Joint Optimization of User Association, Data Delivery Rate and Precoding for Cache-Enabled F-RANs

Tung T. Vu*, Duy T. Ngo*, Lawrence Ong*, Salman Durrani† and Richard H. Middleton*
*School of Electrical Engineering and Computing, The University of Newcastle, Callaghan NSW 2308, Australia
Email: thanhtung.vu@uon.edu.au, {duy.ngo, lawrence.ong, richard.middleton}@newcastle.edu.au
†Research School of Engineering, The Australian National University, Canberra ACT 2601, Australia
Email: salman.durrani@anu.edu.au

***Abstract*—This paper considers the downlink of a cache-enabled fog radio access network (F-RAN) with limited fronthaul capacity, where user association (UA), data delivery rate (DDR) and signal precoding are jointly optimized. We formulate a mixed-integer nonlinear programming problem in which the weighted difference of network throughput and total power consumption is maximized, subject to the predefined DDR requirements and the maximum transmit power at each eRRH. To address this challenging problem, we first apply the ℓ_0-norm approximation and ℓ_1-norm minimization techniques to deal with the UA. After this key step, we arrive at an approximated problem that only involves the joint optimization of DDR and precoding. By using the alternating descent method, we further decompose this problem into a convex subproblem for DDR allocation and a nonconvex subproblem for precoding design. While the former is globally solved by the interior-point method, the latter is solved by a specifically tailored successive convex quadratic programming method. Finally, we propose an iterative algorithm for the original joint optimization that is guaranteed to converge. Importantly, each iteration of the developed algorithm only involves solving simple convex problems. Numerical examples demonstrate that the proposed design significantly improves both throughput and power performances, especially in practical F-RANs with limited fronthaul capacity. Compared to the sole precoder design for a given cache placement, our joint design is shown to improve the throughput by 50% while saving at least half of the total power consumption in the considered examples.**

I. Introduction

A fog radio access network (F-RAN) is recently proposed as an alternative to the cloud radio access network (C-RAN) to support mobile edge computing [1]–[3]. Capable of exploiting the advantages of both local caching and centralized signal processing, this novel network architecture is expected to significantly improve both spectral and energy efficiencies of the fifth-generation (5G) of cellular systems [4]. In an F-RAN, traditional high-cost high-power base stations (BSs) are replaced by low-cost low-power enhanced remote radio heads (eRRHs). Equipped with a finite-storage cache, each eRRH is connected to a central base band unit (BBU) via a fronthaul link. If a user (UE) requests a file that is available at the local caches of its serving eRRHs, the file can be directly retrieved from the caches. Otherwise, the file will be fetched from the BBU to the serving eRRHs before being transferred the UEs via radio access links.

In practical settings, the performance of an F-RAN is constrained by the capacity of its fronthaul links. User association (UA) can help reduce the fronthaul traffic by assigning UEs to appropriate eRRHs and save power by putting the unassigned eRRHs into sleep mode [5]. To minimize the fronthaul traffic and transmission power, the work of [6] considers the joint design of multicast beamforming and BS clustering (which is essentially a UA problem) in cache-enabled C-RANs, where the BSs are assigned to the groups of UEs requesting the same file. To maximize the throughput, the joint design of data delivery rate (DDR) and precoding has been studied in [7]. However, since [7] assigns the UEs to the eRRHs heuristically, it may not exploit the full potential of UA to enhance the system performance.

In this paper, we jointly design UA, DDR and signal precoding for the downlink of a cache-enabled F-RAN with limited fronthaul capacity. We aim to maximize the weighted difference between the network throughput and the total power consumption, the latter of which consists of the operating power and the transmission power in both fronthaul and radio access links. The formulated optimization problem is constrained on meeting the predefined DDR requirements, the limited fronthaul capacity and the maximum transmit power at each eRRH. In this mixed-integer nonlinear program, the strong coupling among the optimizing variables makes it even more challenging to be solved globally.

Here, we first express the UA variables as the functions of the ℓ_0-norm of the precoding matrices. We approximate this ℓ_0-norm with its weighted ℓ_1-norm [8] and update the weight factor iteratively. We then obtain an approximated problem in the variables of DDRs and precoding matrices only. Next, we apply the alternating descent method in [9] to decompose the approximated problem into a convex subproblem for the DDR allocation and a nonconvex subproblem for the precoder design. The former is readily solved by a convex solver, whereas the latter is dealt with by the successive convex quadratic programming framework of [10]. Finally, we propose an iterative algorithm that is proved to converge once initialized from a feasible point. Each iteration of the algorithm corresponds to solving simple convex programs.

In the numerical examples with practical parameter settings, the proposed design demonstrates its capability in substantially improving both throughput and power performances. Compared to solely designing the precoders for a given cache placement, the developed joint design offers 50% throughput gain while consuming only 50% of the total power otherwise required. The performance enhancement is particularly pronounced in cases with limited-capacity fronthaul links.

Notation: Boldfaced symbols are used for vectors and capitalized boldfaced symbols for matrices. $\boldsymbol{X}^H$ is the conjugate transposition of a matrix $\boldsymbol{X}$. $\boldsymbol{I}$ and $\boldsymbol{0}$ are the identity and zero matrix with the appropriate dimensions respectively. $||.||_0$ denotes the ℓ_0-norm. $\langle \boldsymbol{X} \rangle$ means the trace of a matrix $\boldsymbol{X}$.

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

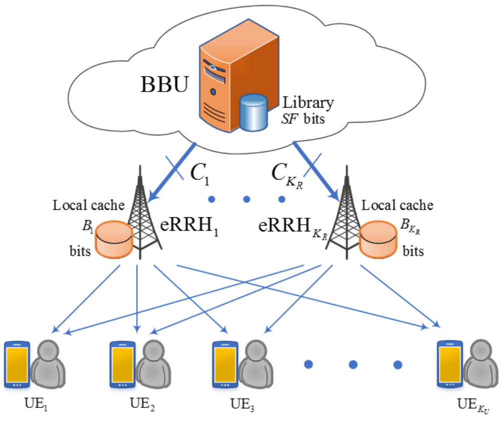


Fig. 1. Illustration of a general F-RAN.

