Article

Energy-Efficient Joint Design of Fronthaul and Edge Links for Cache-Aided C-RAN Systems with Wireless Fronthaul

Junbeom Kim [1], Daesung Yu [1], Seung-Eun Hong [2] and Seok-Hwan Park [1],*

[1] Division of Electronic Engineering, Chonbuk National University, Jeonju 54896, Korea
[2] Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Korea
* Correspondence: seokhwan@jbnu.ac.kr; Tel.: +82-63-270-2357

check for
updates

Abstract: This work addresses the joint design of fronthaul and edge links for a cache-aided cloud radio access network (C-RAN) system with a wireless fronthaul link. Motivated by the fact that existing techniques, such as C-RAN and edge caching, come at the cost of increased energy consumption, an energy efficiency (EE) metric is defined and adopted as the performance metric for optimization. As the fronthaul links can be used to transfer quantized and precoded baseband signals or hard information of uncached files, both soft- and hard-transfer fronthauling strategies are considered. Extensive numerical results validate the impact of edge caching, as well as the advantages of the energy-efficient design over the spectrally-efficient scheme. Additionally, the two fronthauling strategies—the soft- and hard-transfer schemes—are compared in terms of EE.

1. Introduction

It has been envisioned that the cloud radio access network (C-RAN) architecture will be able to meet the ever-increasing traffic demands of future wireless communication systems, by migrating baseband signal processing functionalities from base stations, or remote radio heads (RRHs), to a cloud processor (CP) or baseband processing unit (BBU) [1,2].In particular, improved spectral efficiency is expected to be achieved with the C-RAN architecture, thanks to centralized baseband processing at CPs. One of major challenges of implementing C-RAN systems is the overhead on the fronthaul links connecting CPs and RRHs [3–10]. The overhead becomes particularly serious when one adopts cost-effective wireless fronthaul links (see, e.g., [11–19]).

The overhead, or capacity requirements, of the fronthaul links in C-RAN systems can be alleviated by adopting narrowband IoT (NB-IoT) technology, which supports low-power and low-cost devices [20], or by adding edge caching functionalities to RRHs [21–28]. The idea of the latter is that the RRHs are equipped with local caches at which popular content frequently requested by mobile user equipment (UE) are pre-fetched, such that the amount of data delivered over the fronthaul links can be reduced at the delivery phase. As in [27], we refer to RRHs equipped with caches as edge nodes (ENs). In [27], an information theoretic analysis of cache-aided C-RAN systems was addressed. In [21–26,28] cache-aided C-RAN systems were studied in the aspects of signal processing. Specifically, signal processing optimization for the delivery phase for fixed pre-fetching strategies was discussed in [23–26,28]. We also mention that, as implementing caching at the edge of the network may face drawbacks, due to limited cache sizes and user mobility from one cell to another, a co-operative hierarchical caching approach has been studied in [29]. In this work, we focus on the design of the

delivery phase for a fixed pre-fetching strategy and leave the joint design of co-operative caching and delivery strategies as an interesting future research direction.

All of the techniques discussed above, such as C-RAN and edge caching, come at the cost of increased energy consumption [30,31], which causes increases both in the resulting carbon footprint and system costs. In order to address this issue, it is important to design systems by adopting energy efficiency (EE), instead of spectral efficiency, as the performance metric (as in, e.g., [28,30–34]). The energy-efficient design of C-RAN systems with wired fronthaul links has been addressed in [31,34], and [28] extended such design to the case with edge caching and imperfect channel state information (CSI). The EE metric has also been adopted as the evaluation criterion in [35–37] for the design of wireless powered mobile computing systems, reliable routing for wireless ad hoc networks, and MAC strategies for data dissemination. We also note that the C-RAN technology has been recently used for cloud-aided cognitive ambient back-scatter wireless sensor networks, with the aim of providing energy-efficient communications [38].

In this work, we propose an energy-efficient design of a cache-aided C-RAN system equipped with wireless (instead of wired) fronthaul links. We address the design under two different fronthauling strategies: Soft-transfer and hard-transfer schemes (see, e.g., [26,27]). As illustrated in Figure 1, these two fronthauling strategies differ in the type of information delivered over the fronthaul links. Under the soft-transfer (also referred to as compression-based) fronthauling strategy, as in the traditional C-RAN functional split (see, e.g., [8,13,15,26,34]), the CP performs precoding of the files that are requested by UEs but which are uncached by ENs, and quantizes and compresses the precoded baseband signals. Then, the CP sends the bit-streams describing the quantized results to the ENs over the fronthaul link. The information delivered over the fronthaul link is called soft information, as it is a processed version of the uncached files (see, e.g., [26,27]). In contrast, the hard-transfer (or data sharing-based) fronthauling strategy (see, e.g., [8,17,25,26,34]),the fronthaul link delivers the uncached files in raw form, referred to as hard information, to the requesting ENs, where the files are locally processed jointly with the cached files. Signal processing optimization under these two fronthauling strategies has been tackled in [26] for a cache-aided C-RAN system with wired fronthaul links. Furthermore, ref. [27] compared the two fronthauling strategies, in terms of the tradeoff between the cache size at the ENs and the delivery latency.
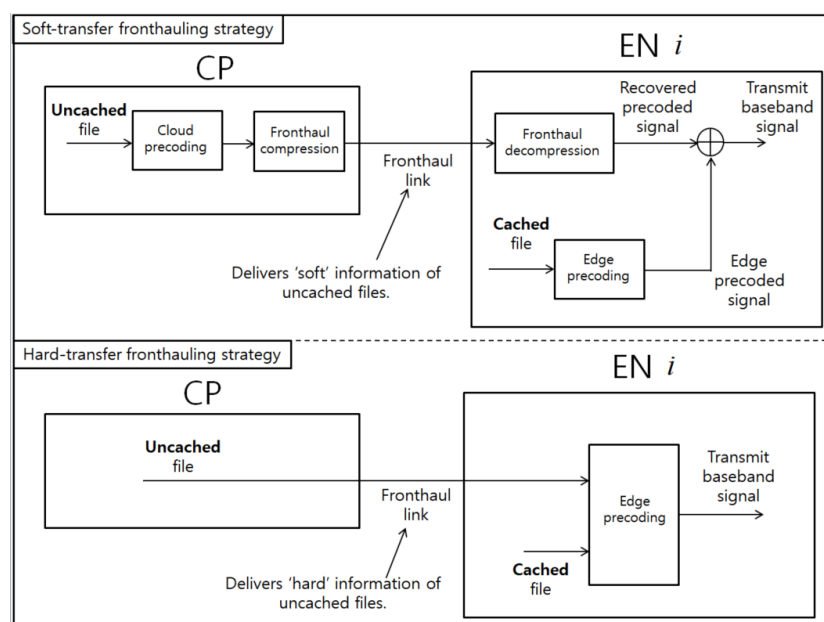

Figure 1. Illustration of soft-transfer and hard-transfer fronthauling strategies.

We tackle the joint design of CP-to-EN fronthaul and EN-to-UE edge links with the goal of maximizing the EE performance while satisfying the CP and per-EN transmission power constraints under the discussed fronthauling strategies. Due to the non-convexity of the formulated problems, we convert the problems into difference-of-convex (DC) problems by change of variables and rank relaxation. To tackle each of the obtained DC problems, we propose a concave-convex procedure (CCCP)-based iterative algorithm which converges to a suboptimal solution. Through analysis of the extensive numerical results, we show that the proposed energy-efficient design provides an EE gain over spectrally-efficient design which increases with the system signal-to-noise ratios (SNRs) and the number of ENs. Furthermore, we show that the hard-transfer fronthauling scheme outperforms the soft-transfer scheme in overall simulated set-ups, as the former is more effective in utilizing the wireless fronthaul resources for multicasting to multiple ENs, which may request overlapping uncached files.

Main contribution of this work is as follows: We tackle, for the first time, the energy-efficient joint design of the fronthaul and edge links for cache-aided C-RAN systems with wireless fronthaul links. Although there are some related works that have studied the impact of edge caching, wireless fronthaul, energy-efficient design and the comparison between the soft- and hard-transfer fronthauling strategies, these factors have not been considered jointly in the literature, as summarized in Table 1. By tackling the challenging optimization problems and through extensive numerical results, we observe the advantages of the energy-efficient design, hard-transfer fronthauling scheme, and the joint design of fronthaul and edge links, in terms of EE performance, in the considered system. Specifically, we show that the hard-transfer fronthauling scheme has the potential for better leveraging the overlapping nature of the uncached files that need to be delivered to the ENs over wireless fronthaul links.

Table 1. Summary of the main differences between this and related studies on C-RAN systems.

Features		References
Edge caching	No edge caching	[4–18,31,34,39]
	With edge caching	[21–28,40], this work
Fronthaul links	Wired fronthaul	[4–10,21,23–28,31,34,39,40]
	Wireless fronthaul	[11–18], this work
Design goal	Max. spectral efficiency	[4–10,13–18,26,39]
	Min. delivery latency	[23,27,40]
	Min. power or network cost	[11,12,24,25,28]
	Max. energy efficiency	[21,31,34], this work
Fronthauling strategy	Soft-transfer	[4,5,7,9,10,12,13,15,39]
	Hard-transfer	[11,14,16–18,21,24,25,40]
	Both	[6,8,23,26–28,31,34], this work

The paper is organized as follows: In Section 2, we describe the system model for a cache-aided C-RAN system with a wireless fronthaul link and define the EE metric. In Section 3, we describe the operations of the CP, ENs, and UEs under the soft-transfer fronthauling strategy; then, the joint optimization of the fronthaul and edge links is tackled. The operations and optimization under the hard-transfer fronthauling scheme are discussed in Section 4. We provide, in Section 5, numerical results validating the advantages of the proposed energy-efficient designs for the soft- and hard-transfer schemes. The paper is concluded in Section 6.

We summarize some notation used throughout the paper, as follows: We denote by $I(X;Y)$ the mutual information between two random variables X and Y. A circularly symmetric complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We define $\mathbb{C}^{M \times N}$ as the set of all $M \times N$ complex matrices, and $\mathrm{E}(\cdot)$ stands for expectation. The Hermitian transpose,

determinant, and trace operations are denoted by $(\cdot)^H$, $\det(\cdot)$, and $\mathrm{tr}(\cdot)$, respectively, and $\mathrm{rank}(\cdot)$ denotes the rank of the input matrix. We denote an $M \times M$ identity matrix and an $M \times N$ zero matrix, whose elements are filled with 0, as $\mathbf{I}_M$ and $\mathbf{0}_{M \times N}$, respectively. We denote the OR operation by $(\cdot | \cdot)$ for binary input variables.

