Two-Level Transmission Scheme for Cache-Enabled Fog Radio Access Networks

Shiwen He [iD], *Member, IEEE*, Chenhao Qi [iD], *Senior Member, IEEE*, Yongming Huang [iD], *Senior Member, IEEE*, Qi Hou, and Arumugam Nallanathan [iD], *Fellow, IEEE*

Abstract—In this paper, we investigate the downlink transmission for cache-enabled fog radio access networks aiming at maximizing the delivery rate under the constraints of fronthaul capacity, maximum transmit power, and size of files. To reduce the delivery latency and the burden on fronthaul links and make full use of the local cache and baseband signal processing capabilities of enhanced remote radio heads (eRRHs), a two-level transmission scheme including cache-level and network-level transmission is proposed. In cache-level transmission, only requested files cached at the local cache are transmitted to the corresponding users. The duration of cache-level transmission is the delay caused by the transfer between the baseband unit (BBU) and eRRHs as well as the signal processing at the BBU. The remaining requested files are jointly transmitted to the corresponding users at network-level transmission. For cache-level transmission, a centralized optimization algorithm is first presented and then a decentralized optimization algorithm is provided to avoid the exchange of signaling among eRRHs. Meanwhile, another centralized optimization algorithm is presented to tackle the optimization problem for network-level transmission. All presented algorithms are proved to converge to the Karush–Kuhn–Tucker solutions of the problems. Numerical results are provided to validate the effectiveness of the proposed transmission scheme as well as evaluating the system performance.

Index Terms—Edge caching, cache-enabled fog radio access network, coordinated beamforming, joint transmission.

S. He is with the School of Information Science and Engineering, Central South University, Changsha 410083, China (e-mail: shiwen.he.hn@csu.edu.cn).

C. Qi and Q. Hou are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: qch@seu.edu.cn; 220150860@seu.edu.cn).

Y. Huang is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: huangym@seu.edu.cn).

A. Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: a.nallanathan@qmul.ac.uk).

I. INTRODUCTION

TO COPE with the continuously increasing number of wireless devices, to satisfy the demand of high data rates, and to meet the stringent quality-of-service (QoS) requirement of the emerging wireless services, various advanced communication technologies have been proposed in recent years [1]. For example, cooperative communication [2], massive multiple-input multiple-output (MIMO) [3], and network densification [4] are key technologies to achieve high capacity, high data rate, and guarantee network coverage in the fifth generation (5G) mobile cellular systems. Especially for network densification, cloud radio access network (C-RAN) is considered as an emerging network architecture that shows significant promises in suppressing interference and coordinately allocating resource for wireless networks [6]–[8]. In C-RAN, the remote radio heads (RRHs) are connected to a central processor, i.e., baseband signal processing unit (BBU), through fronthaul links [9]–[11]. However, performing completely joint processing requires not only signaling overhead but also payload data sharing among all coordinated RRHs, resulting in tremendous burden on fronthaul links [12]–[16]. Thus, the design of transmission scheme needs to take the limitation of the capacity of fronthaul links into account. In [12]–[14], the maximization problem of weighted sum rate was investigated under the constraint of fronthaul capacity. In [15], the data-sharing strategy and compression strategy were studied regarding the energy efficiency of the C-RANs. In [16], joint design of fronthaul and radio access links for C-RANs was investigated. Note that a common assumption in aforementioned literature is that the coordinated RRHs do not have the ability to cache data files.

One of the most challenging requirements in 5G is the provision of connections with low end-to-end latency. In most cases, the latency should be smaller than 100 ms, while sometimes it is smaller than 10 ms [17]. The low latency is demanded by some applications such as augmented reality and vehicular communications [18]. One solution to reduce the end-to-end latency as well as the fronthaul bottleneck in C-RANs is using distributed caches at the edge of the mobile cellular network, referred to as fog radio access networks (F-RANs) [19]–[21]. The resulting RRHs are referred to as enhanced RRHs (eRRHs). The eRRHs can pre-fetch the most frequently requested files to the eRRHs' local caches during off-peak traffic periods so that the fronthaul overhead can be

reduced during the peak traffic periods. In this way, lower latency and higher spectral efficiency can be achieved [21]. In [22], the optimization of hybrid cache placement of data files for coordinated relaying networks was studied. Then the cache placement in F-RANs was investigated by fully considering the flexibility of physical-layer transmission and diverse content preferences of different users [23]. In [24], joint design of cloud and edge processing for the downlink of F-RANs was investigated for maximizing the minimum user delivery rate. In [25], the problem of minimizing the system cost that accounts for fronthaul overhead while satisfying the target signal-to-interference-plus-noise ratio (SINR) was considered. More recently, Xiang *et al.* [26] proposed to use wireless caching to enhance the physical layer security of video streaming in cellular networks with limited backhaul capacity.

In general, the sum rate maximization or power minimization problem subject to QoS and transmit power constraints needs to be centrally solved for interference channels [24]–[26]. Meanwhile, some decentralized optimization algorithms were also developed to reduce the burden of information collection [27]–[31]. In [27], a distributed weighted sum rate maximization (WSRMax) method was proposed via exploiting the relation between the user rate and the minimum mean square error (MMSE) for interfering broadcast channel. At each iteration all mobile terminals need to first estimate the covariance matrices of their received signals, and then compute and feed back some parameters to the eRRHs. To avoid the estimation of the covariance matrices at each mobile terminal, Weeraddana *et al.* [28] proposed a distributed approach to maximize the WSRMax problem with limited signaling exchange among eRRHs. In [29], both centralized and distributed beamforming algorithms were proposed to solve power minimization and SINR balancing problem. In [30], the max-min fairness problem under the total power constraint was tackled from the perspective of transmit power minimization, where a distributed hierarchical iterative algorithm was proposed based on the uplink-downlink duality. In [31], a decentralized optimization method was proposed by exploiting interference temperature [32] to maximize the system energy efficiency. However, the decentralized optimization algorithms in the aforementioned literature need signaling exchange in each iteration during the optimization process, which not only requires strict synchronization but also generates some backhaul or fronthaul burden among coordinated eRRHs. The signaling exchange per iteration between coordinated eRRHs is nevertheless not desirable for low latency communications.

The joint transmission of cached and uncached requested contents in [24]–[26] only used the caching to reduce the burden on fronthaul links without making full use of the caching to reduce the delivery latency. However, one of the main motivation of introducing edge caching is to reduce the delivery latency for satisfying the demands of the latency-sensitive traffic for future wireless communication system [33]. To the best of the authors' knowledge, how to fully exploit the capabilities, i.e., caching and baseband signal processing, of eRRHs to reduce the delivery latency and the overhead

of fronthaul links is still an enormous challenge for wireless transmission in F-RANs. In this paper, we investigate the downlink transmission mechanism for cache-enabled F-RANs. We divide the transmission of the requested files into two-level, i.e., cache-level transmission and network-level transmission. In cache-level transmission, only requested files cached at the local cache are transmitted to the corresponding users. The duration of cache-level transmission is the delay caused by the transfer between the BBU and eRRHs as well as by the signal processing at the BBU. The remaining requested files are jointly transmitted to the corresponding users at network-level transmission. Our objective is to maximize the delivery rate under the constraints of fronthaul capacity, maximum transmit power of each eRRH, and size of cached data files. The corresponding optimization algorithms are presented to address the problems of interest. Specifically, the contribution of this work is summarized as follows:

- A two-level transmission scheme including cache-level and network-level transmission is proposed to reduce the delivery latency and the burden on fronthaul links by making full use of the local cache and the signal processing capabilities of eRRHs.
- For cache-level transmission, coordinated beamforming is adopted to improve the spectral efficiency. A centralized optimization algorithm is firstly designed to maximize the sum of network delivery rates. To avoid the demand of centralized processing and the exchange of signaling among eRRHs, a decentralized algorithm is then presented to maximize the sum of network delivery rates by introducing a new concept of signal-to-interference-leakage-noise ratio (SILNR) for cache-level transmission.
- For network-level transmission, joint transmission is adopted to maximize the system performance such that the inter-user interference can be effectively controlled. At the same time, an optimization algorithm is developed using convex approximation methods.
- All presented algorithms are proved to converge to the Karush-Kuhn-Tucker (KKT) solutions of the problems. The computational complexity of all presented algorithms is also analyzed. Numerical results are further provided to corroborate the effectiveness of proposed algorithms and unveil that how to cache files and how many files to cache are key problems for cache-enabled F-RANs.

The remainder of this paper is organized as follows. The system model is described in Section II. Section III formulates the optimization problem of the proposed two-level transmission scheme. The design of optimization algorithm for cache-level transmission is investigated in Section IV. In Section V, an optimization algorithm for network-level transmission is developed. Numerical results are provided in Section VI. Finally, Section VII concludes the paper.

Notations: Bold lowercase and uppercase symbols represent column vectors and matrices, respectively. The superscripts $(\cdot)^T$ and $(\cdot)^H$ represent the matrix transpose and conjugate transpose, respectively. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. For a complex-valued number x and a matrix $\mathbf{A}$, $|x|$ and $|\mathbf{A}|$ denote the absolute value of x and the determinant of $\mathbf{A}$, respectively. $\text{tr}(\cdot)$ and $\|\cdot\|$ denote the trace and the
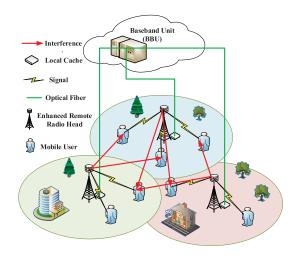
Fig. 1.   Illustration of the system model.