II. System Model and Problem Formulation

Consider the general F-RAN model [7] illustrated in Fig. 1, where there are K_U UEs capable of establishing wireless connections with K_R RRHs. Each UE $k \in \mathcal{K}_U \triangleq \{1, \ldots, K_U\}$ is equipped with N_u antennas, whereas each eRRH $i \in \mathcal{K}_R \triangleq \{1, \ldots, K_R\}$ with N_r antennas. The eRRH $i \in \mathcal{K}_R$ connects to the baseband unit (BBU) in the core network via a fronthaul link of capacity $C_i > 0$ (b/s).

A. Data Request and Pre-fetching

First, in the data request phase each UE $k \in \mathcal{K}_U$ requests a random file f_k from the library $\mathcal{F} \triangleq \{1, ..., F\}$ stored in the BBU. Without loss of generality, assume that all files in the library are of the same size S bits; hence, the total size of the file library is SF bits. Denote by $\mathcal{F}^* \triangleq \bigcup_{k \in \mathcal{K}_U} \{f_k\}$ the set of all the files requested by all K_U users. To bring data contents closer to the UEs, each eRRH $i \in \mathcal{K}_R$ is equipped with a local cache that can store $B_i > 0$ bits.

It is practical to assume that the eRRHs have limited storage capacity and therefore only a subset of the file library is cached at each eRRH. In this paper, we adopt the fractional cache distinct strategy [7]. Each file in the library is split into M subfiles of equal size $\bar{S} = S/M$ bits. Each eRRH $i \in \mathcal{K}_R$ then randomly selects the set $\mathcal{M}^i$ of $\lfloor \frac{B_i}{\bar{S}} \rfloor$ subfiles from the BBU file library to store in its local cache. The cache state information, which shows whether the subfile (f_k, m) is cached at eRRH i or not, $f_k \in \mathcal{F}^*, \ m \in \mathcal{M} \triangleq \{1, \ldots, M\}$ is summarized as

$$c^i_{f_k,m} \triangleq \begin{cases} 1, & \text{if } (f_k, m) \in \mathcal{M}^i \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

With data pre-fetching, the files requested by the UEs can now be retrieved directly from the local cache of the serving eRRHs instead of from the BBU. If the subfile (f_k, m) requested by UE $k \in \mathcal{K}_U$ is not available at eRRH i's local cache, (f_k, m) will be fetched from the BBU via the fronthaul link.

In each data-request duration, the requested files and local caches are known. The cache state information $c^i_{f_k,m}$ $(f_k \in \mathcal{F}^*, \ m \in \mathcal{M}, \ i \in \mathcal{K}_R)$ is thus available at the BBU. In the following, we focus on the data delivery phase from the BBU and/or the eRRHs to the UEs in the transmission block (OTB) interval. Note that we do not optimize across multiple OTBs but instead only in one OTB. A similar optimization scheme can be applied to each OTBs.

B. Data Delivery

In the considered F-RAN, the central BBU allows each eRRH to serve multiple UEs and each UE to be served by multiple eRRHs. We model the associations of eRRHs and UEs by the following binary variables:

$$a_{k,i} \triangleq \begin{cases} 1, & \text{if eRRH } i \text{ serves UE } k \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Let $R_{f_k,m} \leq \bar{S}$ be the data delivery rate (DDR) of subfile (f_k, m). Here, we allow $R_{f_k,m}$ to be transferred to UE $k \in \mathcal{K}_U$ in the considered OTB interval and leave $\bar{S} - R_{f_k,m}$ bits for the next OTB.

First, let us consider the data transmission from the BBU to the eRRHs. At this step, the BBU needs to decide the set of associated eRRHs to which each missing requested subfile (f_k, m) is transferred via the fronthaul links. There are trade-offs between transferring all missing requested subfiles and only transferring selected subfiles to the associated eRRHs. While the former requires more data to be transferred via fronthaul links, it can give more throughput gain as a result of coherence combining. In this paper, we adopt the former approach. The total data rate on the fronthaul link of eRRH $i \in \mathcal{K}_R$ can therefore be expressed as:

$$R^{FH}_i \triangleq \sum_{f_k \in \mathcal{F}^*} a_{k,i} \sum_{m \in \mathcal{M}} (1 - c^i_{f_k,m}) R_{f_k,m}. \tag{3}$$

Next, we consider the data transmission from the eRRHs to the UEs. Let $\boldsymbol{s}_{f_k,m} \in \mathbb{C}^{d \times 1}$ denote the encoded baseband signal of subfile (f_k, m) with d data streams. Assume that $\boldsymbol{s}_{f_k,m} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. The eRRH i precodes $\boldsymbol{s}_{f_k,m}$ by a matrix $\boldsymbol{F}^i_{f_k,m} \in \mathbb{C}^{N_r \times d}$ to obtain the transmitted signal $\boldsymbol{x}_i = \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \boldsymbol{F}^i_{f_k,m} \boldsymbol{s}_{f_k,m}$. Denote by $\boldsymbol{H}_{k,i} \in \mathbb{C}^{N_u \times N_r}$ the flat-fading channel matrix from eRRH i to UE k and by $\boldsymbol{H}_k \triangleq [\boldsymbol{H}_{k,1}, ..., \boldsymbol{H}_{k,K_R}] \in \mathbb{C}^{N_u \times N_R}$ the channel matrix from all eRRHs to UE k, where $N_R \triangleq K_R N_r$. Assume that channel states $\boldsymbol{H}_{k,i}, \ k \in \mathcal{K}_U, \ i \in \mathcal{K}_R$ remain unchanged during each OTB and are made available to the BBU and eRRHs.