2. System Model

We consider the downlink of a cache-aided C-RAN system where, as shown in Figure 2, there is a CP that communicates with N_U UEs through N_E ENs. We denote the numbers of antennas of CP, EN i, and UE k by n_C, $n_{E,i}$, and $n_{U,k}$, respectively. Here, we assume that each EN i uses the same number $n_{E,i}$ of antennas for both reception and transmission on the fronthaul and edge links. However, this discussion can be easily generalized to asymmetric cases. We define $n_E = \sum_{i \in \mathcal{N}_E} n_{E,i}$ and $n_U = \sum_{k \in \mathcal{N}_U} n_{U,k}$, where $\mathcal{N}_E = \{1, 2, \ldots, N_E\}$ and $\mathcal{N}_U = \{1, 2, \ldots, N_U\}$ denote the sets of indices of the ENs and UEs, respectively. Each EN i is equipped with a local cache of size B_i bits, to which popular contents can be pre-fetched to reduce the fronthaul overhead.
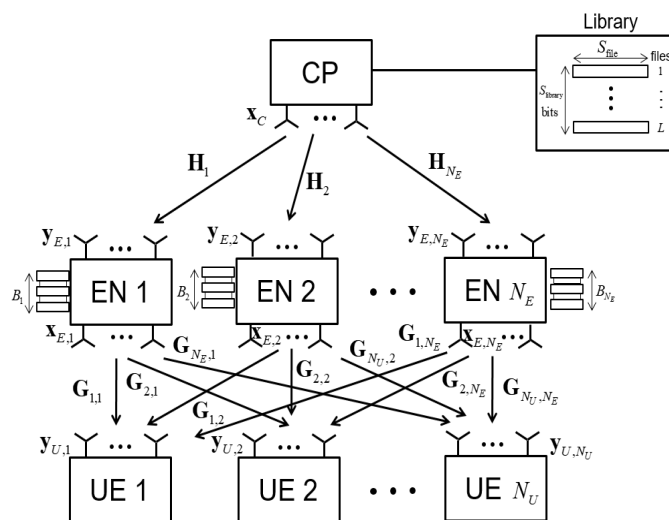

Figure 2. Cache-aided C-RAN system with wireless fronthaul link.

2.1. Content-Based Communication and Multicast Groups

As in [26,39], we consider a content-based multicast scenario. Accordingly, each UE k independently requests a file f_{l_k} from a library $\mathcal{F} = \{f_1, f_2, \ldots f_L\}$, where each file is of S_{file} bits and $l_k \in \mathcal{L} \triangleq \{1, 2, \ldots, L\}$ denotes the index of the file requested by UE k. Therefore, the library $\mathcal{F}$ has $S_{\mathrm{library}} = LS_{\mathrm{file}}$ bits. We assume that the CP can access the library $\mathcal{F}$ with a negligible delay. As the UE requests can overlap each other, we define multicast groups. We first define the set of indices of the requested files as $\mathcal{L}_{\mathrm{req}} = \cup_{k \in \mathcal{N}_U}\{l_k\}$, and $L_{\mathrm{req}} = |\mathcal{L}_{\mathrm{req}}|$ distinct indices in $\mathcal{L}_{\mathrm{req}}$ as $\tilde{l}_1, \tilde{l}_2, \ldots \tilde{l}_{L_{\mathrm{req}}}$. Defining the gth multicast group as the set of UEs that request file $f_{\tilde{l}_g}$ (i.e., $\mathcal{N}_{U,g} = \{k \in \mathcal{N}_U | l_k = \tilde{l}_g\}$), the system has L_{req} multicast groups. We denote by $\mathcal{G}_U = \{1, 2, \ldots, L_{\mathrm{req}}\}$ the set of the indices of the multicast groups. We assume that the L_{req} multicast groups have equal priority and, hence, they are communicated from the CP to the requesting UEs at the same data rate of R bits per symbol.

2.2. Edge Caching

During the off-peak traffic period, each EN i can pre-fetch some popular contents which are frequently requested by UEs to its local cache (of size B_i bits). As the ENs typically have lighter

structures than the CP, it is reasonable to assume that $B_i \leq S_{\text{library}}$. We model the pre-fetching strategy by defining a binary caching variable

$$c_{i,l} = \begin{cases} 1, & \text{if file } f_l \text{ is cached by EN } i \\ 0, & \text{otherwise} \end{cases}.$$ (1)

We note that the caching variables $\mathbf{c} = \{c_{i,l}\}_{i \in \mathcal{N}_E, l \in \mathcal{L}}$ cannot be determined in adaptation to instantaneous file requests or CSI, as the pre-fetching takes place during off-peak traffic periods. In this work, we assume that the caching variables $\mathbf{c}$ are arbitrarily pre-fixed, and the joint design of co-operative pre-fetching and delivery strategies is left as a future work.

2.3. Wireless Channel Models of Fronthaul and Edge Links

We assume that the CP-to-EN fronthaul link is orthogonal to the EN-to-UE edge link. Therefore, the two links do not interfere with each other. Also, we consider flat fading channel models for both the CP-to-EN fronthaul and EN-to-UE edge links. Then, the baseband received signal $\mathbf{y}_{E,i} \in \mathbb{C}^{n_{E,i} \times 1}$ of EN i on the wireless fronthaul link is given as

$$\mathbf{y}_{E,i} = \mathbf{H}_i \mathbf{x}_C + \mathbf{z}_{E,i},$$ (2)

where $\mathbf{x}_C \in \mathbb{C}^{n_C \times 1}$ denotes the transmitted signal of the CP, which is subject to the power constraint $\mathrm{E}||\mathbf{x}_C||^2 \leq P_C$, where $\mathbf{H}_i \in \mathbb{C}^{n_{E,i} \times n_C}$ represents the channel matrix from the CP to EN i and $\mathbf{z}_{E,i} \sim \mathcal{CN}(\mathbf{0}, \sigma_E^2 \mathbf{I}_{n_{E,i}})$ denotes the additive white Gaussian noise (AWGN) vector at EN i.

Each EN i processes the received signal $\mathbf{y}_{E,i}$, generating a baseband signal $\mathbf{x}_{E,i} \in \mathbb{C}^{n_{E,i} \times 1}$ which is transmitted over EN-to-UE edge links. We impose a transmission power constraint on $\mathbf{x}_{E,i}$ as $\mathrm{E}||\mathbf{x}_{E,i}||^2 \leq P_{E,i}$. The received signal $\mathbf{y}_{U,k} \in \mathbb{C}^{n_{U,k} \times 1}$ of UE k on the edge link can be written as

$$\mathbf{y}_{U,k} = \sum_{i \in \mathcal{N}_E} \mathbf{G}_{k,i} \mathbf{x}_{E,i} + \mathbf{z}_{U,k} = \mathbf{G}_k \mathbf{x}_E + \mathbf{z}_{U,k},$$ (3)

where $\mathbf{G}_{k,i} \in \mathbb{C}^{n_{U,k} \times n_{E,i}}$ represents the channel matrix from EN i to UE k, $\mathbf{z}_{U,k} \sim \mathcal{CN}(\mathbf{0}, \sigma_U^2 \mathbf{I}_{n_{U,k}})$ denotes the AWGN vector at UE k, $\mathbf{G}_k = [\mathbf{G}_{k,1} \mathbf{G}_{k,2} \cdots \mathbf{G}_{k,N_E}]$ represents the channel matrix from all ENs to UE k, and $\mathbf{x}_E = [\mathbf{x}_{E,1}^H \mathbf{x}_{E,2}^H \cdots \mathbf{x}_{E,N_E}^H]^H$ is transmitted signal of all ENs.

In this work, we assume that the CP has perfect CSI $\{\mathbf{H}_i\}_{i \in \mathcal{N}_E}$ and $\{\mathbf{G}_{k,i}\}_{k \in \mathcal{N}_U, i \in \mathcal{N}_E}$ and manages the operation of all nodes (i.e., the CP, ENs, and UEs). The analysis of the impact of imperfect CSI and robust design, taking into account the CSI error, are left as future work.

2.4. Energy Efficiency Metric

As illustrated in Section 1, our goal is to address the energy-efficient design for the described cache-aided C-RAN system with wireless fronthaul link. The EE metric, denoted by Θ, is defined as the data rate R that can be supported per unit power [31–34]. Mathematically, we write Θ as

$$\Theta = \frac{R}{\mathrm{E}\,||\mathbf{x}_C||^2 + \eta_C + \sum_{i \in \mathcal{N}_E}\left(\mathrm{E}\,||\mathbf{x}_{E,i}||^2 + \eta_{E,i}\right)},$$ (4)

where η_C and $\eta_{E,i}$ represent the constant circuit powers consumed at the CP and EN i, respectively; regardless of the radio-frequency (RF) transmission. As the constraints on the achievable rate R are formulated differently, depending on the fronthaul transmission mode, we will clarify the constraints under the soft-transfer and hard-transfer fronthauling modes in Sections 3 and 4, respectively.

3. Energy-Efficient Design Under Soft-Transfer Fronthaul Mode

In this section, we discuss the energy-efficient design of the cache-aided C-RAN system described in Section 2 under the soft-transfer fronthauling scheme, whereby the fronthaul links carry quantized and precoded versions of the uncached files. The detailed operation is described in Sections 3.1–3.4, and the optimization will be discussed in Section 3.5.

3.1. Fronthaul SDMA Precoding

The CP needs to first establish communication links to the ENs to send soft information of uncached files (which will be detailed in the next subsection). For communication to ENs, time division multiple access (TDMA) was considered in [14,16], while the works [12,13,15,17] assumed space time division multiple access (SDMA) fronthaul beamforming (or precoding) techniques. In this work, we assume the latter, as it is more efficient when the CP uses sufficiently many antennas.

To elaborate on fronthaul SDMA precoding, we denote by $\mathbf{s}_{E,i} \in \mathbb{C}^{d_{E,i} \times 1}$ the data signal which encodes the soft information to be delivered to EN i. The number $d_{E,i}$ of data streams should satisfy the condition $d_{E,i} \leq \mathrm{rank}(n_C, n_{E,i})$, and we assume a Gaussian channel codebook; that is, $\mathbf{s}_{E,i} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_{E,i}})$. With SDMA fronthaul precoding, the transmitted signal of the CP is given as

$$\mathbf{x}_C = \sum_{i \in \mathcal{N}_E} \mathbf{F}_i \mathbf{s}_{E,i}, \tag{5}$$

where $\mathbf{F}_i \in \mathbb{C}^{n_C \times d_{E,i}}$ represents the precoding matrix for $\mathbf{s}_{E,i}$. With (5), the CP power constraint can be rewritten as

$$\sum_{i \in \mathcal{N}_E} \mathrm{tr}\left(\mathbf{F}_i \mathbf{F}_i^H\right) \leq P_C. \tag{6}$$

If we assume that EN i decodes the signal $\mathbf{s}_{E,i}$ based on the received signal $\mathbf{y}_{E,i}$ without interference decoding, the fronthaul rate, C_i, at which the CP can communicate with EN i in bits per symbol, is constrained as

$$C_i \leq f_{E,i}(\mathbf{F}) \triangleq I(\mathbf{s}_{E,i}; \mathbf{y}_{E,i}) \tag{7}$$

$$= \log_2 \det\left(\sum_{j \in \mathcal{N}_E} \mathbf{H}_i \mathbf{F}_j \mathbf{F}_j^H \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}\right) - \log_2 \det\left(\sum_{j \in \mathcal{N}_E \setminus \{i\}} \mathbf{H}_i \mathbf{F}_j \mathbf{F}_j^H \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}\right),$$

with the notation $\mathbf{F} \triangleq \{\mathbf{F}_i\}_{i \in \mathcal{N}_E}$.

