Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks

Seok-Hwan Park[1], Osvaldo Simeone[2] and Shlomo Shamai (Shitz)[3]
[1]Division of Electronic Engineering, Chonbuk National University, Jeonju, 561-756 Korea
[2]CWCSPR, New Jersey Institute of Technology, 07102 Newark, New Jersey, USA
[3]Department of Electrical Engineering, Technion, Haifa, 32000, Israel
Email: seokhwan@jbnu.ac.kr, osvaldo.simeone@njit.edu, sshlomo@ee.technion.ac.il

***Abstract*—This work studies the joint design of cloud and edge processing for the downlink of a fog radio access network (F-RAN). In an F-RAN, cloud processing is carried out by a baseband processing unit (BBU) that is connected to enhanced remote radio heads (eRRHs) by means of fronthaul links. Edge processing is instead enabled by local caching of popular content at the eRRHs. Focusing on the design of the delivery phase for an arbitrary pre-fetching strategy, a novel superposition coding approach is proposed that is based on the hybrid use of the fronthaul links in both *hard-transfer* and *soft-transfer* modes. With the former, non-cached files are communicated over the fronthaul links to a subset of eRRHs, while, with the latter, the fronthaul links are used to convey quantized baseband signals as in a cloud RAN (C-RAN). The problem of maximizing the delivery rate is tackled under fronthaul capacity and per-eRRH power constraints. Numerical results are provided to validate the performance of the proposed hybrid delivery scheme for different baseline pre-fetching strategies.**

***Index Terms*—Cloud radio access network, fog network, caching, precoding.**

I. Introduction

Cloud radio access network (C-RAN) is an emerging architecture for the fifth-generation (5G) of wireless system, in which a centralized baseband signal processing unit (BBU) implements the baseband processing functionalities of a set of remote radio heads (RRHs), which are connected to the BBU by means of fronthaul links [1][2]. Recently, an evolved network architecture, referred to as *Fog Radio Access Network* (F-RAN), has been proposed, which enhances the C-RAN architecture by allowing the RRHs to be equipped with caching and signal processing functionalities [3]-[5]. The architecture at hand is also referred to as a hybrid of cloud and fog processing in the literature [6]. The resulting RRHs are referred to here as *enhanced RRHs* (eRRHs) (see Fig. 1).

As a cache-aided system, an F-RAN operates in two phases, namely the pre-fetching and the delivery phases [7]-[11] (see also [12][13]). Pre-fetching operates at the large time scale corresponding to the period in which content popularity remains constant. This time scale encompasses multiple transmission intervals. Based on the cached file messages, the delivery phase, instead, operates separately on each transmission interval. The fronthaul-aware design of the pre-fetching or delivery phases was studied in [7]-[11] under the assumption that the fronthaul links in an F-RAN are leveraged in a *hard-transfer mode*, that is, to convey to the eRRHs the requested content that is not present in the local caches.

S.-H. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT&Future Planning) [2015R1C1A1A01051825]. The work of O. Simeone was partially supported by the U.S. NSF through grant 1525629. The work of S. Shamai was partly supported by the Israel Science Foundation (ISF).

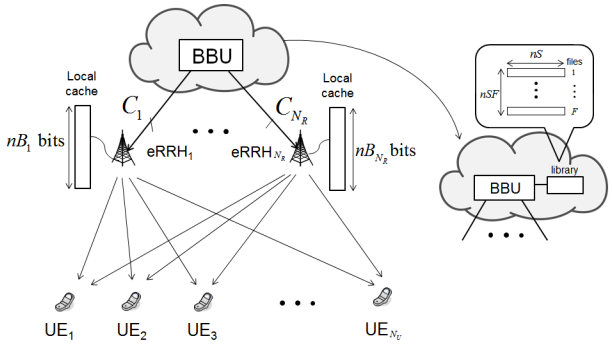

Figure 1. Illustration of an F-RAN, which has both cloud and edge processing capabilities: the BBU, in the "cloud", can perform joint baseband processing and the eRRHs are equipped with local caches.

In contrast, in this work, we propose novel delivery strategies that leverage the *soft-transfer fronthaul mode* that is typical of C-RAN (see, e.g., [1][2]). The most general proposed approach is a hybrid of hard- and soft-transfer modes that is based on fronthaul quantization and superposition coding. Each eRRH transmits the superposition of two signals, one that is locally encoded based on the content stored in the cache or received on the fronthaul link via hard-transfer mode, and another that is encoded at the BBU and quantized for transmission on the fronthaul link. We tackle the problem of optimizing this strategy, and special cases thereof, with the aim of maximizing the delivery rate, while satisfying fronthaul capacity and per-eRRH power constraints. Numerical results are provided to compare the performance of hard-, soft- and hybrid-transfer fronthauling modes for baseline pre-fetching strategies.

