $\chi_{k,1}$, $\forall k\in\mathcal{G}_g$, $\forall g\in\mathcal{G}$, problem (4) can be equivalently reformulated into the following

$$\min \max_{g\in\mathcal{G}} \frac{S}{\mathrm{r}_{g,1}} \tag{13a}$$

$$\text{s.t. } \mathrm{r}_{g,1}\leq \ln\left(1+\overline{\gamma}_{k,1}\right), \forall k\in\mathcal{G}_g, \forall g\in\mathcal{G} \tag{13b}$$

$$\overline{\gamma}_{k,1}\leq \frac{\left|\mathbf{h}_k^H\mathbf{w}_{g_k}\right|^2}{\chi_{k,1}}, \forall k\in\mathcal{K}_{\mathrm{U}} \tag{13c}$$

$$\sum_{g\in\mathcal{G}\setminus\{g_k\}}\left|\mathbf{h}_k^H\mathbf{w}_g\right|^2+\sigma_k^2\leq \chi_{k,1}, \forall k\in\mathcal{K}_{\mathrm{U}} \tag{13d}$$

$$\sum_{g\in\mathcal{G}}\left\|c_{f_g,i}\mathbf{w}_{g,i}\right\|^2\leq P_i, \forall i\in\mathcal{K}_{\mathrm{R}} \tag{13e}$$

where the optimization variables are $\mathbf{w}_g$, $\mathrm{r}_{g,1}$, $\forall g\in\mathcal{G}$, $\overline{\gamma}_{k,1}$, and $\chi_{k,1}$, $\forall k\in\mathcal{K}_{\mathrm{U}}$. At the optimal point of problem (13), the inequality constraints (13c) and (13d) are activated. In (13c), $\overline{\gamma}_{k,1}$ and

$\frac{\left|\mathbf{h}_k^H \mathbf{w}_{g_k}\right|^2}{\chi_{k,1}}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, are convex, respectively. But, constraints (13c) is still non-convex. To deal with the non-convex constraints, we invoke a result of [34]–[36] which shows that if we replace $\frac{\left|\mathbf{h}_k^H \mathbf{w}_{g_k}\right|^2}{\chi_{k,1}}$ by its convex low bound and iteratively solve the resulting problem by judiciously updating the variables until convergence, we can obtain a Karush-Kuhn-Tucker (KKT) point of problem (13). To this end, we approximate problem (13) as follows

$$\min \max_{g \in \mathcal{G}} \frac{S}{\mathrm{r}_{g,1}} \tag{14a}$$

$$\text{s.t. } (13\text{b}), (13\text{d}), (13\text{e}) \tag{14b}$$

$$\overline{\gamma}_{k,1} \leq \overline{\varphi}^{(t)} \left(\mathbf{w}_{g_k}, \chi_{k,1}\right), \forall k \in \mathcal{K}_{\mathrm{U}} \tag{14c}$$

where the optimization variables are $\mathbf{w}_g$, $\mathrm{r}_{g,1}$, $\forall g \in \mathcal{G}$, $\overline{\gamma}_{k,1}$, and $\chi_{k,1}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$. In (14c), $\overline{\varphi}^{(t)}\left(\mathbf{w}, \chi\right)$ is a convex low boundary of function $\frac{\left|\mathbf{h}_k^H \mathbf{w}_{g_k}\right|^2}{\chi_{k,1}}$ and is defined as

$$\overline{\varphi}^{(t)}\left(\mathbf{w}, \chi\right) \triangleq \frac{2\Re\left(\left(\mathbf{w}^{(t)}\right)^{\mathrm{H}} \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} \mathbf{w}\right)}{\chi^{(t)}} - \left(\frac{\left|\mathbf{h}_k^H \mathbf{w}^{(t)}\right|}{\chi^{(t)}}\right)^2 \chi \tag{15}$$

where t denotes the index of iteration, $\mathbf{w}^{(t)}$ and $\chi^{(t)}$ represent the values of variables $\mathbf{w}$ and χ obtained at the t-th iteration, respectively.

Next, we pay our attention to objective function (14a). By introducing auxiliary variable η, we can transform problem (14) equivalently into the following convex form

$$\min \ \eta \tag{16a}$$

$$\text{s.t. } (13\text{b}), (13\text{d}), (13\text{e}), (14\text{c}) \tag{16b}$$

$$\ln\left(S\right) - \ln\left(\eta\right) - \ln\left(\mathrm{r}_{g,1}\right) \leq 0, \forall g \in \mathcal{G} \tag{16c}$$

where the optimization variables are η, $\mathbf{w}_g$, $\mathrm{r}_{g,1}$, $\forall g \in \mathcal{G}$, $\overline{\gamma}_{k,1}$ and $\chi_{k,1}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$. Note that in constraint (16c), we exploit the positive nature of η and $r_{g,1}$, i.e., $\eta > 0$ and $r_{g,1} > 0$. This is because if $r_{g,1} = 0$, the delivery latency is infinite, i.e., problem (16) becomes meaningless. At the $(t+1)$-th iteration, problem (16) is convex and can be easily solved with a classical optimization solver, such as CVX [29], [37]. The detailed steps of solving problem (16) are summarized in Algorithm 1 that converges to a KKT solution of problem (10), please see Appendix A for the detailed proof.

Algorithm 1 Solution of problem (16)

1: Set $t=0$ and $\eta^{(t)}$ to a non-zero value. Initializing $\mathbf{w}_g^{(t)}$ to be a non-zero beamforming vector, $\forall g \in \mathcal{G}$, such that constraint (12c) is satisfied;

2: Compute $\chi_{k,1}^{(t)}$ as follows

$$\chi_{k,1}^{(t)} = \sum_{g \in \mathcal{G} \backslash \left\{f_{g_k}\right\}} \left|\mathbf{h}_k^H \mathbf{w}_g^{(t)}\right|^2 + \sigma_k^2, \forall k \in \mathcal{K}_{\mathrm{U}}; \tag{17}$$

3: Solve problem (16) to obtain $\eta^{(t+1)}$, $\mathbf{w}_g^{(t+1)}$, $\mathrm{r}_{g,1}^{(t+1)}$, $\forall g \in \mathcal{G}$, $\overline{\gamma}_{k,1}^{(t+1)}$ and $\chi_{k,1}^{(t+1)}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$;

4: If $\left|\frac{\eta^{(t+1)}-\eta^{(t)}}{\eta^{(t)}}\right| \leqslant \varepsilon$, stop iteration. Otherwise, set $t \leftarrow t+1$ and go to Step 2.

B. Solving Problem (10) *for PCBT Scheme*

In this subsection, we focus on investigating the solution of problem (10), and propose an effecient optimization method to solve it. Compared to problem (4), solving problem (10) becomes more challenging, because there are additional non-convex fractional item in the objective (10a) and non-convex fronthaul capacity constraints (10d). To overcome these difficulties, we need to leverage some new mathematical methods to transform non-convex problem (10) into convex one. For simplicity, define $\mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, and $\overline{\mathbf{W}}_g = \overline{\mathbf{w}}_g \overline{\mathbf{w}}_g^H$, $\forall g \in \mathcal{G}$. Note that $\overline{\mathbf{W}}_g = \overline{\mathbf{w}}_g \overline{\mathbf{w}}_g^H$ if and only if $\overline{\mathbf{W}}_g = \overline{\mathbf{w}}_g \overline{\mathbf{w}}_g^H \succeq \mathbf{0}$ and $\operatorname{rank}\left(\overline{\mathbf{W}}_g\right) = 1$. Dropping the rank one constraint of $\overline{\mathbf{W}}_g$, $\forall g \in \mathcal{G}$, problem (10) can be rewritten as

$$\min \max_{g \in \mathcal{G}} \frac{S}{\mathrm{r}_{g,2}} + \tau \tag{18a}$$

$$\text{s.t. } \mathrm{r}_{g,2} - \ln\left(\mu_{k,2}\right) + \ln\left(\chi_{k,2}\right) \leq 0, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{18b}$$

$$\sum_{g \in \mathcal{G}} \operatorname{Tr}\left(\mathbf{P}_{g,i}^T(1)\, \overline{\mathbf{W}}_g \mathbf{P}_{g,i}(1)\right) + \operatorname{Tr}\left(\mathbf{\Omega}_i\right) \leq P_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{18c}$$

$$\overline{\mathbf{W}}_g \succeq \mathbf{0}, \forall g \in \mathcal{G}, \mathbf{\Omega}_i \succeq \mathbf{0}, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{18d}$$

$$g_i(\mathcal{V}, \mathcal{O}) \leqslant C_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{18e}$$

where the optimization variables are $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g \in \mathcal{G}$. In (18b), $\mu_{k,2}$ and $\chi_{k,2}$ are defined respectively by

$$\mu_{k,2} = \sum_{g \in \mathcal{G}} \operatorname{Tr}\left(\mathbf{H}_k \overline{\mathbf{W}}_g\right) + \operatorname{Tr}\left(\mathbf{\Omega} \mathbf{H}_k\right) + \sigma_k^2, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{19a}$$

$$\chi_{k,2}=\sum_{g'\in\mathcal{G}\backslash\{g\}}\mathrm{Tr}\left(\mathbf{H}_k\overline{\mathbf{W}}_{g'}\right)+\mathrm{Tr}\left(\mathbf{\Omega}\mathbf{H}_k\right)+\sigma_k^2,\forall k\in\mathcal{G}_g,\forall g\in\mathcal{G}. \tag{19b}$$

In (18c), permutation matrix function $\mathbf{P}_{g,i}\left(x\right)$ is defined as

$$\mathbf{P}_{g,i}\left(x\right)=\left[\mathbf{0}_{N_{\mathrm{t}}\times(i-1)N_{\mathbf{t}}},x\mathbf{I}_{N_{\mathrm{t}}\times N_{\mathrm{t}}},\mathbf{0}_{N_{\mathrm{t}}\times(K_{\mathrm{R}}N_{\mathbf{t}}-iN_{\mathbf{t}})}\right]^T. \tag{20}$$

Problem (18) is non-convex due to the non-convexity of objective function (18a) and constraints (18b) and (18e). Consequently, it is difficult to obtain the global optimal solution of problem (18). In what follows, we aim to relax the optimization conditions in order to provide reasonable design for practical implementation.