3.2. Cloud Precoding and Fronthaul Compression

In order to improve the performance of the multicast transmission on the EN-to-UE edge links, the CP performs cloud precoding of uncached files for each EN i by

$$\tilde{\mathbf{x}}_{E,i} = \sum_{g \in \mathcal{G}_U} \bar{c}_{i,\tilde{l}_g} \mathbf{U}_{i,g} \mathbf{s}_{U,g}, \tag{8}$$

where $\mathbf{U}_{i,g} \in \mathbb{C}^{n_{E,i} \times d_{U,g}}$ is the precoding matrix applied to the signal $\mathbf{s}_{U,g} \in \mathbb{C}^{d_{U,g} \times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_{U,g}})$ which encodes the file $f_{\tilde{l}_g}$ not available at EN i with $c_{i,\tilde{l}_g} = 0$. For a binary variable c, we define $\bar{c} = 1 - c$. The number $d_{U,g}$ of data streams for the gth multicast group is set to satisfy the condition $d_{U,g} = \min\{n_E, \min_{k \in \mathcal{N}_{U,g}} n_{U,k}\}$.

To send the precoded signal $\tilde{\mathbf{x}}_{E,i}$ to EN i over the fronthaul link of capacity C_i given in (7), the CP performs fronthaul quantization and compression on $\tilde{\mathbf{x}}_{E,i}$. As in the related works [4–10] on fronthaul compression, we model the output signal of the compression as

$$\hat{\mathbf{x}}_{E,i} = \tilde{\mathbf{x}}_{E,i} + \mathbf{q}_{E,i}, \tag{9}$$

where $\mathbf{q}_{E,i}$ represents the distortion signal caused by the quantization. Following the rate-distortion theoretic approaches considered in the related works [4–10,13,15], we assume a Gaussian quantization codebook under which the quantization noise signal $\mathbf{q}_{E,i}$ is independent of the source signal $\tilde{\mathbf{x}}_{E,i}$ and distributed as $\mathbf{q}_{E,i} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$. The output signal $\hat{\mathbf{x}}_{E,i}$ can be reliably delivered to EN i if the following condition is met:

$$
\begin{aligned}
g_i(\mathbf{U}, \mathbf{\Omega}) &\triangleq I(\tilde{\mathbf{x}}_{E,i}; \hat{\mathbf{x}}_{E,i}) \\
&= \log_2 \det \left(\sum_{g \in \mathcal{G}_U} \bar{c}_{i,\tilde{l}_g} \mathbf{U}_{i,g} \mathbf{U}_{i,g}^H + \mathbf{\Omega}_i \right) - \log_2 \det(\mathbf{\Omega}_i) \leq C_i,
\end{aligned}
\tag{10}
$$

with the notations $\mathbf{U} \triangleq \{\mathbf{U}_{i,g}\}_{i \in \mathcal{N}_E, g \in \mathcal{G}_U}$ and $\mathbf{\Omega} \triangleq \{\mathbf{\Omega}_i\}_{i \in \mathcal{N}_E}$.

3.3. Edge Precoding and Superposition Coding

Among the requested files $\{f_l\}_{l \in \mathcal{L}_{\text{req}}}$, each EN i can locally process the files $\{f_l\}_{l \in \mathcal{L}_{\text{req}}, c_{i,l}=1}$ that are pre-fetched at its local cache. Therefore, we assume that EN i sends a superposition of the quantized signal $\hat{\mathbf{x}}_{E,i}$, which was received on the wireless fronthaul link, and a locally precoded signal, written as $\sum_{g \in \mathcal{G}_U} c_{i,\tilde{l}_g} \mathbf{V}_{i,g} \mathbf{s}_{U,g}$. Here, $\mathbf{V}_{i,g} \in \mathbb{C}^{n_{E,i} \times d_{U,g}}$ and $\mathbf{s}_{U,g} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_{U,g}})$ represent the precoder matrix and the data signal, respectively, for the file $f_{\tilde{l}_g}$. As a result, the transmitted signal $\mathbf{x}_{E,i}$ of EN i is given as

$$
\begin{aligned}
\mathbf{x}_{E,i} &= \hat{\mathbf{x}}_{E,i} + \sum_{g \in \mathcal{G}_U} c_{i,\tilde{l}_g} \mathbf{V}_{i,g} \mathbf{s}_{U,g} \\
&= \sum_{g \in \mathcal{G}_U} \mathbf{T}_{i,g} \mathbf{s}_{U,g} + \mathbf{q}_{E,i},
\end{aligned}
\tag{11}
$$

where we have defined the effective precoding matrix $\mathbf{T}_{i,g} = \bar{c}_{i,\tilde{l}_g} \mathbf{U}_{i,g} + c_{i,\tilde{l}_g} \mathbf{V}_{i,g}$ for the gth multicast group at EN i. With the precoding model (11), the transmission power constraint at EN i can be rewritten as

$$
\sum_{g \in \mathcal{G}_U} \text{tr}\left(\mathbf{T}_{i,g} \mathbf{T}_{i,g}^H\right) + \text{tr}(\mathbf{\Omega}_i) \leq P_{E,i}.
\tag{12}
$$

3.4. Decoding and Achievable Rate

We assume that a UE k in the gth multicast group (i.e., $k \in \mathcal{N}_{U,g}$) tries to decode its requested content $f_{\tilde{l}_g}$ from the received signal $\mathbf{y}_{U,k}$ without decoding the interference signals from the other multicast groups. In order for the explained decoding to be successful, the rate R should be bounded by

$$
\begin{aligned}
R \leq f_{U,g,k}(\mathbf{T}, \mathbf{\Omega}) &\triangleq I(\mathbf{s}_{U,g}; \mathbf{y}_{U,k}) \\
&= \log_2 \det \left(\sum_{g' \in \mathcal{G}_U} \mathbf{G}_k \mathbf{T}_{g'} \mathbf{T}_{g'}^H \mathbf{G}_k^H + \mathbf{G}_k \bar{\mathbf{\Omega}} \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}} \right) \\
&\quad - \log_2 \det \left(\sum_{g' \in \mathcal{G}_U \setminus \{g\}} \mathbf{G}_k \mathbf{T}_{g'} \mathbf{T}_{g'}^H \mathbf{G}_k^H + \mathbf{G}_k \bar{\mathbf{\Omega}} \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}} \right),
\end{aligned}
\tag{13}
$$

for all $g \in \mathcal{G}_U$ and $k \in \mathcal{N}_{U,g}$, with $\mathbf{T}_g = [\mathbf{T}_{1,g}^H \mathbf{T}_{2,g}^H \cdots \mathbf{T}_{N_E,g}^H]^H$, $\mathbf{T} \triangleq \{\mathbf{T}_g\}_{g \in \mathcal{G}_U}$, and $\bar{\mathbf{\Omega}} = \text{diag}(\mathbf{\Omega}_1, \mathbf{\Omega}_2, \ldots, \mathbf{\Omega}_{N_E})$.

3.5. Optimization

In this subsection, we discuss the joint optimization of the fronthaul SDMA precoding $\mathbf{F}$, the effective cloud and edge precoding $\mathbf{T}$, and the quantization noise covariance matrices $\mathbf{\Omega}$ with the

criterion of maximizing the EE performance Θ, as defined in (4). We can mathematically formulate the mentioned optimization problem as

$$\underset{\mathbf{F},\mathbf{T},\boldsymbol{\Omega},R}{\text{maximize}} \quad \frac{R}{p_{\text{total}}\left(\mathbf{F},\mathbf{T},\boldsymbol{\Omega}\right)} \tag{14a}$$

$$\text{s.t. } R \leq f_{U,g,k}\left(\mathbf{T},\boldsymbol{\Omega}\right), \, g \in \mathcal{G}_U, \, k \in \mathcal{N}_{U,g}, \tag{14b}$$

$$g_i\left(\mathbf{T},\boldsymbol{\Omega}\right) \leq f_{E,i}\left(\mathbf{F}\right), \, i \in \mathcal{N}_E, \tag{14c}$$

$$p_C\left(\mathbf{F}\right) \leq P_C, \tag{14d}$$

$$p_{E,i}\left(\mathbf{T},\boldsymbol{\Omega}\right) \leq P_{E,i}, \, i \in \mathcal{N}_E. \tag{14e}$$

In (14), the function $p_{\text{total}}(\mathbf{F},\mathbf{T},\boldsymbol{\Omega})$ measures the total power consumption at the CP and the ENs, and is defined as

$$p_{\text{total}}(\mathbf{F},\mathbf{T},\boldsymbol{\Omega}) = p_C(\mathbf{F}) + \eta_C + \sum_{i\in\mathcal{N}_E} \left(p_{E,i}(\mathbf{T},\boldsymbol{\Omega}) + \eta_{E,i}\right), \tag{15}$$

where $p_C(\mathbf{F}) \triangleq \sum_{i\in\mathcal{N}_E} \text{tr}(\mathbf{F}_i\mathbf{F}_i^H)$ and $p_{E,i}(\mathbf{T},\boldsymbol{\Omega}) \triangleq \sum_{g\in\mathcal{G}_U} \text{tr}(\mathbf{T}_{i,g}\mathbf{T}_{i,g}^H) + \text{tr}(\boldsymbol{\Omega}_i)$ represent the powers consumed for RF transmissions at the CP and EN i, respectively. The function $g_i(\mathbf{T},\boldsymbol{\Omega})$ in (14c) is obtained by substituting $\mathbf{U}_{i,g} \leftarrow \mathbf{T}_{i,g}$ for all $g \in \mathcal{G}_U$ into the function $g_i(\mathbf{U},\boldsymbol{\Omega})$ defined in (10). We note that the condition (14b), which comes from (13), guarantees successful decoding of the requested files at the UEs, and the constraint (14c), obtained by combining the conditions (7) and (10), imposes reliable decompression of the quantized signals at ENs. The constraints (14d) and (14e) correspond to the transmission power constraints at the CP and ENs, respectively.

It is challenging to solve the problem (14), which is non-convex due to the objective function in (14a) and the constraints (14b) and (14c). We first handle the non-convexity of the objective fractional function by replacing it with a new variable Θ, which is constrained by

$$\Theta \leq \frac{R}{p_{\text{total}}\left(\mathbf{F},\mathbf{T},\boldsymbol{\Omega}\right)}. \tag{16}$$

As the above constraint is still non-convex, we consider the following equivalent constraint obtained by taking the monotonically increasing log function on both sides:

$$\ln\Theta \leq \ln R - \ln p_{\text{total}}\left(\mathbf{F},\mathbf{T},\boldsymbol{\Omega}\right). \tag{17}$$

Another way of handling the non-convexity of (16) is to adopt the fractional programming (FP) approach, as in [41]. We leave the comparison between the proposed decoupling and the FP approaches as a future work.

