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

Seok-Hwan Park, *Member, IEEE*, Osvaldo Simeone, *Fellow, IEEE*, and Shlomo Shamai (Shitz), *Fellow, IEEE*

***Abstract*— This paper studies the joint design of cloud and edge processing for the downlink of a fog radio access network (F-RAN). In an F-RAN, as in cloud-RAN (C-RAN), a baseband processing unit (BBU) can perform joint baseband processing on behalf of the remote radio heads (RRHs) that are connected to the BBU by means of the fronthaul links. In addition to the minimal functionalities of conventional RRHs in C-RAN, the RRHs in an F-RAN may be equipped with local caches, in which frequently requested contents can be stored, as well as with baseband processing capabilities. They are hence referred to as enhanced RRH (eRRH). This paper focuses on the design of the delivery phase for an arbitrary pre-fetching strategy used to populate the caches of the eRRHs. Two fronthauling modes are considered, namely, a *hard-transfer mode*, whereby non-cached files are communicated over the fronthaul links to a subset of eRRHs, and a *soft-transfer mode*, whereby the fronthaul links are used to convey quantized baseband signals as in a C-RAN. Unlike the hard-transfer mode in which baseband processing is traditionally carried out only at the eRRHs, the soft-transfer mode enables both centralized precoding at the BBU and local precoding at the eRRHs based on the cached contents, by means of a novel superposition coding approach. To attain the advantages of both approaches, a hybrid design of soft- and hard-transfer modes is also proposed. The problem of maximizing the delivery rate is tackled under fronthaul capacity and per-eRRH power constraints. Numerical results are provided to compare the performance of hard- and soft-transfer fronthauling modes, as well as of the hybrid scheme, for different baseline pre-fetching strategies.**

***Index Terms*— Fog radio access network, edge caching, pre-fetching, fronthaul compression, beamforming, C-RAN.**

I. Introduction

CLOUD radio access network (C-RAN) is an emerging architecture for the fifth-generation (5G) of wireless systems, in which a centralized baseband signal processing unit (BBU) implements the baseband processing functionalities of a set of remote radio heads (RRHs) (see, e.g., [1]). The RRHs are limited-complexity devices that typically implement only radio-frequency (RF) functionalities and they are connected to the BBU by means of fronthaul links [2]–[4]. In the digital fronthauling approach adopted by the Common Public Radio Interface (CPRI) specification [5], the BBU performs channel coding on behalf of the RRHs, and then quantizes and compresses the encoded baseband signals prior to the transfer to the RRHs (see, e.g., [6]–[9]). The quantization and compression of the baseband signals allow the communication between BBU and RRHs to take place over finite-capacity fronthaul links.

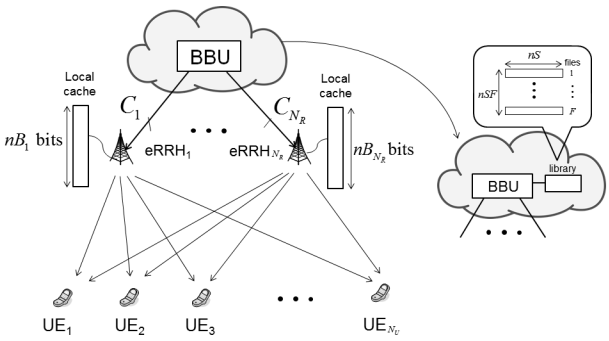

Fig. 1. Illustration of an F-RAN, which has both cloud and edge processing capabilities: the BBU, in the "cloud", can perform joint baseband processing and the eRRHs are equipped with local caches.

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 storage and signal processing functionalities [10]–[12]. The resulting RRHs are referred to here as *enhanced RRHs* (eRRHs).[1] In an F-RAN, edge caching can be performed to pre-fetch the most frequently requested files to the eRRHs' local caches during off-peak traffic periods, e.g., at night, as illustrated in Fig. 1. In this way, fronthaul overhead can be reduced during the peak traffic periods, and higher spectral efficiencies or lower delivery latency can be obtained, see, e.g., [13]. It is emphasized that, unlike C-RAN [14], the goal of the F-RAN architecture is not that of minimizing the deployment and operating costs by means of reduced-complexity edge nodes, but rather that of maximizing the system performance in terms of delivery rate by leveraging both *cloud (BBU) and edge (caching) resources* [15]–[19].[2]

As a cache-aided system, an F-RAN operates in two phases, namely the pre-fetching and the delivery phases [15]–[19]

Manuscript received January 9, 2016; revised May 19, 2016 and August 22, 2016; accepted August 28, 2016. Date of publication September 1, 2016; date of current version November 9, 2016. The work of S.-H. Park was supported by the National Research Foundation of Korea funded by the Korean Government within the Ministry of Science, ICT and Future Planning under Grant 2015R1C1A1A01051825. The work of O. Simeone was supported by U.S. NSF under Grant 1525629. The work of S. Shamai was supported in part by the Israel Science Foundation and in part by the European Research Council Advanced Grant under Grant 694630. The associate editor coordinating the review of this paper and approving it for publication was M. Li.

S.-H. Park is with the Division of Electronic Engineering, Chonbuk National University, Jeonju 54896, South Korea (e-mail: seokhwan@jbnu.ac.kr).

O. Simeone is with the Center for Wireless Information Processing, Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA, NJ 07102 USA (e-mail: osvaldo.simeone@njit.edu).

S. Shamai is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sshlomo@ee.technion.ac.il).

Digital Object Identifier 10.1109/TWC.2016.2605104

[1] In [12], eRRHs are referred to as Radio Remote Systems (RRSs).

[2] See also [20, Sec. D].

(see also [21], [22]). Pre-fetching amounts to the storage of popular content during off-peak traffic periods. Between two such periods, the delivery phase encompasses multiple transmission intervals, which are generally characterized by the different content requests by the users.

A. Related Works

The optimization of the downlink of C-RAN systems entails the design of transfer strategies on the fronthaul and of precoding strategies on the channel between RRHs and user equipments (UEs). This problem has been studied in a number of works, including [6]–[9] and [23]–[27]. For instance, the papers [6]–[9] and [23], [24] deal with the maximization of the (weighted) sum-rate under fronthaul capacity constraints, whereas the problem of minimizing system costs that account for fronthaul overhead while satisfying target signal-to-interference-plusnoise ratio (SINR) constraints was studied in [25]–[27]. Since the resulting optimization programs are non-convex, the typical approach is to obtain numerical solutions by means of state-of-the-art iterative successive convex approximation strategies (see [28] and references therein).

When considering F-RAN systems, the optimization problem is made more complex by the need to optimize, jointly with fronthaul transfer and wireless precoding, also the caching strategy. The problem of optimizing edge caching alone, assuming no fronthaul connections from eRRHs to cloud, has recently received significant attention in the information-theoretic literature [21], [29]–[32]. Generalizing these works, the information-theoretic framework developed in [33] and [34] for the analysis of F-RAN demonstrates the interplay of fronthaul, wireless and caching policies for the minimization of the delivery latency.

A fronthaul-aware design of the caching, or pre-fetching, policy was studied with the aim of minimizing the average delivery latency while satisfying the cache memory constraints in [15]. Since the optimization problem turns out to be a mixed integer nonlinear program, the authors obtained a difference-of-convex (DC) problem by means of smooth approximation and integer relaxation, and proposed a successive convex approximation algorithm. In [16], the authors consider the joint design of cooperative beamforming and eRRH clustering for the delivery phase, under an arbitrary fixed pre-fetching strategy, with the goal of minimizing the network cost, which is defined as the sum of transmit power and backhaul cost, under quality-of-service constraints. A similar problem was tackled in [19] by assuming that coded, instead of uncoded, caching is exploited (see also [35]). In [36], a stochastic geometry-based analysis is provided of a specific hybrid caching strategy (see Sec. [36, Sec. II-B]). Reference [17] proposes a hypergraph-based framework to obtain first-order quantitative insights into the performance of an F-RAN architecture without the need to perform the non-convex optimization considered in [15]–[19]. As an extension of this work, in [37], the problem of minimizing delivery latency across fronthaul and wireless links was studied under fronthaul capacity constraints. Furthermore, building on a preliminary posting of this article, reference [38] aimed at optimizing eRRH clustering and precoding strategies for fixed pre-fetching strategy with the criterion of minimizing the sum of transmission powers of eRRHs under the target SINR and fronthaul capacity constraints. We emphasize that the numerical solutions adopted in all references [15], [16], [19], [23], and [25]–[27] are based on successive convex approximation techniques.

B. Main Contributions

In the references [15]–[17], [25]–[27], [29]–[31], [35], [36] summarized above, the fronthaul links in an F-RAN are leveraged in a *hard-transfer mode* to convey to the eRRHs the requested content that is not present in the local caches. In contrast, in this work, we consider not only the mentioned hard-transfer mode, but also a novel *soft-transfer mode* for the use of the fronthaul links. The proposed approach 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 of the cache and another that is encoded at the BBU and quantized for transmission on the fronthaul link. Specifically, we study the joint design of cloud and edge processing for the delivery phase of an F-RAN for an arbitrary pre-fetching strategy by considering hard-transfer and soft-transfer fronthauling strategies. For both fronthauling modes, we tackle the problem of optimizing cloud and edge processing, i.e., processing at the BBU and at the eRRHs, with the goal of maximizing the delivery rate while satisfying fronthaul capacity and per-eRRH power constraints. Furthermore, to reap the advantages of the two fronthauling approaches, we also propose a hybrid design of hard- and soft-transfer modes. This generalizes the approach of [9], where it was studied in the absence of caching. Numerical results are provided to compare the performance of hard- and soft-transfer fronthauling modes, as well as the hybrid scheme, for baseline pre-fetching strategies.