The first thing of addressing problem (18) is to transfer it into a solvable form by using some mathematical methods. By introducing auxiliary variables η and θ, problem (18) can be equivalently reformulated as

$$\min\ \ \eta+\theta \tag{21a}$$

$$\text{s.t. (18b), (18c), (18d), (18e),} \tag{21b}$$

$$S\leq\eta\mathrm{r}_{g,2},\forall g\in\mathcal{G} \tag{21c}$$

$$\theta=\frac{S}{\min\limits_{i\in\mathcal{K}_{\mathrm{R}}}\left(\ln\left(|\mathbf{A}_i|\right)-\ln\left(|\mathbf{\Omega}_i|\right)\right)} \tag{21d}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g\in\mathcal{G}$. Note that constant τ_0 in objective function (21a) is omitted. Problem (21) can be convexified as problem (22) via some basic mathematical operation and using the PDD and SCA methods [30]–[33], please see Appendix B for the details.

$$\min\ \ \eta+\theta+\frac{1}{2\rho}\sum_{i\in\mathcal{K}_{\mathrm{R}}}\mathbb{1}\left(\sum_{g\in\mathcal{G}}\overline{c}_{f_g,i}\right)\left|\frac{S}{\theta}+\phi\left(\mathbf{\Omega}_i,\mathbf{\Omega}_i^{(t)}\right)-\ln\left(|\mathbf{A}_i|\right)+\rho\lambda\right|^2 \tag{22a}$$

$$\text{s.t. } \mathrm{r}_{g,2}-\ln\left(\mu_{k,2}\right)+\phi\left(\chi_{k,2},\chi_{k,2}^{(t)}\right)\leq 0,\forall k\in\mathcal{K}_{\mathrm{U}}, \tag{22b}$$

$$\sum_{g\in\mathcal{G}}\mathrm{Tr}\left(\mathbf{P}_{g,i}^T\left(1\right)\overline{\mathbf{W}}_g\mathbf{P}_{g,i}\left(1\right)\right)+\mathrm{Tr}\left(\mathbf{\Omega}_i\right)\leq P_i,\forall i\in\mathcal{K}_{\mathrm{R}} \tag{22c}$$

$$\overline{\mathbf{W}}_g\succeq\mathbf{0},\forall g\in\mathcal{G},\mathbf{\Omega}_i\succeq\mathbf{0},\forall i\in\mathcal{K}_{\mathrm{R}} \tag{22d}$$

$$\phi\left(\mathbf{A}_i,\mathbf{A}_i^{(t)}\right)-\ln\left(|\mathbf{\Omega}_i|\right)\leqslant C_i,\forall i\in\mathcal{K}_{\mathrm{R}}, \tag{22e}$$

$$\ln\left(S\right)-\ln\left(\eta\right)-\ln\left(\mathrm{r}_{g,2}\right)\leq 0,\forall g\in\mathcal{G}, \tag{22f}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g \in \mathcal{G}$. In problem (22), λ is the Lagrange multiplier and ρ is a scalar penalty parameter. This penalty parameter improves the robustness compared to other optimization methods for constrained problems (e.g. dual ascent method) and in particular achieves convergence without the need of specific assumptions for the objective function, i.e. strict convexity and finiteness [30]–[33][4].

When Lagrange multiplier λ and penalty parameter ρ are fixed, problem (22) is convex and can be easily solved by a classical optimization solver, such as the CVX [29], [37]. Based on this observation, in the sequel, we adopt an alternative optimization method to address problem (22). In particular, we first solve problem (22) with fixed λ and ρ, and then update Lagrange multiplier λ and penalty parameter ρ according to the constraint violation condition [32]. A step-by-step description for solving problem (22) is given in Algorithm 2, where l and t denote the number of iterations, respectively. ϵ and $\varsigma^{(l)}$ are a stopping threshold and an approximation stopping threshold, respectively. ω is a control parameter. $\zeta^{(t)}$ denotes the objective value of problem (22) at the t-th iteration. According to [32, Corollary 3.1], Algorithm 2 guarantees convergence to a KKT solution of problem (21). In Algorithm 2, Step 4 solves a convex problem, which can be efficiently implemented by the primal-dual interior point method with approximate complexity of $O\left(\left(G\left(2N_{\mathrm{t}}^2+1\right)+2\right)^{3.5}\right)$ [29]. The overall computational complexity of Algorithm 2 is $O\left(\upsilon_2\left(G\left(N_{\mathrm{t}}K_{\mathrm{R}}+4\right)\right)^{3.5}\right)$, where υ_2 denotes the number of the operations of Step 4. Due the influence of the rank relaxation, an optimal solution of problem (22) is not necessary an optimal solution of problem (18). Therefore, we need to adopt a specific method that can be found in Appendix C to obtain the solution of problem (18) from the solution of problem (22). The initialization of Algorithm 2 is finished using the method proposed in Appendix D.

C. Solving Problem (12) *for PCPT Scheme*

In this subsection, we focus on the optimization of problem (12) for the PCPT scheme, under the assumption that partial requested files are cached at the local cache of the network edge and partial requested files need to be fetched from the BBU. Compared to problems (4) and (10), solving problem (12) is more challenging as the pipelined transmission of requested

[4]In constraint (33c), we exploit the positive nature of η and $r_{g,2}$, i.e., $\eta > 0$ and $r_{g,2} > 0$. This is because that if $r_{g,2} = 0$, the delivery latency is infinite, i.e., problem (18) becomes meaningless.

Algorithm 2 Solution of problem (22)

1: Set $\zeta^{(0)}$ to be a non-zero value and initialize non-zero beamforming matrix $\overline{\mathbf{W}}_g^{(0)}$, $\forall g \in \mathcal{G}$, $\mathbf{\Omega}_i^{(0)}$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, such that constraints (18c), (18d) and (18e) are satisfied;

2: Let $l = 0$, initialize $\lambda^{(l)}$ and $\rho^{(l)}$ to be a non-zero value;

3: Let $t = 0$, compute $\chi_{k,2}^{(t)}$ as follows:

$$\chi_{k,2}^{(t)} = \sum_{g \in \mathcal{G} \setminus \{g_k\}} \mathrm{Tr}\left(\mathbf{H}_k \overline{\mathbf{W}}_g^{(t)}\right) + \mathrm{Tr}\left(\mathbf{\Omega}^{(t)} \mathbf{H}_k\right) + \sigma_k^2, \forall k \in \mathcal{K}_{\mathrm{U}}. \tag{23}$$

4: Let $t \leftarrow t + 1$. Solve problem (22) to obtain $\zeta^{(t)}$, $\eta^{(t)}$, $\theta^{(t)}$, $\overline{\mathbf{W}}_g^{(t)}$, $\mathrm{r}_{g,2}^{(t)}$, $\forall g \in \mathcal{G}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, $\mathbf{\Omega}_i^{(t)}$, $\forall i \in \mathcal{K}_{\mathrm{R}}$;

5: If $\left|\frac{\zeta^{(t)} - \zeta^{(t-1)}}{\zeta^{(t-1)}}\right| \leq \epsilon^{(l)}$, go to Step 6. Otherwise, compute $\chi_{k,2}^{(t)}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, and go to Step 4;

6: If $\left|\frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left|\mathbf{A}_i^{(t)}\right|\right) - \ln\left(\left|\mathbf{\Omega}_i^{(t)}\right|\right)\right)\right| \leq \varepsilon$, stop iteration. Otherwise, go to Step 7;

7: If $\left|\frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left|\mathbf{A}_i^{(t)}\right|\right) - \ln\left(\left|\mathbf{\Omega}_i^{(t)}\right|\right)\right)\right| \leq \varsigma^{(l)}$, update λ and ρ as follows

$$\lambda^{(l+1)} = \lambda^{(l)} + \frac{1}{\rho}\left(\frac{S}{\theta^{(t)}} - \min_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left|\mathbf{A}_i^{(t)}\right|\right) - \ln\left(\left|\mathbf{\Omega}_i^{(t)}\right|\right)\right)\right) \tag{24a}$$

$$\rho^{(l+1)} = \rho^{(l)}. \tag{24b}$$

Otherwise, update λ and ρ as follows

$$\lambda^{(l+1)} = \lambda^{(l)} \tag{25a}$$

$$\rho^{(l+1)} = \omega \rho^{(l)}; \tag{25b}$$

8: Let $l \leftarrow l + 1$, $\varsigma^{(l+1)} = \omega \left|\frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left|\mathbf{A}_i^{(t)}\right|\right) - \ln\left(\left|\mathbf{\Omega}_i^{(t)}\right|\right)\right)\right|$, and go to Step 3.