II. System Model

We consider the downlink of an F-RAN, where N_U multi-antenna user equipments (UEs) are served by N_R multi-

antenna eRRHs that are connected to a BBU in the "cloud" through digital fronthaul links. Each eRRH i in an F-RAN is equipped with a cache, which can store nB_i bits, where n is the number of (baud-rate) symbols of each downlink coded transmission block. Furthermore, it also has baseband processing capabilities. Each eRRH i is connected to the BBU with a fronthaul link of capacity C_i bit per symbol of the downlink channel for $i \in \mathcal{N}_R \triangleq \{1, \ldots, N_R\}$. We denote the numbers of antennas of eRRH i and UE k by $n_{R,i}$ and $n_{U,k}$, respectively, and define the notations $n_R \triangleq \sum_{i \in \mathcal{N}_R} n_{R,i}$.

We consider communication for content delivery via the outlined F-RAN system. Accordingly, UEs request contents, or files, from a library of F files, each of size nS bits, which are delivered by the network across a number of transmission intervals. Labeling the files in order of popularity, the probability $P(f)$ of a file f to be selected is defined by Zipf's distribution $P(f) = cf^{-\gamma}$ for $f \in \mathcal{F} \triangleq \{1, \ldots, F\}$, where $\gamma \geq 0$ is a given popularity exponent and $c \geq 0$ is set such that $\sum_{f \in \mathcal{F}} P(f) = 1$ (see, e.g., [7]-[9]). Each UE k requests file $f_k \in \mathcal{F}$ with the probability $P(f = f_k)$, and the requested files f_k are independent across the index k.

Assuming flat-fading channel, the baseband signal $\mathbf{y}_k \in \mathbb{C}^{n_{U,k} \times 1}$ received by UE k in each transmission interval is given as

$$\mathbf{y}_k = \sum_{i \in \mathcal{N}_R} \mathbf{H}_{k,i}\mathbf{x}_i + \mathbf{z}_k = \mathbf{H}_k\mathbf{x} + \mathbf{z}_k, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{C}^{n_{R,i} \times 1}$ is the baseband signal transmitted by eRRH i in a given downlink discrete channel use, or symbol; $\mathbf{H}_{k,i} \in \mathbb{C}^{n_{U,k} \times n_{R,i}}$ denotes the channel response matrix from eRRH i to UE k; $\mathbf{z}_k \in \mathbb{C}^{n_{U,k} \times 1}$ is the additive noise distributed as $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I})$; $\mathbf{H}_k \triangleq [\mathbf{H}_{k,1} \ldots \mathbf{H}_{k,N_R}] \in \mathbb{C}^{n_{U,k} \times n_R}$ collects the channel matrices $\mathbf{H}_{k,i}$ from each eRRH i to any UE k; and $\mathbf{x} \triangleq [\mathbf{x}_1; \ldots; \mathbf{x}_{N_R}] \in \mathbb{C}^{n_R \times 1}$ is the signal transmitted by all the eRRHs. We assume that each eRRH i is subject to the average transmit power constraint stated as $\mathbb{E}\|\mathbf{x}_i\|^2 \leq P_i$. Furthermore, the channel matrices $\{\mathbf{H}_{k,i}\}_{k \in \mathcal{N}_U, i \in \mathcal{N}_R}$ are assumed to remain constant during each transmission interval and to be known to the BBU and eRRHs.

The system operates in two phases, namely *pre-fetching* and *delivery* (see, e.g., [12]). Pre-fetching operates at a large time scale corresponding to the period in which file popularity remains constant. This time scale encompasses multiple transmission intervals. The delivery phase operates separately on each transmission interval. Satisfying each vector of users' requests may generally require *multiple* transmission intervals and, as detailed in Sec. IV, in this paper we focus on one such transmission interval with the aim of maximizing the *delivery rate*. Then, new requests $\{f_k\}_{k \in \mathcal{N}_U}$ are considered and the corresponding files are transmitted.