Define $\bar{\boldsymbol{F}}_{f_k,m} \triangleq \left[(\boldsymbol{F}^1_{f_k,m})^H, (\boldsymbol{F}^2_{f_k,m})^H, \ldots (\boldsymbol{F}^{K_R}_{f_k,m})^H \right]^H \in \mathbb{C}^{N_R \times d}$. Note that since $\boldsymbol{F}^i_{f_k,m} = \boldsymbol{0}$ when eRRH i does not serve UE k. Then, the received signal $\boldsymbol{y}_k \in \mathbb{C}^{N_u \times 1}$ at UE k for the requested file f_k is:

$$\begin{aligned} \boldsymbol{y}_k \triangleq\, & \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_k,m} \boldsymbol{s}_{f_k,m} + \sum_{q \in \mathcal{M} \setminus \{m\}} \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_k,q} \boldsymbol{s}_{f_k,q} \\ & + \sum_{f_\ell \in \mathcal{F}^* \setminus \{f_k\}} \sum_{m \in \mathcal{M}} \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_\ell,m} \boldsymbol{s}_{f_\ell,m} + \boldsymbol{n}_k, \end{aligned} \tag{4}$$

where $\boldsymbol{n}_k \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Sigma}_k)$ is the additive noise term.

To deal with the interference expressed in the second and third terms on the right-hand side of (4), we assume each UE k performs the successive interference cancellation (SIC) decoding for the subfiles with the order $\boldsymbol{s}_{f_k,1} \rightarrow ... \rightarrow \boldsymbol{s}_{f_k,M}$. After applying the SIC scheme, the achievable data rate $R_{f_k,m}$ can be bounded as [7], [11]:

$$R_{f_k,m} \leq g_{f_k,m}(\bar{\boldsymbol{F}}) \triangleq \log \left| \boldsymbol{I}_{N_{U,k}} + \boldsymbol{\Pi}_{f_k,m} \boldsymbol{\Pi}^H_{f_k,m} \boldsymbol{\Xi}^{-1}_{f_k,m} \right|, \tag{5}$$

where $\bar{\boldsymbol{F}} \triangleq \{\bar{\boldsymbol{F}}_{f_\ell,m}\}_{f_\ell \in \mathcal{F}^*, m \in \mathcal{M}}$, $\boldsymbol{\Pi}_{f_k,m} \triangleq \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_k,m}$, and

$$\begin{aligned} \boldsymbol{\Xi}_{f_k,m} \triangleq & \sum_{q=m+1}^{M} \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_k,q} \bar{\boldsymbol{F}}^H_{f_k,q} \boldsymbol{H}^H_k + \\ & \sum_{f_\ell \in \mathcal{F}^* \setminus \{f_k\}} \sum_{q \in \mathcal{M}} \boldsymbol{H}_k \bar{\boldsymbol{F}}_{f_\ell,q} \bar{\boldsymbol{F}}^H_{f_\ell,q} \boldsymbol{H}^H_k + \boldsymbol{\Sigma}_k\,. \end{aligned} \tag{6}$$

The network throughput is then defined as the following sum rate:

$$R_{\text{sum}} \triangleq \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} R_{f_k,m}. \tag{7}$$

C. Power Consumption Model

To see the effect of the joint design on the power performance, we adopt a practical power consumption model in [12]. Specifically, the per-OTB power consumption by eRRH $i \in \mathcal{K}_R$ is modeled as:

$$P_i^{eRRH} \triangleq \begin{cases} \beta_i P_i^{tx} + P_{i,a}, & \text{if } 0 < P_i^{tx} \leq P_i \\ P_{i,s}, & \text{if } P_i^{tx} = 0 \end{cases}, \tag{8}$$

where constant $\beta_i > 0$, $i \in \mathcal{K}_R$ reflects the power amplifier efficiency, feeder loss and other loss factors due to power supply and cooling for eRRH i [12]; P_i^{tx} is the transmit power required to deliver all requested files from eRRH i as

$$P_i^{tx} \triangleq \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle, \tag{9}$$

in which $\bar{\boldsymbol{E}}_i \in \mathbb{C}^{N_R \times N_r}$ is zero everywhere except an identity matrix of size N_r from row $(i-1)N_r + 1$ to row iN_r; $P_{i,a}$ is the power required to support eRRH i in the active mode; and $P_{i,s} < P_{i,a}$ is the power consumption in the sleep mode.

The fronthaul link from the BBU to eRRH $i \in \mathcal{K}_R$ is modeled as a set of communication channels with a total capacity C_i and total power dissipation $P_{i,\max}^{FH}$. Its power consumption is given by

$$P_i^{FH} \triangleq \frac{R_i^{FH}}{C_i} P_{i,\max}^{FH} = \alpha_i R_i^{FH}, \tag{10}$$

where $\alpha_i \triangleq P_{i,\max}^{FH}/C_i$ and R_i^{FH} is defined in (3).

From (8) and (10), the total network power consumption is:

$$\begin{aligned} P_{\text{total}} &\triangleq \sum_{i \in \mathcal{K}_R} (P_i^{eRRH} + P_i^{FH}) \\ &= \sum_{i \in \mathcal{K}_R} \big(\beta_i P_i^{tx} + {}_{\{P_i^{tx}\}} P_{i,\Delta} + \alpha_i R_i^{FH}\big) + P_s, \end{aligned} \tag{11}$$

where $P_{i,\Delta} \triangleq P_{i,a} - P_{i,s}$, $P_s \triangleq \sum_{i \in \mathcal{K}_R} P_{i,s}$ and

$${}_{\{P_i^{tx}\}} \triangleq \begin{cases} 1, & \text{if } P_i^{tx} > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