files. Following the similar procedure used for problem (10), problem (12) can be reformulated as

$$\min \max_{g \in \mathcal{G}} \frac{S - \tau \mathrm{r}_{g,1}}{\mathrm{r}_{g,2}} + \tau \tag{26a}$$

$$\text{s.t. } (18\text{b}), (18\text{c}), (18\text{d}), (18\text{e}) \tag{26b}$$

$$\mathrm{r}_{g,1} - \ln\left(\mu_{k,1}\right) + \ln\left(\chi_{k,1}\right) \leq 0, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{26c}$$

$$\sum_{g \in \mathcal{G}} \mathrm{Tr}\left(\mathbf{P}_{g,i}^T\left(c_{f_g,i}\right)\mathbf{W}_g\mathbf{P}_{g,i}\left(c_{f_g,i}\right)\right) \leq P_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{26d}$$

$$\tau \mathrm{r}_{g,1} \leq S, \forall g \in \mathcal{G} \tag{26e}$$

where the optimization variables are $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$. In (26c), $\mu_{k,1}$ and $\chi_{k,1}$ are defined respectively as

$$\mu_{k,1} = \sum_{g \in \mathcal{G}} \mathrm{Tr}\left(\mathbf{H}_k\mathbf{W}_g\right) + \sigma_k^2, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{27a}$$

$$\chi_{k,1} = \sum_{g' \in \mathcal{G}\setminus\{g\}} \mathrm{Tr}\left(\mathbf{H}_k\mathbf{W}_{g'}\right) + \sigma_k^2, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{27b}$$

where $\mathbf{W}_g = \mathbf{w}_g\mathbf{w}_g^H$, $\forall g \in \mathcal{G}$. $\mathbf{W}_g = \mathbf{w}_g\mathbf{w}_g^H$ if and only if $\mathbf{W}_g \succeq \mathbf{0}$ and $\mathrm{rank}\left(\mathbf{W}_g\right) = 1$. In (26), the rank one constraints are omitted. Similarly, problem (26) can be approximated as convex upper bound problem (28), please see Appendix E for the details.

$$\min \ \ \eta + \theta + \frac{1}{2\rho}\sum_{i \in \mathcal{K}_{\mathrm{R}}} \mathbb{1}\left(\sum_{g \in \mathcal{G}} \overline{c}_{f_g,i}\right)\left|\frac{S}{\theta} + \phi\left(\mathbf{\Omega}_i, \mathbf{\Omega}_i^{(t)}\right) - \ln\left(|\mathbf{A}_i|\right) + \rho\lambda\right|^2 \tag{28a}$$

$$\text{s.t. } (22\text{b}), (22\text{c}), (22\text{d}), (22\text{e}), (26\text{d}), \tag{28b}$$

$$\mathrm{r}_{g,1} - \ln\left(\mu_{k,1}\right) + \phi\left(\chi_{k,1}, \chi_{k,1}^{(t)}\right) \leq 0, \forall k \in \mathcal{K}_{\mathrm{U}}, \tag{28c}$$

$$S - \tau_0\mathrm{r}_{g,1} - \psi_g - \kappa_g \leq 0, \forall g \in \mathcal{G} \tag{28d}$$

$$\phi\left(\tau_0 + \theta, \tau_0 + \theta^{(t)}\right) + \phi\left(\mathrm{r}_{g,1}, \mathrm{r}_{g,1}^{(t)}\right) - \ln\left(S\right) \leq 0 \tag{28e}$$

$$\phi\left(\psi_g, \psi_g^{(t)}\right) - \ln\left(\theta\right) - \ln\left(\mathrm{r}_{g,1}\right) \leq 0, \forall g \in \mathcal{G} \tag{28f}$$

$$\phi\left(\kappa_g, \kappa_g^{(t)}\right) - \ln\left(\eta\right) - \ln\left(\mathrm{r}_{g,2}\right) \leq 0, \forall g \in \mathcal{G}, \tag{28g}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$. Note that in problem (28), if the requested file f_g is not stored at any eRRH, i.e., $\sum_{i \in \mathcal{K}_{\mathrm{R}}} c_{f_g,i} = 0$, $r_{g,1} = 0$ and constraints (28d), (28e), (28f), and (28g) are replaced with constraint (22f) for group g. If $S = \tau\mathrm{r}_{g,1}$ holds for group $g \in \mathcal{G}$, the constraint corresponding to group g in (28d), (28f), and (28g) are removed. It is not difficult to see that problem (28) is convex and can be solved with a classical convex optimization solver [29], [37].

Our proposed algorithm for solving problem (26) consists of two loops. Specifically, we update the Lagrange multiplier λ and the scalar penalty parameter ρ according to certain criteria

with fixed other optimization variables in the outer loop. While, in the inner loop, we use the classical optimization solver to solve problem (28). The inner loop and the outer loop are alternative implemented until a certain stop criterion is satisfied. The detailed description for solving problem (28) is given in Algorithm 3, where l and t denotes the number of iterations, respectively, $\zeta^{(t)}$ represents the objective value of problem (28) at the t-th iteration and $0 < \nu < 1$. The analysis of the convergence and computational complexity is similar to that for Algorithm 2 and is omitted here. In addition, the initialization of Algorithm 3 can be realized using the method described in Appendix D.

V. Numerical Results

In this section, we present numerical results to evaluate the performance of the proposed transmission schemes for cache-enabled multigroup multicasting RANs. For simplicity, we consider that all eRRHs have the same maximum transmit power and fronthaul capacity, i.e., $P_i = P$ and $C_i = C$, $\forall i \in \mathcal{K}_{\mathrm{R}}$. In the cache-enabled multigroup multicasting RAN system, the positions of eRRHs and users are uniformly distributed within a circular cell of radius 500 m, as illustrated in Fig. 4. The channel vector $\mathbf{h}_{k,i}$ from eRRH i to user k is modeled as $\mathbf{h}_{k,i} = \sqrt{\varrho_{k,i}}\widetilde{\mathbf{h}}_{k,i}$, where the channel power $\varrho_{k,i}$ is given as $\varrho_{k,i} = 1/\left(1 + (d_{k,i}/d_0)^{\alpha}\right)$ and the elements of $\widetilde{\mathbf{h}}_{k,i}$ are independent and identically distributed (i.i.d.) with $\mathcal{CN}(0, 1)$. All users have the same noise variance, i.e., $\sigma_k^2 = \sigma^2$, $k \in \mathcal{K}_{\mathrm{U}}$. The eRRHs are equipped with caches of equal size, i.e., $B_i = B = \lfloor \xi SF \rfloor$, $i \in \mathcal{K}_{\mathrm{R}}$, where ξ denotes the fractional caching proportion. The cache states $c_{f,i}$, $f \in \mathcal{F}$, $i \in \mathcal{K}_{\mathrm{R}}$, are randomly generated. To illustrate the effectiveness of the proposed schemes, we include the numerical performance of the transmission scheme which aims to maximize the minimum delivery rate, labeled as "JCEO Scheme" [19]. If not stated otherwise, the simulation is performed with the parameters given in Table II.

Fig. 5 and Fig. 6 illustrate the convergence trajectory of Algorithm 1 and Algorithm 3 for different random channel realization (RCR) with $S = 1.5$ nats/Hz, $P = 20$ dB, $C = 2$ nats/Hz/s, $K_{\mathrm{U}} = 3$, and $N_{\mathrm{t}} = 1$. In the right subfigure of Fig. 6, the approximation error is defined as $\left|\theta - \frac{S}{\min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln(|\mathbf{A}_i|) - \ln(|\mathbf{\Omega}_i|)\right)}\right|$. Fig. 5 demonstrates that a non-increasing sequence is generated with the running of Algorithm 1. The inner loop and the outer loop of Algorithm 3 also generate a non-increasing sequence, respectively, as illustrated in Fig. 6. Recalling the bounded properties of the objective of problems (16) and (47), the convergence of Algorithm 1 and Algorithm 3

Algorithm 3 Solution of problem (48)

1: Set $\zeta^{(0)}$ to be a non-zero value and initialize non-zero beamforming matrix $\mathbf{W}_g^{(0)}$, $\overline{\mathbf{W}}_g^{(0)}$, $\forall g \in \mathcal{G}$, $\mathbf{\Omega}_i^{(0)}$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, such that constraints (18c), (18d), (18e), and (26d) are satisfied;

2: Let $l = 0$, initialize $\lambda^{(l)}$ and $\rho^{(l)}$ to be a non-zero value;