In the **pre-fetching phase**, each eRRH i downloads and stores up to nB_i bits from the library of files, which is of size nSF bits (see Fig. 1). We define the *fractional caching capacity* μ_i of eRRH i as

$$\mu_i \triangleq \frac{B_i}{SF}. \tag{2}$$

Accordingly, each eRRH can potentially store a fraction μ_i of each file (see [11]-[13]). Different standard pre-fetching policies will be considered as detailed in Sec. III. Note that pre-fetching strategies cannot be adapted to the channel matrices or requested file profile $\{f_k\}_{k \in \mathcal{N}_U}$ in each transmission interval.

In the **delivery phase**, the eRRHs transmit in the downlink in order to deliver the requested files $\mathcal{F}_{\text{req}} \triangleq \cup_{k \in \mathcal{N}_U}\{f_k\}$ to the UEs. The transmitted signal $\mathbf{x}_i$ of each eRRH i is obtained as a function of the information stored in its local cache, as well as of the information received from the BBU on the fronthaul link.

III. Pre-Fetching Phase

The pre-fetching policy chooses nB_i bits out of the library of nSF bits to be stored in the cache of eRRH i. The pre-fetching strategy is determined based only on long-term state information about the popularity distribution $P(f)$, as well as on the cache memory sizes $\{B_i\}_{i \in \mathcal{N}_R}$, file size nS and the fronthaul capacities $\{C_i\}_{i \in \mathcal{N}_R}$.

In this paper, as in [8][10][12], we limit our attention to uncoded strategies. To this end, for the sake of generality, we assume that each file f is split into L subfiles $(f, 1), \ldots, (f, L)$ such that each subfile (f, l) is of size nS_l bits with $\sum_{l \in \mathcal{L}} S_l = S$ and $\mathcal{L} \triangleq \{1, \ldots, L\}$ (see, e.g., [12, Sec. III]). Then, the pre-fetching strategy can be modeled by defining binary caching variables $\{c^i_{f,l}\}_{f \in \mathcal{F}, l \in \mathcal{L}, i \in \mathcal{N}_R}$ as

$$c^i_{f,l} = \begin{cases} 1, & \text{if subfile } (f,l) \text{ is cached by eRRH } i \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

while satisfying the cache memory constraint at eRRH i as

$$\sum_{f \in \mathcal{F}} \sum_{l \in \mathcal{L}} c^i_{f,l} S_l \leq B_i = \mu_i FS, \text{ for all } i \in \mathcal{N}_R. \tag{4}$$

While the problem formulation to be given in later sections applies to any choice of pre-fetching variables (3), the following subsections discuss three explicit standard pre-fetching strategies that will be considered in Sec. V for numerical performance evaluation. For the rest of this section, we set $\mu_i = \mu$ for $i \in \mathcal{N}_R$ to avoid a more cumbersome notation.

A. Cache Most Popular

We first consider a pre-fetching strategy in which all eRRHs cache the same N_C most popular files, namely $f = 1, \ldots, N_C$, where N_C is given as $N_C = \lfloor \mu F \rfloor$ in order to satisfy the cache constraints. This approach, which was also considered in [8, Sec. V], is expected to be a good choice when the parameter γ of the distribution $P(f)$ is large, i.e., when only a few popular files are frequently requested by UEs. We obtain it by setting $L = 1$, $c^i_{f,l} = 1$ if $f \leq N_C$ and $c^i_{f,l} = 0$ otherwise. We refer to this strategy as Cache Most Popular (CMP).

B. Cache Distinct

When the parameter γ is small, it may be advantageous to store as many distinct files as possible in the caches. Thus, we also consider a pre-fetching strategy where eRRH 1 stores files $1, N_R+1, \ldots$; eRRH 2 stores files $2, N_R+2, \ldots$; and so on, until caches are full. This pre-fetching strategy, referred to as Cache Distinct (CD), is obtained by choosing $L = 1$, $c^i_{f,l} = 1$ if $i = \mathrm{mod}(f-1, N_R)+1$ and $c^i_{f,l} = 0$ otherwise. The number N_C of files that can be stored in each cache is again $N_C = \lfloor \mu F \rfloor$.