Euclidean norm, respectively. $\mathbf{0}_{N \times N}$ and $\mathbf{I}_{N \times N}$ denote the $N \times N$ zero matrix and $N \times N$ identity matrix, respectively. The circularly symmetric complex Gaussian distribution with mean μ and covariance matrix $\mathbf{R}$ is denotes by $\mathcal{CN}(\mu, \mathbf{R})$. $\mathrm{Diag}(\mathbf{a})$ denotes a diagonal matrix with the main diagonal given by $\mathbf{a}$.

II. SYSTEM MODEL

As illustrated in Fig. 1, the considered cache-enabled F-RANs include a BBU connected to K_{R} eRRHs serving totally K_{U} single-antenna users. The index sets of the eRRHs and the users are denoted as $\mathcal{K}_{\mathrm{R}} \triangleq \{1, \cdots, K_{\mathrm{R}}\}$ and $\mathcal{K}_{\mathrm{U}} \triangleq \{1, \cdots, K_{\mathrm{U}}\}$, respectively. The eRRH i equipped with $N_{\mathrm{t},i}$ antennas is connected to the BBU through an error-free fronthaul link with the capacity of C_i bits per symbol. The eRRH i is also equipped with a cache, which can store $nB_i > 0$ bits, $\forall i \in \mathcal{K}_{\mathrm{R}}$, where n is the number of symbols transmitted in each downlink transmission interval. The BBU in the "cloud" can perform signal processing globally, while the eRRHs at the edge of the "cloud" can only perform signal processing locally.

Assume that each user $k \in \mathcal{K}_{\mathrm{U}}$ requests contents or files from library $\mathcal{F} = \{1, \cdots, F\}$ stored in the BBU. Without loss of generality, we assume that all files in library $\mathcal{F}$ are of the same size of nS bits. For simplicity, in this paper, we further assume that the files cannot be partitioned and different eRRH caches different files. The eRRH $i \in \mathcal{K}_{\mathrm{R}}$ pre-stores $\lfloor B_i/S \rfloor$ files in its cache based on the user behavior, the delay-sensitive requirement, the long-term information of the popularity distribution, the fronthaul capacity, and etc [22]–[25]. The state of cached file f, $f \in \mathcal{F}$, can be modeled by defining binary variables $c_{f,i}$, $f \in \mathcal{F}$, $i \in \mathcal{K}_{\mathrm{R}}$, as

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

At the same time, $c_{f,i}$ satisfies the conditions $\sum\limits_{i \in \mathcal{K}_{\mathrm{R}}} c_{f,i} = 0$ or $\sum\limits_{i \in \mathcal{K}_{\mathrm{R}}} c_{f,i} = 1$, indicating that different eRRH caches

different files. For simplicity, in this work, we assume that the state information $c_{f,i}$, $f \in \mathcal{F}$, $i \in \mathcal{K}_{\mathrm{R}}$, is predetermined. Thus, if file f_k requested by user k is cached in the local cache of its serving eRRH, it can be retrieved directly from the eRRH without downloading from the BBU.[1] Otherwise, the information of file f_k needs to be transferred to its serving eRRH from the BBU via the fronthaul link.

III. PROBLEM FORMULATIONS

In order to make full use of the functionalities of cache as well as the signal processing capability of each eRRH, in this paper, the data transmission is divided into two-level transmission. The first level transmission is cache-level transmission where all eRRHs coordinately transmit cached files to scheduled users via coordinated beamforming. The duration of cache-level transmission is the delay caused by the signal processing at the BBU and the information exchange between the eRRH and the BBU. The second level transmission is network-level transmission where a quantized version of the precoded signals of the untransmitted requested files is sent to the eRRHs via fronthaul links and then all eRRHs jointly transmit it to K_{U} users via joint transmission.

A. Cache-Level Transmission

To reduce the latency incurred by the propagation of the fronthaul link and the signal processing at the BBU, eRRH i firstly transmits data to users, denoted as $\mathcal{K}_i \subseteq \mathcal{K}_{\mathrm{U}}$, whose requested files are stored in the local cache in cache-level transmission. Furthermore, $\mathcal{K}_i \cap \mathcal{K}_j \neq \emptyset$, $i \neq j \in \mathcal{K}_{\mathrm{R}}$, and $\cup_{i \in \mathcal{K}_{\mathrm{R}}} \mathcal{K}_i \subseteq \mathcal{K}_{\mathrm{U}}$. As a numerical example, user set $\mathcal{K}_i$ can be given by

$$\mathcal{K}_i = \left\{ k : \left| \{k' : c_{f_{k'},i} \|\mathbf{h}_{k',i}\| > c_{f_k,i} \|\mathbf{h}_{k,i}\|\} \right| < N_{\mathrm{s}} \\ \text{and } c_{f_k,i} == 1 \right\} \quad (2)$$

where N_{s} is a scalable parameter with $N_{\mathrm{s}} \leq N_{\mathrm{t},i}$ and f_k is the requested file index of user k. Here we assume that each user requests a single file. For cache-level transmission, the signal transmitted by the eRRH i can be expressed as

$$\mathbf{x}_i^{\mathrm{c}} = \sum_{k \in \mathcal{K}_i} c_{f_k,i} \mathbf{g}_{k,i} s_{f_k}. \quad (3)$$

where $\mathbf{g}_{k,i} \in \mathbb{C}^{N_{\mathrm{t},i} \times 1}$ is the beamforming vector and $\mathbf{g}_{k,i} = \mathbf{0}$ for $k \notin \mathcal{K}_i$. The signal received at user k, $\forall k \in \mathcal{K}_{\mathrm{U}}$, can be expressed as:

$$y_k^{\mathrm{c}} = \sum_{i \in \mathcal{K}_{\mathrm{R}}} \mathbf{h}_{k,i}^H \mathbf{x}_i^{\mathrm{c}} + n_k, \quad (4)$$

where $\mathbf{h}_{k,i} \in \mathbb{C}^{N_{\mathrm{t},i} \times 1}$ denotes the channel coefficients between user k and eRRH i and $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN). Note that in (4), we assume that baseband signal s_{f_k} and noise n_k are independent, $\forall i \in \mathcal{K}_{\mathrm{R}}, k \in \mathcal{K}_{\mathrm{U}}$. Thus, the achievable rate of user

---

[1]Each eRRH can retrieve the information of requested files by learning-based methods and is out of the scope of this work. The developed method in this work can be easily extended to the case where each eRRH caches multiple different files and the case where the file can be partitioned into multiple subfiles.

$$\gamma_{k,i}^{\mathrm{c}} = \frac{\left| c_{f_k,i} \mathbf{h}_{k,i}^H \mathbf{g}_{k,i} \right|^2}{\displaystyle\sum_{k' \in \mathcal{K}_i \setminus \{k\}} \left| c_{f_{k'},i} \mathbf{h}_{k,i}^H \mathbf{g}_{k',i} \right|^2 + \sum_{j \in \mathcal{K}_{\mathrm{R}} \setminus \{i\}} \sum_{k' \in \mathcal{K}_j} \left| c_{f_{k'},j} \mathbf{h}_{k,j}^H \mathbf{g}_{k',j} \right|^2 + \sigma^2}. \tag{5}$$

$k \in \mathcal{K}_{\mathrm{U}}$ served by eRRH i is calculated as $\mathrm{R}_{k,i}^{\mathrm{c}} = \log_2\left(1 + \gamma_{k,i}^{\mathrm{c}}\right)$, where $\gamma_{k,i}^{\mathrm{c}}$ is the signal-to-interference-plus-noise ratio (SINR) that is calculated as (5), at the top of this page.

The goal is to maximize the sum of delivery rate by optimizing the transmit beamforming vectors at each level transmission. Thus, for cache-level transmission, the optimization of $\mathbf{g}_{k,i}$, $k \in \mathcal{K}_i$, $i \in \mathcal{K}_{\mathrm{R}}$, can be realized by solving the following problem[2]:

$$\max \ \tau \sum_{i \in \mathcal{K}_{\mathrm{R}}} \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^{\mathrm{c}}, \tag{6a}$$

$$s.t. \ \mathrm{r}_{k,i}^{\mathrm{c}} \le \mathrm{R}_{k,i}^{\mathrm{c}}, \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_{\mathrm{R}}, \tag{6b}$$

$$\tau \mathrm{r}_{k,i}^{\mathrm{c}} \le S, \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_{\mathrm{R}}, \tag{6c}$$

$$\sum_{k \in \mathcal{K}_i} \left\| c_{f_k,i} \mathbf{g}_{k,i} \right\|^2 \le P_i, \quad \forall i \in \mathcal{K}_{\mathrm{R}}, \tag{6d}$$

where the variable is $\left\{ \mathbf{g}_{k,i}, \mathrm{r}_{k,i}^{\mathrm{c}} \right\}_{k \in \mathcal{K}_i, i \in \mathcal{K}_{\mathrm{R}}}$. $\mathrm{r}_{k,i}^{\mathrm{c}}$ denotes the delivery rate of user k served by eRRH i, $\forall k \in \mathcal{K}_i$, $\forall i \in \mathcal{K}_{\mathrm{R}}$, for cache-level transmission. In problem (6), constraint (6b) shows that the delivery rate $\mathrm{r}_{k,i}^{\mathrm{c}}$ is bounded by the achievable data rate $\mathrm{R}_{k,i}^{\mathrm{c}}$. Constraint (6c) denotes that the number of delivery data cannot exceed the cached file size S. Constraint (6d) means the maximum transmit power constraint on the beamforming vectors. Delay τ is a summation of the signal processing time at the BBU and the information exchange time between eRRH i and the BBU.[3] From the description of problem (6), one can see that the duration of cache-level transmission is also the delay τ.