With the described manipulation, we obtain the following problem equivalent to (14):

$$\underset{\mathbf{F},\mathbf{T},\boldsymbol{\Omega},R,\Theta}{\text{maximize}} \quad \Theta \tag{18a}$$

$$\text{s.t. } \ln\Theta \leq \ln R - \ln p_{\text{total}}\left(\mathbf{F},\mathbf{T},\boldsymbol{\Omega}\right), \tag{18b}$$

$$R \leq f_{U,g,k}\left(\mathbf{T},\boldsymbol{\Omega}\right), \, g \in \mathcal{G}_U, \, k \in \mathcal{N}_{U,g}, \tag{18c}$$

$$g_i\left(\mathbf{T},\boldsymbol{\Omega}\right) \leq f_{E,i}\left(\mathbf{F}\right), \, i \in \mathcal{N}_E, \tag{18d}$$

$$p_C\left(\mathbf{F}\right) \leq P_C, \tag{18e}$$

$$p_{E,i}\left(\mathbf{T},\boldsymbol{\Omega}\right) \leq P_{E,i}, \, i \in \mathcal{N}_E. \tag{18f}$$

Although (18) is still non-convex, we can obtain a DC problem by change of variables $\tilde{\mathbf{F}}_i = \mathbf{F}_i\mathbf{F}_i^H$ and $\tilde{\mathbf{T}}_g = \mathbf{T}_g\mathbf{T}_g^H$ and by relaxing the constraints $\text{rank}(\tilde{\mathbf{F}}_i) \leq d_{E,i}$ and $\text{rank}(\tilde{\mathbf{T}}_g) \leq d_{U,g}$. It was reported in, for example, [4,13,15,25,26], that an efficient solution to a DC problem can be found by using the CCCP approach. The main idea of CCCP is to iteratively solve the convex problems that are obtained

by linearly approximating the terms that induce the non-convexity of the problem. The approximated convex problem changes over the iterations, as the reference point used for the linear approximation at each step is set to the solution of the convex problem in the previous iteration. The detailed algorithm was derived, in a similar way to those in [4,13,15,25,26], as Algorithm 1. After the CCCP based iterative algorithm converges, the obtained quadratic matrices $\tilde{\mathbf{F}}_i$ and $\tilde{\mathbf{T}}_g$ may not satisfy the rank constraints. We propose to obtain each CP precoding matrix $\mathbf{F}_i$ with the standard projection approach: $\mathbf{F}_i \leftarrow \mathbf{Q}_{d_{E,i}}(\tilde{\mathbf{F}}_i)\,\mathbf{S}_{d_{E,i}}(\tilde{\mathbf{F}}_i)^{1/2}$, where $\mathbf{Q}_d(\cdot)$ takes the leading d eigenvectors of the input matrix as the column vectors and $\mathbf{S}_d(\cdot)$ is a diagonal matrix, whose diagonal elements are the d leading eigenvalues of the input matrix. Similarly, each effective cloud and edge precoding matrix $\mathbf{T}_g$ is obtained as $\mathbf{T}_g \leftarrow \mathbf{Q}_{d_{U,g}}(\tilde{\mathbf{T}}_g)\,\mathbf{S}_{d_{U,g}}(\tilde{\mathbf{T}}_g)^{1/2}$.

Algorithm 1 CCCP-based algorithm for problem (18).

---

1. Initialize the variables $\Theta', \tilde{\mathbf{F}}', \tilde{\mathbf{T}}'$ and Ω' that satisfy the constraints of the problem (18).
2. Update $\Theta'', \tilde{\mathbf{F}}'', \tilde{\mathbf{T}}''$ and Ω'' as an optimal solution of the (convex) problem (A1) in Appendix A.
3. Stop if a convergence criterion is satisfied. Otherwise, set $\Theta' \leftarrow \Theta'', \tilde{\mathbf{F}}' \leftarrow \tilde{\mathbf{F}}'', \tilde{\mathbf{T}}' \leftarrow \tilde{\mathbf{T}}''$ and $\Omega' \leftarrow \Omega''$ and go back to Step 2.

The proposed CCCP algorithm is an instance of the successive convex approximation (SCA) approach [42], whose worst-case order of complexity is given as $\mathcal{O}(N_{\mathrm{itr}}\sqrt{N_{\mathrm{const}}}\log(N_{\mathrm{const}}/\epsilon))$ [43]. Here, N_{itr} denotes the maximum number of iterations, N_{const} is the number of constraints of the convex problem (A1) in Appendix A, and ϵ indicates the desired error tolerance. Using simulation, we checked that the algorithm converges within a few tens of iterations for all simulated cases. We will show, in Section 5, the convergence behavior of the algorithm. Furthermore, the number N_{const} of constraints of (A1) is equal to $N_{\mathrm{const}} = N_U + 2N_E + 2$.

4. Energy-Efficient Design Under Hard-Transfer Fronthaul Mode

In the soft-transfer fronthauling scheme in Section 3, we used fronthaul SDMA precoding to create orthogonal fronthaul links across ENs, over which the quantized signals $\{\hat{\mathbf{x}}_{E,i}\}_{i \in \mathcal{N}_E}$ are communicated. However, this approach may not be efficient, in the sense that the overlapping nature of the files requested by different ENs is not sufficiently leveraged. Motivated by this observation, in this section, we discuss the energy-efficient design under hard-transfer fronthauling mode, whereby the fronthaul links carry hard information of the uncached files.

4.1. Connectivity Level and Fronthaul Multicasting

We assume that, under the hard-transfer fronthauling mode, each UE k is served by the union of the closest α ENs, which are denoted as $\mathcal{N}_{E,k}$, and that the ENs that cache the content f_{l_k} requested by UE k (i.e., $c_{i,l_k} = 1$). This means that each UE is served by at least α ENs, where we refer to α as the connectivity level. We note that increasing α has conflicting impacts on the system performance (see also, e.g., [40]): With larger α, the overhead of the wireless fronthaul link will increase, as the ENs need to receive more files from the CP. On the other hand, the inter-group interference signals which occur on the edge link will be better managed when the ENs co-operate with a larger connectivity level α. This suggests that the connectivity level α should be carefully chosen in adaptation to the system environment, such as the channel states $\{\mathbf{H}_i\}_{i \in \mathcal{N}_E}$ and $\{\mathbf{G}_{k,i}\}_{k \in \mathcal{N}_U, i \in \mathcal{N}_E}$. In this section, we discuss optimization for fixed α; however, in Section 5, we will show the performance when the optimal value of α is chosen.

Under the hard-transfer fronthauling mode, each EN i needs to receive the files that are requested by the nearby UEs k with $i \in \mathcal{N}_{E,k}$ and not cached by EN i from the CP on the fronthaul link. We denote the set of the indices of those files as

$$\mathcal{L}_{E,i} = \left\{ l_k, k \in \mathcal{N}_U \,\middle|\, i \in \mathcal{N}_{E,k} \text{ and } c_{i,l_k} = 0 \right\}. \tag{19}$$

Furthermore, we define the binary transfer variables $u_{i,l}$ as

$$
u_{i,l} = \begin{cases} 1, & \text{if } l \in \mathcal{L}_{E,i} \\ 0, & \text{otherwise} \end{cases}.
\tag{20}
$$

Equation (20) means that the variable $u_{i,l}$ takes a value of 1 if the file $f_{i,l}$ needs to be delivered to EN i over the fronthaul link and 0 otherwise (i.e., if the file $f_{i,l}$ is pre-fetched at EN i or $i \notin \mathcal{N}_{E,k}$).

As there could be overlaps among the sets $\mathcal{L}_{E,1}, \mathcal{L}_{E,2}, \ldots, \mathcal{L}_{E,N_E}$, we consider a multicast fronthaul transmission from the CP to the ENs. We define the set $\mathcal{L}_E = \cup_{i \in \mathcal{N}_E} \mathcal{L}_{E,i}$ of the indices of all the files which are multicast over the fronthaul link. We denote the $L_E = |\mathcal{L}_E|$ distinct indices in $\mathcal{L}_E$ as $\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_{L_E}$ (i.e., $\mathcal{L}_E = \{\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_{L_E}\}$). Therefore, we have L_E multicast groups for the fronthaul multicast transmission. We denote the set of the ENs corresponding to the gth fronthaul multicast group which need to receive the file $f_{\hat{l}_g}$ over the fronthaul link as $\mathcal{N}_{E,g} = \{i \in \mathcal{N}_E | \hat{l}_g \in \mathcal{L}_{E,i}\}$. Furthermore, we define the set of the indices of the fronthaul multicast groups as $\mathcal{G}_E = \{1, 2, \ldots, L_E\}$.

4.2. Fronthaul Multicast Precoding

In this subsection, we describe the fronthaul multicast transmission from the CP to the ENs. The CP performs channel encoding on the file $f_{\hat{l}_g}$ for each fronthaul multicast group $g \in \mathcal{G}_E$ and obtains a baseband signal $s_{E,g} \in \mathbb{C}^{d_{E,g} \times 1}$, distributed as $\mathbf{s}_{E,g} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_{E,g}})$. We set the number $d_{E,g}$ of data streams, such that $d_{E,g} \leq \min\{n_C, \min_{i \in \mathcal{N}_{E,g}} n_{E,i}\}$ is satisfied.

The CP precodes the encoded signals $\{s_{E,g}\}_{g \in \mathcal{G}_E}$ so that its transmitted signal $\mathbf{x}_C$ is given as

$$
\mathbf{x}_C = \sum_{g \in \mathcal{G}_E} \mathbf{A}_g \mathbf{s}_{E,g},
\tag{21}
$$

where $\mathbf{A}_g \in \mathbb{C}^{n_C \times d_{E,g}}$ is the precoding matrix for the gth fronthaul multicast group. With (21), the CP power constraint can be rewritten as

$$
\sum_{g \in \mathcal{G}_E} \mathrm{tr}\left(\mathbf{A}_g \mathbf{A}_g^H\right) \leq P_C.
\tag{22}
$$

We assume that each EN i performs a symbol-by-symbol decoding to recover the files $\{f_l\}_{l \in \mathcal{L}_{E,i}}$ from the received signal $\mathbf{y}_{E,i}$ without decoding the interference signals. Furthermore, since the received signal $\mathbf{y}_{E,i}$ may contain the signals that encode the cached files of EN i, the EN can exploit its cached contents for known interference cancellation. Under this assumption, the achievable rate for the gth fronthaul multicast group, denoted by $R_{E,g}$, is constrained as

$$
\begin{aligned}
R_{E,g} &\leq f_{E,g,i}(\mathbf{A}) \\
&\triangleq I\left(\mathbf{s}_{E,g}; \mathbf{y}_{E,i} \middle| \left\{\mathbf{s}_{E,m}, m \in \mathcal{G}_E | c_{i,\hat{l}_m} = 1\right\}\right) \\
&= \log_2 \det\left(\sum_{m \in \mathcal{G}_E} \bar{c}_{i,\hat{l}_m} \mathbf{H}_i \mathbf{A}_m \mathbf{A}_m^H \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}\right) \\
&\quad - \log_2 \det\left(\sum_{m \in \mathcal{G}_E \setminus \{g\}} \bar{c}_{i,\hat{l}_m} \mathbf{H}_i \mathbf{A}_m \mathbf{A}_m^H \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}\right),
\end{aligned}
\tag{23}
$$

for all $i \in \mathcal{N}_{E,g}$, where we have used the notation $\mathbf{A} \triangleq \{\mathbf{A}_g\}_{g \in \mathcal{G}_E}$.