The rest of the paper is organized as follows. We describe the system model in Sec. II and review some baseline pre-fetching strategies in Sec. III. We discuss the design of delivery phase under hard-transfer fronthaul mode in Sec. IV and then propose a novel soft-transfer strategy in Sec. V. A hybrid design of hard- and soft-transfer modes is studied in Sec. VI, and extensive numerical results are presented in Sec. VII. We close the paper with some concluding remarks in Sec. VIII.

C. Notation

We adopt standard information-theoretic definitions for the mutual information $I(X;Y)$ between the random variables X and Y [39]. The circularly symmetric complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$. The set of all $M \times N$ complex matrices is denoted by $\mathbb{C}^{M\times N}$, and $\mathbb{E}(\cdot)$ represents the expectation operator. The operation $(\cdot)^\dagger$ denotes Hermitian transpose of a matrix or vector, and $\bar{a}$ is defined as $1-a$ for a binary variable $a \in \{0,1\}$. For a scalar x, $\lfloor x \rfloor$ denotes the largest integer not larger than x.

II. System Model

As illustrated in Fig. 1, we consider the downlink of an F-RAN, where N_U multi-antenna UEs are served

by N_R multi-antenna eRRHs that are connected to a BBU in the "cloud" through digital fronthaul links. The model generalizes the C-RAN set-up studied in, e.g., [6]–[9]. In fact, in addition to the functionalities performed by conventional RRHs in C-RAN, such as upconversion and RF transmission, 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. 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 $n_R \triangleq \sum_{i=1}^{N_R} n_{R,i}$ and $n_U \triangleq \sum_{k=1}^{N_U} n_{U,k}$.

Each eRRH i is connected to the BBU with a fronthaul link of capacity C_i bit per symbol of the downlink channel for $i \in \mathcal{N}_R \triangleq \{1, \ldots, N_R\}$. As such, in each transmission interval of duration T, any eRRH i can receive TC_i bits from the BBU prior to transmission on the wireless channel.

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. The assumption of equal file-sizes is standard (see, e.g., [40]) and can be justified because, in practice, what is cached and requested by users are chunks of videos, e.g., fragments of a given duration, which may be safely assumed to be of the same length. Labeling the files in order of popularity, the probability $P(f)$ of a file f to be selected is defined by Zipf's distribution (see, e.g., [15]–[17])

$$P(f) = cf^{-\gamma} \tag{1}$$

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 $\sum_{f \in \mathcal{F}} P(f) = 1$. Note that, as the exponent γ increases, the popularity distribution becomes more skewed towards the most popular files. Each UE k requests file $f_k \in \mathcal{F}$ with the probability (1), 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

$$\mathbf{y}_k = \sum_{i \in \mathcal{N}_R} \mathbf{H}_{k,i}\mathbf{x}_i + \mathbf{z}_k = \mathbf{H}_k\mathbf{x} + \mathbf{z}_k, \tag{2}$$

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 $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, \Sigma_{\mathbf{z}_k})$ for some covariance matrix $\Sigma_{\mathbf{z}_k}$; $\mathbf{H}_k \triangleq [\mathbf{H}_{k,1} \ldots \mathbf{H}_{k,N_R}] \in \mathbb{C}^{n_{U,k} \times n_R}$ 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

$$\mathbb{E}\,\|\mathbf{x}_i\|^2 \leq P_i. \tag{3}$$

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 robust design with imperfect CSI or via alternating distributed optimization [41] is out of the scope of this work.

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

$$\mu_i \triangleq \frac{B_i}{SF}. \tag{4}$$

Accordingly, each eRRH can potentially store a fraction μ_i of each file (see [21], [22]). 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. We consider two different approaches depending on the type of the information transferred on the fronthaul links: *hard-transfer fronthauling* and *soft-transfer fronthauling*. In the former, the fronthaul links are used for the transfer of hard information regarding the missing files that are not cached by the eRRHs as in [15]–[17]; while, with the soft-transfer mode, the fronthaul links transfer quantized version of the precoded signals for the missing files, in line with the C-RAN paradigm. Soft- and hard-mode fronthauling strategies were compared for C-RAN systems, i.e., with no caching, in terms of achievable rates under an ergodic fading channel model in [42] and in terms of energy expenditure in [43]. In the next sections, we detail separately the pre-fetching and delivery phases. Moreover, for the delivery phase, we will consider separately operations with hard- and soft-transfer fronthauling, and also with a hybrid scheme that combines the advantages of the two fronthauling approaches. As we will explain in Sec. IV-VI, the performance metric of interest is the downlink transmission rate in each transmission interval. The symbols described in this section are summarized in Table 1.

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. Different policies for caching can be considered, including coded caching [19], [35]. 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 [16], [17], and [21], 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., [21, Sec. III]). Then, the pre-fetching strategy can be modeled by defining binary caching variables

TABLE I
TABLE SUMMARIZING IMPORTANT SYMBOLS USED THROUGHOUT THE PAPER

Symbol	Meaning
N_R, N_U	Numbers of eRRHs and UEs
$\mathcal{N}_R$, $\mathcal{N}_U$	Sets of eRRHs and UEs
$n_{R,i}$, $n_{U,k}$	Numbers of antennas of eRRH i and UE k
P_i	Transmission power of eRRH i
C_i	Capacity of fronthaul link to eRRH i
F	Number of files in the library
$P(f)$	Probability of file f to be selected
n	Number of symbols of each coded block
nS	Size of each file f
nB_i	Cache size of eRRH i
μ_i	Fractional caching capacity
f_k	File requested by UE k
$\mathcal{F}_{\text{req}}$	Set of all requested files
$\mathbf{y}_k$	Signal received by UE k
$\mathbf{x}_i$	Signal transmitted by eRRH i
$\mathbf{H}_{k,i}$	Channel matrix from eRRH i to UE k

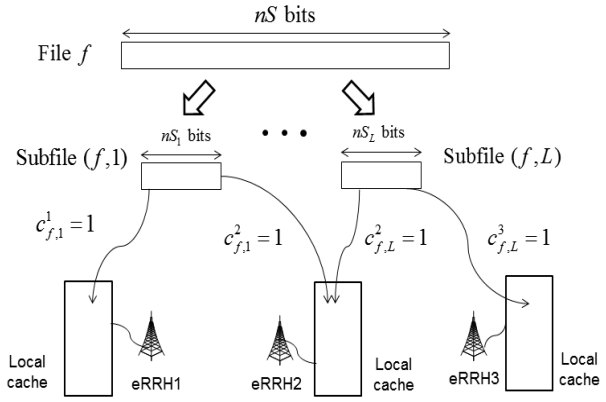

Fig. 2. Illustration of the pre-fetching phase for an example with $N_R = 3$ eRRHs.

$\{c^i_{f,l}\}_{f\in\mathcal{F},l\in\mathcal{L},i\in\mathcal{N}_R}$ as

$$c^i_{f,l} = \begin{cases} 1, & \text{if subfile } (f,l) \text{ is cached by eRRH } i \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

while satisfying the cache memory constraint at eRRH i as

$$\sum_{f\in\mathcal{F}}\sum_{l\in\mathcal{L}} c^i_{f,l} S_l \le B_i = \mu_i F S, \quad \text{for all } i \in \mathcal{N}_R. \tag{6}$$

Fig. 2 illustrates an example.

While the problem formulation to be given in later sections applies to any choice of pre-fetching variables (5), the following subsections discuss three explicit standard pre-fetching strategies that will be considered in Sec. VII for numerical performance evaluation. For the rest of this section, we set $\mu_i = \mu$ for $i \in \mathcal{N}_R$ in order 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 [16, 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 $L = 1$ and

$$c^i_{f,l} = \begin{cases} 1, & \text{if } f \le N_C \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

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 $L = 1$ and

$$c^i_{f,l} = \begin{cases} 1, & \text{if } i = \text{mod}(f-1, N_R) + 1 \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

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 randomized Fractional Cache Distinct (FCD) pre-fetching strategy, where each file f is split into multiple subfiles, i.e., $L > 1$, and distributed over the eRRHs so that each eRRH stores up to $\lfloor \mu L \rfloor$ fragments of each file that are randomly chosen without replacement. This policy can be implemented by setting the caching variables $c^i_{f,l}$ to

$$c^i_{f,l} = \begin{cases} 1, & \text{if } l \in \mathcal{L}^i_f \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{L}^i_f$ is a set of $\lfloor \mu L \rfloor$ random numbers randomly picked from $\mathcal{L}$, which are independent across the file f and eRRH indices i. Randomized caching was also considered in [16, Sec. V] without file splitting, i.e., with $L = 1$.

We close this section with two remarks. First, a hybrid CMP and FCD caching policy was proposed in [36], whereby part of the cache of each eRRH is used to cache the same most popular files and the rest is instead leveraged to store distinct fragments of less popular files. The second remark is that the optimization of pre-fetching strategy based on long-term state information could be addressed by adopting stochastic optimization techniques (see, e.g., [44]), but here we leave this challenging aspect as an interesting open problem.