B. Network-Level Transmission

For network-level transmission, all eRRHs coordinately transmit data to all scheduled users by using joint transmission [2]. Although joint transmission can effectively control the intra- and inter-eRRH interference and improve the quality of service for each user, the BBU needs to collect all channel state information (CSI) and allocates the processed data files to each eRRH resulting in a considerable large amount of overhead on the fronthaul links. For network-level transmission,

the signal transmitted by eRRH i is expressed as[4]:

$$\mathbf{x}_i^{\mathrm{n}} = \mathbf{x}_i + \mathbf{q}_i = \sum_{k \in \mathcal{K}_{\mathrm{U}}} \mathbf{u}_{k,i} s_{f_k} + \mathbf{q}_i, \tag{7}$$

where $\mathbf{u}_{k,i} \in \mathbb{C}^{N_{\mathrm{t},i} \times 1}$ is the beamforming vector for file f_k used at eRRH i. The quantization noise $\mathbf{q}_i \in \mathbb{C}^{N_{\mathrm{t},i} \times 1}$ is assumed to be independent of the transmitting signal and distributed as $\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$. The signals $\mathbf{x}_i$ and $\mathbf{x}_j$ for different eRRHs $i \ne j$ are quantified independently so that the quantization noise signals $\mathbf{q}_i$ and $\mathbf{q}_j$ are independent [24]. For ease of presentation, let $\mathbf{\Omega}$ be the covariance matrix of quantization noise, i.e., $\mathbf{\Omega} = \mathrm{Diag}(\mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_{K_{\mathrm{R}}})$. For network-level transmission, the signal received at user k, $\forall k \in \mathcal{K}_{\mathrm{U}}$, can be expressed as:

$$y_k^{\mathrm{n}} = \sum_{i \in \mathcal{K}_{\mathrm{R}}} \mathbf{h}_k^H \mathbf{x} + n_k, \tag{8}$$

where $\mathbf{h}_k = \left[\mathbf{h}_{k,1}^H, \cdots, \mathbf{h}_{k,K_{\mathrm{R}}}^H \right]^H$ and $\mathbf{x} = \left[(\mathbf{x}_i^{\mathrm{n}})^H, \cdots, \left(\mathbf{x}_{K_{\mathrm{R}}}^{\mathrm{n}}\right)^H \right]^H$. For the joint transmission, we implicitly assume that the stringent synchronization of timing and frequency among eRRHs has been finished. Similarly, the achievable rate of user $k \in \mathcal{K}_{\mathrm{U}}$ is calculated as $\mathrm{R}_k^{\mathrm{n}} = \log_2(1 + \gamma_k^{\mathrm{n}})$, where SINR γ_k^{n} is given by:

$$\gamma_k^{\mathrm{n}} = \frac{\left| \mathbf{h}_k^H \mathbf{u}_k \right|^2}{\displaystyle\sum_{k' \in \mathcal{K}_{\mathrm{U}}, k' \ne k} \left| \mathbf{h}_k^H \mathbf{u}_{k'} \right|^2 + \mathbf{h}_k^H \mathbf{\Omega} \mathbf{h}_k + \sigma^2}, \tag{9}$$

where $\mathbf{u}_k = \left[\mathbf{u}_{k,1}^{\mathrm{H}}, \cdots, \mathbf{u}_{k,K_{\mathrm{R}}}^{\mathrm{H}} \right]^{\mathrm{H}}$.

Similarly, for network-level transmission, the transmit beamforming vectors $\mathbf{u}_k$, $k \in \mathcal{K}_{\mathrm{U}}$, are optimized centrally at the BBU by addressing the following problem:

$$\max \ \sum_{k \in \mathcal{K}_{\mathrm{U}}} \mathrm{r}_k^{\mathrm{n}}, \tag{10a}$$

$$s.t. \ \mathrm{r}_k^{\mathrm{n}} \le \mathrm{R}_k^{\mathrm{n}}, \quad \forall k \in \mathcal{K}_U, \tag{10b}$$

$$\sum_{k \in \mathcal{K}_{\mathrm{U}}} \left\| \mathbf{B}_i \mathbf{u}_k \right\|^2 + \mathrm{tr}(\mathbf{\Omega}_i) \le P_i, \quad \forall i \in \mathcal{K}_R, \tag{10c}$$

$$\mathrm{r}_k^{\mathrm{n}} \le \max\left(S - \tau \overline{\mathrm{r}}_k^{\mathrm{c}}, 0\right), \quad \forall k \in \mathcal{K}_U, \tag{10d}$$

$$\log_2\left(\left|\overline{\mathbf{A}}_i\right|\right) - \log_2\left(\left|\mathbf{\Omega}_i\right|\right) \le C_i, \quad \forall i \in \mathcal{K}_R, \tag{10e}$$

where the variables are $\{\mathbf{u}_k, \mathrm{r}_k^{\mathrm{n}}, \mathbf{\Omega}_i\}_{k \in \mathcal{K}_{\mathrm{U}}, i \in \mathcal{K}_{\mathrm{R}}}$. Also note that $\mathrm{r}_k^{\mathrm{n}}$ denotes the delivery rate of user k during network-level transmission, $\forall k \in \mathcal{K}_{\mathrm{U}}$, and $\overline{\mathbf{A}}_i =$

---

[2]The idea of this work can be extended to the scenario where multiple users request the same file via multi-cast transmission as well as the scenario where the user equipped with multiple antennas requests multiple files via spatial multiplexing. What we need to do is to adjust the formulation of problems (6) and (10) according to the parameters of the corresponding scenario. However, it is left as our future work.

[3]For simplicity, we assume that the delay τ is the same for all eRRHs. For full caching case, i.e., all requested files are stored in local cache, the value of the delay τ is set be unit in optimization problem (6) and network-level transmission is not need.

[4]Note that point-to-point compression algorithm is adopted to quantize the precoded signals for each eRRH during the network-level transmission. As pointed out in [8], we can adopt some advanced compression algorithm to quantize the precoded signals such that the system performance can be further improved. Therefore, we would like to study the performance of the proposed two-level transmission scheme with advanced compression algorithm in our future work.

$\mathbf{B}_i \sum_{k \in \mathcal{K}_U} (1 - c_{f_k,i}) \mathbf{u}_k \mathbf{u}_k^H \mathbf{B}_i^H + \mathbf{\Omega}_i$ with matrix $\mathbf{B}_i \in \mathbb{C}^{N_{t,i} \times N_t}$ being defined as

$$\mathbf{B}_i = \left[\mathbf{0}_{N_{t,i} \times N_{\mathbf{B}_i}^1}, \mathbf{I}_{N_{t,i} \times N_{t,i}}, \mathbf{0}_{N_{t,i} \times N_{\mathbf{B}_i}^2} \right]$$

with $N_{\mathbf{B}_i}^1 = \sum_{j=1}^{i-1} N_{t,j}$ and $N_{\mathbf{B}_i}^2 = N_{tt} - \sum_{j=1}^{i} N_{t,j}$, where $N_{tt} = \sum_{i \in \mathcal{K}_R} N_{t,i}$. $\overline{\mathrm{r}}_k^c = \max_{i \in \mathcal{K}_R} \mathrm{r}_{k,i}^c$ denotes the maximum transmission rate of user k at cache-level transmission.[5] In problem (10), (10d) constrains the size of network-level transmission file not to exceed the remaining file size $\max(S - \tau \overline{\mathrm{r}}_k^c, 0)$.[6] Constraint (10e) is due to the fronthaul capacity constraints ensuring the signal $\mathbf{x}_i^n$ can be reliably recovered by eRRH i [34, Ch. 3]. Note that when the delay τ is zero, problem (10) becomes the optimization problem of precoding matrices for traditional cloud-based coordinated joint transmission system under the constraints of the fronthaul capacity, the size of data files, and the maximum transmit power of eRRHs.

From the aforementioned analysis, one can see that the minimum delivery content of our proposed two-level transmission scheme is the sum of the minimum delivery rate achieved by cache-level transmission and that of network-level transmission, i.e., $\mathcal{T}_P = \tau \sum_{i \in \mathcal{K}_R} \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^c + \sum_{k \in \mathcal{K}_U} \mathrm{r}_k^n$. The total minimum delivery content is $\mathcal{T}_T = \sum_{k \in \mathcal{K}_U} \mathrm{r}_k^n$ for the traditional cloud-based coordinated joint transmission system.

Remark 1: Note that in network-level transmission, all eRRHs transmit quantized precoding signals received from the BBU. However, in [21], all eRRHs transmit a superposition of precoded signals received from the BBU and locally precoded signals based on the caching files. The coupling of the precoding matrices used at all eRRHs make the optimization problem more difficult such that a central controller is used to optimize the precoding matrices [21]. Thus, the precoding matrices used at each eRRH need to be transmitted to the corresponding eRRH. In our proposed scheme, we use the caching files to simultaneously reduce the delivery latency and the burden on fronthaul links. Namely, in cache-level transmission, the required files which are cached at eRRHs are first transmitted to the corresponding users during the duration of the delay τ. Meanwhile, the burden on fronthaul links is also reduced under the constraints of (10d) and (10e).