4.3. Edge Multicast Precoding

Once the fronthaul multicast transmission is finished, each EN i can process the files f_l which have been pre-fetched to its cache (i.e., $c_{i,l} = 1$) or have been received from the CP (i.e., $u_{i,l} = 1$) for the edge multicast transmission. We assume that EN i obtains its transmitted signal $\mathbf{x}_{E,i}$ on the edge link by performing an edge multicast precoding as

$$\mathbf{x}_{E,i} = \sum_{g \in \mathcal{G}_U} (c_{i,\tilde{l}_g} | u_{i,\tilde{l}_g}) \mathbf{D}_{i,g} \mathbf{s}_{U,g}, \tag{24}$$

where $\mathbf{D}_{i,g} \in \mathbb{C}^{n_{E,i} \times d_{U,g}}$ denotes the precoding matrix for the gth edge multicast group at EN i and $\mathbf{s}_{U,g} \in \mathbb{C}^{d_{U,g} \times 1}$ represents the baseband signal encoding $f_{\tilde{l}_g}$ with $d_{U,g} \leq \min\{n_E, \min_{k \in \mathcal{N}_{U,g}} n_{U,k}\}$ of data streams, and is distributed as $\mathbf{s}_{U,g} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_{U,g}})$. The transmission power constraint for EN i under (24) can be stated as

$$\sum_{g \in \mathcal{G}_U} (c_{i,\tilde{l}_g} | u_{i,\tilde{l}_g}) \operatorname{tr}\left(\mathbf{D}_{i,g} \mathbf{D}_{i,g}^H\right) \leq P_{E,i}. \tag{25}$$

To simplify the notation, we define the effective edge multicast precoding matrix $\mathbf{W}_{i,g} = (c_{i,\tilde{l}_g} | u_{i,\tilde{l}_g}) \mathbf{D}_{i,g}$ which is constrained by

$$\operatorname{tr}\left(\mathbf{W}_{i,g} \mathbf{W}_{i,g}^H\right) = 0, \text{ if } c_{i,\tilde{l}_g} = u_{i,\tilde{l}_g} = 0. \tag{26}$$

Using (26), we can write the total transmitted signal $\mathbf{x}_E = [\mathbf{x}_{E,1}^H \cdots \mathbf{x}_{E,N_E}^H]^H$ of all the ENs on the wireless edge link as

$$\mathbf{x}_E = \sum_{g \in \mathcal{G}_U} \mathbf{W}_g \mathbf{s}_{U,g}, \tag{27}$$

with the effective edge precoding matrix $\mathbf{W}_g = [\mathbf{W}_{1,g}^H \cdots \mathbf{W}_{N_E,g}^H]^H$ for the gth edge multicast group. We note that the ith submatrix $\mathbf{W}_{i,g}$ of $\mathbf{W}_g$ can be expressed as $\mathbf{W}_{i,g} = \mathbf{E}_i^H \mathbf{W}_g$, where the shaping matrix $\mathbf{E}_i \in \mathbb{C}^{n_E \times n_{E,i}}$ is defined as

$$\mathbf{E}_i = \left[\mathbf{0}_{(\sum_{j=1}^{i-1} n_{E,j}) \times n_{E,i}}^H \quad \mathbf{I}_{n_{E,i}} \quad \mathbf{0}_{(\sum_{j=i+1}^{N_E} n_{E,j}) \times n_{E,i}}^H \right]^H. \tag{28}$$

We assume that each UE k in the gth edge multicast group performs single-user decoding to obtain the requested file $f_{\tilde{l}_g}$, based on the received signal $\mathbf{y}_{U,k}$. The achievable rate R is, hence, limited by

$$R \leq f_{U,g,k}(\mathbf{W}) \triangleq I\left(\mathbf{s}_{U,g}; \mathbf{y}_{U,k}\right) \tag{29}$$

$$= \log_2 \det\left(\sum_{g' \in \mathcal{G}_U} \mathbf{G}_k \mathbf{W}_{g'} \mathbf{W}_{g'}^H \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}}\right)$$

$$- \log_2 \det\left(\sum_{g' \in \mathcal{G}_U \setminus \{g\}} \mathbf{G}_k \mathbf{W}_{g'} \mathbf{W}_{g'}^H \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}}\right),$$

for all $g \in \mathcal{G}_U$ and $k \in \mathcal{N}_{U,g}$, where we have used the notation $\mathbf{W} \triangleq \{\mathbf{W}_g\}_{g \in \mathcal{G}_U}$.

We note that each gth edge multicast file $f_{\tilde{l}_g}$ needs to also be reliably communicated over the fronthaul link if it belongs to the set of the fronthaul multicast files; that is, $\tilde{l}_g \in \mathcal{L}_E$. Therefore, the rate R of the edge multicast files has the following additional constraint:

$$R \leq R_{E,g'}, \text{ if } \tilde{l}_g = \hat{l}_{g'}, \ g \in \mathcal{G}_U, \ g' \in \mathcal{G}_E. \tag{30}$$

4.4. Optimization

We now address the joint optimization of the fronthaul multicast precoding **A** and the effective edge multicast precoding matrices **W**, with the aim of maximizing the EE performance under the constraints on the transmision powers at the CP and the ENs. We can state the problem as

$$\underset{\mathbf{A},\mathbf{W},R,\mathbf{R}_E}{\text{maximize}} \frac{R}{p_{\text{total}}(\mathbf{A},\mathbf{W})} \tag{31a}$$

$$\text{s.t. } R \leq f_{U,g,k}(\mathbf{W}), \ g \in \mathcal{G}_U, \ k \in \mathcal{N}_{U,g}, \tag{31b}$$

$$R_{E,g} \leq f_{E,g,i}(\mathbf{A}), \ g \in \mathcal{G}_E, \ i \in \mathcal{N}_{E,g}, \tag{31c}$$

$$R \leq R_{E,g'}, \text{ if } \tilde{l}_g = \hat{l}_{g'}, \ g \in \mathcal{G}_U, \ g' \in \mathcal{G}_E, \tag{31d}$$

$$\text{tr}(\mathbf{E}_i^H \mathbf{W}_g \mathbf{W}_g^H \mathbf{E}_i) = 0, \text{ if } c_{i,\tilde{l}_g} = u_{i,\tilde{l}_g} = 0, \ i \in \mathcal{N}_E, \ g \in \mathcal{G}_U, \tag{31e}$$

$$p_C(\mathbf{A}) \leq P_C, \tag{31f}$$

$$p_{E,i}(\mathbf{W}) \leq P_{E,i}, \ i \in \mathcal{N}_E, \tag{31g}$$

with $\mathbf{R}_E \triangleq \{R_{E,g}\}_{g \in \mathcal{G}_E}$. In (31), the function $p_{\text{total}}(\mathbf{A},\mathbf{W})$ is equal to the total power consumption of the network and is defined as

$$p_{\text{total}}(\mathbf{A},\mathbf{W}) = p_C(\mathbf{A}) + \eta_C + \sum_{i \in \mathcal{N}_E} (p_{E,i}(\mathbf{W}) + \eta_{E,i}), \tag{32}$$

with $p_C(\mathbf{A}) = \sum_{g \in \mathcal{G}_E} \text{tr}(\mathbf{A}_g \mathbf{A}_g^H)$ and $p_{E,i}(\mathbf{W}) = \sum_{g \in \mathcal{G}_U} \text{tr}(\mathbf{E}_i^H \mathbf{W}_g \mathbf{W}_g^H \mathbf{E}_i)$ being the RF transmision powers at the CP and EN i, respectively. The constraint (31b), which is equivalent to (29), imposes that the requested files are reliably decoded at the UEs and the constraint (31c) guarantees that all the multicast messages on the wireless fronthaul link are successfully decoded by the requesting ENs. As the rate of each requested file cannot exceed the rate at which the file is communicated over the fronthaul link, the constraint (31d) is imposed. The condition (31e) indicates that each EN can precode only the files that are stored in its cache or received from the CP over the fronthaul link. The constraints (31f) and (31g) stand for the CP and per-EN transmision power constraints.

The formulated problem (31) is non-convex. However, since it has a similar form to that of the problem (14) defined for the soft-transfer scheme, we can tackle (31) in a similar manner to that discussed in Section 3.5: We replace the objective fractional function in (31a) with a new variable Θ, which is constrained by

$$\ln \Theta \leq \ln R - \ln p_{\text{total}}(\mathbf{A},\mathbf{W}). \tag{33}$$

This yields the following equivalent problem:

$$\underset{\mathbf{A},\mathbf{W},R,\mathbf{R}_E,\Theta}{\text{maximize}} \Theta \tag{34a}$$

$$\text{s.t. } \ln \Theta \leq \ln R - \ln p_{\text{total}}(\mathbf{A},\mathbf{W}), \tag{34b}$$

$$(31b)\text{--}(31g). \tag{}$$

As in Section 3.5, from (34), we can obtain a DC problem by defining a change of variables $\tilde{\mathbf{A}}_g = \mathbf{A}_g \mathbf{A}_g^H$ and $\tilde{\mathbf{W}}_g = \mathbf{W}_g \mathbf{W}_g^H$ and relaxing the non-convex constraints $\text{rank}(\tilde{\mathbf{A}}_g) \leq d_{E,g}$ and $\text{rank}(\tilde{\mathbf{W}}_g) \leq d_{U,g}$. Thus, we propose to tackle the DC problem by deriving a CCCP-based iterative algorithm, which is detailed in Algorithm 2, followed by the projections $\mathbf{A}_g \leftarrow \mathbf{Q}_{d_{E,g}}(\tilde{\mathbf{A}}_g) \mathbf{S}_{d_{E,g}}(\tilde{\mathbf{A}}_g)^{1/2}$ and $\mathbf{W}_g \leftarrow \mathbf{Q}_{d_{U,g}}(\tilde{\mathbf{W}}_g) \mathbf{S}_{d_{U,g}}(\tilde{\mathbf{W}}_g)^{1/2}$. Similar to Algorithm 1, the complexity of Algorithm 2 can be expressed as $\mathcal{O}(N_{\text{itr}} \sqrt{N_{\text{const}}} \log(N_{\text{const}}/\epsilon))$, where the number N_{const} of constraints for the convex problem (A2) is randomly given, depending on the request profiles of the UEs and the cached contents of the ENs. It is guaranteed that N_{const} is bounded as $N_{\text{const}} \leq N_U^2 + N_E N_U + 2N_E + N_U + 2$, where the upper bound is approximated, for sufficiently large N_U and N_E, as $N_U(N_U + N_E)$.