IV. DELIVERY PHASE WITH HARD-TRANSFER FRONTHAULING

For a given pre-fetching strategy, in this section, we consider the design of the delivery phase in each transmission interval

under the hard-transfer fronthaul mode, where the fronthaul links are used to transfer hard information of subfiles that are not cached by eRRHs. This mode was also considered in [15]–[17], [25]–[27], [29]–[31], [35], and [36]. The formulation considered here is akin to that of [16], with the difference that in this paper we study the maximization of the delivery rate under fronthaul capacity constraints, rather than the minimization of a compound cost function that includes both downlink power and fronthaul capacity as in [16]. The analysis of hard-mode fronthaul is included here mostly for the purpose of comparison with the soft-transfer mode.

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. Our goal is that of maximizing the minimum of transmission rates $R_{f,l}$ that can be supported in any transmission interval. The minimum rate metric is argued in [45] to account for different quality-of-service criteria. A related optimization under the alternative performance criterion of delivery latency across fronthaul and wireless segments in a given transmission interval can be found in [37], and an information-theoretic study of the delivery latency, allowing also for pipelined fronthaul-edge transmission, is presented in [33].

Hard-mode fronthauling requires the determination of the set of eRRHs to which each subfile (f, l) is transferred on the fronthaul link. We do this by defining the binary variable $d^i_{f,l}$ as

$$d^i_{f,l} = \begin{cases} 1, & \text{if subfile } (f,l) \text{ is transferred to eRRH } i \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The fronthaul capacity constraint for each eRRH i is stated as

$$\sum_{f\in\mathcal{F}} \sum_{l\in\mathcal{L}} d^i_{f,l} R_{f,l} \leq C_i. \tag{11}$$

In this work, we assume that the fronthaul transfer variables $\{d^i_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$ are predetermined. As an example, in the numerical results in Sec. VII, we will assume that the variables $\{d^i_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$ 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 not stored the subfile and have the largest channel gains $||\mathbf{H}_{k,i}||^2_F$ to the UE, where $N_F \leq N_R$ is a parameter that defines the scheme, i.e.,

$$d^i_{f_k,l} = \begin{cases} 1, & i \in \left\{ i : \left| \left\{ i' : \left\| \mathbf{H}_{k,i'} \right\|^2_F > \left\| \mathbf{H}_{k,i} \right\|^2_F \right\} \right| < N_F \right\} \\ & \quad \text{and } c^i_{f_k,l} = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Note that this implies that the cluster of eRRHs that cooperates for the transmission of any subfile for the hard-transfer mode is of size N_F plus the number of eRRHs that cache that subfile.

Based on the cached or transferred subfiles (f, l) with $c^i_{f,l} = 1$ or $d^i_{f,l} = 1$, respectively, each eRRH i performs channel encoding to produce the encoded baseband signal $\mathbf{x}_i$. Denoting as $\mathcal{N}_{F,i} \triangleq \{(f,l) | c^i_{f,l} = 1 \text{ or } d^i_{f,l} = 1\}$ the set of subfiles available at eRRH i, the eRRH performs linear precoding as in [16] to obtain the transmitted signal $\mathbf{x}_i$ as

$$\mathbf{x}_i = \sum_{(f,l)\in\mathcal{N}_{F,i}} \mathbf{V}^i_{f,l} \mathbf{s}_{f,l} = \sum_{f\in\mathcal{F}} \sum_{l\in\mathcal{L}} (1 - \bar{c}^i_{f,l} \bar{d}^i_{f,l}) \mathbf{V}^i_{f,l} \mathbf{s}_{f,l}, \tag{13}$$

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 $\mathbf{s}_{f,l} \in \mathbb{C}^{n_{S,f,l} \times 1}$ that encodes the subfile (f, l) and is distributed as $\mathbf{s}_{f,l} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

With (13), the received signal $\mathbf{y}_k$ in (2) can be written as

$$\mathbf{y}_k = \sum_{l\in\mathcal{L}} \mathbf{H}_k \bar{\mathbf{V}}_{f_k,l} \mathbf{s}_{f_k,l} + \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}} \sum_{l\in\mathcal{L}} \mathbf{H}_k \bar{\mathbf{V}}_{f,l} \mathbf{s}_{f,l} + \mathbf{z}_k, \tag{14}$$

where the aggregated precoding matrix $\bar{\mathbf{V}}_{f,l} \in \mathbb{C}^{n_R \times n_{S,f,l}}$ for subfile (f, l) is defined as

$$\bar{\mathbf{V}}_{f,l} \triangleq \begin{bmatrix} (1 - \bar{c}^1_{f,l} \bar{d}^1_{f,l}) \mathbf{V}^1_{f,l} \\ (1 - \bar{c}^2_{f,l} \bar{d}^2_{f,l}) \mathbf{V}^2_{f,l} \\ \vdots \\ (1 - \bar{c}^{N_R}_{f,l} \bar{d}^{N_R}_{f,l}) \mathbf{V}^{N_R}_{f,l} \end{bmatrix}. \tag{15}$$

In (14), the first term is the desired signal to be decoded by the receiving UE k, and the second term is the superposition of the interference signals encoding the files requested by the other UEs.

We assume that, based on (14), each UE k performs successive interference cancellation (SIC) decoding. Without loss of generality, we consider the decoding order $\mathbf{s}_{f_k,1} \to \ldots \to \mathbf{s}_{f_k,L}$ so that the rate $R_{f_k,l}$ of the subfile (f_k, l) is bounded as

$$\begin{aligned} R_{f_k,l} &\leq q_{k,l}\left(\bar{\mathbf{V}}\right) \\ &\triangleq I\left(\mathbf{s}_{f_k,l}; \mathbf{y}_k | \mathbf{s}_{f_k,1}, \ldots, \mathbf{s}_{f_k,l-1}\right) \\ &= \log\det\left(+ \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}} \sum_{m\in\mathcal{L}} \mathbf{H}_k \bar{\mathbf{V}}_{f,m} \bar{\mathbf{V}}^\dagger_{f,m} \mathbf{H}^\dagger_k + \Sigma_{\mathbf{z}_k} \right) \\ &\quad - \log\det\left(+ \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}} \sum_{m\in\mathcal{L}} \mathbf{H}_k \bar{\mathbf{V}}_{f,m} \bar{\mathbf{V}}^\dagger_{f,m} \mathbf{H}^\dagger_k + \Sigma_{\mathbf{z}_k} \right), \end{aligned} \tag{16}$$

where we defined the notation $\bar{\mathbf{V}} \triangleq \{\bar{\mathbf{V}}_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$.

A. Problem Definition and Optimization

We aim at maximizing the minimum-user rate $R_{\min}$ defined as $R_{\min} \triangleq \min_{f\in\mathcal{F}_{\text{req}}} R_f$ while satisfying per-eRRH fronthaul capacity and power constraints, where $R_f \triangleq \sum_{l\in\mathcal{L}} R_{f,l}$ denotes the achievable delivery rate for file f. 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. We also refer to [45] for a discussion of the relevance of the criterion $R_{\min}$ to

guarantee quality-of-experience performance metrics for video delivery. The problem is stated as

$$\underset{\bar{\mathbf{V}}, R_{\min}, \mathbf{R}}{\text{maximize}} \; R_{\min} \tag{17a}$$

$$\text{s.t. } R_{\min} \le \sum_{l \in \mathcal{L}} R_{f,l}, \quad f \in \mathcal{F}_{\text{req}}, \tag{17b}$$

$$R_{f_k,l} \le q_{k,l}\left(\bar{\mathbf{V}}\right), \quad l \in \mathcal{L}, \; k \in \mathcal{N}_U, \tag{17c}$$

$$\sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} d^i_{f,l} R^i_{f,l} \le C_i, \quad i \in \mathcal{N}_R, \tag{17d}$$

$$R_{f,l} \le S_l, \quad f \in \mathcal{F}_{\text{req}}, \; l \in \mathcal{L}, \tag{17e}$$

$$\sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} (1 - \bar{c}^i_{f,l} \bar{d}^i_{f,l}) \text{tr}\left(\mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l} \bar{\mathbf{V}}_{f,l}^\dagger \mathbf{E}_i\right) \le P_i, \quad i \in \mathcal{N}_R, \tag{17f}$$

where we define 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 $\mathbf{R} \triangleq \{R_{f,l}\}_{f \in \mathcal{F}_{\text{req}}, l \in \mathcal{L}}$. In the problem, the constraint (17e) imposes that the rate $R_{f,l}$ of each subfile be limited by the subfile size S_l, and the constraint (17f) is equivalent to the per-eRRH power constraints (3) within the precoding model (13). We emphasize that in (17), the pre-fetching variables (5) and the fronthaul transfer variables (10) are fixed.

The solution of problem (17) is made difficult by the non-convexity in the constraint (17c). Here, noting that the left-hand side of (17c) has the DC structure when stated in terms of the covariance matrices $\mathbf{W}_{f,l} \triangleq \bar{\mathbf{V}}_{f,l} \bar{\mathbf{V}}_{f,l}^\dagger \succeq \mathbf{0}$, as in [7], [8], and [16], we adopt the concave-convex procedure (CCCP) for tackling (17). Specifically, we address problem (17) 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 $\text{rank}(\mathbf{W}_{f,l}) \le n_{S,f,l}$.