C. Fractional Cache Distinct

Unlike CMP, CD does not enable cooperative transmission from multiple eRRHs based only on the content of the caches, since each file cannot be stored by multiple eRRHs. To address this issue, which can be significant if the fronthaul capacities C_i are small, we consider a Fractional Cache Distinct (FCD) pre-fetching strategy, where each file f is split into N_R disjoint subfiles, i.e., $L = N_R$, and distributed over eRRHs chosen randomly without replacement. To this end, the sizes of the files are set to $S_l = \mu S$ with $\mu = 1/N_R$ for $l \in \mathcal{N}_R$, and the FCD with general μ was discussed in [14, Sec. III-C]. This policy can be implemented by setting the caching variables $c^i_{f,l}$ to $c^i_{f,l} = 1$ if $l = i_{f,l}$ and $c^i_{f,l} = 0$ otherwise, where $i_{f,1}, \ldots, i_{f,L}$ are obtained as random permutations of the numbers $1, \ldots, N_R$, which are independent across the index f. Randomized caching was also considered in [8, Sec. V] without file splitting, i.e., with $L = 1$.

IV. Delivery Phase With Hybrid Fronthauling

For a given pre-fetching strategy, in this section, we consider the design of the delivery phase in each transmission interval under a hybrid fronthauling mode, whereby the capacity of each fronthaul link is used to carry both hard and soft information about the uncached files. Note that a hybrid scheme was also considered in [15] but for a system with no caching. Furthermore, as a special case of the proposed hybrid scheme, we obtain a novel soft-transfer mode strategy that was not considered in [7]-[9].

To elaborate, we define $\tilde{C}_i \leq C_i$ as the rate used on the ith fronthaul for the soft-transfer mode, that is, for transferring quantized version of the precoded signals for the missing files, in line with the C-RAN paradigm. The rest of the frontahul link of $C_i - \tilde{C}_i$ bit/symbol is instead used for the hard-transfer mode, i.e., for transferring hard information of subfiles that are not cached by the eRRHs. As in [7]-[9], hard-mode fronthauling requires the determination of the set of eRRHs to which each subfile (f,l) is transferred on the fronthaul link. This is done by defining the binary variable $d^i_{f,l}$ as

$$d^i_{f,l} = \begin{cases} 1, & \text{if subfile } (f,l) \text{ is transferred to eRRH } i \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Accounting for both soft- and hard-transfer fronthauling, the fronthaul capacity constraint for each eRRH i is stated as

$$\sum_{f\in\mathcal{F}} \sum_{l\in\mathcal{L}} d^i_{f,l} R_{f,l} + \tilde{C}_i \leq C_i. \tag{6}$$

The signal $\mathbf{x}_i$ transmitted by eRRH i on the downlink channel is given as the superposition of a locally encoded signal and of a BBU-encoded and quantized signal as

$$\mathbf{x}_i = \sum_{f\in\mathcal{F}_{\mathrm{req}}} \sum_{l\in\mathcal{L}} \left(1 - \bar{c}^i_{f,l}\bar{d}^i_{f,l}\right) \mathbf{V}^i_{f,l}\mathbf{s}_{f,l} + \hat{\mathbf{x}}_i, \tag{7}$$

where $\mathbf{V}^i_{f,l} \in \mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix for the baseband signal $\mathbf{s}_{f,l} \in \mathbb{C}^{n_{S,f,l}\times 1}$, which encodes the subfile (f,l) available at the eRRH and is distributed as $\mathbf{s}_{f,l} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, while $\hat{\mathbf{x}}_i$ represents the quantized baseband signal received from the BBU on the fronthaul link. Note that the contribution of subfile (f,l) to the first term in (7) is non-zero if the subfile (f,l) is available at the eRRH by caching or via hard-mode fronthauling, i.e., with $c^i_{f,l} = 1$ or $d^i_{f,l} = 1$, respectively.

We now elaborate on the BBU-encoded signal $\hat{\mathbf{x}}_i$. The BBU precodes the subfiles (f,l) that are not available at eRRH i, i.e., with $\bar{c}^i_{f,l}\bar{d}^i_{f,l} = 1$, producing the signal

$$\tilde{\mathbf{x}}_i = \sum_{f\in\mathcal{F}_{\mathrm{req}}} \sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}\mathbf{s}_{f,l}, \tag{8}$$

where $\mathbf{U}^i_{f,l} \in \mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix for the baseband signal $\mathbf{s}_{f,l}$ that encodes the fragment (f,l) not available at eRRH i. The signal $\tilde{\mathbf{x}}_i$ is quantized, obtaining the signal $\hat{\mathbf{x}}_i$ in the right-hand side of (7) as