Algorithm 2 CCCP-based algorithm for problem (34).

---
1. Initialize the variables Θ', $\tilde{\mathbf{A}}'$, and $\tilde{\mathbf{W}}'$ that satisfy the constraints of the problem (34).
2. Update Θ'', $\tilde{\mathbf{A}}''$ and $\tilde{\mathbf{W}}''$ as an optimal solution of the (convex) problem (A2) in Appendix A.
3. Stop if a convergence criterion is satisfied. Otherwise, set $\Theta' \leftarrow \Theta''$, $\tilde{\mathbf{A}}' \leftarrow \tilde{\mathbf{A}}''$ and $\tilde{\mathbf{W}}' \leftarrow \tilde{\mathbf{W}}''$ and go back to Step 2.

---

5. Numerical Results

In this section, we validate the effectiveness of the proposed energy-efficient designs under the soft-transfer and hard-transfer fronthauling strategies proposed in Sections 3 and 4, respectively, through numerical results. Throughout the section, we assume that the positions of the ENs and UEs are uniformly distributed within a circular region of radius 100 m and the CP is located at the center. Furthermore, we assume Rayleigh fading for all channel elements of the fronthaul $\{\mathbf{H}_i\}_{i\in\mathcal{N}_E}$ and edge links $\{\mathbf{G}_{k,i}\}_{k\in\mathcal{N}_U, i\in\mathcal{N}_E}$, and adopt the path-loss model $1/(1 + (\text{distance}/D_0)^\beta)$ considered in [13,15,39], where we set $D_0 = 30$ m and $\beta = 3$. We define the SNRs of the fronthaul and edge links as P_C/σ_E^2 and P_E/σ_U^2, respectively; where we assume that every EN uses the same transmission power P_E, i.e., $P_{E,i} = P_E$ for all $i \in \mathcal{N}_E$. Also, we set the circuit powers η_C and $\eta_{E,i}$ of the CP and each EN to be 10 dB larger than the noise powers, σ_E^2 and σ_U^2, of the corresponding links.

The number of files in the library $\mathcal{F}$ is set to $L = 10$, and the Zipf's distribution is considered for the popularity of the files. That is, the probability $\Pr[l_k = l]$ that UE k requests the file f_l is given as $\Pr[l_k = l] = cl^{-\gamma}$ with $c = 1/(\sum_{l=1}^{L} l^{-\gamma})$. We set the constant γ (which controls the skewness of the popularity among the files) as $\gamma = 1$. Assuming that $B_i = B$ for all $i \in \mathcal{N}_E$, we define the fractional cache size as $\mu = B/S_{\text{library}} \in [0, 1]$: $\mu = 0$ means that the ENs do not have caching functionality, and $\mu = 1$ indicates that the full library $\mathcal{F}$ is available at all the ENs. For a partial caching case, with $0 < \mu < 1$, we consider a random caching strategy, in which each EN pre-fetches $\lfloor \mu L \rfloor$ files, randomly chosen from the L files in the library $\mathcal{F}$.

We first observe the convergence behavior of the proposed CCCP-based iterative algorithm for the soft-transfer fronthauling scheme in Section 3 (i.e., Algorithm 1) by plotting, in Figure 3, the average EE performance Θ with respect to the number of iterations for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 10, 20\}$ dB. The figure shows that, regardless of SNR values, the algorithm converged within a few tens of iterations.
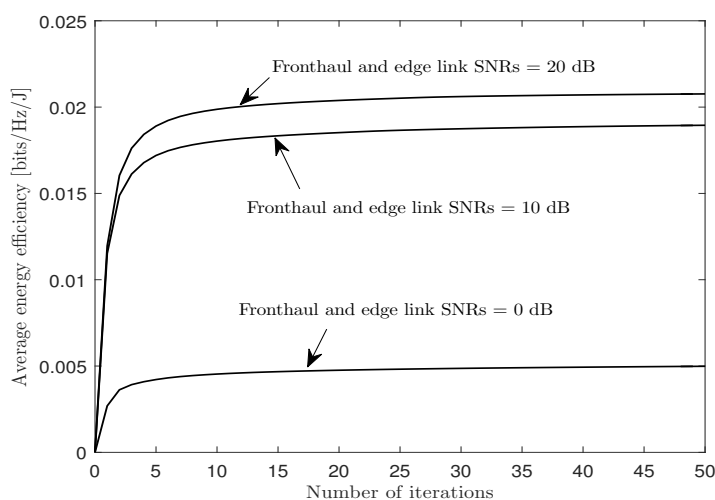

Figure 3. Average EE Θ versus the number of iterations with the soft-transfer fronthauling strategy for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 10, 20\}$ dB.

In Figure 4, we plot the average EE performance Θ of the proposed energy-efficient design with the soft-transfer fronthauling strategy proposed in Section 3 versus the fractional cache size μ for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $P_C/\sigma_E^2 \in \{0, 5, 10, 15, 20\}$ dB, and $P_E/\sigma_U^2 = 20$ dB. The figure shows that, as the ENs could pre-fetch more popular contents to the local caches, the overhead on the wireless fronthaul link was reduced, which led to better EE performance of the overall network. Furthermore, it is noted that the impact of the cache size was more pronounced when the fronthaul SNR P_C/σ_E^2 was smaller, owing to the fact that, with lower fronthaul SNR level, the fronthaul overhead became a performance bottleneck. Hence, equipping ENs with caches will be more helpful. In a similar vein, for the full caching case ($\mu = 1$), varying the fronthaul SNR did not affect the performance.
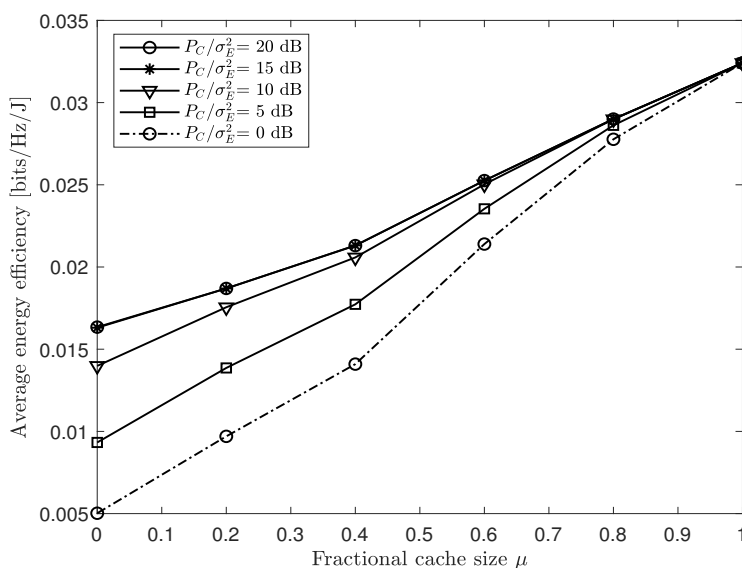
Figure 4. Average EE Θ versus the fractional cache size μ with the soft-transfer fronthauling strategy (Section 3) for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $P_C/\sigma_E^2 \in \{0, 5, 10, 15, 20\}$ dB, and $P_E/\sigma_U^2 = 20$ dB.

In Figure 5, we show the average EE Θ, as well as the spectral efficiency (SE) R, of the proposed energy-efficient soft-transfer scheme in Section 3 with respect to the edge link SNR P_E/σ_U^2 for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 \in \{0, 20\}$ dB. To validate the importance of the energy-efficient design, we also plot the EE and SE performance of the spectrally-efficient design, which solves the problem (14) replacing the EE objective function with the spectral efficiency R. From Figure 5, we observe that, when the fronthaul and edge links had small SNRs, the energy- and spectrally-efficient schemes provided similar EE and SE performances, which means that the energy-efficient solution tended to use full transmission power at the CP and the ENs to maximize the SE metric. In contrast, when the fronthaul or edge link had a sufficiently large SNR, it would be better to use only partial transmission power at the CP or the ENs to achieve a better EE performance, at the cost of the SE value.
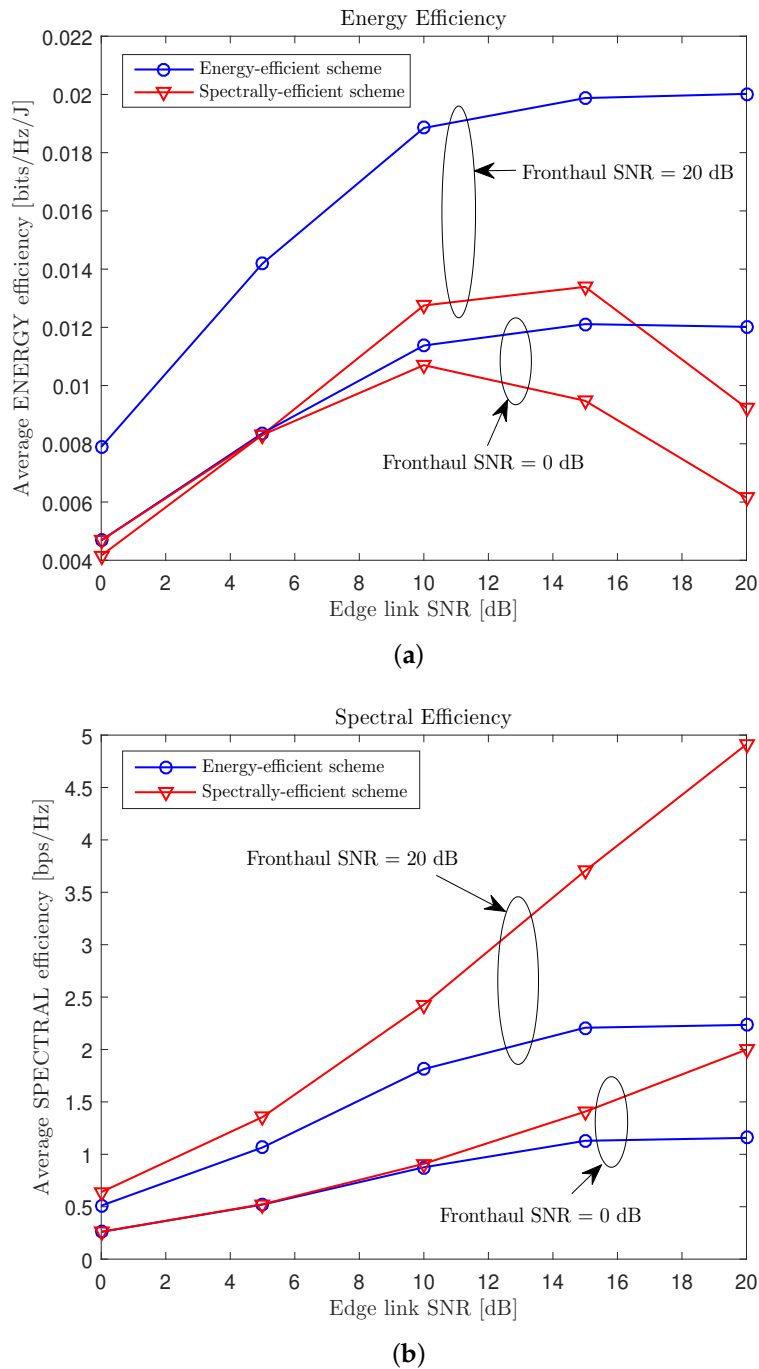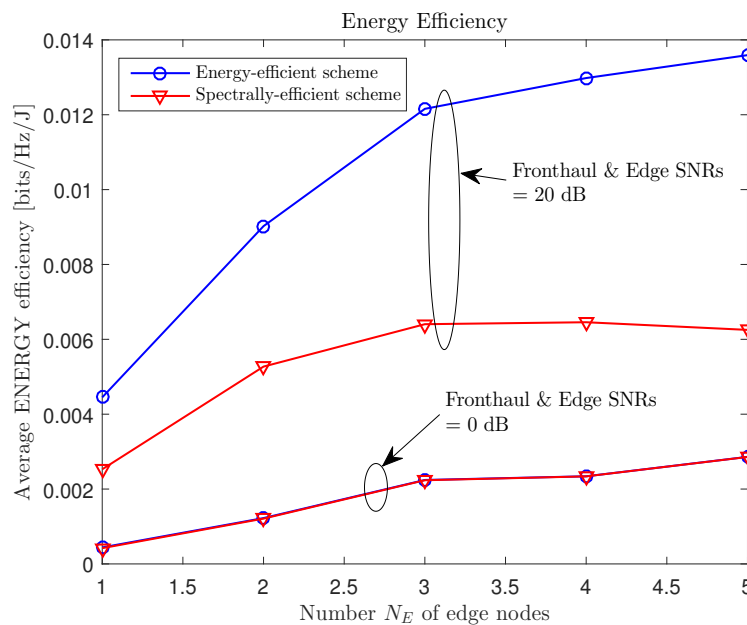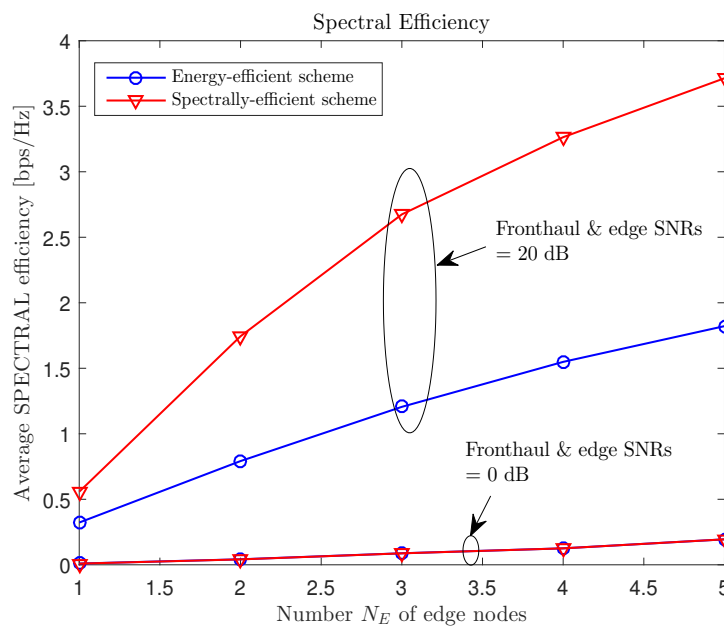
Energy Efficiency