IV. DESIGN OF OPTIMIZATION ALGORITHM FOR CACHE-LEVEL TRANSMISSION

In this section, we first focus on addressing problem (6) by designing a centralized optimization method. Then, to avoid the exchange of signaling between coordinated eRRHs, a novel concept of signal-to-interference-leakage-noise ratio (SILNR) is introduced to maximize the intra-eRRH delivery rate while

suppressing inter-eRRH interference. Finally, we design a decentralized optimization algorithm to obtain the coordinated beamforming vectors for cache-level transmission.

A. Centralized Optimization for Cache-Level Transmission

For coordinated interference networks, the existing of intra- and inter-eRRH interference leads the user rate $\mathrm{R}_{k,i}^c$ to be a nonconcave function. In other words, it is generally difficult to find the global optimal solution of (6). Therefore, this requires the relaxations of the optimization conditions in order to provide reasonable design for practical implementations. The first thing of addressing problem (6) is to transfer it into a solvable form by using some mathematical methods. In the sequel, we introduce some auxiliary variables $\epsilon_{k,i}$, $\forall k \in \widetilde{\mathcal{K}}_i$, $\forall i \in \mathcal{K}_R$, which denote the interference leakage limitation [35]. Thus, problem (6) is equivalently reformulated as:

$$\max \sum_{i \in \mathcal{K}_R} \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^c, \tag{11a}$$

$$s.t. \ \mathrm{r}_{k,i}^c \leq \widehat{\mathrm{R}}_{k,i}^c, \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_R, \tag{11b}$$

$$\sum_{k \in \mathcal{K}_i} \left| c_{f_k,i} \mathbf{h}_{k',i}^H \mathbf{g}_{k,i} \right|^2 \leqslant \epsilon_{k',i}, \quad \forall k' \in \widetilde{\mathcal{K}}_i, \ \forall i \in \mathcal{K}_R, \tag{11c}$$

$$\tau \mathrm{r}_{k,i}^c \leq S, \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_R, \tag{11d}$$

$$\sum_{k \in \mathcal{K}_i} \left\| c_{f_k,i} \mathbf{g}_{k,i} \right\|^2 \leq P_i, \quad \forall i \in \mathcal{K}_R, \tag{11e}$$

where the variables are $\left\{ \mathbf{g}_{k,i}, \mathrm{r}_{k,i}^c \right\}_{k \in \mathcal{K}_i, i \in \mathcal{K}_R}$ and $\{\epsilon_{k,i}\}_{k \in \widetilde{\mathcal{K}}_i, i \in \mathcal{K}_R}$. The set $\widetilde{\mathcal{K}}_i$ is given by

$$\widetilde{\mathcal{K}}_i = \left\{ k' : k' \in \mathcal{K}_U \setminus \mathcal{K}_i \text{ and } c_{f_{k'},j} = 1, \forall j \in \mathcal{K}_R \setminus \{i\} \right\}.$$

and $\widehat{\mathrm{R}}_{k,i}^c = \log_2 \left(1 + \widehat{\gamma}_{k,i}^c \right)$ with $\widehat{\gamma}_{k,i}^c$ redefined as

$$\widehat{\gamma}_{k,i}^c = \frac{\left| c_{f_k,i} \mathbf{h}_{k,i}^H \mathbf{g}_{k,i} \right|^2}{\sum_{k' \in \mathcal{K}_i \setminus \{k\}} \left| c_{f_{k'},i} \mathbf{h}_{k,i}^H \mathbf{g}_{k',i} \right|^2 + \overline{\sigma}_{k,i}^2} \tag{12}$$

where $\overline{\sigma}_{k,i}^2 = \sum_{j \in \mathcal{K}_R \setminus \{i\}} \epsilon_{k,j} + \sigma^2$. Constraint (11c) is activated at the optimal point of problem (11). Unfortunately, problem (11) is still a nonconvex due to the existing of inter-user interference in the denominator of $\widehat{\gamma}_{k,i}^c$ defined in (12). To overcome this difficulty, we firstly reformulate problem (11) into the following equivalent form:

$$\max \sum_{i \in \mathcal{K}_R} \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^c, \tag{13a}$$

$$s.t. \ \mathrm{r}_{k,i}^c \leq \log_2 \left(1 + \overline{\gamma}_{k,i}^c \right), \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_R, \tag{13b}$$

$$\overline{\gamma}_{k,i}^c \leqslant \frac{\left| c_{f_k,i} \mathbf{h}_{k,i}^H \mathbf{g}_{k,i} \right|^2}{\chi_{k,i}}, \quad \forall k \in \mathcal{K}_i, \ \forall i \in \mathcal{K}_R, \tag{13c}$$

$$\sum_{k \in \mathcal{K}_i} \left| c_{f_k,i} \mathbf{h}_{k',i}^H \mathbf{g}_{k,i} \right|^2 \leqslant \epsilon_{k',i}, \quad \forall k' \in \widetilde{\mathcal{K}}_i, \ \forall i \in \mathcal{K}_R, \tag{13d}$$

---

[5] We assume that the BBU can obtain the values of $\mathrm{r}_{k,i}^c$, $\forall k \in \mathcal{K}_U$, $\forall i \in \mathcal{K}_R$, according to the cache state information $\{c_{f,i}\}_{f \in \mathcal{F}, i \in \mathcal{K}_R}$ and the transmission scheme adopted by each eRRH, i.e., (6).

[6] Note that different from [12]–[15], [24], and [25], the size of transmitted files in cache-level transmission is included in the constraints (10d). This provides a chance to increase transmit rate of untransmitted files, i.e., improve the system delivery rate.

$$\sum_{k'\in\mathcal{K}_i\setminus\{k\}}\left|c_{f_{k'},i}\mathbf{h}_{k,i}^H\mathbf{g}_{k',i}\right|^2+\overline{\sigma}_{k,i}^2\leq\chi_{k,i},\quad\forall k\in\mathcal{K}_i,$$
$$\forall i\in\mathcal{K}_\mathrm{R}, \tag{13e}$$

$$\tau\mathrm{r}_{k,i}^\mathrm{c}\leq S,\quad\forall k\in\mathcal{K}_i,\ \forall i\in\mathcal{K}_\mathrm{R}, \tag{13f}$$

$$\sum_{k\in\mathcal{K}_i}\|c_{f_k,i}\mathbf{g}_{k,i}\|^2\leq P_i,\quad\forall i\in\mathcal{K}_\mathrm{R}, \tag{13g}$$

where the variables are $\left\{\mathbf{g}_{k,i},\mathrm{r}_{k,i}^\mathrm{c},\overline{\gamma}_{k,i},\chi_{k,i}\right\}_{k\in\mathcal{K}_i,i\in\mathcal{K}_\mathrm{R}}$ and $\{\epsilon_{k,i}\}_{k\in\widetilde{\mathcal{K}}_i,i\in\mathcal{K}_\mathrm{R}}$. Note that in problem (13), constraints (13c), (13d), and (13e) are activated at the optimal point of problem (13). Although the function $\frac{\left|c_{f_k,i}\mathbf{h}_{k,i}^H\mathbf{g}_{k,i}\right|^2}{\chi_{k,i}}$ in (13c) is convex function, constraint (13c) is non-convex. Therefore, problem (13) is still non-convex. If $\widetilde{\mathcal{K}}_i=\emptyset$, the corresponding constraint given in (13d) can be removed, implying that the interference suppression among eRRH i, $\forall i\in\mathcal{K}_\mathrm{R}$ is not needed.

In what follows, we try to iteratively convexify the non-convex constraint (13c) such that an iterative optimization method is designed to address problem (13). In light of [36]–[39], we replace the right side of the inequality in (13c) by its lower bound, which now can be obtained by the first order approximation due to the convexity of $\frac{\left|c_{f_k,i}\mathbf{h}_{k,i}^H\mathbf{g}_{k,i}\right|^2}{\chi_{k,i}}$. In particular, we have

$$\frac{\left|c_{f_k,i}\mathbf{h}_{k,i}^H\mathbf{g}_{k,i}\right|^2}{\chi_{k,i}}\geq\Phi_{k,i}^{(I)}\left(\mathbf{g}_{k,i},\chi_{k,i}\right),\quad\forall k\in\mathcal{K}_i,\ \forall i\in\mathcal{K}_\mathrm{R}, \tag{14}$$

where $\Phi_{k,i}^{(I)}\left(\mathbf{g}_{k,i},\chi_{k,i}\right)$ is defined as

$$\Phi_{k,i}^{(I)}\left(\mathbf{g}_{k,i},\chi_{k,i}\right)\triangleq\frac{2\Re\left(c_{f_k,i}\left(\mathbf{g}_{k,i}^{(I)}\right)^{\mathrm{H}}\mathbf{h}_{k,i}\mathbf{h}_{k,i}^{\mathrm{H}}\mathbf{g}_{k,i}\right)}{\chi_{k,i}^{(I)}}$$
$$-\left(\frac{\left|c_{f_k,i}\mathbf{h}_{k,i}^H\mathbf{g}_{k,i}^{(I)}\right|}{\chi_{k,i}^{(I)}}\right)^2\chi_{k,i}. \tag{15}$$

In (15), $\mathbf{g}_{k,i}^{(I)}$ and $\chi^{(I)}$ denote the values of variables $\mathbf{g}_{k,i}$ and χ obtained at the I-th iteration, respectively. Thus, we resort to solving iteratively problem (16) to obtain the solution to problem (11) instead of solving problem (13) directly.

$$\max\ \sum_{i\in\mathcal{K}_\mathrm{R}}\sum_{k\in\mathcal{K}_i}\mathrm{r}_{k,i}^\mathrm{c}, \tag{16a}$$
$$s.t.\ (13\mathrm{b}),(13\mathrm{d}),(13\mathrm{e}),(13\mathrm{f}),(13\mathrm{g}), \tag{16b}$$
$$\overline{\gamma}_{k,i}^\mathrm{c}\leq\Phi_{k,i}^{(I)}\left(\mathbf{g}_{k,i},\chi_{k,i}\right),\quad\forall k\in\mathcal{K}_i,\ \forall i\in\mathcal{K}_\mathrm{R}, \tag{16c}$$

where the variables are $\left\{\mathbf{g}_{k,i},\mathrm{r}_{k,i}^\mathrm{c},\overline{\gamma}_{k,i}^\mathrm{c},\chi_{k,i}\right\}_{k\in\mathcal{K}_i,i\in\mathcal{K}_\mathrm{R}}$ and $\{\epsilon_{k,i}\}_{k\in\widetilde{\mathcal{K}}_i,i\in\mathcal{K}_\mathrm{R}}$. Problem (16) is a convex problem that can be easily solved with classical convex optimization solver, e.g., CVX [40]. The detailed steps used to solve problem (11) are summarized as Algorithm 1 where η is a predefined stop threshold.