The resulting algorithm is described in Algorithm 1, where the function $\tilde{q}_{k,l}(\mathbf{W}^{(t+1)}, \mathbf{W}^{(t)})$ is defined as

$$\tilde{q}_{k,l}\left(\mathbf{W}^{(t+1)}, \mathbf{W}^{(t)}\right) \triangleq \log\det\left(\begin{array}{c}\sum_{m=l}^{L} \mathbf{H}_k \mathbf{W}^{(t+1)}_{f_k,m} \mathbf{H}_k^\dagger \\ + \sum_{f \in \mathcal{F}_{\text{req}} \setminus \{f_k\}} \sum_{m \in \mathcal{L}} \mathbf{H}_k \mathbf{W}^{(t+1)}_{f,m} \mathbf{H}_k^\dagger \\ + \Sigma_{\mathbf{z}_k}\end{array}\right) - \varphi\left(\begin{array}{c}\sum_{m=l+1}^{L} \mathbf{H}_k \mathbf{W}^{(t+1)}_{f_k,m} \mathbf{H}_k^\dagger \\ + \sum_{f \in \mathcal{F}_{\text{req}} \setminus \{f_k\}} \sum_{m \in \mathcal{L}} \mathbf{H}_k \mathbf{W}^{(t+1)}_{f,m} \mathbf{H}_k^\dagger \\ + \Sigma_{\mathbf{z}_k}, \\ \sum_{m=l+1}^{L} \mathbf{H}_k \mathbf{W}^{(t)}_{f_k,m} \mathbf{H}_k^\dagger \\ + \sum_{f \in \mathcal{F}_{\text{req}} \setminus \{f_k\}} \sum_{m \in \mathcal{L}} \mathbf{H}_k \mathbf{W}^{(t)}_{f,m} \mathbf{H}_k^\dagger \\ + \Sigma_{\mathbf{z}_k}\end{array}\right), \tag{18}$$

with the notation $\varphi(\mathbf{A}, \mathbf{B}) \triangleq \log\det(\mathbf{B}) + \text{tr}(\mathbf{B}^{-1}(\mathbf{A} - \mathbf{B}))$. After the convergence of the algorithm, each precoding matrix $\bar{\mathbf{V}}_{f,l}$ is obtained as $\bar{\mathbf{V}}_{f,l} \leftarrow \mathbf{V}_{n_{S,f,l}}(\mathbf{W}_{f,l}) \text{diag}(\mathbf{d}_{n_{S,f,l}}(\mathbf{W}_{f,l}))^{1/2}$, where $\mathbf{V}_N(\mathbf{A})$ takes the N leading eigenvectors of the matrix $\mathbf{A}$ as its columns, $\mathbf{d}_N(\mathbf{A})$ is a vector whose elements are given as the corresponding eigenvalues, and each precoding matrix $\mathbf{V}^i_{f,l}$ for eRRH i can be obtained as $\mathbf{V}^i_{f,l} \leftarrow (1 - \bar{c}^i_{f,l} \bar{d}^i_{f,l}) \mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}$.

Algorithm 1 CCCP Algorithm for Problem (17)

1. Initialize the matrices $\mathbf{W}^{(1)}$ to arbitrary positive semidefinite matrices that satisfy the per-eRRH power constraints (17f) and set $t = 1$.

2. Update the matrices $\mathbf{W}^{(t+1)}$ as a solution of the following convex problem:

$$\underset{\mathbf{W}^{(t+1)} \succeq \mathbf{0}, R_{\min}, \mathbf{R}}{\text{maximize}} \; R_{\min} \tag{19a}$$

$$\text{s.t. } R_{\min} \le \sum_{l \in \mathcal{L}} R_{f,l}, \quad f \in \mathcal{F}_{\text{req}}, \tag{19b}$$

$$R_{f_k,l} \le \tilde{q}_{k,l}\left(\mathbf{W}^{(t+1)}, \mathbf{W}^{(t)}\right), \quad l \in \mathcal{L}, \; k \in \mathcal{N}_U, \tag{19c}$$

$$\sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} d^i_{f,l} R^i_{f,l} \le C_i, \quad i \in \mathcal{N}_R, \tag{19d}$$

$$R_{f,l} \le S_l, \quad f \in \mathcal{F}_{\text{req}}, \; l \in \mathcal{L}, \tag{19e}$$

$$\sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} (1 - \bar{c}^i_{f,l} \bar{d}^i_{f,l}) \text{tr}\left(\mathbf{E}_i^\dagger \mathbf{W}^{(t+1)}_{f,l} \mathbf{E}_i\right) \le P_i, \quad i \in \mathcal{N}_R. \tag{19f}$$

3. Stop if a convergence criterion is satisfied. Otherwise, set $t \leftarrow t + 1$ and go back to Step 2.

We refer to [16] for a discussion of known results on the convergence of CCCP. We also note that, an alternative approach, not based on rank relaxation, would be to use successive convex approximation methods [28] based on lower bounds obtained from Fenchel duality (see, e.g., [46]).

V. Delivery Phase With Soft-Transfer Fronthauling

Unlike the hard-transfer mode that uses the fronthaul links to transfer hard information on missing files, in the soft-transfer mode typical of C-RAN, the fronthaul links are used to transfer a quantized version of the precoded signals of the missing files. Accordingly, the signal $\mathbf{x}_i$ transmitted by eRRH i on the downlink channel is given as the superposition of two signals, one that is locally encoded based on the content in the cache and another that is encoded at the BBU and quantized for transmission on the fronthaul link. This yields

$$\mathbf{x}_i = \sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} c^i_{f,l} \mathbf{V}^i_{f,l} \mathbf{s}_{f,l} + \hat{\mathbf{x}}_i, \tag{20}$$

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 $\mathbf{s}_{f,l}$ encoding the cached file (f, l), while $\hat{\mathbf{x}}_i$ represents the quantized baseband signal received from the BBU on the fronthaul link. Note that in a C-RAN, the transmitted signal would be given solely by the quantized signal $\hat{\mathbf{x}}_i$, which is discussed next.

The BBU precodes the subfiles that are not stored in each eRRH i producing the signal

$$\tilde{\mathbf{x}}_i = \sum_{f \in \mathcal{F}_{\text{req}}} \sum_{l \in \mathcal{L}} \bar{c}^i_{f,l} \mathbf{U}^i_{f,l} \mathbf{s}_{f,l}, \tag{21}$$

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 (21) is quantized, obtaining the signal $\hat{\mathbf{x}}_i$ as

$$\hat{\mathbf{x}}_i = \tilde{\mathbf{x}}_i + \mathbf{q}_i, \tag{22}$$

where $\mathbf{q}_i$ denotes the quantization noise independent of $\tilde{\mathbf{x}}_i$ and distributed as $\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$ with the 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 [6].[3] Using standard information theoretic results (see, e.g., [39, Ch. 3]), the signal $\hat{\mathbf{x}}_i$ can be reliably recovered by eRRH i if the condition

$$\begin{aligned} g_i\left(\mathbf{U}, \mathbf{\Omega}\right) &\triangleq I\left(\tilde{\mathbf{x}}_i; \hat{\mathbf{x}}_i\right) \\ &= \log\det\left(\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\mathbf{U}^i_{f,l}\mathbf{U}^{i\dagger}_{f,l} + \mathbf{\Omega}_i\right) \\ &\quad - \log\det\left(\mathbf{\Omega}_i\right) \leq C_i \end{aligned} \tag{23}$$

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

With (20), the signal $\mathbf{y}_k$ received by UE k in (2) can be written as

$$\begin{aligned} \mathbf{y}_k = &\sum_{l\in\mathcal{L}} \mathbf{H}_k\bar{\mathbf{V}}_{f_k,l}\mathbf{s}_{f_k,l} + \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{l\in\mathcal{L}} \mathbf{H}_k\bar{\mathbf{V}}_{f,l}\mathbf{s}_{f,l} \\ &+ \mathbf{H}_k\mathbf{q} + \mathbf{z}_k, \end{aligned} \tag{24}$$

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 $\bar{\mathbf{V}}^i_{f,l} \triangleq c^i_{f,l}\mathbf{V}^i_{f,l} + \bar{c}^i_{f,l}\mathbf{U}^i_{f,l}$ 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 $\bar{\mathbf{\Omega}} \triangleq \text{diag}(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_{N_R})$. Similar to the case with hard-transfer fronthauling, we assume that UE k performs SIC decoding based on (24) with the decoding order $\mathbf{s}_{f_k,1} \rightarrow \ldots \rightarrow \mathbf{s}_{f_k,L}$, so that the rate $R_{f_k,l}$ of the subfile (f_k, l) is bounded as

$$\begin{aligned} R_{f_k,l} &\leq q_{k,l}\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right) \\ &\triangleq I\left(\mathbf{s}_{f_k,l}; \mathbf{y}_k | \mathbf{s}_{f_k,1}, \ldots, \mathbf{s}_{f_k,l-1}\right) \\ &= \log\det\left(+\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{m=l}^{L}\sum_{m\in\mathcal{L}} \mathbf{H}_k\bar{\mathbf{V}}_{f_k,m}\bar{\mathbf{V}}^\dagger_{f_k,m}\mathbf{H}^\dagger_k + \mathbf{H}_k\bar{\mathbf{\Omega}}\mathbf{H}^\dagger_k + \mathbf{\Sigma}_{\mathbf{z}_k}\right) \\ &\quad - \log\det\left(+\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{m=l+1}^{L}\sum_{m\in\mathcal{L}} \mathbf{H}_k\bar{\mathbf{V}}_{f_k,m}\bar{\mathbf{V}}^\dagger_{f_k,m}\mathbf{H}^\dagger_k + \mathbf{H}_k\bar{\mathbf{\Omega}}\mathbf{H}^\dagger_k + \mathbf{\Sigma}_{\mathbf{z}_k}\right), \end{aligned} \tag{25}$$

where we defined the notation $\bar{\mathbf{V}} \triangleq \{\bar{\mathbf{V}}_{f,l}\}_{f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}}$.