3: Let $t = 0$, and compute $\mu_{k,p}^{(t)}$ and $\chi_{k,p}^{(t)}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, $\forall p \in \mathcal{P}$ with $\mathbf{W}_{g_k}^{(t)}$, $\overline{\mathbf{W}}_{g_k}^{(t)}$, and $\mathbf{\Omega}^{(t)}$. Let

$$\theta^{(t)} = \frac{S}{\min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left| \mathbf{A}_i^{(t)} \right| \right) - \ln\left(\left| \mathbf{\Omega}_i^{(t)} \right| \right) \right)} \tag{29a}$$

$$\mathrm{r}_{g,p}^{(t)} = \nu \min\left(\min_{k \in \mathcal{G}_g} \frac{\ln\left(\mu_{k,p}^{(t)}\right)}{\ln\left(\chi_{k,p}^{(t)}\right)}, \frac{S}{\tau_0 + \theta^{(t)}} \right) \tag{29b}$$

$$\psi_g^{(t)} = \theta^{(t)} \mathrm{r}_{g,1}^{(t)} \tag{29c}$$

$$\kappa_g^{(t)} = S - \left(\tau_0 + \theta^{(t)}\right) \mathrm{r}_{g,1}^{(t)}; \tag{29d}$$

4: Let $t \leftarrow t + 1$. Solve problem (48) to obtain $\zeta^{(t)}$, $\eta^{(t)}$, $\theta^{(t)}$, $\kappa_g^{(t)}$, $\psi_g^{(t)}$, $\mathbf{W}_g^{(t)}$, $\overline{\mathbf{W}}_g^{(t)}$, $\forall g \in \mathcal{G}$, $\mathrm{r}_{g,p}^{(t)}$, $\forall p \in \mathcal{P}$, $\mathbf{\Omega}_i^{(t)}$, $\forall i \in \mathcal{K}_{\mathrm{R}}$;

5: If $\left| \frac{\zeta^{(t)} - \zeta^{(t-1)}}{\zeta^{(t-1)}} \right| \leq \epsilon^{(l)}$, go to Step 6. Otherwise, compute $\chi_{k,p}^{(t)}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, $\forall p \in \mathcal{P}$, and go to Step 4;

6: If $\left| \frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left| \mathbf{A}_i^{(t)} \right| \right) - \ln\left(\left| \mathbf{\Omega}_i^{(t)} \right| \right) \right) \right| \leq \varepsilon$, stop iteration. Otherwise, go to Step 7;

7: If $\left| \frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left| \mathbf{A}_i^{(t)} \right| \right) - \ln\left(\left| \mathbf{\Omega}_i^{(t)} \right| \right) \right) \right| \leq \varsigma^{(l)}$, update λ and ρ via (24). Otherwise, update λ and ρ via (25);

8: Let $l \leftarrow l + 1$, $\varsigma^{(l+1)} = \omega \left| \frac{S}{\theta^{(t)}} - \min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(\left| \mathbf{A}_i^{(t)} \right| \right) - \ln\left(\left| \mathbf{\Omega}_i^{(t)} \right| \right) \right) \right|$, and go to Step 3.

can be guaranteed [34][5].

Fig. 7 shows the delivery latency versus the fractional caching proportion ξ with $S = 1.5$ nats/Hz, $P = 20$ dB, $C = 2$ nats/Hz/s, $K_{\mathrm{U}} = 6$, $G = 3$, and $N_{\mathrm{t}} = 1$. It can be observed that the FCBT scheme achieves the best performance in terms of delivery latency. The transmission

[5]As Algorithm 2 is similar to Algorithm 3, we only present the convergence trajectory of Algorithm 3. The convergence of Algorithm 2 can be guaranteed as well.

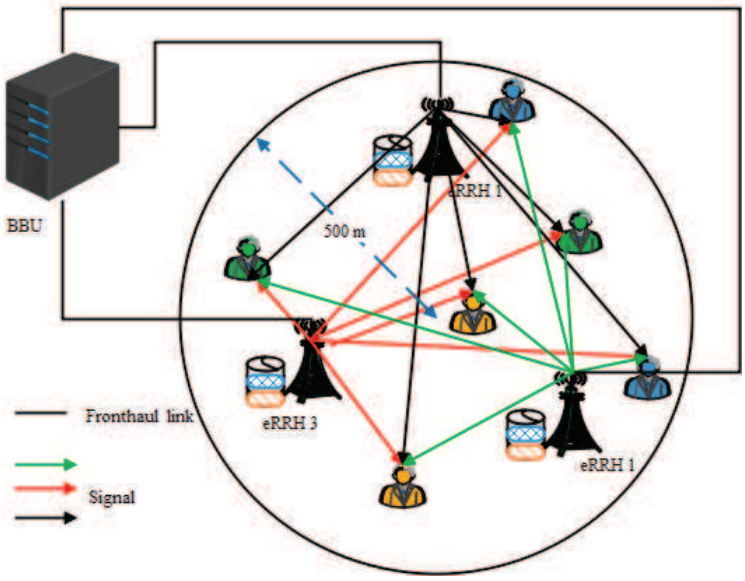

Fig. 4. Simulation Model, $K_{\mathrm{R}} = 3$ and $K_{\mathrm{U}} = 6$.

TABLE II: Simulation parameters

Symbol	Value	Symbol	Value	Symbol	Value
K_{R}	3	K_{U}	3/6	N_{t}	1/4
ξ	0.5	σ^2	1	d_0	50 m
F	10	α	3	τ_0	10 ms
ε	10^{-5}	$\varsigma^{(0)}$	10^{-3}	$\epsilon^{(0)}$	10^{-3}
$\lambda^{(0)}$	0.5	$\rho^{(0)}$	0.5	ω	0.6
ν	0.1	δ	0.5	–	–

scheme without caching (TSWC) scheme achieves the largest delivery latency. This is because all requested files need to be fetched from the BBU such that the limited capacity of fronthaul links has significant negative impact on the system performance. The delivery latency of the other two transmission schemes decreases as the fractional caching proportion ξ increases. The larger the fractional caching proportion ξ, the greater the probability of the requested files that are cached at the local cache. Thus, the impact of the capacity of fronthaul links on the system performance is reduced. In addition, the PCPT scheme outperforms the PCBT scheme, because the PCPT scheme takes advantage of the delay τ interval to transmit cached requested files before the uncached requested files arrive at the eRRHs.

Fig. 8 illustrates the delivery latency versus the fronthaul capacity C with $S = 1.2$ nats/Hz,

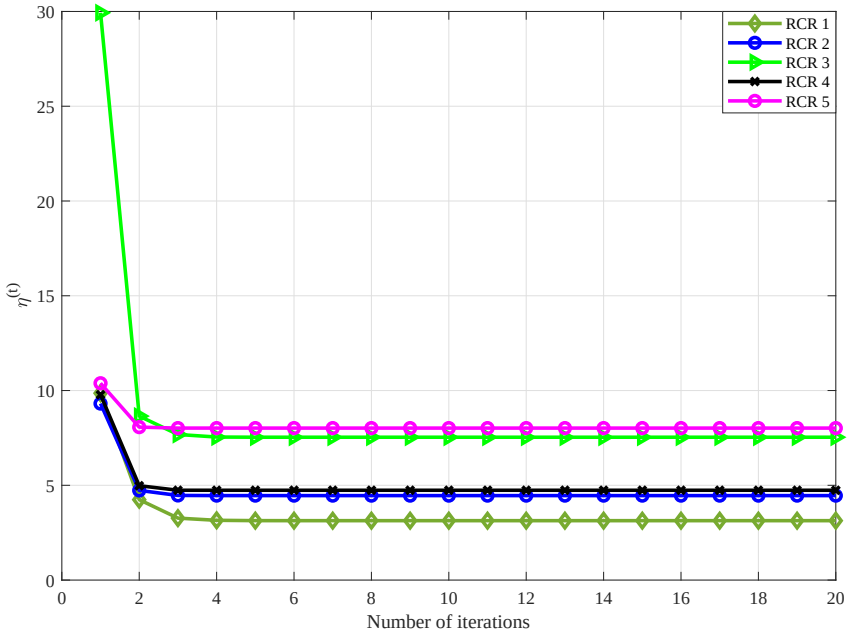

Fig. 5. Convergence trajectories of Algorithm 1 for different RCRs.