Let r^c denote the objective value of problem (16). If we replace $\mathbf{g}_{k,i}$, $\mathrm{r}_{k,i}^\mathrm{c}$, $\overline{\gamma}_{k,i}^\mathrm{c}$, $\chi_{k,i}$, and $\epsilon_{k',i}$ with $\mathbf{g}_{k,i}^{(I)}$, $\left(\mathrm{r}_{k,i}^\mathrm{c}\right)^{(I)}$,

---

Algorithm 1 Beamformer Optimization for Cache-Level Transmission

1: Set $I=0$. Initialize non-zero $\mathbf{g}_{k,i}^{(I)}$ such that the power constraint is satisfied and $\mathrm{r}_{k,i}^{c(I)}=0,\forall k\in\mathcal{K}_i,i\in\mathcal{K}_\mathrm{R}$. Compute $\overline{\gamma}_{k,i}^{c(I)}$, $\chi_{k,i}^{(I)}$, and $\epsilon_{k',i}^{(I)}$ with $\mathbf{g}_{k,i}^{(I)}$, $k\in\mathcal{K}_i$, $k'\in\widetilde{\mathcal{K}}_i$, $i\in\mathcal{K}_\mathrm{R}$.

2: Solve (16) obtaining $\mathbf{g}_{k,i}^{(I+1)}$, $\mathrm{r}_{k,i}^{c(I+1)}$, $\overline{\gamma}_{k,i}^{c(I+1)}$, $\chi_{k,i}^{(I+1)}$, and $\epsilon_{k',i}^{(I+1)}$, $k\in\mathcal{K}_i$, $k'\in\widetilde{\mathcal{K}}_i$, $i\in\mathcal{K}_\mathrm{R}$ for the given $\mathbf{g}_{k,i}^{(I)}$ and $\chi_{k,i}^{(I)}$.

3: If $\left|\sum_{i\in\mathcal{K}_\mathrm{R}}\sum_{k\in\mathcal{K}_i}\mathrm{r}_{k,i}^{c(I+1)}-\sum_{i\in\mathcal{K}_\mathrm{R}}\sum_{k\in\mathcal{K}_i}\mathrm{r}_{k,i}^{c(I)}\right|\leqslant\eta$, stop iteration. Otherwise, set $I\leftarrow I+1$ and go to step 2.

$\left(\overline{\gamma}_{k,i}^c\right)^{(I)}$, $\chi_{k,i}^{(I)}$, and $\epsilon_{k',i}^{(I)}$, $k'\in\widetilde{\mathcal{K}}_i$, $k\in\mathcal{K}_i$, $i\in\mathcal{K}_\mathrm{R}$, respectively, all constraints are still satisfied, which means that the optimal solution of the I-th iteration is feasible point of problem (16) in the $(I+1)$-th iteration. This is because of the approximation in (16c) [37]. Thus, the objective obtained in the $(I+1)$-th iteration is no smaller than that in the I-th iteration, i.e., $(r^\mathrm{c})^{(I+1)}\geq(r^\mathrm{c})^{(I)}$. In other words, Algorithm 1 generates a nondecreasing sequence of objective values $(r^\mathrm{c})^{(I)}$. Moreover, the problem has an upper bound due to the power constraints. Therefore, the convergence of Algorithm 1 can be guaranteed by the monotonic boundary theorem [38]. Furthermore, following the same arguments as those in [37, Th. 1], we can prove that Algorithm 1 converges to a Karush-Kuhn-Tucker (KKT) solution of problem (11).

In Algorithm 1, Step 2 solves a convex problem, which can be efficiently implemented by primal-dual interior point method with approximate complexity of $O\left(\left(\overline{K}_\mathrm{U}\left(N_\mathrm{tm}+4\right)\right)^{3.5}\right)$, where $N_\mathrm{tm}=\max_{i\in\mathcal{K}_\mathrm{R}}N_{\mathrm{t},i}$, $\overline{K}_\mathrm{U}=\sum_{i\in\mathcal{K}_\mathrm{R}}|\mathcal{K}_i|$, and $O\left(\cdot\right)$ stands for the big-O notation [41]. The overall computational complexity is $O\left(\kappa_1\left(K_\mathrm{U}\left(N_\mathrm{tm}+4\right)\right)^{3.5}\right)$, where κ_1 denotes the number of the iteration of Algorithm 1.

B. Decentralized Optimization for Cache-Level Transmission

Though problem (6) can be solved by Algorithm 1, it needs to be implemented centrally at BBU, which results in large burden on fronthaul links. Note that the existing decentralized optimization methods that are designed to address the optimization problem for coordinated multiple point network need signaling exchange among BSs, which results in burden on backhaul link and incurs a certain delivery latency [27]–[31]. From the perspective of reducing the delivery latency caused by information exchange and making the best use of the available signal processing capability of each eRRH, Algorithm 1 and the existing decentralized optimization algorithms [27]–[31] are not suitable for cache-enabled F-RANs.

For F-RANs, the objective is to reduce delivery latency and the overhead of fronthaul links. With the help of the cached contents, each eRRH can design its transmitted signal based

$$\widetilde{\gamma}_{k,i}^{\mathrm{c}} = \frac{\left| c_{f_k,i} \mathbf{h}_{k,i}^H \mathbf{g}_{k,i} \right|^2}{\sum\limits_{k' \in \mathcal{K}_i \setminus \{k\}} \left| c_{f_{k'},i} \mathbf{h}_{k,i}^H \mathbf{g}_{k',i} \right|^2 + \sum\limits_{k' \in \widetilde{\mathcal{K}}_i} \sum\limits_{k \in \mathcal{K}_i} \left| c_{f_k,i} \mathbf{h}_{k',i}^H \mathbf{g}_{k,i} \right|^2 + \sigma^2}. \tag{17}$$

only on its local cached content. In addition, the signalling exchange between eRRHs results in a large potential latency and requires stringent synchronization between eRRHs. Therefore, the optimization of transmit beamformers is expected to be independently implemented at each eRRH for cache-level transmission. The denominator of the right item in (3) implies that eRRH i cannot independently optimize $\mathbf{g}_{k,i}$, $k \in \mathcal{K}_i$, due to the existing of inter-eRRH interference. In order to release the coupling among eRRHs and to achieve autonomous optimization of the beamformers, we introduce a new concept that is defined as the signal-to-interference-leakage-plus-noise ratio (SILNR), i.e., (17), at the top of this page. In the denominator of (17), the first term denotes the intra-eRRH interference and the second term denotes the total power leaked from user $k \in \mathcal{K}_i$ to all other user $k' \in \widetilde{\mathcal{K}}_i$. It is not difficult to see that the SILNR $\widetilde{\gamma}_{k,i}^{\mathrm{c}}$ only depends on the local CSI, i.e., the channel coefficient from eRRH i to all users, which can be easily obtained, $\forall i \in \mathcal{K}_{\mathrm{R}}$. To design an autonomous optimization, we replace the achievable rate $\mathrm{R}_{k,i}^{\mathrm{c}}$ in constraints (6b) with an approximated achievable rate $\widetilde{\mathrm{R}}_{k,i}^{\mathrm{c}} = \log_2 \left(1 + \widetilde{\gamma}_{k,i}^{\mathrm{c}} \right)$, $\forall k \in \mathcal{K}_i$, $\forall i \in \mathcal{K}_{\mathrm{R}}$. Thus, instead of directly addressing problem (6), the problem of beamformer optimization for the cache-level transmission is formulated as a series of parallel decentralized optimization problems given by (18) at eRRH i.