A. Problem Definition and Optimization

As in Sec. IV-A, we aim at maximizing the minimum-user rate $R_{\min} \triangleq \min_{f\in\mathcal{F}_{\text{req}}} R_f$ subject to per-eRRH fronthaul capacity and transmit power constraints. The problem is stated as

$$\underset{\bar{\mathbf{V}}, \mathbf{\Omega}, R_{\min}, \mathbf{R}}{\text{maximize}} \; R_{\min} \tag{26a}$$

$$\text{s.t. } R_{\min} \leq \sum_{l\in\mathcal{L}} R_{f,l}, \quad f \in \mathcal{F}_{\text{req}}, \tag{26b}$$

$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right), \quad l \in \mathcal{L}, \; k \in \mathcal{N}_U, \tag{26c}$$

$$g_i\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right) \leq C_i, \quad i \in \mathcal{N}_R, \tag{26d}$$

$$R_{f,l} \leq S_l, \quad f \in \mathcal{F}_{\text{req}}, \; l \in \mathcal{L}, \tag{26e}$$

$$\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \text{tr}\left(\mathbf{E}^\dagger_i\bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}^\dagger_{f,l}\mathbf{E}_i\right) + \text{tr}\left(\mathbf{\Omega}_i\right) \leq P_i, \quad i \in \mathcal{N}_R, \tag{26f}$$

where the function $g_i(\bar{\mathbf{V}}, \mathbf{\Omega})$ is defined, with a small abuse of notation, from (23), as

$$\begin{aligned} g_i\left(\bar{\mathbf{V}}, \mathbf{\Omega}\right) \triangleq &\log\det\left(\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\mathbf{E}^\dagger_i\bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}^\dagger_{f,l}\mathbf{E}_i + \mathbf{\Omega}_i\right) \\ &- \log\det\left(\mathbf{\Omega}_i\right), \end{aligned} \tag{27}$$

given that, if $\bar{c}^i_{f,l} = 1$, then $\mathbf{E}^\dagger_i\bar{\mathbf{V}}_{f,l} = \mathbf{U}^i_{f,l}$.

As for problem (17), we tackle (26) by means of the CCCP approach as applied to a rank-relaxed version of (17), where the optimization variables are given as $\mathbf{W}_{f,l} \triangleq \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}^\dagger_{f,l}$ and the rank constraints $\text{rank}(\mathbf{W}_{f,l}) \leq n_{S,f,l}$ are relaxed. The resulting algorithm is detailed in Algorithm 2, where we

Algorithm 2 CCCP Algorithm for Problem (26)

1. Initialize the matrices $\mathbf{W}^{(1)}$ and $\mathbf{\Omega}^{(1)}$ to arbitrary positive semidefinite matrices that satisfy the per-eRRH fronthaul capacity constraints (26d) and power constraints (26f) and set $t = 1$.

2. Update the matrices $\mathbf{W}^{(t+1)}$ and $\mathbf{\Omega}^{(t+1)}$ as a solution of the following convex problem:

$$\underset{\mathbf{W}, \mathbf{\Omega}\succeq\mathbf{0}, R_{\min}, \mathbf{R}}{\text{maximize}} \; R_{\min} \tag{30a}$$

$$\text{s.t. } R_{\min} \leq \sum_{l\in\mathcal{L}} R_{f,l}, \quad f \in \mathcal{F}_{\text{req}}, \tag{30b}$$

$$R_{f_k,l} \leq \tilde{q}_{k,l}\left(\mathbf{W}, \mathbf{\Omega}, \mathbf{W}^{(t)}, \mathbf{\Omega}^{(t)}\right), \quad l \in \mathcal{L}, \quad k \in \mathcal{N}_U, \tag{30c}$$

$$\tilde{g}_i\left(\mathbf{W}, \mathbf{\Omega}, \mathbf{W}^{(t)}, \mathbf{\Omega}^{(t)}\right) \leq C_i, \quad i \in \mathcal{N}_R, \tag{30d}$$

$$R_{f,l} \leq S_l, \quad f \in \mathcal{F}_{\text{req}}, \; l \in \mathcal{L}, \tag{30e}$$

$$\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \text{tr}\left(\mathbf{E}^\dagger_i\mathbf{W}_{f,l}\mathbf{E}_i\right) + \text{tr}\left(\mathbf{\Omega}_i\right) \leq P_i, \quad i \in \mathcal{N}_R, \tag{30f}$$

3. Stop if a convergence criterion is satisfied. Otherwise, set $t \leftarrow t + 1$ and go back to Step 2.

[3]The multivariate compression method proposed in [7] allows the signals $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ for different eRRHs $i \neq j$ to be jointly quantized, hence obtaining correlated quantization noises. We do not further pursue the application of multivariate compression here, although its inclusion in the analysis could be carried out in a similar manner.

defined the functions

$$\tilde{q}_{k,l}\left(\mathbf{W}^{(t+1)},\boldsymbol{\Omega}^{(t+1)},\mathbf{W}^{(t)},\boldsymbol{\Omega}^{(t)}\right) \triangleq \log\det\left(\begin{array}{c}\sum_{m=l}^{L}\mathbf{H}_k\mathbf{W}_{f_k,m}^{(t+1)}\mathbf{H}_k^{\dagger}\\ +\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{m\in\mathcal{L}}\mathbf{H}_k\mathbf{W}_{f,m}^{(t+1)}\mathbf{H}_k^{\dagger}\\ +\mathbf{H}_k\bar{\boldsymbol{\Omega}}^{(t+1)}\mathbf{H}_k^{\dagger}+\Sigma_{\mathbf{z}_k}\end{array}\right) - \varphi\left(\begin{array}{c}\sum_{m=l+1}^{L}\mathbf{H}_k\mathbf{W}_{f_k,m}^{(t+1)}\mathbf{H}_k^{\dagger}\\ +\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{m\in\mathcal{L}}\mathbf{H}_k\mathbf{W}_{f,m}^{(t+1)}\mathbf{H}_k^{\dagger}\\ +\mathbf{H}_k\bar{\boldsymbol{\Omega}}^{(t+1)}\mathbf{H}_k^{\dagger}+\Sigma_{\mathbf{z}_k},\\ \sum_{m=l+1}^{L}\mathbf{H}_k\mathbf{W}_{f_k,m}^{(t)}\mathbf{H}_k^{\dagger}\\ +\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{m\in\mathcal{L}}\mathbf{H}_k\mathbf{W}_{f,m}^{(t)}\mathbf{H}_k^{\dagger}\\ +\mathbf{H}_k\bar{\boldsymbol{\Omega}}^{(t)}\mathbf{H}_k^{\dagger}+\Sigma_{\mathbf{z}_k}\end{array}\right), \quad (28)$$

and

$$\tilde{g}_i\left(\mathbf{W}^{(t+1)},\boldsymbol{\Omega}^{(t+1)},\mathbf{W}^{(t)},\boldsymbol{\Omega}^{(t)}\right) \triangleq \varphi\left(\begin{array}{c}\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\bar{c}_{f,l}^{i}\mathbf{E}_i^{\dagger}\mathbf{W}_{f,l}^{(t+1)}\mathbf{E}_i+\boldsymbol{\Omega}_i^{(t+1)},\\ \sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\bar{c}_{f,l}^{i}\mathbf{E}_i^{\dagger}\mathbf{W}_{f,l}^{(t)}\mathbf{E}_i+\boldsymbol{\Omega}_i^{(t)}\end{array}\right) - \log\det\left(\boldsymbol{\Omega}_i^{(t+1)}\right). \quad (29)$$

After the convergence of the algorithm, each precoding matrix $\bar{\mathbf{V}}_{f,l}$ is obtained as $\bar{\mathbf{V}}_{f,l} \leftarrow \mathbf{V}_{n_{S,f,l}}(\mathbf{W}_{f,l})\text{diag}\left(\mathbf{d}_{n_{S,f,l}}(\mathbf{W}_{f,l})\right)^{1/2}$ as in Sec. IV-A.