$P = 20$ dB, $K_{\mathrm{U}} = 6$, and $N_{\mathrm{t}} = 4$. Except for the FCBT scheme that is not limited by the capacity C of the fronthaul links, the delivery latency achieved by the other three transmission schemes decreases with an increasing capacity C of fronthaul links. This implies that caching at the network edge helps to reduce the burden on the fronthaul links, i.e., the amount of requested files being fetched from the BBS decreases. Therefore, the impact of constraints (10d) on the performance of the PCBT and PCPT schemes decreases. In addition, the second item of (9) may reduce with fronthaul capacity C increases. As a consequence, the delivery latency reduces. When the system performance is limited to fronthaul capacity C or the achievable rate R_k, $k \in \mathcal{K}_{\mathrm{U}}$, i.e., is not limited to file size S, the PCBT and JCEO schemes achieve the same delivery latency. This is because they make full use of all resources to maximize the minimum delivery rate and the rate on the fronthaul links, i.e., minimize the delivery latency defined in (10a) which takes into account the fairness for all multicast groups. Compared to the TSWC, PCBT, and JCEO schemes, the performance achieved by the PCPT scheme is closer to that of the FCBT scheme. Except for exploiting delay τ to transmit requested files, an advantage of the PCPT scheme is to increase the degree of freedom for power allocation in each transmission phase and to reduce

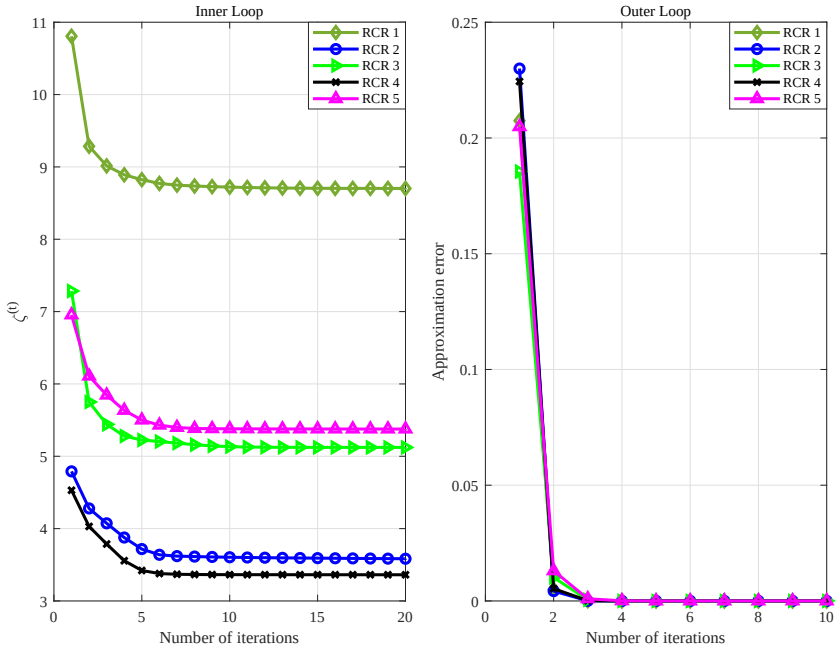

Fig. 6. Convergence trajectories of Algorithm 3 for different RCRs.

the inter-group interference for cache-enabled multigroup multicasting RANs. As a consequence, the system throughput can be increased and the delivery latency is reduced.

Fig. 9 shows the delivery latency versus file size S with $C = 1.5$ nats/Hz/s, $P = 20$ dB, $K_{\mathrm{U}} = 6$, and $N_{\mathrm{t}} = 4$. Given the channel statistics and the capacity C of the fronthaul links, the time to transmit all requested files increases as file size S increases from the objective function of problems (4), (10), and (9), respectively. At the same time, the burden on the fronthaul links also increases with increasing file size S. Results illustrated in Fig. 9 demonstrate that the delivery latency of all transmission schemes increases as file size S increases. Compared to the other non-FCBT transmission schemes, the PCPT scheme obtains the minimum delivery latency, since it effectively exploits the delay interval incurred by fetching the uncached requested files from the BBU and by the baseband signal processing at the BBU to transmit requested files. When file size S is smaller, the proposed PCBT and PCPT schemes outperform the JCEO scheme in terms of delivery latency. This is because the delivery rate of the JCEO scheme is limited by the file size. However, when the system performance is not limited to file size S, i.e., file size S is sufficiently large, the PCBT and JCEO schemes achieve the same delivery latency

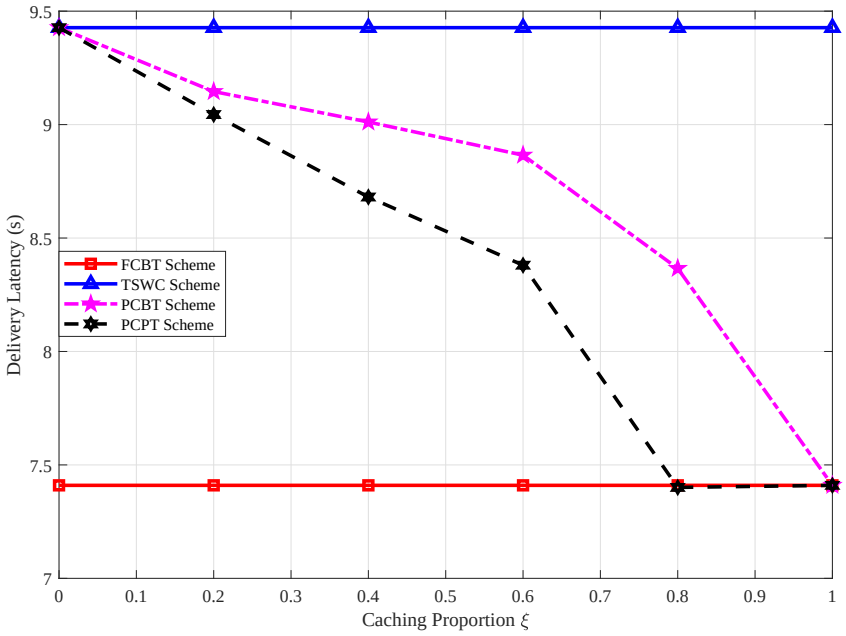

Fig. 7. Delivery latency versus the caching proportion ξ.

since they make full use of all resources to maximize the minimum delivery rate and the rate on the fronthaul links. It also means that the PCBT and JCEO schemes guarantee the fairness among all multicast groups. The delivery latency of the TSWC scheme increases rapidly as file size S increases as the fronthaul links become more congested. The results indicate that the network edge caching becomes more and more important as file size S increases, especially for the delay-sensitive data traffic.

VI. Conclusions

In this paper, according to the caching capability of each eRRH, we investigated three transmission schemes to minimize the delivery latency for cache-enabled multigroup multicasting RANs. Correspondingly, three delivery latency minimization problems were formulated. The formulated optimization problems are non-convex and difficult to obtain directly the global optimum solutions. We further developed an efficient algorithm to address each delivery latency minimization problem. Finally, numerical results were provided to valuate the effectiveness of the proposed algorithms and shown that the PCPT scheme outperforms the PCBT scheme in terms

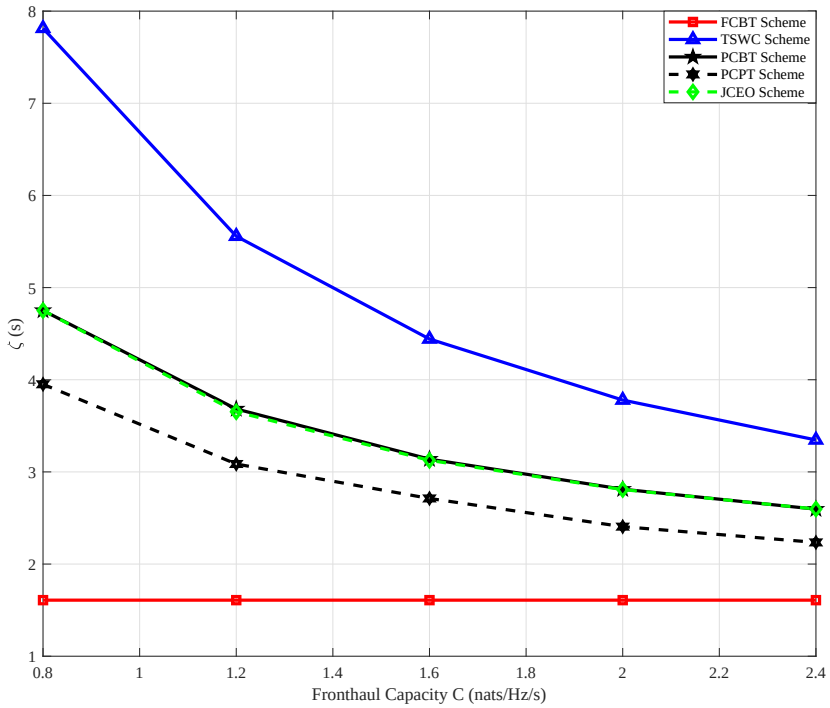

Fig. 8. Delivery latency versus the fronthaul capacity C.

of delivery latency. It can be observed that if the transmission scheme is designed carefully, caching frequently requested content at the network edge can effectively reduce the delivery latency while reducing the burden on the fronthaul links.