$$\max \ \tau \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^{\mathrm{c}}, \tag{18a}$$

$$s.t. \ \mathrm{r}_{k,i}^{\mathrm{c}} \leq \widetilde{\mathrm{R}}_{k,i}^{\mathrm{c}}, \quad \forall k \in \mathcal{K}_i, \tag{18b}$$

$$\tau \mathrm{r}_{k,i}^{\mathrm{c}} \leq S, \quad \forall k \in \mathcal{K}_i, \tag{18c}$$

$$\sum_{k \in \mathcal{K}_i} \| c_{f_k,i} \mathbf{g}_{k,i} \|^2 \leq P_i, \tag{18d}$$

where the variable is $\left\{ \mathbf{g}_{k,i}, \mathrm{r}_{k,i}^{\mathrm{c}} \right\}_{k \in \mathcal{K}_i}$. The goal of problem (18) is to maximize the local delivery rate, while taking the inter-eRRH interference suppression into account, i.e., maximizing the total delivery rate of cache-level transmission as much as possible. Similarly, problem (18) can be equivalently transformed into the following form:

$$\max \ \sum_{k \in \mathcal{K}_i} \mathrm{r}_{k,i}^{\mathrm{c}}, \tag{19a}$$

$$s.t. \ \mathrm{r}_{k,i}^{\mathrm{c}} \leq \log_2 \left(1 + \overline{\gamma}_{k,i}^{\mathrm{c}} \right), \quad \forall k \in \mathcal{K}_i, \tag{19b}$$

$$\overline{\gamma}_{k,i}^{\mathrm{c}} \leqslant \frac{\left| c_{f_k,i} \mathbf{h}_{k,i}^H \mathbf{g}_{k,i} \right|^2}{\chi_{k,i}}, \quad \forall k \in \mathcal{K}_i, \tag{19c}$$

$$\sum_{k \in \mathcal{K}_i} \left\| c_{f_k,i} \mathbf{h}_{k',i}^H \mathbf{g}_{k,i} \right\|^2 \leqslant \epsilon_{k',i}, \quad \forall k' \in \widetilde{\mathcal{K}}_i, \tag{19d}$$

$$\sum_{k' \in \mathcal{K}_i \setminus \{k\}} \left| c_{f_{k'},i} \mathbf{h}_{k,i}^H \mathbf{g}_{k',i} \right|^2 + \widetilde{\sigma}_{k,i}^2 \leq \chi_{k,i}, \quad \forall k \in \mathcal{K}_i, \tag{19e}$$

$$\tau \mathrm{r}_{k,i}^{\mathrm{c}} \leq S, \quad \forall k \in \mathcal{K}_i, \tag{19f}$$

$$\sum_{k \in \mathcal{K}_i} \| c_{f_k,i} \mathbf{g}_{k,i} \|^2 \leq P_i, \tag{19g}$$

where the variables are $\left\{ \mathbf{g}_{k,i}, \mathrm{r}_{k,i}^{\mathrm{c}}, \overline{\gamma}_{k,i}^{\mathrm{c}}, \chi_{k,i} \right\}_{k \in \mathcal{K}_i}$ and $\{\epsilon_{k',i}\}_{k' \in \widetilde{\mathcal{K}}_i}$. $\widetilde{\sigma}_{k,i}^2 = \sum\limits_{k' \in \widetilde{\mathcal{K}}_i} \epsilon_{k',i} + \sigma^2$. Furthermore, problem (19) can be independently solved at each eRRH with similar procedure as Algorithm 1 without any signaling exchange among eRRHs. The overall computational complexity is $O \left(\kappa_{i,2} \left(|\mathcal{K}_i| \left(N_{\mathrm{tm}} + 3 \right) + K_{\mathrm{U}} \right)^{3.5} \right)$, where $\kappa_{i,2}$ denotes the number of the iterations of decentralized optimization algorithm of problem (19).

Remark 2: Note that to obtain the performance gain, the centralized optimization algorithm needs to collect all channel state information (CSI) and data file to the central processing unit. However, the decentralized optimization algorithm needs only partial CSI that can be obtained by the training sequence of the uplink and optimize the beamforming vector without sharing data files. In time division duplex (TDD) system, each BS can directly estimate the CSI of its own users without any information exchange among eRRHs, by exploiting uplink and downlink reciprocity. In addition, each eRRH can estimate the crosstalk channels to other users served by other eRRHs from the reverse link.

V. DESIGN OF OPTIMIZATION ALGORITHM FOR NETWORK-LEVEL TRANSMISSION

In this section, we focus on addressing problem (10) with centralized way at the BBU. Different from the signal processing at each eRRHs, due to the fact that the BBU has the whole CSI and data files of all users, the BBU can centrally design the precoders for all users and jointly transmit data files to each user. Thus, the intra- and inter-eRRH interference can be efficiently controlled to maximize the sum of delivery data rate. Following a similar procedure used to solve problem (11), we firstly introduce some auxiliary variables ι_k, μ_k, $\forall k \in \mathcal{K}_{\mathrm{U}}$, and then reformulate equivalently problem (10) into the following form:

$$\max \ \sum_{k \in \mathcal{K}_{\mathrm{U}}} \mathrm{r}_k^{\mathrm{n}}, \tag{20a}$$

$$s.t. \ \mathrm{r}_k^{\mathrm{n}} \leq \log_2 \left(1 + \iota_k \right), \quad \forall k \in \mathcal{K}_U, \tag{20b}$$

$$\iota_k \leq \frac{\left| \mathbf{h}_k^H \mathbf{u}_k \right|^2}{\mu_k}, \quad \forall k \in \mathcal{K}_U, \tag{20c}$$

$$\sum_{k' \in \mathcal{K}_{\mathrm{U}}, k' \neq k} \left| \mathbf{h}_k^H \mathbf{u}_{k'} \right|^2 + \mathbf{h}_k^H \boldsymbol{\Omega} \mathbf{h}_k^H + \sigma^2 \leq \mu_k,$$

$$\forall k \in \mathcal{K}_U, \tag{20d}$$

$$\sum_{k \in \mathcal{K}_U} \|\mathbf{B}_i \mathbf{u}_k\|^2 + \mathrm{tr}\,(\mathbf{\Omega}_i) \leq P_i, \quad \forall i \in \mathcal{K}_R, \tag{20e}$$

$$\mathrm{r}_k^n \leq \max\left(S - \tau \overline{\mathrm{r}}_k^c, 0\right), \quad \forall k \in \mathcal{K}_U, \tag{20f}$$

$$\log_2\left(|\overline{\mathbf{A}}_i|\right) - \log_2\left(|\mathbf{\Omega}_i|\right) \leqslant C_i, \quad \forall i \in \mathcal{K}_R, \tag{20g}$$

where the variable is $\{\mathbf{u}_k, \mathrm{r}_k^n, \iota_k, \mu_k, \mathbf{\Omega}_i\}_{k \in \mathcal{K}_U, i \in \mathcal{K}_R}$. Note that the constraints (20c) and (20d) are activated at the optimal point of problem (20). The difficulties of solving problem (20) lie in (20c) and (20g), as those contraints are non-convex. To overcome the non-convexity of the right hands of the inequalities in (20c), we resort to the following inequality [36]

$$\frac{\left|\mathbf{h}_k^H \mathbf{u}_k\right|^2}{\mu_k} \geq \psi_k^{(I)}\left(\mathbf{u}_k, \mu_k\right), \quad \forall k \in \mathcal{K}_U, \tag{21}$$

where $\psi_k^{(I)}\left(\mathbf{u}_k, \mu_k\right)$ is defined as

$$\psi_k^{(I)}\left(\mathbf{u}_k, \mu_k\right) \triangleq \frac{2\Re\left(\left(\mathbf{u}_k^{(I)}\right)^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{u}_k\right)}{\mu_k^{(I)}} - \left(\frac{\left|\mathbf{h}_k^H \mathbf{u}_k^{(I)}\right|}{\mu_k^{(I)}}\right)^2 \mu_k. \tag{22}$$

To convexify the left side of constraint (20g), we need to transform the first function $\log_2(\cdot)$ into a convex form. According to the concavity property of function $\log_2(\cdot)$, we have

$$\log_2\left(|\overline{\mathbf{A}}_i|\right) \leq \varphi\left(\overline{\mathbf{A}}_i, \overline{\mathbf{B}}_i\right) \tag{23}$$

where $\overline{\mathbf{B}}_i = \mathbf{B}_i \sum_{k \in \mathcal{K}_U} \mathbf{u}_k^{(I)} \left(\mathbf{u}_k^{(I)}\right)^H \mathbf{B}_i^H + \mathbf{\Omega}_i^{(I)}$ and $\varphi(\mathbf{A}, \mathbf{B}) = \log_2\left(|\mathbf{B}|\right) + \frac{1}{\ln(2)} \mathrm{tr}\left(\mathbf{B}^{-1}\left(\mathbf{A} - \mathbf{B}\right)\right)$. Now, instead of directly solving problem (20), we iteratively solve the following problem to obtain its solution by replacing the right hands of the inequalities in (20c) with $\psi_k^{(I)}\left(\mathbf{u}_k, \mu_k\right), \forall k \in \mathcal{K}_U$ and $\log_2\left(|\overline{\mathbf{A}}_i|\right)$ with the right side of (23).

$$\max \sum_{k \in \mathcal{K}_U} \mathrm{r}_k^n, \tag{24a}$$

$$s.t.\ (20b), (20d), (20e), (20f), \tag{24b}$$

$$\iota_k \leq \psi_k^{(I)}\left(\mathbf{u}_k, \mu_k\right), \quad \forall k \in \mathcal{K}_U, \tag{24c}$$

$$\varphi\left(\overline{\mathbf{A}}_i, \overline{\mathbf{B}}_i\right) - \log_2\left(|\mathbf{\Omega}_i|\right) \leqslant C_i, \quad \forall i \in \mathcal{K}_R, \tag{24d}$$

where the variable is $\{\mathbf{u}_k, \mathrm{r}_k^n, \iota_k, \mu_k, \mathbf{\Omega}_i\}_{k \in \mathcal{K}_U, i \in \mathcal{K}_R}$. After applying the convexity operations, problem (20) is transformed into approximated convex problem (24) that can be easily solved with convex optimization tools. A step-by-step summary of the beamforming design is given in Algorithm 2 for network-level transmission. Note that when τ equals to zero, problem (10) is equivalent to the optimization of beamforming vectors in traditional cloud-based coordinated joint transmission, which means that Algorithm 2 can also be used for the optimization of traditional cloud-based coordinated joint transmission systems.