VI. Delivery Phase With Hybrid Fronthauling

In this section, we consider the design of a hybrid hard- and soft-transfer mode fronthauling scheme, whereby, unlike the strategies discussed in Sec. IV and Sec. V, the capacity of each fronthaul link is generally used to carry both hard and soft information about the uncached files. In this scheme, as a hybrid of (13) or (20), the signal $\mathbf{x}_i$ transmitted by eRRH i on the downlink channel is given as

$$\mathbf{x}_i = \sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\left(1-\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i}\right)\mathbf{V}_{f,l}^{i}\mathbf{s}_{f,l}+\hat{\mathbf{x}}_i, \quad (31)$$

where, as for (20), $\mathbf{V}_{f,l}^{i}\in\mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix applied by eRRH i on the baseband signal $\mathbf{s}_{f,l}$ encoding the subfile (f,l), and $\hat{\mathbf{x}}_i$ represents the quantized baseband signal received from the BBU on the fronthaul link. Similar to (20), the first term for subfile (f,l) is non-zero if the subfile (f,l) is available at the eRRH by caching or via hard-mode fronthauling, i.e., with $c_{f,l}^{i}=1$ or $d_{f,l}^{i}=1$, respectively. The fronthaul transfer variables $\{d_{f,l}^{i}\}_{f\in\mathcal{F}_{\text{req}},l\in\mathcal{L}}$ of the hybrid fronthauling strategy is set to those of the hard-transfer mode discussed in Sec. IV with N_F giving the best performance.

The BBU precodes the subfiles (f,l) that are not available at eRRH i, i.e., with $\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i}=1$, producing the signal

$$\tilde{\mathbf{x}}_i = \sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i}\mathbf{U}_{f,l}^{i}\mathbf{s}_{f,l}, \quad (32)$$

where $\mathbf{U}_{f,l}^{i}\in\mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix for the baseband signal $\mathbf{s}_{f,l}$. The quantized signal $\hat{\mathbf{x}}_i$ in the right-hand side of (31) is given as (22) which can be reliably recovered by eRRH i if the condition

$$g_i\left(\mathbf{U},\boldsymbol{\Omega}\right) \triangleq I\left(\tilde{\mathbf{x}}_i;\hat{\mathbf{x}}_i\right) = \log\det\left(\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i}\mathbf{U}_{f,l}^{i}\mathbf{U}_{f,l}^{i\dagger}+\boldsymbol{\Omega}_i\right) - \log\det\left(\boldsymbol{\Omega}_i\right)\le\tilde{C}_i \quad (33)$$

is satisfied, where we recall that $\boldsymbol{\Omega}_i$ denotes the covariance matrix of the quantization noise in (22), and we defined $\tilde{C}_i\le C_i$ as the rate used on the ith fronthaul for the soft-transfer mode. The rest of the frontahul link of $C_i-\tilde{C}_i$ bit/symbol can be used for the hard-transfer mode, i.e., for transferring the subfiles (f,l) with $d_{f,l}^{i}=1$. Accounting for both soft- and hard-transfer fronthauling, the fronthaul capacity constraint for each eRRH i is then stated as

$$\sum_{f\in\mathcal{F}}\sum_{l\in\mathcal{L}}d_{f,l}^{i}R_{f,l}+\tilde{C}_i\le C_i. \quad (34)$$

With (31), the signal $\mathbf{y}_k$ received by UE k in (2) can be written as (24), with the only difference that the aggregated precoding matrix $\bar{\mathbf{V}}_{f,l}\triangleq[\bar{\mathbf{V}}_{f,l}^{1};\ldots;\bar{\mathbf{V}}_{f,l}^{N_R}]$ for subfile (f,l) consists of the submatrices $\bar{\mathbf{V}}_{f,l}^{i}\triangleq(1-\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i})\mathbf{V}_{f,l}^{i}+\bar{c}_{f,l}^{i}\bar{d}_{f,l}^{i}\mathbf{U}_{f,l}^{i}$. Assuming the SIC decoding with the same decoding order, the rate $R_{f_k,l}$ of the subfile (f_k,l) is achievable if the condition (25) is satisfied.

A. Problem Definition and Optimization

We aim at optimizing the precoding matrices $\mathbf{V}$ and $\mathbf{U}$ applied at the eRRHs and the BBU, 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, as in Sec. IV-A and Sec. V-A, while satisfying the fronthaul capacity (34) and per-eRRH power constraints (3). The problem can be formulated as

$$\underset{\bar{\mathbf{V}},\boldsymbol{\Omega},R_{\min},\mathbf{R},\tilde{\mathbf{C}}}{\text{maximize}} \quad R_{\min} \quad (35a)$$

$$\text{s.t. } R_{\min}\le\sum_{l\in\mathcal{L}}R_{f,l},\quad f\in\mathcal{F}_{\text{req}}, \quad (35b)$$

$$R_{f_k,l}\le q_{k,l}\left(\bar{\mathbf{V}},\boldsymbol{\Omega}\right),\quad l\in\mathcal{L},\ k\in\mathcal{N}_U, \quad (35c)$$

$$g_i\left(\bar{\mathbf{V}},\boldsymbol{\Omega}\right)\le\tilde{C}_i,\quad i\in\mathcal{N}_R, \quad (35d)$$

$$\sum_{f\in\mathcal{F}}\sum_{l\in\mathcal{L}}d_{f,l}^{i}R_{f,l}+\tilde{C}_i\le C_i,\quad i\in\mathcal{N}_R, \quad (35e)$$

$$R_{f,l}\le S_l,\quad f\in\mathcal{F}_{\text{req}},\ l\in\mathcal{L}, \quad (35f)$$

$$\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}}\text{tr}\left(\mathbf{E}_i^{\dagger}\bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^{\dagger}\mathbf{E}_i\right)+\text{tr}\left(\boldsymbol{\Omega}_i\right)\le P_i,\quad i\in\mathcal{N}_R. \quad (35g)$$

As for problems (17) and (26), we can apply the CCCP approach to a rank-relaxed version of the problem (35), where the rank constraints $\text{rank}(\mathbf{W}_{f,l})\le n_{S,f,l}$ are removed. The procedure follows in the same manner as for Algorithms 1 and 2, and will not be detailed here.

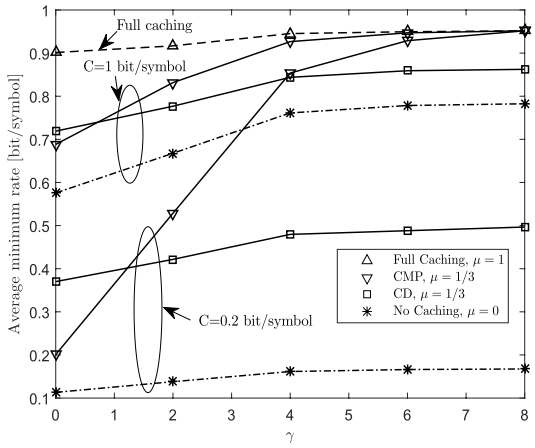

Fig. 3. Average minimum rate $R_{\min}$ versus the parameter γ of the Zipf's distribution in (1) for a F-RAN downlink under soft-transfer fronthauling mode ($\mu = 0, 1/3, 1$, $F = 3$, $S = 1$, $C = 0.2$ and 1 and $P/N_0 = 20$ dB).

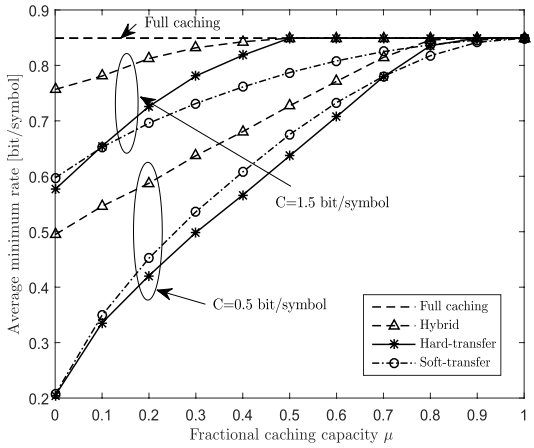

Fig. 4. Average minimum rate $R_{\min}$ versus the fractional caching capacity μ for an F-RAN downlink under FCD pre-fetching ($C = 0.5$ and 1.5, $\gamma = 0.5$, $F = 6$, $S = 1$ and $P/N_0 = 20$ dB).

VII. Numerical Results

In this section, we present some numerical results that compare the performance of hard-transfer and soft-transfer fronthauling modes, as well as of the hybrid scheme, with the pre-fetching strategies discussed in Sec. III. We consider an F-RAN system where the positions of eRRHs and UEs are uniformly distributed within a circular cell of radius 500m. The channel $\mathbf{H}_{k,i}$ from eRRH i to UE k is modeled as $\mathbf{H}_{k,i} = \sqrt{\rho_{k,i}}\tilde{\mathbf{H}}_{k,i}$, where the channel power $\rho_{k,i}$ is given as $\rho_{k,i} = 1/(1 + (d_{k,i}/d_0)^{\alpha})$ and the elements of $\tilde{\mathbf{H}}_{k,i}$ are independent and identically distributed (i.i.d.) as $\mathcal{CN}(0, 1)$. We set the parameters $d_0 = 50$m and $\alpha = 3$. We consider a symmetric setting where the covariance matrix $\Sigma_{\mathbf{z}_k}$ is given as $\Sigma_{\mathbf{z}_k} = N_0\mathbf{I}$ for all UEs $k \in \mathcal{N}_U$, and 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., $B_i = B$ and $\mu_i = \mu$ for $i \in \mathcal{N}_R$. If not stated otherwise, we set $N_R = N_U = 3$ and $n_{R,i} = n_{U,k} = 1$.