APPENDIX

A. Convergence and Computational Complexity Analysis

1) Convergence analysis of Algorithm 1: Consider the $(t+1)$-th iteration of Algorithm 1 that solves the optimization problem (16). If we replace η, $\mathbf{w}_g$, $\mathrm{r}_{g,1}$, $\forall g \in \mathcal{G}$, $\overline{\gamma}_{k,1}$ and $\chi_{k,1}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$ with $\eta^{(t)}$, $\mathbf{w}_g^{(t)}$, $\mathrm{r}_{g,1}^{(t)}$, $\forall g \in \mathcal{G}$, $\overline{\gamma}_{k,1}^{(t)}$ and $\chi_{k,1}^{(t)}$, $\forall k \in \mathcal{K}_{\mathrm{U}}$, respectively, all constraints are still satisfied, which means that the solution of the t-th iteration is feasible point of problem (16) in the $(t+1)$-th iteration. Thus, the objective function obtained in the $(t+1)$-th iteration is not larger than that in the t-th iteration, i.e., $\eta^{(t+1)} \leq \eta^{(t)}$. That is to say, Algorithm 1 generates a non-increasing sequence of objective value $\eta^{(t)}$. Moreover, the problem is bounded due to the power

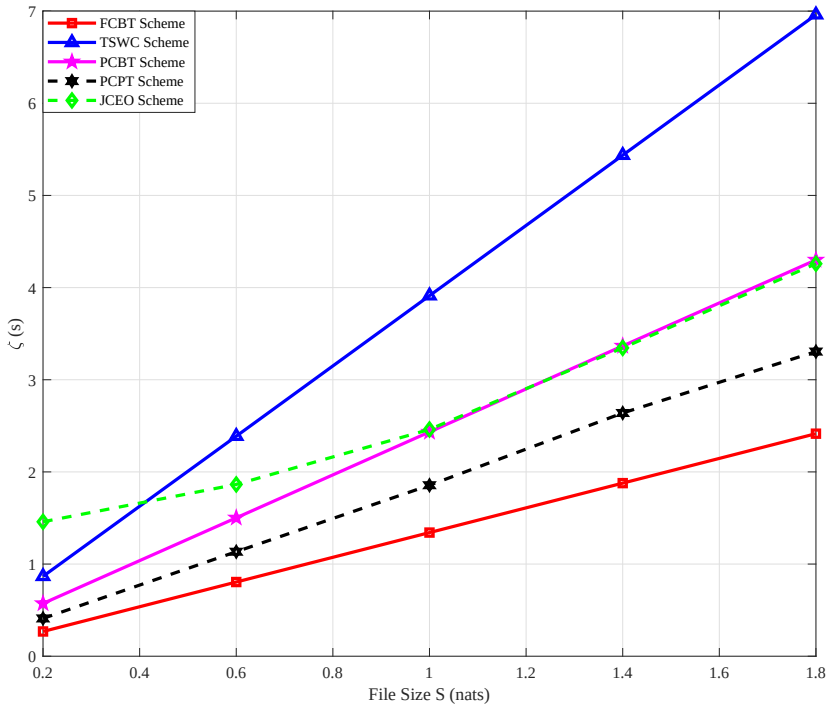

Fig. 9. Delivery latency versus the file size S.

constraints. Therefore, the convergence of Algorithm 1 can be guaranteed by the monotonic boundary theorem [38]. Furthermore, based on the arguments presented in [34, Theorem 1], we can prove that Algorithm 1 converges to a KKT solution of problem (10).

*2) Computational complexity analysis of Algorithm 1:*In Algorithm 1, Step 3 solves a convex problem, which can be efficiently implemented by the primal-dual interior point method with approximate complexity of $O\left(\left(G\left(N_{\mathrm{t}}K_{\mathrm{R}}+4\right)\right)^{3.5}\right)$, where $O\left(\cdot\right)$ stands for the big-O notation [29]. Letting υ_1 be the number of iterations in Algorithm 1, the overall computational complexity is $O\left(\upsilon_1\left(G\left(N_{\mathrm{t}}K_{\mathrm{R}}+4\right)\right)^{3.5}\right)$.

B. Convexity of Problem (22)

In this subsection, we address the difficulties of solving problem (18) step-by-step. First, we convexify non-convex constraints (18b) and (18e). Then, we convexify objective function (18a).

According to the concave property of function $\ln(\cdot)$, constraint (18b) can be convexified as

$$\mathrm{r}_{g,2} - \ln(\mu_{k,2}) + \phi\left(\chi_{k,2}, \chi_{k,2}^{(t)}\right) \leq 0, \forall k \in \mathcal{K}_{\mathrm{U}}. \tag{30}$$

Similarly, constraint (18e) can be approximated with the following convex form

$$\phi\left(\mathbf{A}_i, \mathbf{A}_i^{(t)}\right) - \ln(|\mathbf{\Omega}_i|) \leqslant C_i, \forall i \in \mathcal{K}_{\mathrm{R}}. \tag{31}$$

In (31), $\mathbf{A}_i$ is redefined as $\mathbf{A}_i = \sum_{g\in\mathcal{G}} \mathbf{P}_{g,i}^T\left(\overline{c}_{f_g,i}\right)\overline{\mathbf{W}}_g\mathbf{P}_{g,i}\left(\overline{c}_{f_g,i}\right) + \mathbf{\Omega}_i$, In (30) and (31), $\phi(\mathbf{A},\mathbf{B})$ is defined as

$$\phi(\mathbf{A},\mathbf{B}) = \ln(|\mathbf{B}|) + \mathrm{tr}\left(\mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\right). \tag{32}$$

In problem (18), if we replace constraints (18b) and (18e) with (30) and (31), respectively, all constraints in (18) are transformed into convex forms. Thus, problem (22) can be reformulated as

$$\min \ \ \eta + \theta \tag{33a}$$

$$\text{s.t. } (18c), (18d), (30), (31) \tag{33b}$$

$$\ln(S) - \ln(\eta) - \ln(\mathrm{r}_{g,2}) \leq 0, \forall g \in \mathcal{G} \tag{33c}$$

$$\theta = \frac{S}{\min\limits_{i\in\mathcal{K}_{\mathrm{R}}}\left(\ln(|\mathbf{A}_i|) - \ln(|\mathbf{\Omega}_i|)\right)} \tag{33d}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g \in \mathcal{G}$. Note that constant τ_0 in objective function (33a) is omitted. In constraint (33c), we exploit the positive nature of η and $r_{g,2}$, i.e., $\eta > 0$ and $r_{g,2} > 0$. This is because if $r_{g,2} = 0$, the delivery latency is infinite, i.e., problem (18) becomes meaningless. The new challenge of solving problem (33) is the introduction of equality constraint (33d) which is non-convex. To remove the equality constraint, we use the Lagrangian duality method to address problem (33). The partial augmented Lagrangian function of problem (33) is given by

$$\min \ \ \eta + \theta + \frac{1}{2\rho}\left|\frac{S}{\theta} - \min_{i\in\mathcal{K}_{\mathrm{R}}}\left(\ln(|\mathbf{A}_i|) - \ln(|\mathbf{\Omega}_i|)\right) + \rho\lambda\right|^2 \tag{34a}$$

$$\text{s.t. } (18c), (18d), (30), (31), (33c) \tag{34b}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g \in \mathcal{G}$. In problem (34), λ is the Lagrange multiplier and ρ is a scalar penalty parameter. This penalty parameter improves the robustness compared to other optimization methods for constrained problems (e.g. dual ascent

method) and in particular achieves convergence without the need of specific assumptions for the objective function, i.e. strict convexity and finiteness [30]–[33]. The smaller the value of ρ, the greater the probability of equality (33d) holds.

It is still challenging to address problem (34) directly, since the third term in objective function (34a) is difficult to tackle. To further obtain a tractable form of problem (34), we relax problem (34) to the following form

$$\min \ \eta+\theta+\frac{1}{2\rho}\sum_{i\in\mathcal{K}_{\mathrm{R}}}\mathbb{1}\left(\sum_{g\in\mathcal{G}}\overline{c}_{f_g,i}\right)\left|\frac{S}{\theta}+\ln\left(|\mathbf{\Omega}_i|\right)-\ln\left(|\mathbf{A}_i|\right)+\rho\lambda\right|^2 \tag{35a}$$

$$\text{s.t. } (18\text{c}),(18\text{d}),(30),(31),(33\text{c}), \tag{35b}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g\in\mathcal{G}$. In (35), indicator function $\mathbb{1}\left(x\right)$ is defined as follows:

$$\mathbb{1}\left(x\right)=\begin{cases}0, & \text{if } x=0,\\ 1, & \text{if } x\neq 0.\end{cases} \tag{36}$$

When all requested files are stored at eRRH i, the value of $\ln\left(|\mathbf{A}_i|\right)-\ln\left(|\mathbf{\Omega}_i|\right)$ should be a very large constant value. Therefore, without affecting the solution of (35), in (35a), the indication function $\mathbb{1}\left(x\right)$ is introduced to move away from the item related to eRRH i which stores all requested files at its local cache. At the optimal point of problem (35), there is at least an eRRH i such that equality constraint (33d) holds. Using again the SCA method, problem (35) can be further convexified as the following convex upper bound problem

$$\min \ \eta+\theta+\frac{1}{2\rho}\sum_{i\in\mathcal{K}_{\mathrm{R}}}\mathbb{1}\left(\sum_{g\in\mathcal{G}}\overline{c}_{f_g,i}\right)\left|\frac{S}{\theta}+\phi\left(\mathbf{\Omega}_i,\mathbf{\Omega}_i^{(t)}\right)-\ln\left(|\mathbf{A}_i|\right)+\rho\lambda\right|^2 \tag{37a}$$

$$\text{s.t. } (18\text{c}),(18\text{d}),(30),(31),(33\text{c}), \tag{37b}$$

where the optimization variables are η, θ, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,2}$, $\forall g\in\mathcal{G}$.