D. Problem Formulation

In this paper, we aim to jointly design the UA, DDR and precoding in order to improve the network sum rate in (7) as well as to reduce the total power consumption in (11). Let us define $\boldsymbol{a} \triangleq \{a_{k,i}\}_{k \in \mathcal{K}_U, i \in \mathcal{K}_R}$ and $\boldsymbol{R} \triangleq \{R_{f_k,m}\}_{f_k \in \mathcal{F}^*, m \in \mathcal{M}}$. For a given cache state information $\{c_{f_k,m}^i\}_{i \in \mathcal{K}_R, f_k \in \mathcal{F}^*, m \in \mathcal{M}}$, the design problem is formulated as:

$$\max_{\boldsymbol{a},\boldsymbol{R},\bar{\boldsymbol{F}}} \ \mathcal{P}_1(\boldsymbol{R}, \bar{\boldsymbol{F}}) \triangleq R_{\text{sum}} - \eta P_{\text{total}} \tag{13a}$$

$$\text{s.t. } R_{\text{QoS}} \leq R_{f_k,m} \leq \bar{S}, \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{13b}$$

$$\sum_{f_k \in \mathcal{F}^*} a_{k,i} \sum_{m \in \mathcal{M}} (1 - c_{f_k,m}^i) R_{f_k,m} \leq C_i, \forall i \in \mathcal{K}_R \tag{13c}$$

$$R_{f_k,m} \leq g_{f_k,m}(\bar{\boldsymbol{F}}), \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, , \tag{13d}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R \tag{13e}$$

$$\text{and (2)}, \ \forall k \in \mathcal{K}_U, \ i \in \mathcal{K}_R. \tag{13f}$$

Here, the weight $\eta > 0$ specifies the relative importance between the sum rate and the total power. Constraint (13b) imposes the minimum rate $R_{\text{QoS}} \geq 0$ and maximum rate $\bar{S}$ for each subfile [see the second paragraph of Sec. II-B]. Constraint (13c) expresses the bottleneck at fronthaul link $i \in \mathcal{K}_R$ with the limited backhaul capacity $C_i \geq 0$ [see (3)]. Constraint (13d) is indeed (5). Finally, constraint (13e) requires that the total transmit power at each eRRH $i \in \mathcal{K}_R$ must not exceed the predefined budget $P_i \geq 0$. While problem (13) is already a difficult nonconvex nonsmooth optimization problem, the strong coupling among the optimizing variables makes it even more challenging to be solved globally.

III. Proposed Joint Optimization Algorithm

First, we will deal with the nonsmooth nature of (13f) and (12). We begin by expressing (13f) as:

$$a_{k,i} = \begin{cases} 0, & \text{iff } \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} = \boldsymbol{0}, \forall m \in \mathcal{M} \\ 1, & \text{otherwise} \end{cases} \tag{14}$$

$\forall k \in \mathcal{K}_U, i \in \mathcal{K}_R$. To see this, note that if eRRH i does not serve UE k, all corresponding precoders $\boldsymbol{F}_{f_k,m}^i = \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m}$ must be $\boldsymbol{0}$ and then $a_{k,i} = 0$. Otherwise, $a_{k,i} = 1$ and there exists at least one of the corresponding precoders $\boldsymbol{F}_{f_k,m}^i = \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \neq \boldsymbol{0}$. Therefore, without loss of optimality, $a_{k,i}$ can be further rewritten as:

$$a_{k,i} = \Big\| \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle \Big\|_0. \tag{15}$$

Similarly, (12) can also be replaced by:

$${}_{\{P_i^{tx}\}} = \big\|P_i^{tx}\big\|_0 = \Big\| \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle \Big\|_0. \tag{16}$$

We respectively approximate the nonconvex ℓ_0-norms (15) and (16) by their reweighted ℓ_1-norms as [12]

$$a_{k,i} = \mu_{k,i} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle, \tag{17}$$

$${}_{\{P_i^{tx}\}} = \theta_i \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle, \tag{18}$$

where weights $\mu_{k,i}$ and θ_i are iteratively updated according to

$$\mu_{k,i} = \frac{c_1}{\sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle + \tau_1}, \tag{19}$$

$$\theta_i = \frac{c_2}{\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle + \tau_2}. \tag{20}$$

Here, c_1 and c_2 are constant numbers whereas τ_1 and τ_2 are constant regularization factors. With (17), variable $\boldsymbol{a}$ can now be expressed in terms of $\bar{\boldsymbol{F}}$ and thus be removed from (13). With (18), we can rewrite P_{total} as:

$$P_{\text{total}} = \sum_{i \in \mathcal{K}_R} \sum_{f_k \in \mathcal{F}^*} \tau_{f_k}^i \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle + P_s \tag{21}$$

where $\tau_{f_k}^i \triangleq \upsilon_i + \alpha_i \vartheta_{f_k}^i$, $\vartheta_{f_k}^i \triangleq \mu_{k,i} \sum_{m \in \mathcal{M}} (1 - c_{f_k,m}^i) R_{f_k,m}$ and $\upsilon_i \triangleq \beta_i + \theta_i P_{i,\Delta}$.

The above result allows us to transform (13) to the following approximated problem:

$$\max_{\boldsymbol{R},\bar{\boldsymbol{F}}} \ \mathcal{P}_2(\boldsymbol{R}, \bar{\boldsymbol{F}}) \triangleq R_{\text{sum}} - \eta P_{\text{total}} \tag{22a}$$

$$\text{s.t. } R_{QoS} \leq R_{f_k,m} \leq \bar{S}, \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{22b}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \vartheta_{f_k}^i \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle \leq C_i, \forall i \in \mathcal{K}_R \tag{22c}$$

$$R_{f_k,m} \leq g_{f_k,m}(\bar{\boldsymbol{F}}), \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{22d}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}_{f_k,m}^H \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R. \tag{22e}$$