We first study the impact of the file popularity on the F-RAN performance. To this end, in Fig. 3, we plot the average minimum rate $R_{\min}$ versus the parameter γ of the Zipf's distribution in (1), 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 $F = 3$, $S = 1$, $C = 0.2$ and $C = 1$ and $P/N_0 = 20$ dB. 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). Note that full caching is equivalent to the MIMO broadcast part of the cut-set upper bound [47, Th. 14.10.1]. 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 approach, which suffers from a significant number of cache misses, particularly for low values of C. We also note that, when γ is sufficiently large, the performance of CMP approaches that of full caching scheme even with a small fronthaul capacity, due to the high probability that cooperative transmission across all eRRHs is possible based only on the cached contents.

In Fig. 4, we investigate the effect of the fractional caching capacity μ on the average minimum rate in two regimes of fronthaul capacity, namely low, here, $C = 0.5$ bit/symbol, and moderate, here, $C = 1.5$ bit/symbol. We adopt the FCD strategy and compare the performance of soft- and hard-transfer fronthauling modes with the hybrid mode proposed in Sec. VI. For the hard-transfer mode, we adopted the value of the parameter N_F corresponding to the largest rate $R_{\min}$ for any value of μ. We set $L = 10$ for the FCD pre-fetching strategy so that each file f is split into 10 disjoint subfiles, each of equal size $S_l = S/10$, and each cache stores up to $\lfloor 10\mu \rfloor$ randomly chosen fragments of each file. The performance was evaluated for the values of $\mu = 0, 0.1, 0.2, \ldots, 1.0$, and interpolated linearly since the curves can be made continuous by increasing the parameter L. Note that all schemes provide the same performance for $\mu = 1$, since every eRRH has access to the requested contents. The plot emphasizes the different relative behavior of the soft and hard fronthauling strategies in different fronthaul and caching set-ups. In particular, the soft-transfer fronthauling strategy is seen to offer potentially large gains for low fronthaul and intermediate caching capacities. This suggests that, if the eRRHs have moderate caching capabilities, soft-transfer fronthauling provides the best way to use low-capacity fronthaul links. Conversely, if the fronthaul capacity is large enough as compared to the minimum delivery rate, and if the caching capacity is sufficiently large, hard fronthauling can offer some performance gains over soft-mode fronthauling. Finally, the hybrid scheme is seen to outperform the soft- and hard-transfer modes, particularly at lower caching capacities.

We then further study the role of the fronthaul capacity by plotting in Fig. 5 the average minimum rate $R_{\min}$ versus the fronthaul capacity C for an F-RAN system with the

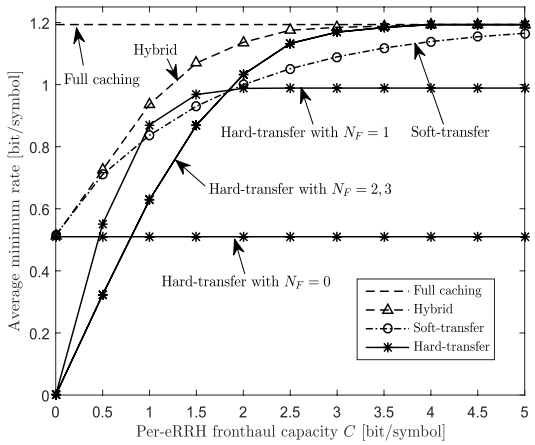

Fig. 5. Average minimum rate $R_{\min}$ versus the fronthaul capacity C for an F-RAN downlink under FCD pre-fetching ($\mu = 1/3$ and 1, $F = 6$, $S = 2$, $\gamma = 0.2$ and $P/N_0 = 20$ dB).

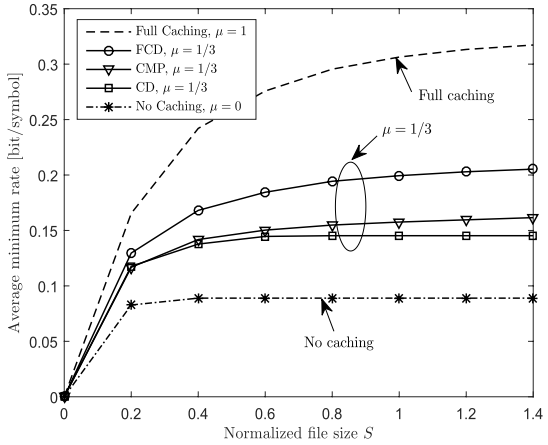

Fig. 6. Average minimum rate $R_{\min}$ versus the normalized file size S for an F-RAN downlink under soft-transfer mode fronthauling ($\mu = 0, 1/3$ and 1, $F = 6$, $C = 0.5$, $\gamma = 0.5$ and $P/N_0 = 10$ dB).

FCD pre-fetching, and with $\mu = 1/3$ and 1, $F = 6$, $S = 2$, $\gamma = 0.2$ and $P/N_0 = 20$ dB. 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, for small fronthaul capacity C, it is desirable to reduce the cluster size, and hence N_F, for hard-transfer fronthauling, since a larger cluster size requires the transfer of each subfile to more eRRHs on the fronthaul links of small capacity, which limits the rate of the subfile. The figure confirms the observation in Fig. 4 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, we note that the hybrid scheme 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.

We now examine the impact of the file size S on the optimal caching policy. In Fig. 6, we show the average minimum

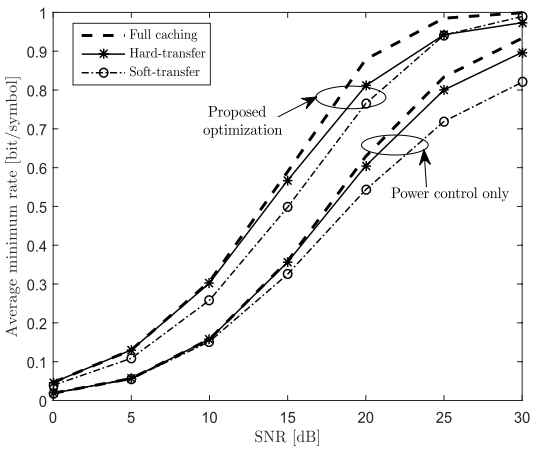

Fig. 7. Average minimum rate $R_{\min}$ versus the SNR P/N_0 for an F-RAN downlink under FCD pre-fetching ($\mu = 1/3$ and 1, $F = 6$, $C = 1.5$, $\gamma = 0.5$ and $S = 1$).

rate $R_{\min}$ versus the normalized file size S for an F-RAN downlink with soft-transfer mode fronthauling. We set the parameters $F = 6$, $C = 0.5$, $\gamma = 0.5$ and $P/N_0 = 10$ dB. The figure suggests that, for all pre-fetching strategies, the minimum rate $R_{\min}$ increases with a larger S in the regime of small file sizes, in which the performance is limited by the file size S rather than the fronthaul capacity C. Moreover, the performance gain of the FCD strategy compared to the CMP and CD is more pronounced for larger S, since the partitioning of a file into multiple fragments becomes more advantageous for the purpose of caching as the file size S increases.

Finally, Fig. 7 plots the average minimum rate $R_{\min}$ versus the SNR P/N_0 for an F-RAN downlink with the FCD pre-fetching and parameters set as $\mu = 1/3$ and 1, $F = 6$, $C = 1.5$, $\gamma = 0.5$ and $S = 1$. In order to quantify the impact of the proposed precoding optimization algorithms discussed in Sec. IV and V, we also plot the performance achieved under the assumption that no precoding, but only power control is carrid out at each eRRH. This amounts to setting the covariance matrix $\mathbf{W}_{f,l} = \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^{\dagger}$ of the baseband signal that encodes each subfile (f, l) as a diagonal matrix. As in Fig. 4, the performance of the hard-transfer mode was evaluated with the parameter N_F giving the largest rate $R_{\min}$. It can be seen from the figure that, in the intermediate SNR regimes, the proposed algorithms that optimize precoding provide more than 4 dB SNR gains, and rate gains of around 70% at SNR = 10 dB, for both the hard- and soft-transfer fronthauling modes as compared to the optimized power control schemes. We can also see that, with the proposed optimization, soft-transfer fronthauling improves over the hard-transfer scheme at sufficiently large SNRs similar to the case of low fronthaul capacity C.

VIII. Conclusion

In this work, we have studied joint design of cloud and edge processing for an F-RAN architecture in which each edge node is equipped not only with the functionalities of standard RRHs in C-RAN, but also with local cache and baseband processing capabilities. For any given pre-fetching strategy, we considered the optimization of the delivery phase with

the goal of maximizing the minimum delivery rate of the requested files while satisfying the fronthaul capacity and per-eRRH power constraints. We considered two basic fronthauling modes, namely hard- and soft-transfer fronthauling, as well as a hybrid mode. Specifically, with the hard-transfer mode, the fronthaul links are used to transmit the requested files that are not in the local caches, while the soft-transfer mode employs the fronthaul links following the C-RAN principle of transferring quantized baseband signals. We compared the performance of hard-, soft- and hybrid-transfer fronthauling modes with different baseline pre-fetching strategies.