Similarly, the sequence generated by Algorithm 2 is also nondecreasing. In other words, Algorithm 2 also converges to a KKT solution of problem (10). In Algorithm 2, Step 2 solves a convex optimization problem, which can be efficiently implemented by the primal-dual interior point method with approximate complexity of $O\left(\left(K_U\left(N_{tt} + 3\right) + K_R N_{tm}^2\right)^{3.5}\right)$ [41].

Algorithm 2 Beamformer Optimization for Network-Level Transmission

1: Set $I = 0$. Initialize non-zero $\mathbf{u}_k^{(I)}$ such that power constraint is satisfied and $\mathrm{r}_k^{n(I)} = 0, \forall k \in \mathcal{K}_U$. Compute $\iota_k^{(I)}$ and $\mu_k^{(I)}$ with $\mathbf{u}_k^{(I)}, k \in \mathcal{K}_U$.

2: Solve (24) obtaining $\mathbf{u}_k^{(I+1)}$, $\mathrm{r}_k^{n(I+1)}$, $\iota_k^{(I+1)}$, and $\mu_k^{(I+1)}$, $k \in \mathcal{K}_U$ for the given $\mathbf{u}_k^{(I)}$ and $\mu_k^{(I)}$.

3: If $\left|\sum_{k \in \mathcal{K}_U} \mathrm{r}_k^{n(I+1)} - \sum_{k \in \mathcal{K}_i} \mathrm{r}_k^{n(I)}\right| \leqslant \eta$, then stop iteration. Otherwise, set $I \leftarrow I + 1$ and go to step 2.

TABLE I

SIMULATION PARAMETERS

Symbol	Value	Symbol	Value
K_R	3	K_U	5
N_t	4	F	10
B	$3S$	σ^2	1
d_0	50 m	α	3
η	10^{-3}	N_s	2

Suppose Algorithm 2 takes totally κ_3 number of the iteration to converge. The overall computational complexity is $O\left(\kappa_3\left(K_U\left(N_{tt} + 3\right) + K_R N_{tm}^2\right)^{3.5}\right)$. It can be observed that among our developed algorithms, the computational complexity of network-level optimization algorithm is higher than that of cache-level optimization algorithm and is similar with the proposed algorithms in [24].

VI. NUMERICAL RESULTS

In this section, we present some numerical results to evaluate the performance of proposed algorithms. We consider a cache-enabled F-RAN where the positions of eRRHs and UEs are uniformly distributed within a circular cell of radius of 500 m. The channel vector $\mathbf{h}_{k,i}$ from eRRH i to the 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 $\varrho_{k,i} = 1/\left(1 + \left(d_{k,i}/d_0\right)^\alpha\right)$ and the elements of $\widetilde{\mathbf{h}}_{k,i}$ are independent and identically distributed (i.i.d.) with zero mean and unit variance. For simplicity, we assume that the eRRHs have the same number of transmit antennas, the same transmit power, the same fronthaul capacity, and the same size of cache, i.e., $N_{t,i} = N_t$, $P_i = P$, $C_i = C$ and $B_i = B$, $\forall i \in \mathcal{K}_R$. If not stated elsewhere, the simulation is performed with the parameters given in Table I.

In the figures, "CO-CEHTM" and "DO-CEHTM" denote respectively the centralized optimization and decentralized optimization for cache-level transmission in our developed scheme. "TC-CJTM" denotes the traditional cloud-based coordinated joint transmission mechanism, i.e., $\tau = 0$ in Algorithm 2. We assume that all files have the same degree of popularity. Namely, the cache state information $c_{f,i}, f \in \mathcal{F}$, $i \in \mathcal{K}_R$, is randomly given subject to the cache capacity constraint and satisfies the conditions $\sum_{i \in \mathcal{K}_R} c_{f,i} = 0$ or $\sum_{i \in \mathcal{K}_R} c_{f,i} = 1$ while $\sum_{f \in \mathcal{F}} c_{f,i} = \lfloor B/S \rfloor$. The requested file index f_k of user k is random selected from set $\mathcal{F}$. For cache-level transmission
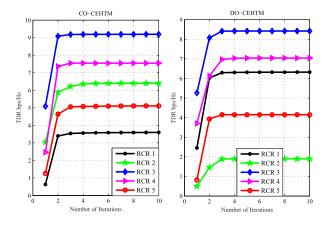
Fig. 2. Convergence trajectory of presented algorithms, $C = 5$ bits/symbol, $S = 10$, and $P = 20$ dB.
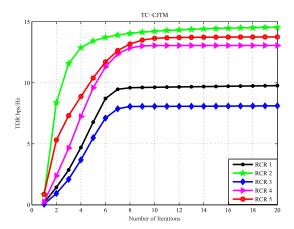

Fig. 3. Convergence trajectory of presented algorithms, $C = 5$ bits/symbol, $S = 10$, and $P = 20$ dB.

Fig. 4. Comparison of the average TDR and TAR of centralized and decentralized optimization algorithms for cache-level transmission, $\tau = 0.01$ s, $C = 5$ bit/symbol, and $S = 10$.

of full caching, the user set served by each eRRH does not intersect each other, i.e., each user is served by only one eRRH. All simulation results are obtained by averaging over 1000 random channel realizations.

Fig. 2 and Fig. 3 illustrates respectively the convergence trajectory of the above mentioned three algorithms for a few randomly channel realizations (RCR) where $\tau = 1$ ms for CO-CEHTM and DO-CEHTM and $\tau = 0$ ms for TC-CJTM. It is seen that all algorithms generate a non-decreasing total delivery rate (TDR) sequence. Recalling that all the objective functions of problem (11), (20), and (20) are boundary function due to the limitation of maximum allowable transmission power. The convergence of all algorithms can be guaranteed by the monotonic boundary theorem [38]. Furthermore, it is observed that all algorithms can fast converge to a stationary point within limited number iterations, e.g., 10 iterations.

Fig. 4 shows the comparison of average total delivery rate (TDR) and total achievable data rate (TAR) for centralized and decentralized optimization algorithms for cache-level transmission, respectively. It is shown that the performance of the decentralized and centralized algorithms is very close. The decentralized optimization algorithm can obtain more than 97% of the performance obtained by the centralized optimization algorithm in terms of TDR and TAR, respectively. The performance gain of the centralized optimization algorithm is
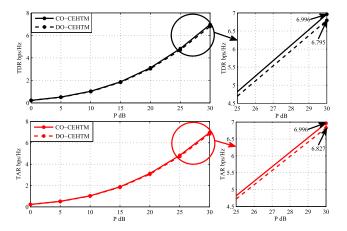
obtained at the cost of large number of backhaul overhead and CSI collection. In other words, the decentralized optimization algorithm is effective for delay sensitive data service and communication system with limited backhaul/fronthaul link.

Fig. 5 illustrates the average TDR versus the fronthaul capacity C for cache-enabled F-RANs with different delays. From this figure, we can observe that the average TDR increases with a larger capacity C of fronthaul link in the region of small to middle fronthaul capacity, in which the performance is limited by the capacity of fronthaul link rather than the file size and the achievable data rate. However, for large capacity of fronthaul link, the average TDR performance is limited by other factors instead of the fronthaul capacity C and becomes saturated. In addition, the centralized and decentralized optimization for cache-level transmission generate very close performance. Furthermore, increasing the delay results in increasing average TDR. The reason is that the duration of independent transmission at cache-level transmission or network level transmission increases with an increasing delay. In addition, the optimization of transmit beamforming vectors is transferred from eRRHs to the BBU such that the intra-cell interference and inter-cell interference are more effectively controlled. One can also note that the proposed two-level transmission scheme outperforms the algorithm proposed in [24] due to the delay τ which is used to transmit requested files.

In Fig. 6, we further evaluate the impact of the caching capacity B on the average TDR versus the fronthaul capacity C for cache-enabled F-RANs. When the fronthaul capacity is sufficiently small, full caching provides a better way for cache-enabled F-RANs. We note that the proposed two-level transmission scheme outperforms the TC-CJTM due to the fact that the proposed two-level transmission scheme exploits fully the duration of the delay τ to transmit data to partial users. Together with Fig. 5, it is shown that even if the delay τ is 10 ms, the proposed two-level transmission scheme still obtain a certain performance gain compared to the TC-CJTM. In addition, one can see that different caching capacity B have different impact on the TDR performance of the proposed two-level transmission scheme. It means that how to cache files and
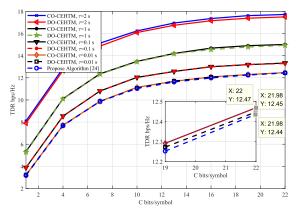
Fig. 5.   Comparison of the average TDR of hybrid transmission with centralized and decentralized optimization for cache-level transmission, $P = 20$ dB and $S = 10$.
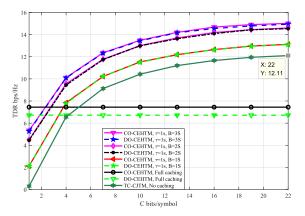

Fig. 6.   Comparison of the average TDR of hybrid transmission with centralized and decentralized optimization for cache-level transmission, $P = 20$ dB and $S = 10$.

Fig. 7.   Comparison of the average TDR of hybrid transmission with centralized and decentralized optimization for cache-level transmission, $P = 20$ dB and $C = 5$ bits/symbol.