C. Gaussian Randomization

Due to the rank relaxation, in general, the solution to problem (22), denoted as $\eta^{(o)}$, $\theta^{(o)}$, $\overline{\mathbf{W}}_g^{(o)}$, $\mathbf{\Omega}^{(o)}$, $\mathrm{r}_{g,2}^{(o)}$, $\forall g\in\mathcal{G}$, may not comprise only rank-one matrices $\overline{\mathbf{W}}_g^{(o)}$, $\forall g\in\mathcal{G}$. Hence, the optimum beamforming vectors cannot be directly extracted from the obtained $\overline{\mathbf{W}}_g^{(o)}$. If $\overline{\mathbf{W}}_g^{(o)}$, $\forall g\in\mathcal{G}$, is of rank one, we can write $\overline{\mathbf{W}}_g^{(o)}=\overline{\mathbf{w}}_g^{(o)}\overline{\mathbf{w}}_g^{(o)H}$ and $\overline{\mathbf{w}}_g^{(o)}$ will be a feasible solution

to problem (18). If the rank of $\overline{\mathbf{W}}_g^{(o)}$ is larger than 1, we can use the Gaussian randomization technique to generate candidate beamformer vector from $\overline{\mathbf{W}}_g^{(o)}$. In the randomization technique, we eigendecompose $\overline{\mathbf{W}}_g^{(o)} = \overline{\mathbf{U}}_g \mathbf{\Lambda}_g \overline{\mathbf{U}}_g^H$ and choose $\overline{\mathbf{w}}_g^{(o)} = \sqrt{p_g}\overline{\mathbf{U}}_g \mathbf{\Lambda}_g^{1/2} \mathbf{e}_g$, where $\mathbf{e}_g \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ and p_g denotes the sought power boost (or reduction) factor for multicast group g [39]–[42]. The specific value of p_g, $\forall g \in \mathcal{G}$, can be obtained by solving the following problem

$$\min \sum_{g \in \mathcal{G}} p_g \tag{38a}$$

$$\text{s.t. } \mathrm{r}_{g,2}^{(o)} \leq \mathrm{R}_{k,2}^{(o)}, \forall k \in \mathcal{G}_g, \forall g \in \mathcal{G} \tag{38b}$$

$$\sum_{g \in \mathcal{G}} \left\| \mathbf{P}_{g,i}^T(1) \, \overline{\mathbf{w}}_g^{(o)} \right\|^2 + \mathrm{Tr}\left(\mathbf{\Omega}_i^{(o)} \right) \leq P_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{38c}$$

$$g_i\left(\mathcal{V}^{(o)}, \mathcal{O}^{(o)} \right) \leqslant C_i, \forall i \in \mathcal{K}_{\mathrm{R}} \tag{38d}$$

where the optimization variables are p_g, $\forall g \in \mathcal{G}$. In (38b), $\mathrm{R}_{k,2}^{(o)} = \ln\left(1 + \gamma_{k,2}^{(o)}\right)$ where $\gamma_{k,2}^{(o)}$ is calculated with (8) and $\overline{\mathbf{w}}_g^{(o)}$, $\forall g \in \mathcal{G}$. In (38d), $\mathcal{V}^{(o)} \triangleq \left\{ \mathbf{P}_{g,i}^T(1) \, \overline{\mathbf{w}}_g^{(o)} \right\}_{g \in \mathcal{G}, i \in \mathcal{K}_{\mathrm{R}}}$ and $\mathcal{O}^{(o)} \triangleq \left\{ \mathbf{\Omega}_i^{(o)} \right\}_{i \in \mathcal{K}_{\mathrm{R}}}$.

D. Initialization of Algorithm 2 and 3

In Algorithm 2 and 3, the initialization of $\mathbf{w}_{g,i}$, $\mathbf{u}_{g,i}$, $\mathbf{v}_{g,i}$, $\forall g \in \mathcal{G}$ and $\mathbf{\Omega}_i$, $\forall i \in \mathcal{K}_{\mathrm{R}}$ is finished via two steps. First, the initial values of $\mathbf{w}_{g,i}$, $\mathbf{u}_{g,i}$, $\forall g \in \mathcal{G}$ are randomly chosen and the initial values of $\mathbf{v}_{g,i}$, $\forall g \in \mathcal{G}$ and $\mathbf{\Omega}_i$, $\forall i \in \mathcal{K}_{\mathrm{R}}$ are given by:

$$\mathbf{v}_{g,i} = \begin{cases} \sqrt{\delta \frac{e^{C_i} - 1}{\sum\limits_{g \in \mathcal{G}} \overline{c}_{f_g,i}}} \mathbf{v}_i, & \sum\limits_{g \in \mathcal{G}} \overline{c}_{f_g,i} \neq 0 \\ \mathbf{0}, & \sum\limits_{g \in \mathcal{G}} \overline{c}_{f_g,i} = 0 \end{cases} \tag{39a}$$

$$\mathbf{\Omega}_i = \mathbf{I}_{N_t} \tag{39b}$$

where $\mathbf{v}_i$ is a normalized random vector and $0 < \delta \leq 1$. Then, they are normalized such that power constraint (26d) is satisfied. It is easy to observe that fronthaul capacity constraint (18e) is satisfied when the values of $\mathbf{v}_{g,i}$, $\forall g \in \mathcal{G}$ and $\mathbf{\Omega}_i$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, are given by (39).

E. Convexity of Problem (26)

In this subsection, we focus on convexifying problem (26) via the PPD method and SCA method. Similarly, constraint (26c) can be approximated by

$$\mathrm{r}_{g,1} - \ln\left(\mu_{k,1}\right) + \phi\left(\chi_{k,1}, \chi_{k,1}^{(t)}\right) \leq 0, \forall k \in \mathcal{K}_{\mathrm{U}}. \tag{40}$$

Now, we turn our attention to transform objective (26a) and constraint (26e) into a convex form. Introducing auxiliary variables η and θ, problem (26) can be equivalently rewritten as

$$\min \ \ \eta + \theta \tag{41a}$$

$$\text{s.t. } (18\text{c}), (18\text{d}), (30), (31), (26\text{d}), (40) \tag{41b}$$

$$\left(\tau_0 + \theta\right) \mathrm{r}_{g,1} \leq S, \forall g \in \mathcal{G}, \tag{41c}$$

$$S - \tau \mathrm{r}_{g,1} \leq \eta \mathrm{r}_{g,2} \tag{41d}$$

$$\theta = \frac{S}{\min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(|\mathbf{A}_i|\right) - \ln\left(|\mathbf{\Omega}_i|\right)\right)} \tag{41e}$$

where the optimization variables are η, θ, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$. Note that constant τ_0 in the objective function (41) is omitted. Further, introducing auxiliary κ_g, ψ_g, $\forall g \in \mathcal{G}$, after some basic mathematical operation, problem (41) is equivalently reformulated as

$$\min \ \ \eta + \theta \tag{42a}$$

$$\text{s.t. } (18\text{c}), (18\text{d}), (30), (31), (26\text{d}), (40) \tag{42b}$$

$$\left(\tau_0 + \theta\right) \mathrm{r}_{g,1} \leq S, \forall g \in \mathcal{G} \tag{42c}$$

$$\psi_g \leq \theta \mathrm{r}_{g,1}, \forall g \in \mathcal{G} \tag{42d}$$

$$\kappa_g \leq \eta \mathrm{r}_{g,2}, \forall g \in \mathcal{G} \tag{42e}$$

$$S - \tau_0 \mathrm{r}_{g,1} - \psi_g - \kappa_g \leq 0, \forall g \in \mathcal{G} \tag{42f}$$

$$\theta = \frac{S}{\min\limits_{i \in \mathcal{K}_{\mathrm{R}}} \left(\ln\left(|\mathbf{A}_i|\right) - \ln\left(|\mathbf{\Omega}_i|\right)\right)} \tag{42g}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$. At the optimal point of problem (42), constraints (42d), (42e), and (42f) are activated. Using the monotonic property of function $\ln(\cdot)$, constraints (42c), (42d), and (42e) can be transformed into

a difference of convex functions form, i.e.,

$$\ln\left(\tau_0+\theta\right)+\ln\left(\mathrm{r}_{g,1}\right)-\ln\left(S\right)\leq 0 \tag{43a}$$

$$\ln\left(\psi_g\right)-\ln\left(\theta\right)-\ln\left(\mathrm{r}_{g,1}\right)\leq 0,\forall g\in\mathcal{G} \tag{43b}$$

$$\ln\left(\kappa_g\right)-\ln\left(\eta\right)-\ln\left(\mathrm{r}_{g,2}\right)\leq 0.\forall g\in\mathcal{G} \tag{43c}$$

Further, exploiting the concavity of function $\ln\left(\cdot\right)$, (43) can be approximated to the following convex form

$$\phi\left(\tau_0+\theta,\tau_0+\theta^{(t)}\right)+\phi\left(\mathrm{r}_{g,1},\mathrm{r}_{g,1}^{(t)}\right)-\ln\left(S\right)\leq 0 \tag{44a}$$

$$\phi\left(\psi_g,\psi_g^{(t)}\right)-\ln\left(\theta\right)-\ln\left(\mathrm{r}_{g,1}\right)\leq 0,\forall g\in\mathcal{G} \tag{44b}$$

$$\phi\left(\kappa_g,\kappa_g^{(t)}\right)-\ln\left(\eta\right)-\ln\left(\mathrm{r}_{g,2}\right)\leq 0.\forall g\in\mathcal{G}. \tag{44c}$$

Replacing constraints (42c), (42d), and (42e) with inequalities (44a), (44b), and (44c), respectively, problem (42) can be further approximated to

$$\min\ \ \eta+\theta \tag{45a}$$

$$\text{s.t. (42b), (42f), (44)} \tag{45b}$$

$$\theta=\frac{S}{\min\limits_{i\in\mathcal{K}_{\mathrm{R}}}\left(\ln\left(\left|\mathbf{A}_i\right|\right)-\ln\left(\left|\mathbf{\Omega}_i\right|\right)\right)} \tag{45c}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g\in\mathcal{G}$, $\forall p\in\mathcal{P}$. Similar to problem (33), the main obstacle of solving problem (45) is equality constraint (45c).