$$\hat{\mathbf{x}}_i = \tilde{\mathbf{x}}_i + \mathbf{q}_i, \tag{9}$$

where $\mathbf{q}_i$ denotes the quantization noise, which is independent of $\tilde{\mathbf{x}}_i$ and distributed as $\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$, with covariance matrix $\mathbf{\Omega}_i \succeq \mathbf{0}$. The signals $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ for different eRRHs $i \neq j$ are quantized independently so that the quantization noise signals $\mathbf{q}_i$ and $\mathbf{q}_j$ are independent [16] (see also [17] for more general strategies). Using standard information theoretic results (see, e.g., [18, Ch. 3]), the signal $\hat{\mathbf{x}}_i$ can be reliably recovered by eRRH i if the condition

$$\begin{aligned} &g_i\left(\mathbf{U}, \mathbf{\Omega}\right) \triangleq I\left(\tilde{\mathbf{x}}_i; \hat{\mathbf{x}}_i\right) \\ &= \log\left|\sum_{f\in\mathcal{F}_{\mathrm{req}}} \sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}\mathbf{U}^{i\dagger}_{f,l} + \mathbf{\Omega}_i\right| - \log|\mathbf{\Omega}_i| \leq \tilde{C}_i \end{aligned} \tag{10}$$

is satisfied, where we define the notations $\mathbf{U} \triangleq \{\mathbf{U}^i_{f,l}\}_{f\in\mathcal{F}_{\mathrm{req}}, l\in\mathcal{L}, i\in\mathcal{N}_R}$ and $\mathbf{\Omega} \triangleq \{\mathbf{\Omega}_i\}_{i\in\mathcal{N}_R}$.

With (7), the signal (1) received by UE k can be written as

$$\mathbf{y}_k = \sum_{f\in\mathcal{F}_{\mathrm{req}}} \sum_{l\in\mathcal{L}} \mathbf{H}_k\bar{\mathbf{V}}_{f,l}\mathbf{s}_{f,l} + \mathbf{H}_k\mathbf{q} + \mathbf{z}_k, \tag{11}$$

where we defined the aggregated precoding matrix $\bar{\mathbf{V}}_{f,l} \triangleq [\bar{\mathbf{V}}^1_{f,l}; \ldots; \bar{\mathbf{V}}^{N_R}_{f,l}]$ for subfile (f,l) with $\bar{\mathbf{V}}^i_{f,l} \triangleq (1 - \bar{c}^i_{f,l}\bar{d}^i_{f,l})\mathbf{V}^i_{f,l} + \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}$ and the quantization noise vector

$\mathbf{q} \triangleq [\mathbf{q}_1; \ldots; \mathbf{q}_{N_R}]$ distributed as $\mathbf{q} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{\Omega}})$ with $\bar{\mathbf{\Omega}} \triangleq \text{diag}(\{\mathbf{\Omega}_i\}_{i\in\mathcal{N}_R})$.

We assume that, based on (11), each UE k performs successive interference cancellation decoding while treating the interference signals as noise. Without loss of generality, due to the equivalence of the subfiles of any given file, we consider the decoding order $\mathbf{s}_{f_k,1} \to \ldots \to \mathbf{s}_{f_k,L}$ so that the rate $R_{f_k,l}$ at which subfile (f_k, l) can be reliably transmitted is bounded as

$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right) \triangleq I\left(\mathbf{s}_{f_k,l}; \mathbf{y}_k | \mathbf{s}_{f_k,1}, \ldots, \mathbf{s}_{f_k,l-1}\right) \quad (12)$$
$$= \Phi\left(\begin{array}{c} \mathbf{H}_k \bar{\mathbf{V}}_{f_k,l} \bar{\mathbf{V}}_{f_k,l}^\dagger \mathbf{H}_k^\dagger, \\ \sum_{m=l+1}^{L} \mathbf{H}_k \bar{\mathbf{V}}_{f_k,m} \bar{\mathbf{V}}_{f_k,m}^\dagger \mathbf{H}_k^\dagger + \\ \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}} \sum_{m\in\mathcal{L}} \mathbf{H}_k \bar{\mathbf{V}}_{f,m} \bar{\mathbf{V}}_{f,m}^\dagger \mathbf{H}_k^\dagger \\ + \mathbf{H}_k \bar{\mathbf{\Omega}} \mathbf{H}_k^\dagger + N_0 \mathbf{I} \end{array}\right),$$

where we defined the notation $\bar{\mathbf{V}} \triangleq \{\bar{\mathbf{V}}_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$ and the function $\Phi(\mathbf{A}, \mathbf{B}) \triangleq \log|\mathbf{A}+\mathbf{B}| - \log|\mathbf{B}|$.