Fig. 8.   Comparison of the average TDR of hybrid transmission with centralized and decentralized optimization for cache-level transmission, $C = 5$ bits/symbol, and $S = 10$.

how many files to cache at each eRRH are also key problems for cache-enabled F-RANs and caching all files at eRRHs does not necessarily obtain the optimal system performance.

In Fig. 7, we show the average TDR versus the file size S for cache-enabled F-RANs. Similar to the observation in Fig. 5, the performance of TDR improves with a larger file size S in the region of small fize sizes, in which the performance is limited by the file size S rather than other factors, such as the fronthaul capacity C, the transmit power P, and the achievable data rates. In the region of medium to large file sizes, the TDR becomes saturated where the TDR performance is limited by the fronthaul capacity C, the transmit power P, or the achievable data rates. Note that for the full caching, compared to Fig. 4, the performance gap between the centralized and decentralized optimization for the cache-level transmission increases due to the fact that the inter-eRRH interference suppression is limited to the number of the simultaneously served users. Numerical results further reveal that only when the file size is very small, the TDR performance of TC-CJTM is better than that of CO-CEHTM with full caching. It is due to the fact that the performance of TC-CJTM is constrained by the fronthaul capacity C, which does not affect the performance of CO-CEHTM. The proposed two-level transmission scheme with proper latency outperforms the transmission schemes with full or no caching.
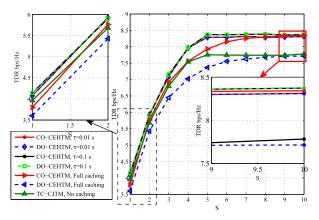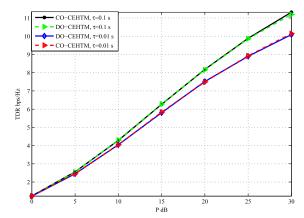
Fig. 8 demonstrates the average performance of TDR versus maximum allowable transmit power P for the downlink cache-enabled F-RANs. Numerical results also show that the performance of the centralized and decentralized optimization of cache-level transmission is very close. Furhtermore, the performance of the TDR increases with an increasing maximum transmit power P. In other words, in a certain power range, the TDR performance is not saturated.

VII. CONCLUSIONS

In this paper, a two-level transmission scheme including cache-level and network-level transmission has been proposed for cache-enabled F-RANs under the constraints of the fronthaul capacity and the maximum allowable transmit power of eRRHs. Centralized optimization algorithms has been developed to address the problem of interest for cache-level and network-level transmission, respectively. To avoid the signaling exchange among eRRHs, a decentralized optimization algorithm has been further developed for cache-level transmission. Finally, numerical results have been provided to evaluate the performance of our developed algorithms.

REFERENCES

[1] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Area Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[2] D. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[3] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 2, no. 2, pp. 114–123, Feb. 2016.

[4] J. Wang, W. Guan, Y. Huang, R. Schober, and X. You, "Distributed optimization of hierarchical small cell networks: A GNEP framework," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 249–264, Feb. 2017.

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

[6] I. V. Loumiotis, E. F. Adamopoulou, K. P. Demestichas, T. A. Stamatiadi, and M. E. Theologou, "Dynamic backhaul resource allocation: An evolutionary game theoretic approach," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 691–698, Feb. 2014.

[7] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

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

[9] A. Checko *et al.*, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2014.

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

[11] Y.-J. Ku *et al.*, "5G radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Commun. Mag.*, vol. 4, no. 1, pp. 46–52, Apr. 2017.

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

[13] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "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.

[14] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, Aug. 2016.

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

[16] S.-H. Park, K.-J. Lee, C. Song, and I. Lee, "Joint design of fronthaul and access links for C-RAN with wireless fronthauling," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1657–1661, Nov. 2016.

[17] R. Hattachi and J. Erfanian. (2015). *5G White Paper-Next Generation Mobile Networks (NGMN) Alliance*. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[18] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[19] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 4, no. 1, pp. 54–61, Apr. 2017.

[20] X. Hou, Y. Li, M. Chen, D. Wu, D. Jin, and S. Chen, "Vehicular fog computing: A viewpoint of vehicles as the infrastructures," *IEEE Trans. Veh. Technol.*, vol. 65, no. 6, pp. 3860–3873, Jun. 2016.

[21] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[22] 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.

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

[24] S.-H. 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.

[25] 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.

[26] L. Xiang, D. W. K. Ng, R. Schober, and V. W. S. 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.

[27] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[28] P. C. Weeraddana, M. Codreanu, M. Latva-Aho, and A. Ephremides, "Multicell MISO downlink weighted sum-rate maximization: A distributed approach," *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 556–570, Feb. 2013.

[29] 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.

[30] Y. Huang, G. Zheng, M. Bengtsson, K.-K. Wong, L. Yang, and B. Ottersten, "Distributed multicell beamforming with limited intercell coordination," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 728–738, Feb. 2011.

[31] Y. Huang, S. He, S. Jin, and W. Chen, "Decentralized energy-efficient coordinated beamforming for multicell systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4302–4314, Nov. 2014.

[32] R. Zhang and S. Cui, "Cooperative interference management with MISO beamforming," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5450–5458, Oct. 2010.

[33] 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.

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

[35] S. He, Y. Huang, H. Wang, S. Jin, and L. Yang, "Leakage-aware energy-efficient beamforming for heterogeneous multicell multiuser systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1268–1281, Jun. 2014.

[36] K. G. Nguyen, L. N. Tran, O. Tervo, Q. D. Vu, and M. Juntti, "Achieving energy efficiency fairness in multicell MISO downlink," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1426–1429, Aug. 2015.

[37] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, pp. 681–683, Jul. 1978.

[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] 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, May 2010.

[40] M. Grant and S. Boyd. (Jun. 2015). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1a*. [Online]. Available: http://cvxr.com/cvx/doc/CVX.pdf

[41] F. A. Potra and S. J. Wright, "Interior-point methods," *J. Comput. Appl. Math.*, vol. 124, nos. 1–2, pp. 281–302, Jan. 2000.

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

Shiwen He (M'14) received the M.S. degree from the Chengdu University of Technology, Chengdu, China, in 2009, and the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, China, in 2013.

From 2013 to 2015, he was a Post-Doctoral Researcher with the State Key Laboratory of Millimeter Waves, Department of Radio Engineering, Southeast University. From 2015 to 2018, he was with the School of Information Science and Engineering, Southeast University. Since 2018, he has been a Professor with the School of Information Science and Engineering, Central South University, Changsha, China. He has authored or co-authored more than 100 technical publications, held 60 granted invention patents, and submitted more than 20 technical proposals to IEEE standards. His research interests include multiuser MIMO communication, cooperative communications, energy efficient communications, millimeter wave communication, wireless local area network, mobile edge computing, and optimization theory. He has served as an Associate Editor for *EURASIP Journal on Wireless Communications*.

Chenhao Qi (S'06–M'10–SM'15) received the B.S. and Ph.D. degrees in signal processing from Southeast University, Nanjing, China, in 2004 and 2010, respectively.

From 2008 to 2010, he visited the Department of Electrical Engineering, Columbia University, New York, NY, USA. Since 2010, he has been with the faculty of the School of Information Science and Engineering, Southeast University, where he is currently an Associate Professor. His research interests include sparse signal processing and wireless communications.

Dr. Qi is an Associate Editor of the IEEE COMMUNICATIONS LETTERS and the IEEE ACCESS.

Yongming Huang (M'10–SM'16) received the B.S. and M.S. degrees from Nanjing University, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, China, in 2007.

Since 2007, he has been a Faculty Member with the School of Information Science and Engineering, Southeast University, where he is currently a Full Professor. From 2008 to 2009, he was visiting the Signal Processing Laboratory, Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden. He has published over 200 peer-reviewed papers, holds over 60 invention patents, and submitted over 10 technical contributions to IEEE standards. His current research interests include MIMO wireless communications, cooperative wireless communications, and millimeter wave wireless communications. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE WIRELESS COMMUNICATIONS LETTERS, *EURASIP Journal on Advances in Signal Processing*, and *EURASIP Journal on Wireless Communications and Networking*.

Qi Hou received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2015. She is currently pursuing the master's degree with Southeast University, Nanjing. Her research interests include green communications and millimeter wave communications.

Arumugam Nallanathan (S'97–M'00–SM'05–F'17) was with the Department of Informatics, Kings College London, from 2007 to 2017, where he was a Professor of wireless communications from 2013 to 2017, and also a Visiting Professor from 2017. He was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from 2000 to 2007. He has been a Professor of wireless communications and the Head of the Communication Systems Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. He has published nearly 400 technical papers in scientific journals and international conferences. His research interests include 5G wireless networks, Internet of Things, and molecular communications. He is a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 (ICC'2016), the IEEE Global Communications Conference 2017 (GLOBECOM'2017), and the IEEE Vehicular Technology Conference 2018 (VTC'2018). He is an IEEE Distinguished Lecturer. He has been selected as a Web of Science Highly Cited Researcher in 2016.

Dr. Nallanathan was a recipient of the IEEE Communications Society SPCE Outstanding Service Award 2012 and the IEEE Communications Society RCC Outstanding Service Award 2014. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of the IEEE Communications Society and Technical Program Chair and a member of the Technical Program Committees in numerous IEEE conferences. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2006–2011), the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (2006–2017), the IEEE WIRELESS COMMUNICATIONS LETTERS, and the IEEE SIGNAL PROCESSING LETTERS.