(**a**)

Spectral Efficiency

(**b**)

Figure 5. Average EE Θ and SE R versus the edge link SNR P_E/σ_U^2 with the "soft-transfer" fronthauling strategy (Section 3) for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 \in \{0, 20\}$ dB ((**a**) EE; (**b**) SE).

Figure 6 compares the average EE Θ and SE R of the energy- and spectrally-efficient schemes for various numbers N_E of ENs, with $n_C = 4$, $N_U = 4$, $n_{E,i} = 2$, $n_{U,k} = 1$, $\mu = 0.3$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 20\}$ dB. We can see that, regardless of the number of edge nodes, the energy-efficient and spectrally-efficient schemes showed almost same EE and SE performance when the SNRs of the fronthaul and edge links were low. However, when the SNRs were sufficiently large, the energy-efficient scheme showed a notable gain, which increased with the number of ENs, over the spectrally-efficient scheme. This suggests that, when there were many ENs and the SNRs were large,

the spectrally-efficient scheme encouraged all of the ENs to use full transmission power, while the energy-efficient scheme allocated only the necessary level of transmission power to the ENs.

(a)

(b)

Figure 6. Average EE Θ and SE R versus the number N_E of edge nodes with the "soft-transfer" fronthauling strategy (Section 3) for a cache-aided C-RAN system with $n_C = 4$, $N_U = 4$, $n_{E,i} = 2$, $n_{U,k} = 1$, $\mu = 0.3$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 20\}$ dB ((**a**) EE; (**b**) SE).

Figure 7 plots the average EE performance Θ of the proposed energy-efficient hard-transfer scheme in Section 4 versus the edge link SNR P_E/σ_U^2 for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 = 20$ dB. In the figure, we include the performance for the cases of all possible connectivity levels $\alpha \in \{1, \ldots, N_E\}$, as well as for the case when the optimal α is chosen, corresponding to the largest EE for each channel realization. As expected, in the regime

of low edge SNRs, the overall performance was limited by the edge link, rather than the fronthaul link. Therefore, it is desirable to use large connectivity level α to maximize the co-operation gain of the ENs. On the other hand, when the edge SNR is large enough, the limitation at the fronthaul link becomes dominant; hence, it would be better to decrease the connectivity level α to reduce the fronthaul overhead. We also note that using the best connectivity level α in adaptation to the instantaneous CSI yields further improvement, particularly in the intermediate edge SNR levels.
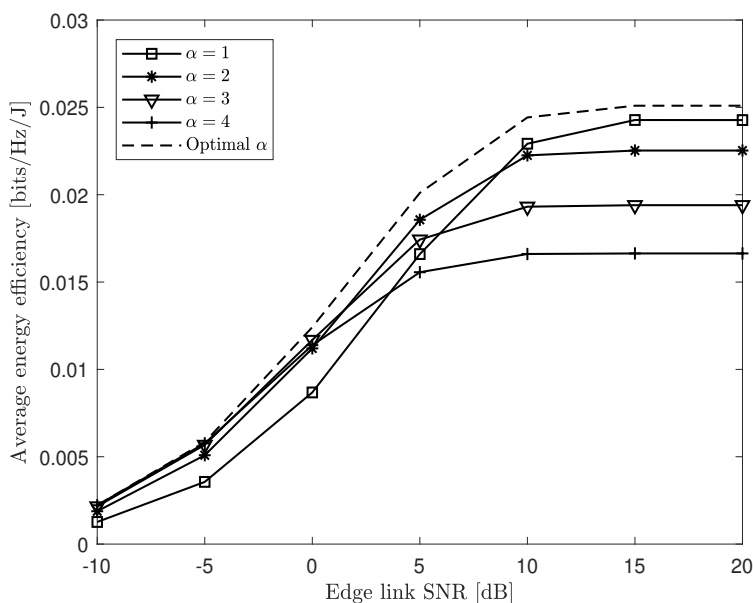
Figure 7. Average EE Θ versus the edge link SNR P_E/σ_U^2 with the proposed "hard-transfer" fronthauling strategy (Section 4) for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 \in 20$ dB.

In Figure 8, we compare the average EE performance Θ of the soft-transfer and hard-transfer fronthauling schemes for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 = 20$ dB. For the hard-transfer scheme, we used the optimal connectivity level α which gave the best performance for each channel sample. We can observe, from the graph, that the hard-transfer fronthauling scheme achieved a better EE performance than the soft-transfer scheme in overall system environments. This supports the fact that the hard-transfer scheme is more effective in utilizing the multicasting opportunity in the wireless fronthaul link, while the soft-transfer scheme starts with orthogonalizing the wireless fronthaul links across the ENs by means of the SDMA fronthaul precoding. However, it should be noted that, when the fronthaul link SNR was significantly smaller than that of the edge link, the soft-transfer scheme (whereby the fronthaul link carried compressed information) showed a better EE performance than the hard-transfer scheme. In Figure 8, we also compare the performance of the proposed joint design of fronthaul and edge links with those of simpler separate design methods. The separate schemes, first, design the fronthaul-related variables with the goal of maximizing the EE of the fronthaul transmission and, then, optimize the remaining variables related to the edge link transmission. The figure shows that the performance gain of the joint design was significant for both the soft-transfer and hard-transfer fronthauling strategies and grew with the edge link SNR.
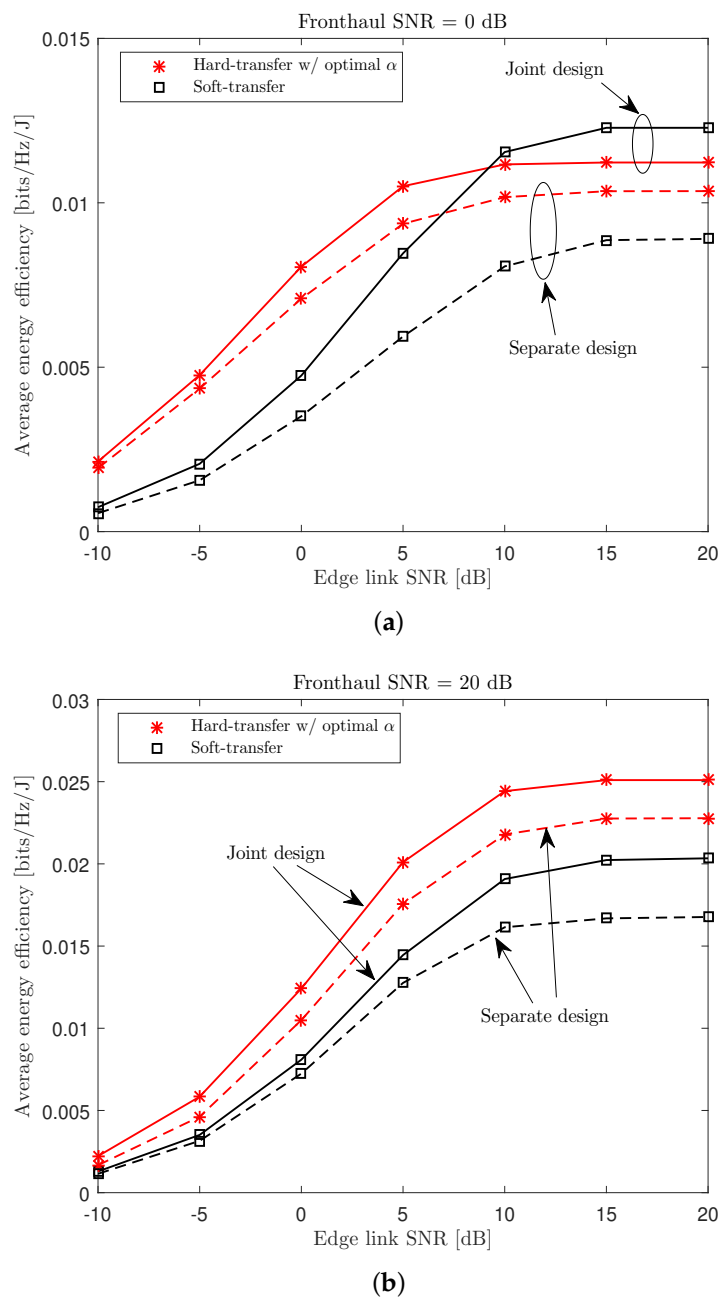
(a)

(b)

Figure 8. Average EE Θ versus the edge link SNR P_E/σ_U^2 with the soft- and hard-transfer fronthauling strategies for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, $\mu = 0.3$, and $P_C/\sigma_E^2 \in \{0, 20\}$ dB ((**a**) $P_C/\sigma_E^2 = 0$ dB; (**b**) $P_C/\sigma_E^2 = 20$ dB).