We allow any subfile (f, l) to be delivered to the UE at a rate $R_{f,l} \leq S_l$, so that $nR_{f,l} \leq nS_l$ bits are transmitted to the UE in the given transmission interval. The remaining $nS_l - nR_{f,l}$ bits can then be sent in the following transmission intervals by solving a similar optimization problem.

A. Problem Definition and Optimization

We aim at optimizing the precoding matrices $\mathbf{V}$ and $\mathbf{U}$ applied at the eRRHs and the BBU and the quantization noise covariance matrices $\mathbf{\Omega}$, along with the capacities $\tilde{\mathbf{C}} \triangleq \{\tilde{C}_i\}_{i\in\mathcal{N}_R}$ used for soft-transfer fronthauling, with the goal of maximizing the minimum-user rate $R_{\min} \triangleq \min_{f\in\mathcal{F}_{\text{req}}} R_f$, where $R_f \triangleq \sum_{l\in\mathcal{L}} R_{f,l}$ denotes the achievable delivery rate for file f, while satisfying the fronthaul capacity (6) and per-eRRH power constraints. We recall from our discussion above that maximizing $R_{\min}$ is instrumental in reducing the number of transmission intervals needed to deliver all the files $\mathcal{F}_{\text{req}}$ to the requesting UEs. The problem is stated as

$$\underset{\bar{\mathbf{V}},\mathbf{\Omega},R_{\min},\mathbf{R},\tilde{\mathbf{C}}}{\text{maximize}} \quad R_{\min} \quad (13a)$$
$$\text{s.t.} \quad R_{\min} \leq \sum_{l\in\mathcal{L}} R_{f,l}, \ f \in \mathcal{F}_{\text{req}}, \quad (13b)$$
$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right), \ l \in \mathcal{L}, \ k \in \mathcal{N}_U, \quad (13c)$$
$$g_i\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right) \leq \tilde{C}_i, \ i \in \mathcal{N}_R, \quad (13d)$$
$$\sum_{f\in\mathcal{F}_{\text{req}}} \sum_{l\in\mathcal{L}} d_{f,l}^i R_{f,l} + \tilde{C}_i \leq C_i, \ i \in \mathcal{N}_R, \quad (13e)$$
$$R_{f,l} \leq S_l, \ f \in \mathcal{F}_{\text{req}}, l \in \mathcal{L}, \quad (13f)$$
$$\sum_{f\in\mathcal{F}_{\text{req}}} \sum_{l\in\mathcal{L}} \text{tr}\left(\mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l} \bar{\mathbf{V}}_{f,l}^\dagger \mathbf{E}_i + \mathbf{\Omega}_i\right) \leq P_i, \ i \in \mathcal{N}_R. \quad (13g)$$

where we defined the matrix $\mathbf{E}_i \in \mathbb{C}^{n_R \times n_{R,i}}$ containing zero entries except for the rows from $\sum_{j=1}^{i-1} n_{R,j}+1$ to $\sum_{j=1}^{i} n_{R,j}$ containing the identity matrix of size $n_{R,i}$, and the notation $\mathbf{R} \triangleq \{R_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$. The function $g_i(\bar{\mathbf{V}}, \mathbf{\Omega})$ in (13d) is defined, with a small abuse of notation, by substituting $\mathbf{U}_{f,l}^i = \mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}$ into $g_i(\mathbf{U}, \mathbf{\Omega})$ in (10). In the problem, the constraint (13f) imposes that the rate $R_{f,l}$ of each subfile be limited by the subfile size S_l, and the constraint (13g) is equivalent to the per-eRRH power constraints within the precoding model (7). We emphasize that in (13), the pre-fetching variables (3) and the fronthaul transfer variables (5) are fixed.

The solution of problem (13) is made difficult by the non-convexity in the constraints (13c) and (13d). Here, noting that the right- and left-hand sides of (13c) and (13d) have the difference-of-convex (DC) structure when stated in terms of the covariance matrices $\mathbf{W}_{f,l} \triangleq \bar{\mathbf{V}}_{f,l} \bar{\mathbf{V}}_{f,l}^\dagger \succeq \mathbf{0}$ and $\mathbf{\Omega}$, as in [8][17], we adopt the concave-convex procedure (CCCP). Specifically, we address problem (13) with optimization variables $\mathbf{W} \triangleq \{\mathbf{W}_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$ by relaxing the rank constraints $\text{rank}(\mathbf{W}_{f,l}) \leq n_{S,f,l}$. The algorithm follows immediately from the standard CCCP approach (see [14] for details).