It was concluded, by means of extensive numerical results, that soft-transfer provides a more effective way to use fronthaul resources than the hard-transfer mode in most operating regimes except for very low SNR regime and moderate fronthaul capacity. In such regimes, hard-transfer fronthauling with a carefully selected cluster size can provide minor gains. It is emphasized that these results hold under the assumptions of information-theoretically optimal point-to-point compression for communication on the fronthaul links. While it is known that point-to-point compression can be improved upon [8], the comparison between the two modes should be revisited in the presence of less effective compression or even only quantization (see also [42] for further discussion in the context of C-RAN). Moreover, the numerical results highlighted the trade-off between fronthaul and caching resources, whereby a smaller fronthaul capacity can be compensated for by a larger cache, particularly for more skewed popularity distributions.

Among open problems, we mention here the analysis in the presence of imperfect CSI and the design of a practical symbol-by-symbol, instead of block, fronthaul quantization algorithms [48].

REFERENCES

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

[2] A. Checko *et al.*, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surv. Tuts.*, vol. 17, no. 1, pp. 405–426, 1st.quart., 2015.

[3] 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.

[4] 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.

[5] E. AB, H. Technologies, N. Corporation, A. Lucent, and N. S. Networks, "Common public radio interface (CPRI): interface specification,"CPRI specification v5.0, Sep. 2011.[doubt]

[6] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," EURASIP J. Adv. Signal Process., vol. 2009, no. 3, pp. 1-10, Feb. 2009.

[7] 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.

[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] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. IEEE Inf. Theory Appl. Workshop*, San Diego, CA, USA, Feb. 2014, pp. 1–6.

[10] M. Peng, S. Yan, K. Zhang, and C. Wang. (Jun. 2015). "Fog computing based radio access networks: Issues and Challenges." [Online]. Available: http://arxiv.org/abs/1506.04233

[11] S. Bi, R. Zhang, Z. Ding, and S. Cui. (Aug. 2016). "Wireless communications in the era of big data." [Online]. Available: http://arxiv.org/abs/1508.06369

[12] H. Jinri and Y. Yannan, "Next generation fronthaul interface," China Mobile, Beijing, China, White Paper, Version 1.0, Oct. 2015.

[13] 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.

[14] C. Kuilin and D. Ran, "C-RAN: The road towards green RAN," China Mobile, Beijing, China, White Paper, Version 2.5, Oct. 2011.

[15] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief. (Sep. 2015). "Backhaul-aware caching placement for wireless networks." [Online]. Available: https://arxiv.org/abs/1509.00558

[16] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, to be published.

[17] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered co-operative beamforming with application to edge caching," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 84–87, Feb. 2016.

[18] R. Tandon and O. Simeone, "Fog radio access networks: Fundamental latency trade-offs," in *Proc. IEEE Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Jan. 2016, pp. 1–5.

[19] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," in *Proc. IEEE Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Jan. 2016, pp. 1–5.

[20] M. Chiang. (Jan. 2016). "Fog networking: An overview on research opportunities." [Online]. Available: http://arxiv.org/abs/1601.00835

[21] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 2015, pp. 809–813.

[22] A. Sengupta, R. Tandon, and O. Simeone. (Dec. 2015). "Cache aided wireless networks: Tradeoffs between storage and latency." [Online]. Available: http://arxiv.org/abs/1512.07856

[23] L. Liu and W. Yu. (Jun. 2016). "Cross-layer design for downlink multi-hop cloud radio access networks with network coding." [Online]. Available: http://arxiv.org/abs/1606.08950

[24] L. Liu, P. Patil, and W. Yu. (Jun. 2016). "An uplink-downlink duality for cloud radio access network." [Online]. Available: http://arxiv.org/abs/1606.09131

[25] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[26] 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.

[27] A. Alameer and A. Sezgin, (Jul. 2016). "Green cloud radio access networks." [Online]. Available: https://arxiv.org/abs/1607.06231

[28] G. Scutari, F. Facchinei, L. Lampariello, and P. Song. (Oct. 2014). " Parallel and distributed methods for nonconvex optimization-part I: Theory." [Online]. Available: http://arxiv.org/abs/1410.4754

[29] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr. (Feb. 2016). "Fundamental limits of cache-aided interference management." [Online]. Available: https://arxiv.org/abs/1602.04207

[30] J. Hachem, U. Niesen, and S. Diggavi. (May 2016). "A layered caching architecture for the interference channel." [Online]. Available: https://arxiv.org/abs/1605.01668

[31] F. Xu, K. Liu, and M. Tao, "Cooperative Tx/Rx caching in interference channels: A storage-latency tradeoff study," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 2034–2038.

[32] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 2029–2033.

[33] A. Sengupta, R. Tandon, and O. Simeone. (May 2016). "Cloud and cache-aided wireless networks: Fundamental latency trade-offs." [Online]. Available: https://arxiv.org/abs/1605.01690

[34] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.

[35] V. Bioglio, F. Gabry, and I. Land. (Aug. 2015). "Optimizing MDS codes for caching at the edge." [Online]. Available: http://arxiv.org/abs/1508.05753

[36] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris. (Jan. 2016). "Cooperative caching and transmission design in cluster-centric small cell networks." [Online]. Available: https://arxiv.org/abs/1601.00321

[37] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint cloud and edge processing for latency minimization in fog radio access networks," in *Proc. IEEE Int. Workshop Signal Adv. Wireless Comm. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–5.
[38] D. Chen, S. Schedler, and V. Kuehn, "Backhaul traffic balancing and dynamic content-centric clustering for the downlink of fog radio access network," in *Proc. 17th IEEE Int. Workshop Signal Adv. Wireless Comm. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–5.
[39] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[40] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen. (Nov. 2013). "Online coded caching." [Online]. Available: https://arxiv.org/abs/1311.3646
[41] A. M. Fouladgar, O. Simeone, S.-H. Park, O. Sahin, and S. Shamai (Shitz), "Signal and interference alignment via message passing for MIMO interference channels," *Trans. Emerg. Telecomm. Technol.*, vol. 27, no. 3, pp. 392–407, Mar. 2016.
[42] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Fronthaul compression and precoding design for C-Rans over ergodic fading channel," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5022–5032, Jul. 2016.
[43] B. Dai and W. Yu. (Jan. 2016). "Energy efficiency of downlink transmission strategies for cloud radio access networks." [Online]. Available: https://arxiv.org/abs/1601.01070
[44] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo. (Jul. 2013). "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks." [Online]. Available: http://arxiv.org/abs/1307.4457
[45] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.
[46] J. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Oprimitzation Theory and Examples*. New York, NY, USA: Springer Verlag, 2006.
[47] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications), 1st ed. New York, NY, USA: Wiley, 1991.
[48] W. Lee, O. Simeone, J. Kang, and S. Shamai (Shitz), "Multivariate fronthaul quantization for downlink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5025–5037, Oct. 2016.

Seok-Hwan Park (M'11) received the B.Sc. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. He was with the Agency for Defense Development, Daejeon, South Korea, the New Jersey Institute of Technology, Newark, NJ, USA, and Samsung Electronics, Suwon, South Korea. Since 2015, he has been with Chonbuk National University, Jeonju, South Korea, as an Assistant Professor. His research interests include mathematical analysis and optimization of signal processing algorithms for physical-layer wireless communication systems.

Dr. Park received the Best Paper Award at the 2006 Asia-Pacific Conference on Communications and an Excellent Paper Award at the IEEE Student Paper Contest in 2006.

Osvaldo Simeone (F'16) received the M.Sc. degree (Hons.) and the Ph.D. degree in information engineering from the Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Information Processing, New Jersey Institute of Technology, Newark, where he is a Professor. His current research interests concern wireless communications, information theory, and machine learning.

Dr. Simeone is a co-recipient of the 2015 IEEE Communication Society Best Tutorial Paper Award and the best paper awards of the IEEE SPAWC 2007 and the IEEE WRECOM 2007. He currently serves as an Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Shlomo Shamai (Shitz) (F'94) received the B.Sc., M.Sc., and Ph.D. degrees from the Technion–Israel Institute of Technology, in 1975, 1981, and 1986 respectively, all in electrical engineering.

From 1975 to 1985, he was with the Communications Research Laboratory as a Senior Research Engineer. Since 1986, he has been with the Department of Electrical Engineering, Technion–Israel Institute of Technology, where he is currently a Technion Distinguished Professor, and holds the William Fondiller Chair of Telecommunications. His research interests include a wide spectrum of topics in information theory and statistical communications.

He is a member of the Israeli Academy of Sciences and Humanities and a Foreign Member of the U.S. National Academy of Engineering. He received the 2011 Claude E. Shannon Award and the 2014 Rothschild Prize in mathematics/computer sciences and engineering.

Dr. Shamai received the 1999 van der Pol Gold Medal of the Union Radio Scientifique Internationale, the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003 and the 2004 Joint IT/COM Societies Paper Award, the 2007 IEEE Information Theory Society Paper Award, the 2009 and 2015 European Commission FP7, Network of Excellence in Wireless COMmunications (NEWCOM++, NEWCOM#) Best Paper Awards, the 2010 Thomson Reuters Award for International Excellence in Scientific Research, the 2014 EURASIP Best Paper Award for the *EURASIP Journal on Wireless Communications and Networking*, and the 2015 IEEE Communications Society Best Tutorial Paper Award. He received 1985 Alon Grant for Distinguished Young Scientists and the 2000 Technion Henry Taub Prize for Excellence in Research. He has served as an Associate Editor of the Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY, and served on the Board of Governors of the Information Theory Society. He has served on the Executive Editorial Board of the IEEE TRANSACTIONS ON INFORMATION THEORY.