Algorithm 1 Joint design of UA, data delivery rate and precoding for F-RANs

```
Initialization: Use Alg. 2 to find a feasible initial point $(\bar{\boldsymbol{F}}^{(1)}, \{\mu_{k,i}, \theta_i\}, \boldsymbol{R}^{(1)})$. Set error tolerances $\epsilon_1, \epsilon_2, \epsilon_3 > 0$. Set $p := 1$.
repeat
    Set $p := p + 1$ and $\kappa := 1$
    repeat
        Set $\kappa := \kappa + 1$
        For given $\boldsymbol{R}^{(\kappa-1)}$, use $\bar{\boldsymbol{F}}^{(\kappa-1)}$ as the initial feasible point and set $n := 1$
        repeat
            Update $n := n + 1$
            Find an optimal solution $\bar{\boldsymbol{F}}^*$ by solving convex problem (26)
            Update $\bar{\boldsymbol{F}}^{(n)} := \bar{\boldsymbol{F}}^*$
        until $\left|\frac{\mathcal{P}_3(\bar{\boldsymbol{F}}^{(n)}) - \mathcal{P}_3(\bar{\boldsymbol{F}}^{(n-1)})}{\mathcal{P}_3(\bar{\boldsymbol{F}}^{(n)})}\right| \leq \epsilon_3$
        Update $\bar{\boldsymbol{F}}^{(\kappa)} := \bar{\boldsymbol{F}}^*$
        Find an optimal solution $\boldsymbol{R}^*$ by solving convex problem (23)
        Update $\boldsymbol{R}^{(\kappa)} := \boldsymbol{R}^*$
    until $\left|\frac{\mathcal{P}_2(\bar{\boldsymbol{F}}^{(\kappa)}, \boldsymbol{R}^{(\kappa)}) - \mathcal{P}_2(\bar{\boldsymbol{F}}^{(\kappa-1)}, \boldsymbol{R}^{(\kappa-1)})}{\mathcal{P}_2(\bar{\boldsymbol{F}}^{(\kappa-1)}, \boldsymbol{R}^{(\kappa-1)})}\right| \leq \epsilon_2$
    Update $\bar{\boldsymbol{F}}^{(p)} := \bar{\boldsymbol{F}}^*$ and $\boldsymbol{R}^{(p)} := \boldsymbol{R}^*$
    Update $\{\mu_{k,i}, \theta_i\}$ according to (19) and (20)
until $\left|\frac{\mathcal{P}_1(\bar{\boldsymbol{F}}^{(p)}, \boldsymbol{R}^{(p)}) - \mathcal{P}_1(\bar{\boldsymbol{F}}^{(p-1)}, \boldsymbol{R}^{(p-1)})}{\mathcal{P}_1(\bar{\boldsymbol{F}}^{(p-1)}, \boldsymbol{R}^{(p-1)})}\right| \leq \epsilon_1$
```

Problem (22) is still difficult due to the strong coupling between $\boldsymbol{R}$ and $\bar{\boldsymbol{F}}$. As such, we propose using an alternating method [9] that deals with one variable at a time and repeats until convergence. Specifically, for a given $\bar{\boldsymbol{F}}$, we solve the following convex subproblem to update $\boldsymbol{R}$:

$$\max_{\boldsymbol{R}} \quad \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} R_{f_k,m} - \eta \left(b + \sum_{i \in \mathcal{K}_R} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} q^i_{f_k,m} R_{f_k,m} \right) \tag{23a}$$

$$\text{s.t.} \quad R_{QoS} \leq R_{f_k,m} \leq \bar{S}, \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{23b}$$

$$\sum_{f_k \in \mathcal{F}^*} q^i_{f_k} \sum_{m \in \mathcal{M}} R_{f_k,m} \leq C_i, \forall i \in \mathcal{K}_R, \tag{23c}$$

$$R_{f_k,m} \leq g_{f_k,m}(\bar{\boldsymbol{F}}), \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{23d}$$

where $q^i_{f_k} \triangleq \alpha_i \mu_{k,i} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle (1 - c^i_{f_k,m})$ and $b \triangleq P_s + \sum_{i \in \mathcal{K}_R} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \upsilon_i \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle$.

Next, for a given $\boldsymbol{R}$, we solve the following subproblem to update $\bar{\boldsymbol{F}}$:

$$\min_{\bar{\boldsymbol{F}}} \sum_{i \in \mathcal{K}_R} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \eta \tau^i_{f_k} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \tag{24a}$$

$$\text{s.t.} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \vartheta^i_{f_k} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq C_i, \forall i \in \mathcal{K}_R, \tag{24b}$$

$$R_{f_k,m} \leq g_{f_k,m}(\bar{\boldsymbol{F}}), \forall f_k \in \mathcal{F}^*, m \in \mathcal{M},, \tag{24c}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R\,. \tag{24d}$$

The key difficulty with subproblem (24) is due to the nonconvex contraint (24c). Let $\boldsymbol{\Phi}_{f_k,m} \triangleq \boldsymbol{\Pi}_{f_k,m} \boldsymbol{\Pi}^H_{f_k,m} + \boldsymbol{\Xi}_{f_k,m}$. Using the first-order Taylor series expansion, we approximate the nonconcave part $g_{f_k,m}(\bar{\boldsymbol{F}})$ of (24c) by its concave lower bound $\Gamma^{(n)}_{f_k,m}(\bar{\boldsymbol{F}})$ defined in (25). This fact can be proved similarly as in [10], and the detailed proof is omitted due to limited space. The nonconvex subproblem (24) can thus be transformed into the following *convex* quadratic program:

Algorithm 2 Finding an initial feasible point for Alg. 1

```
Initialization: Set error tolerance $\epsilon_4 > 0$. Find $\bar{\boldsymbol{F}}^{(1)}$ as $\bar{\boldsymbol{F}}^{(1)}_{f_k,m} := \sqrt{\frac{\bar{P}}{\langle \bar{\boldsymbol{E}}_7^H \boldsymbol{F}_{\text{ran}} \boldsymbol{F}^H_{\text{ran}} \bar{\boldsymbol{E}}_7 \rangle}} \boldsymbol{F}_{\text{ran}}, \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}$ with a random matrix $\boldsymbol{F}_{\text{ran}} \in \mathbb{C}^{K_R N_r \times d}$ and $\bar{P} = \frac{P_i}{M K_U}$ until (28b) and (28c) are satisfied $\forall i \in \mathcal{K}_R$. Set $n := 1$.
For given $\bar{\boldsymbol{F}}^{(1)}$, update $\{\mu_{k,i}, \theta_i\}$ by (19) and (20), and find $\boldsymbol{R}$ by solving (23)
repeat
    Update $n := n + 1$
    Find an optimal solution $\bar{\boldsymbol{F}}^*$ by solving (28)
    Update $\bar{\boldsymbol{F}}^{(n)} := \bar{\boldsymbol{F}}^*$
until $\left|\frac{\mathcal{P}_4(\bar{\boldsymbol{F}}^{(n)}) - \mathcal{P}_4(\bar{\boldsymbol{F}}^{(n-1)})}{\mathcal{P}_4(\bar{\boldsymbol{F}}^{(n)})}\right| \leq \epsilon_4$
```

$$\min_{\bar{\boldsymbol{F}}} \ \mathcal{P}_3(\bar{\boldsymbol{F}}) \triangleq \sum_{i \in \mathcal{K}_R} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \eta \tau^i_{f_k} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \tag{26a}$$

$$\text{s.t.} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \vartheta^i_{f_k} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq C_i, \forall i \in \mathcal{K}_R, \tag{26b}$$

$$R_{f_k,m} \leq \Gamma^{(n)}_{f_k,m}(\bar{\boldsymbol{F}}), \forall f_k \in \mathcal{F}^*, m \in \mathcal{M}, \tag{26c}$$

$$\sum_{f \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R\,. \tag{26d}$$

We are now ready to present Alg. 1 to solve the original problem (13). First, to deal with precoding design, the inner loop finds the locally optimal $\bar{\boldsymbol{F}}$ for a given $\boldsymbol{R}$ by iteratively solving the convex problem (26). For the DDR design, the middle loop alternates between solving (24) to find a locally optimal $\bar{\boldsymbol{F}}$ for a given $\boldsymbol{R}$ and solving (23) to find an optimal $\boldsymbol{R}$ for a given $\bar{\boldsymbol{F}}$. Finally, the outer loop updates $\{\mu_{k,i}, \theta_i\}$ according to (19) and (20) for UA. Alg. 1 terminates when there is no improvement in the objective value $\mathcal{P}_1(\boldsymbol{R}, \bar{\boldsymbol{F}})$ of (13).

Remark 1. Let N_p, N_κ and N_n respectively be the numbers of iterations of the outer loop, the middle loop and the inner loop when Alg. 1 converges. It can be observed that Alg. 1 solves problem (26) $N_p N_\kappa N_n$ times and problem (23) $N_p N_\kappa$ times.

Alg. 1 requires an initial feasible point that satisfies constraints (24b), (24c) and (24d). For this, we consider the following problem:

$$\max_{\bar{\boldsymbol{F}}} \ \min_{m \in \mathcal{M}, f_k \in \mathcal{F}^*} \left\{ \frac{g_{f_k,m}(\bar{\boldsymbol{F}})}{R_{f_k,m}} \right\} \tag{27a}$$

$$\text{s.t.} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \vartheta^i_{f_k,m} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq C_i, \forall i \in \mathcal{K}_R, \tag{27b}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R\,. \tag{27c}$$

Problem (27) can further be transformed into the following *concave* quadratic program:

$$\max_{\bar{\boldsymbol{F}}} \ \mathcal{P}_4(\bar{\boldsymbol{F}}) \triangleq \min_{m \in \mathcal{M}, f_k \in \mathcal{F}^*} \left\{ \frac{\Gamma_{f_k,m}(\bar{\boldsymbol{F}})}{R_{f_k,m}} \right\} \tag{28a}$$

$$\text{s.t.} \sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \vartheta^i_{f_k,m} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq C_i, \forall i \in \mathcal{K}_R, \tag{28b}$$

$$\sum_{f_k \in \mathcal{F}^*} \sum_{m \in \mathcal{M}} \langle \bar{\boldsymbol{E}}_i^H \bar{\boldsymbol{F}}_{f_k,m} \bar{\boldsymbol{F}}^H_{f_k,m} \bar{\boldsymbol{E}}_i \rangle \leq P_i, \forall i \in \mathcal{K}_R, \tag{28c}$$

the solution of which can be found by Alg. 2.

Convergence Analysis: For the inner loop of Alg. 1, the optimal solution $\bar{\boldsymbol{F}}^{(n)}$ of convex problem (26) is feasible to

$$\Gamma_{f_k,m}^{(n)}(\bar{\boldsymbol{F}}) = g_{f_k,m}(\bar{\boldsymbol{F}}^{(n)}) + 2\Re\left\{\left\langle \left(\left(\boldsymbol{\Phi}_{f_k,m}^{(n)} - \boldsymbol{\Pi}_{f_k,m}^{(n)}(\boldsymbol{\Pi}_{f_k,m}^{(n)})^H\right)^{-1}\boldsymbol{\Pi}_{f_k,m}^{(n)}\right)^H (\boldsymbol{\Pi}_{f_k,m}(\bar{\boldsymbol{F}}_{f_k,m}) - \boldsymbol{\Pi}_{f_k,m}^{(n)})\right\rangle\right\}$$

$$- \left\langle \left(\left(\boldsymbol{\Phi}_{f_k,m}^{(n)} - \boldsymbol{\Pi}_{f_k,m}^{(n)}(\boldsymbol{\Pi}_{f_k,m}^{(n)})^H\right)^{-1} - (\boldsymbol{\Phi}_{f_k,m}^{(n)})^{-1}\right)^H (\boldsymbol{\Phi}_{f_k,m}(\bar{\boldsymbol{F}}) - \boldsymbol{\Phi}_{f_k,m}^{(n)})\right\rangle \quad (25)$$