V. Numerical Results

In this section, we present some numerical results that compare the performance of hard-, soft- and hybrid-transfer fronthauling modes with the pre-fetching strategies discussed in Sec. III. Hard-transfer fronthauling is obtained by setting $\tilde{C}_i = 0$ for all $i \in \mathcal{N}_R$ and soft-transfer frontauling is given via the choice $\tilde{C}_i = C_i$ for all $i \in \mathcal{N}_R$ in (13). We consider an F-RAN system where the positions of eRRHs and UEs are uniformly distributed within a circular cell of radius 500m (see [14, Sec. VII] for more details on the path-loss model). We consider a symmetric setting where the eRRHs have the same transmit power and fronthaul capacity, i.e., $P_i = P$ and $C_i = C$ for $i \in \mathcal{N}_R$ and are equipped with caches of equal size, i.e., $\mu_i = \mu$ for $i \in \mathcal{N}_R$. For hard- and hybrid-transfer fronthauling, the fronthaul transfer variables $\{d_{f,l}^i\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$ are set such that the subfile (f_k, l) requested by UE k is transferred on the fronthaul links to the N_F eRRHs that have the largest channel gains $||\mathbf{H}_{k,i}||_F^2$ to the UE and have not stored the subfile, where $N_F \leq N_R$ is a parameter. We focus on the case with $N_R = N_U = 3$, $n_{R,i} = n_{U,k} = 1$ and $P/N_0 = 20$ dB.

We first study the impact of the file popularity on the F-RAN performance by plotting in Fig. 2 the average minimum rate $R_{\min}$ versus the parameter γ of the Zipf's distribution, where the average is taken with respect to the channel, UEs' requests and the system geometry, for an F-RAN downlink with soft-transfer fronthauling. We set the parameters $F = 3$, $S = 1$ and $C \in \{0.2, 1\}$. We compare the performance of CMP and CD pre-fetching with $\mu = 1/3$ with the case of full ($\mu = 1$) and no ($\mu = 0$) caching (FCD is not shown here to avoid clutter). It is observed from the figure that the performance gain of the CMP pre-fetching strategy with a larger γ, and hence with an increased bias towards the most popular files, is more pronounced for lower values of the fronthaul capacity C. This is because, in the regime of small C, cooperative transmission by means of cloud processing, as in C-RAN, cannot compensate for the lack of cooperation opportunities on the cached files that affects the CD approach. In contrast, when γ is sufficiently small, the CD strategy outperforms CMP

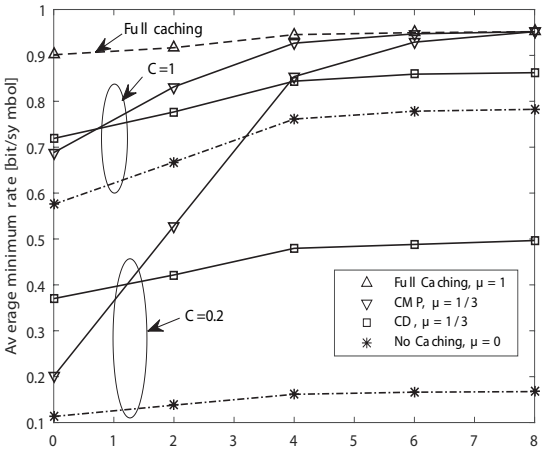

Figure 2. Average minimum rate $R_{\min}$ versus the parameter γ of the Zipf's distribution for an F-RAN downlink under soft-transfer fronthauling mode ($\mu = 0, 1/3, 1$, $F = 3$, $S = 1$ and $C = 0.2$ and 1).