In the following, we resort to the penalty dual decomposition method to solve problem (45). The partial augmented Lagrangian function of problem (45) is given by

$$\min\ \ \eta+\theta+\frac{1}{2\rho}\left|\frac{S}{\theta}-\min_{i\in\mathcal{K}_{\mathrm{R}}}\left(\ln\left(\left|\mathbf{A}_i\right|\right)-\ln\left(\left|\mathbf{\Omega}_i\right|\right)\right)+\rho\lambda\right|^2 \tag{46a}$$

$$\text{s.t. (42b), (42f), (44)} \tag{46b}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g\in\mathcal{G}$, $\forall p\in\mathcal{P}$. In problem (46), λ is the Lagrange multiplier and ρ is a scalar penalty parameter. Following a procedure similar to that used for problem (34), to further obtain a tractable form of problem (46), we relax problem (46) to the following form

$$\min\ \ \eta+\theta+\frac{1}{2\rho}\sum_{i\in\mathcal{K}_{\mathrm{R}}}\mathbb{1}\left(\sum_{g\in\mathcal{G}}\overline{c}_{f_g,i}\right)\left|\frac{S}{\theta}+\ln\left(\left|\mathbf{\Omega}_i\right|\right)-\ln\left(\left|\mathbf{A}_i\right|\right)+\rho\lambda\right|^2 \tag{47a}$$

$$\text{s.t. } (42b), (42f), (44) \tag{47b}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$. Further, using the SCA method to convexify objective (47a), problem (47) can be approximated as the following convex upper bound problem

$$\min \ \eta + \theta + \frac{1}{2\rho} \sum_{i \in \mathcal{K}_{\mathrm{R}}} \mathbb{1} \left(\sum_{g \in \mathcal{G}} \overline{c}_{f_g,i} \right) \left| \frac{S}{\theta} + \phi \left(\mathbf{\Omega}_i, \mathbf{\Omega}_i^{(t)} \right) - \ln \left(|\mathbf{A}_i| \right) + \rho \lambda \right|^2 \tag{48a}$$

$$\text{s.t. } (42b), (42f), (44) \tag{48b}$$

where the optimization variables are η, θ, κ_g, ψ_g, $\mathbf{W}_g$, $\overline{\mathbf{W}}_g$, $\mathbf{\Omega}$, $\mathrm{r}_{g,p}$, $\forall g \in \mathcal{G}$, $\forall p \in \mathcal{P}$.

REFERENCES

[1] A. Gupta and R. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206-1232, Jul. 2015.

[2] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks: A technology overview," *IEEE Commun. Survery & Tutor.*, vol. 17, no. 1, pp. 405-426, First Quarter 2015.

[3] J. Ren, H. Guo, C. Xu and Y. Zhang, "Serving at the Edge: A Scalable IoT Architecture based on Transparent Computing" *IEEE Netw.*, vol. 31, no. 5, pp. 96-105, Aug. 2017.

[4] D. Zhang, L. Tan, J. Ren, M. K. Awad, S. Zhang, Y. Zhang and P. Wan, "Near-optimal and Truthful Online Auction for Computation Offloading in Green Edge-Computing Systems" *IEEE Trans. on Mobile Comput.*, doi: 10.1109/TMC.2019.2901474, 2019, to appear.

[5] S. Zhang, W. Quan, J. Li, W. Shi, P. Yang, X. Shen, "Air-Ground Integrated Vehicular Network Slicing With Content Pushing and Caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2114-2127, Sept. 2018.

[6] R. Pedarsani, M. Maddah-Ali, and U. Niesen, "Online coded caching." *IEEE/ACM Trans. on Net.*, vol. 24, issue 2, pp. 836-845, Apr. 2016.

[7] C. Mouradian , D. Naboulsi, S. Yangui, R. Glitho, M. Morrow, and P. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Survery & Tutor.*, vol. 20, no. 1, pp. 416-464, First Quarter 2018.

[8] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82-89, Aug. 2014.

[9] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," *IEEE Netw.*, pp. 46-53, Jul. 2016.

[10] Y. Shih, W Chu.ng, A. Pang, T. Chiu, and H. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, pp. 52-158, Feb. 2017.

[11] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory.*, vol. 63, no. 10, pp. 6650-6678, Oct. 2017.

[12] J. Liu, B. Bai, J. Zhang, and K. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039-7051, Nov. 2017.

[13] G. Zheng, H. Suraweera, and I. Krikidis, "Optimization of hybrid cache placement for collaborative relaying," *IEEE Commun. Lett.,* vol. 21, no. 2, pp. 442-445, Feb. 2017.

[14] Y. Zhu , G. Zheng , L. Wang , K. Wong, and L. Zhao, "Content placement in cache-enabled sub-6 GHz and millimeter-wave multi-antenna dense small cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2843-2856, May 2018.

[15] J. Kwak , Y. Kim, L. 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.

[16] Y. Cui , Z. Wang, Y. Yang, F. Yang, L. Ding, and L. Qian, "Joint and competitive caching designs in large-scale multi-tier wireless multicasting networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3108-3121, Jul. 2018.

[17] W. Wen, Y. Cui , F. Zheng , S. Jin , and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2809-2825, Jul. 2018.

[18] Z. Zheng , L. Song, Z. Han, G. Li, and H. Poor, "A stackelberg game approach to proactive caching in large-scale mobile edge networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5198-5121, Aug. 2018.

[19] S. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[20] S. He, Y. Wu, J. Ren, Y. Huang, R. Schober, and Y. Zhang, "Hybrid precoder design for cache-enabled millimeter wave radio access networks," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 3, pp. 1707-1722, Mar. 2019.

[21] S. He, C. Qi, Y. Huang, Q. Hou, and A. Nallanathan, "Two-level transmission scheme for cache-enabled fog radio access networks," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 445-456, Jan. 2019.

[22] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118-6131, Sep. 2016.

[23] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907-922, Apr. 2016.

[24] S. Hajri and M. Assaad, "Energy efficiency in cache-enabled small cell networks with adaptive user clustering," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 955-968, Feb. 2018.

[25] L. Xiang, D. Kwan Ng, R. Schober, and V. Wong, "Cache-enabled physical layer security for video streaming in backhaul-limited cellular networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 736-751, Feb. 2018.

[26] F. Zhuang and V. K. N. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684-693, Feb. 2014.

[27] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," *IEEE Trans. Veh. Technol*. vol. 65, no. 7, pp. 5022-5032, Jul. 2016.

[28] A. Gamal and Y. Kim, Network Information Theory. *Cambridge, U.K.: Cambridge Univ. Press*, 2011.

[29] F. Potra and S. Wright, "Interior-point methods," *J. Comput. Appl. Math.*, vol. 124, no. 1, pp. 281-302, Jan. 2000.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1-122, Jan. 2011.

[31] C. 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.

[32] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Signal Process.*, vol. 12, no. 13, pp. 455-468, Jun. 2018.

[33] T. Sun , H. Jiang, L. Cheng, and W. Zhu, "Iteratively linearized reweighted alternating direction method of multipliers for a class of nonconvex problems," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5380-5391, Oct. 2018.

[34] B. Marks and G. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operat. Res.*, vol. 26, no. 4, pp. 681-683, Jul./Aug. 1978.

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

[36] L. Tran, M. Hanif, A. Tölli, and M. Juntti, "Fast converging algorithm for weighted sum rate maximization in multicell MISO downlink," *IEEE Signal Proc. Letters*, vol. 19, no. 12, pp. 872-875, Dec. 2012.

[37] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," *version 2.1a. Jun. 2015. [Online] Available: http://cvxr.com/cvx/doc/CVX.pdf*.

[38] J. Bibby, "Axiomatisations of the average and a further generalisation of monotonic sequences," *Glasgow Math. J.*, vol. 15, no. 1, pp. 63-65, 1974.

[39] N. D. Sidiropoulos, T. N. Davidson, and Z. Q. Luo, "Transmit beamforming for physical layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239-2251, Jun. 2006.

[40] L. Vandenberghe and S. Boyd, "Semidefinite relaxation of quadratic optimization problems," *SIAM Rev.*, vol. 38, no. 1, pp. 49-95, 1996.

[41] E. Karipidis, N. Sidiropoulos, and Z. Luo, "Quality of service and max-min-fair transmit beamforming to multiple co-channel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268-1279, Mar. 2008.

[42] O. Tervo, H. Pennanen, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Distributed optimization for coordinated beamforming in multicell multigroup multicast systems: Power minimization and SINR balancing," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 171-185, Jan. 2018.