TABLE I
SYSTEM PARAMETERS USED IN SIMULATIONS

Distance between adjacent eRRHs	0.3 km
Total bandwidth	10 MHz
Std. deviation of log-normal shadowing	10 dB
Path loss at distance d (km)	$140.7 + 36.7\log_{10}(d)$ dB
Noise variance $\sigma_k^2 = \sigma^2$	-174 dBm/Hz
Maximum eRRH transmit power	24 dBm

TABLE II
EXAMPLE CACHE STATE INFORMATION USED IN SIMULATIONS

$c_{f_k,m}^i$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$
$(f_1, 1), (f_1, 2)$	$1, 1$	$0, 0$	$0, 1$	$0, 0$	$1, 0$	$0, 1$	$0, 0$
$(f_2, 1), (f_2, 2)$	$1, 1$	$0, 1$	$0, 0$	$1, 0$	$0, 0$	$0, 1$	$0, 1$
$(f_3, 1), (f_3, 2)$	$1, 1$	$0, 0$	$1, 0$	$0, 0$	$1, 0$	$1, 0$	$0, 1$

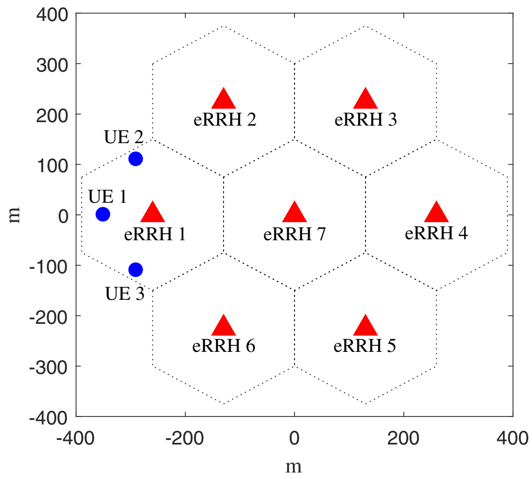


Fig. 2. Network scenario used in simulations

the nonconvex problem (24) and is also better than $\bar{\boldsymbol{F}}^{(n-1)}$, i.e., $\mathcal{P}_3(\bar{\boldsymbol{F}}^{(n)}) \leq \mathcal{P}_3(\bar{\boldsymbol{F}}^{(n-1)})$. Once initialized from a feasible point $\bar{\boldsymbol{F}}^{(0)}$ by using Alg. 2, the inner loop generates a sequence $\{\bar{\boldsymbol{F}}^{(n)}\}$ of improved feasible solutions for the nonconvex program (24), and it will eventually converge to a locally optimal solution of (24). The middle loop utilizes the alternating optimization framework that solves a series of convex problems (23) and (26), and is guaranteed to converge [13]. By choosing $c_1 = \frac{1}{\ln(1+\tau_1^{-1})}$ and $c_2 = \frac{1}{\ln(1+\tau_1^{-2})}$, the outer loop can be proven to be a special case of the majorization-minimization algorithm [12] and is thus guaranteed to converge.

IV. NUMERICAL RESULTS

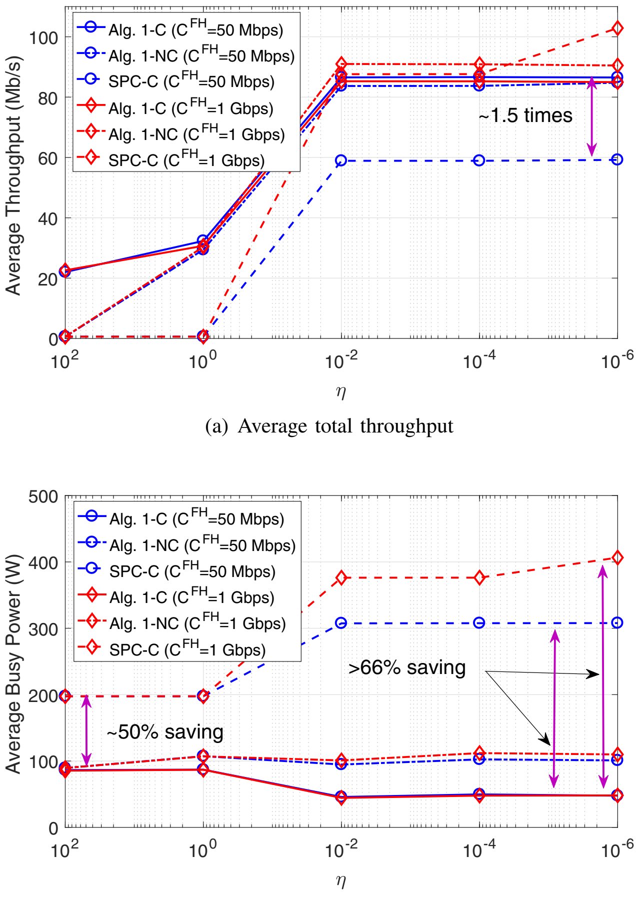
We consider the network scenario in Fig. 2 where the locations of the $K_R = 7$ eRRHs and $K_U = 3$ UEs are fixed. In our simulations, the standard parameters of LTE systems including the channel information are used and expressed in Table I [14]. Assume that each eRRH is equipped with $N_r = 5$ antennas and each UE with $N_u = 2$ antennas. We set $P_i = P$, $C_i = C^{FH}$ $\forall i \in \mathcal{K}_R$ and $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}, \forall k \in \mathcal{K}_U$. At each eRRH, the active mode and the sleep mode consume 84W and 56W, respectively [12]. Constant regulation factors are set as $\tau_1 = 10^{-5}, \tau_2 = 10^{-3}$. The slope of transmit power is $\beta_i = \beta = 2.8$ and $\alpha_i = \alpha = 5$ $\forall i \in \mathcal{K}_R$ [12]. The error tolerances for the proposed algorithms are taken as $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = \epsilon_3 = \epsilon_4 = 10^{-2}$. We set $R_{\text{QoS}} = 0.1$ Mbps, $\bar{S} = 40$ Mbps and $M = 2$. The numerical result is obtained by averaging over 100 independent channel realizations.