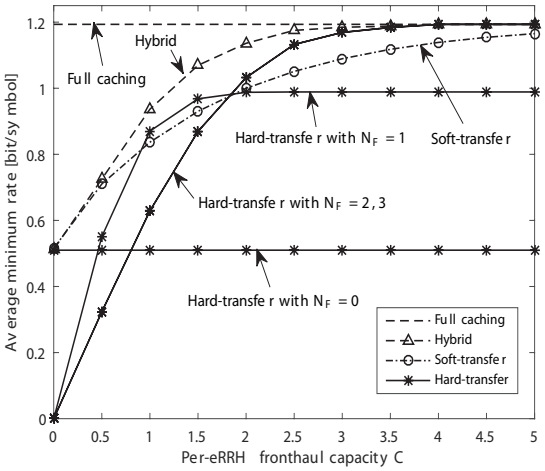

Figure 3. Average minimum rate $R_{\min}$ versus the fronthaul capacity C for an F-RAN downlink under FCD pre-fetching ($\mu = 1/3$ and 1, $F = 6$, $S = 2$ and $\gamma = 0.2$).

approach, which suffers from a significant number of cache misses, particularly for low values of C.

We then study the performance comparison among different delivery strategies by plotting in Fig. 3 the average minimum rate $R_{\min}$ versus the fronthaul capacity C for an F-RAN system with FCD pre-fetching, and with $\mu = 1/3$, $F = 6$, $S = 2$ and $\gamma = 0.2$. From the figure, we observe that the partial caching capacity of the eRRHs, here with $\mu = 1/3$, can be compensated by a larger fronthaul capacity C. For instance, the soft-transfer fronthauling mode with $\mu = 1/3$ needs a fronthaul capacity of $C = 3.38$ bit/symbol to achieve the full-caching upper bound within 5%. Also, it is seen that, if the fronthaul capacity C is sufficiently large, the hard-transfer mode can provide some performance gains over soft-transfer fronthauling, as long as the cooperative cluster size is properly selected. Furthermore, the proposed hybrid scheme, whose performance is here shown for optimized values of N_F, has the capability to improve over both soft- and hard-mode fronthauling, except for very low- and very high-fronthaul capacity regime, in which it reverts to the soft- and hard-mode schemes, respectively.

VI. Concluding Remarks

In this work, we have studied the joint design of cloud and edge processing for an F-RAN architecture in which each edge node is equipped with local cache and baseband processing capabilities. Focusing on the metric of the minimum delivery rate across all UEs, it was concluded that soft-transfer fronthauling, akin to the C-RAN operation, provides a more effective way to use fronthaul resources than the more conventional hard-transfer mode in most operating regimes. Moreover, the proposed hybrid mode based on superposition coding is seen to have the potential to strictly outperform both soft- and hard-transfer modes.

References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Comm. Surveys Tutorials*, vol. 17, no. 1, pp. 405-426, First Quart. 2015.

[2] O. Simeone, A. Maeder, M. Peng, O. Sahin and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," arXiv:1512.07743, Dec. 2015.

[3] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog computing based radio access networks: Issues and Challenges," arXiv:1506.04233, Jun. 2015.

[4] S. Bi, R. Zhang, Z. Ding and S. Cui, "Wireless communications in the era of big data," arXiv:1508.06369, Aug. 2015.

[5] China Mobile, "Next generation fronhtaul interface," White Paper, Oct. 2015.

[6] M. Chiang, "Fog networking: An overview on research opportunities," arXiv:1601.00835, Jan. 2016.

[7] X. Peng, J.-C. Shen, J. Zhang and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," arXiv:1509.00558, Sep. 2015.

[8] M. Tao, E. Chen, H. Zhou and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," arXiv:1512.06938, 2015.

[9] Y. Ugur, Z. H. Awan and A. Sezgin, "Cloud radio access networks with coded caching," arXiv:1512.02385, Dec. 2015.

[10] B. Azari, O. Simeone, U. Spagnolini and A. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching," to appear in *IEEE Wireless Comm. Letters*.

[11] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," submitted, Jan. 2016.

[12] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," *Proc. IEEE Intern. Symp. on Inf. Theory (ISIT) 2015*, Hong Kong, China, Jun. 2015.

[13] A. Sengupta, R. Tandon and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," arXiv:1512.07856, 2015.

[14] S.-H. Park, O. Simeone and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," arXiv:1601.02460, Jan. 2016.

[15] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," *Proc. IEEE Inf. Theory and Application Workshop 2014*, San Diego, CA, USA, Feb. 2014.

[16] O. Simeone, O. Somekh, H. V. Poor and S. Shamai (Shitz), "Downlink multicell processing with limited backhaul capacity," *EURASIP Journal on Advances in Signal Processing*, 2009.

[17] 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 Sig. Processing Mag.*, vol. 31, no. 6, pp. 69-79, Nov. 2014.

[18] A. E. Gamal and Y.-H. Kim, *Network information theory*, Cambridge University Press, 2011.