In the simulation, each eRRH's caching capacity is set as $B_i = B = \xi SF$, $\forall i \in \mathcal{K}_R$ where ξ indicates the fractional caching capacity. Each eRRH can store a maximum of $\lfloor \xi FM \rfloor$ subfiles randomly chosen from the file library. Each UE randomly and independently requests one file from a library of $F = 6$ files. We pick one specific set $\mathcal{F}^*$ of the requested files. The cache state information $\{c_{f_k,m}^i\}$ in (1) are recorded in Table II by checking $\mathcal{F}^*$ against the local caches of eRRHs. Since we do not consider the caching problem, we are allowed to selectively place the UEs close to eRRH 1 where all the requested files are cached. This arrangement helps us reveal the potential gains resulting from an intelligent caching strategy.

To demonstrate the advantages of the joint design offered by Alg. 1, we evaluate the performance of Alg. 1 with local caching (referred to as Alg. 1-C) and Alg. 1 without local caching (referred to as Alg. 1-NC). We also implement the sole precoder design with local caching (referred to as SPD-C) by modifying Alg. 1 such that each UE is always connected to all eRRHs.

Fig. 3(a) compares the throughput performance among Alg. 1-C, Alg. 1-NC and SPD-C. Attention should be paid to the region of small η where problem (13) prioritizes throughput maximization. When the fronthaul capacity is limited ($C^{FH} = 50$ Mbps), the sum rate by Alg. 1-C and Alg. 1-NC are approximately 1.5 times higher than that by the SPD-C scheme. This is because the joint designs show their advantages over the SPD-C in terms of reducing the bottleneck in the fronthaul links [see (13c)] to improve the throughput. However, for the ample fronthaul capacity of $C^{FH} = 1$ Gbp, the local caching at the eRRHs or the fronthaul traffic offload via selective UE-eRRH associations offers almost no throughput advantage. In this case, there is virtually no bottleneck in transferring data traffic from the BBU to each eRRH. When $\eta = 10^{-6}$, the SPD-C even offers higher throughput than that offered by the Alg. 1-C and Alg. 1-NC since it better exploits the coherence combining gain.

Fig. 3(b) demonstrates the power consumption incurred by the three considered schemes. To have meaningful comparisons, we consider the "busy" power by discounting the fixed eRRH



(a) Average total throughput

(b) Average total busy power

Fig. 3. Effects of joint design, local caching and limited backhaul capacity on network performance

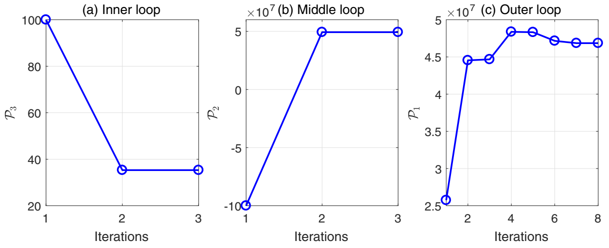


Fig. 4. Convergence behavior of Alg. 1

sleeping power from the total power as $P_{\text{busy}} \triangleq P_{\text{total}} - P_s$. From the figure, Alg. 1-C and Alg. 1-NC are the two best performers, where they save approximate 50% of the power consumption for large η and more than 66% for small η over the SPD-C scheme. This is because the proposed joint design can effectively save power by: (i) using fewer active eRRHs and less transmission power in the fronthaul links by assigning UEs to appropriate eRRHs, and (ii) reducing the transmission power in the access links by designing effective precoders.

Fig. 4 illustrates the convergence behavior of Alg. 1-C with $C^{\text{FH}} = 50$ Mbps. Figs. 4(a), 4(b) and 4(c) plot the convergence of the objective functions of the inner, the middle and the outer loops, respectively. In total, the proposed algorithm requires fewer than 50 iterations to converge, where each iteration corresponds to solving at most two simple convex programs (23) and (26).

V. Conclusions

This paper has studied the joint design of user association, data delivery rate and signal precoding in the downlink of a cache-enabled F-RAN with limited fronthaul capacity. An optimization problem has been formulated with the objective of maximizing the weighted difference of network throughput and total power consumption. The requirements on data delivery rates and maximum eRRH transmit powers have also been included in the design. Applying a range of optimization techniques, we have solved this challenging optimization problem and proposed an iterative algorithm that is guaranteed to converge to obtain a locally optimal solution. Numerical results have shown that our joint design markedly improves both network throughput and power consumption performances of the considered F-RAN.

Acknowledgment

This work is supported in part by an ECR-HDR scholarship from The University of Newcastle and in part by the Australian Research Council Discovery Project grant DP170100939.

References

[1] H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang, "Mobile edge cloud system: Architectures, challenges, and approaches," *IEEE Systems J.*, no. 99, pp. 1–14, 2017.

[2] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[3] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

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

[5] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2016.

[6] 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.

[7] S. H. Park, O. Simeone, and S. S. 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.

[8] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[9] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, Sep. 1999.

[10] H. H. M. Tam, H. D. Tuan, and D. T. Ngo, "Successive convex quadratic programming for quality-of-service management in full-duplex MU-MIMO multicell networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2340–2353, Jun. 2016.

[11] 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.

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

[13] H. H. M. Tam, H. D. Tuan, D. T. Ngo, T. Q. Duong, and H. V. Poor, "Joint load balancing and interference management for small-cell heterogeneous networks with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 872–884, Feb. 2017.

[14] 3GPP TS 36.814 V9.0.0, "3GPP technical specification group radio access network, evolved universal terrestrial radio access (E-UTRA): Further advancements for E-UTRA physical layer aspects (release 9)," 2010.