Lastly, in Figure 9, we compare the average EE performance Θ of the schemes considered in Figure 8, with respect to the fractional cache size μ for a cache-aided C-RAN with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 20\}$ dB. As the soft- and hard-transfer schemes differ only in terms of the fronthaul usage, the EE values of both schemes approached the same value as μ increased (i.e., ENs pre-fetched more content and, hence, the amount of traffic over the fronthaul was reduced). In a similar vein, the advantage of the proposed joint design became minor for sufficiently large μ.
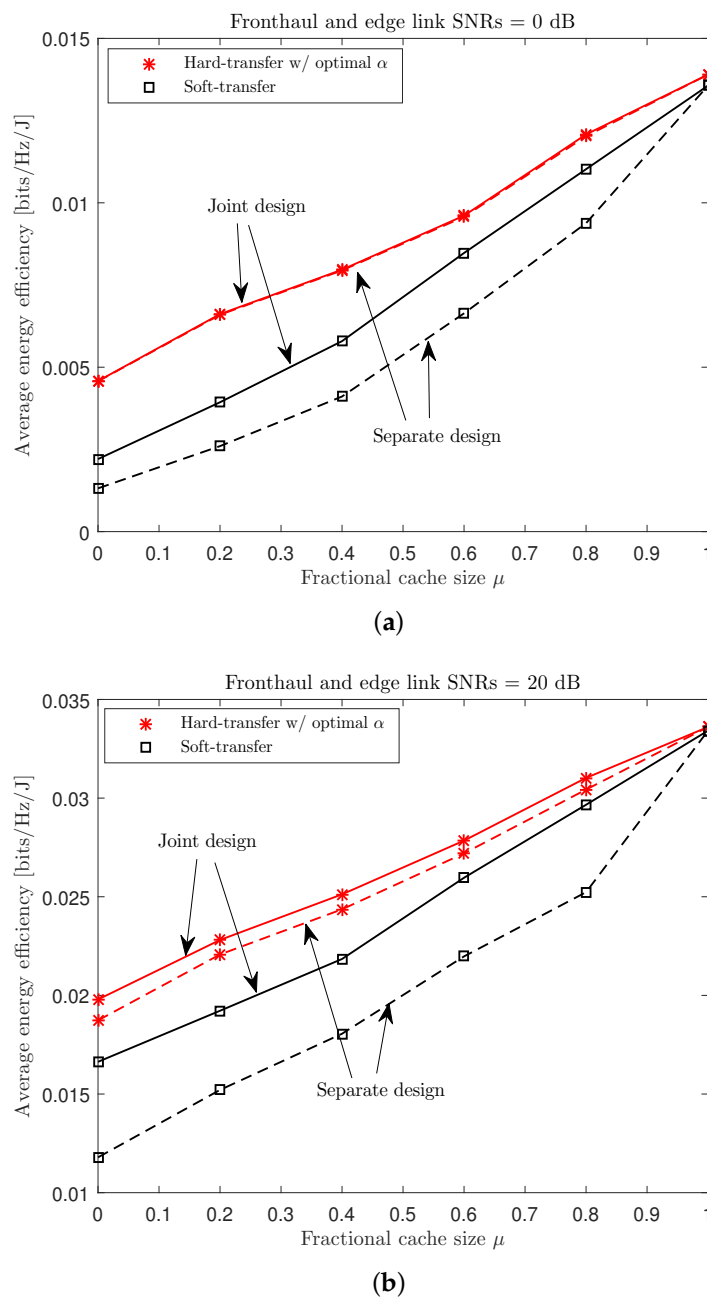
(**a**)

(**b**)

Figure 9. Average EE Θ versus the fractional cache size μ with the soft- and hard-transfer fronthauling strategies for a cache-aided C-RAN system with $n_C = 8$, $N_E = N_U = 4$, $n_{E,i} = n_{U,k} = 2$, and $P_C/\sigma_E^2 = P_E/\sigma_U^2 \in \{0, 20\}$ dB ((**a**) $P_C/\sigma_E^2 = P_E/\sigma_U^2 = 0$ dB; (**b**) $P_C/\sigma_E^2 = P_E/\sigma_U^2 = 20$ dB).

6. Conclusions

We have discussed the energy-efficient joint design of the fronthaul and edge transmission strategies for a cache-aided C-RAN system with a wireless fronthaul link. Specifically, we have tackled the problem of maximizing the EE metric under both the soft- and hard-transfer fronthauling strategies. We have converted the formulated non-convex optimization problems into DC problems by means of change of variables and rank relaxation and tackled the resulting problems using the CCCP approach. Through numerical results, we have validated the impact of caching functionality and the advantages of an energy-efficient design over a spectrally-efficient scheme, particularly at high fronthaul and edge SNRs. It was also observed that, in the overall simulated cases, the hard-transfer scheme can

better utilize the multicasting opportunity of the wireless fronthaul transmission, as compared to the soft-transfer scheme. Furthermore, we have verified that the importance of the joint design of the fronthaul and edge links is more significant when the fronthaul and edge links have a larger SNR, or when the ENs pre-fetch less content.

Appendix A

The convex problem that is solved at Step 2 of Algorithm 1 is defined as

$$\underset{\tilde{\mathbf{F}},\tilde{\mathbf{T}},\boldsymbol{\Omega},R,\Theta}{\text{maximize}} \quad \Theta \tag{A1}$$

$$\text{s.t. } \tilde{\psi}\left(\Theta,\tilde{\mathbf{F}},\tilde{\mathbf{T}},\boldsymbol{\Omega},\Theta',\tilde{\mathbf{F}}',\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right) \leq \ln R,$$

$$R \leq \tilde{f}_{U,g,k}\left(\tilde{\mathbf{T}},\boldsymbol{\Omega},\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right), \ g \in \mathcal{G}_U, \ k \in \mathcal{N}_{U,g},$$

$$\tilde{g}_i\left(\tilde{\mathbf{T}},\boldsymbol{\Omega},\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right) \leq \tilde{f}_{E,i}\left(\tilde{\mathbf{F}},\tilde{\mathbf{F}}'\right), \ i \in \mathcal{N}_E,$$

$$p_C\left(\tilde{\mathbf{F}}\right) \leq P_C,$$

$$p_{E,i}\left(\tilde{\mathbf{T}},\boldsymbol{\Omega}\right) \leq P_{E,i}, \ i \in \mathcal{N}_E,$$

where we define the functions

$$\tilde{\psi}\left(\Theta,\tilde{\mathbf{F}},\tilde{\mathbf{T}},\boldsymbol{\Omega},\Theta',\tilde{\mathbf{F}}',\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right)$$
$$\triangleq \varphi\left(\Theta,\Theta'\right) + \varphi\left(p_{\text{total}}\left(\tilde{\mathbf{F}},\tilde{\mathbf{T}},\boldsymbol{\Omega}\right), p_{\text{total}}\left(\tilde{\mathbf{F}}',\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right)\right),$$
$$\tilde{f}_{U,g,k}\left(\tilde{\mathbf{T}},\boldsymbol{\Omega},\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right)$$
$$\triangleq \log_2 \det\left(\sum_{g' \in \mathcal{G}_U} \mathbf{G}_k \tilde{\mathbf{T}}_{g'} \mathbf{G}_k^H + \mathbf{G}_k \bar{\boldsymbol{\Omega}} \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}}\right)$$
$$- \varphi\left(\begin{array}{c} \sum_{g' \in \mathcal{G}_U \setminus \{g\}} \mathbf{G}_k \tilde{\mathbf{T}}_{g'} \mathbf{G}_k^H + \mathbf{G}_k \bar{\boldsymbol{\Omega}} \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}}, \\ \sum_{g' \in \mathcal{G}_U \setminus \{g\}} \mathbf{G}_k \tilde{\mathbf{T}}'_{g'} \mathbf{G}_k^H + \mathbf{G}_k \bar{\boldsymbol{\Omega}}' \mathbf{G}_k^H + \sigma_U^2 \mathbf{I}_{n_{U,k}} \end{array}\right),$$
$$\tilde{f}_{E,i}\left(\tilde{\mathbf{F}},\tilde{\mathbf{F}}'\right) \triangleq \log_2 \det\left(\sum_{j \in \mathcal{N}_E} \mathbf{H}_i \tilde{\mathbf{F}}_j \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}\right)$$
$$- \varphi\left(\begin{array}{c} \sum_{j \in \mathcal{N}_E \setminus \{i\}} \mathbf{H}_i \tilde{\mathbf{F}}_j \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}}, \\ \sum_{j \in \mathcal{N}_E \setminus \{i\}} \mathbf{H}_i \tilde{\mathbf{F}}'_j \mathbf{H}_i^H + \sigma_E^2 \mathbf{I}_{n_{E,i}} \end{array}\right),$$
$$\tilde{g}_i\left(\tilde{\mathbf{T}},\boldsymbol{\Omega},\tilde{\mathbf{T}}',\boldsymbol{\Omega}'\right)$$
$$\triangleq \varphi\left(\begin{array}{c} \sum_{g \in \mathcal{G}_U} \bar{c}_{i,\tilde{I}_g} \mathbf{E}_i^H \tilde{\mathbf{T}}_g \mathbf{E}_i + \boldsymbol{\Omega}_i, \\ \sum_{g \in \mathcal{G}_U} \bar{c}_{i,\tilde{I}_g} \mathbf{E}_i^H \tilde{\mathbf{T}}'_g \mathbf{E}_i + \boldsymbol{\Omega}'_i \end{array}\right) - \log_2 \det\left(\boldsymbol{\Omega}_i\right),$$

with the notation $\varphi(\mathbf{A}, \mathbf{B}) \triangleq \log_2 \det(\mathbf{B}) + \text{tr}(\mathbf{B}^{-1}(\mathbf{A} - \mathbf{B}))$.

Also, at Step 2 of Algorithm 2, we solve the following convex problem:

$$
\underset{\tilde{\mathbf{A}},\tilde{\mathbf{W}},R,\mathbf{R}_E}{\text{maximize}}\ \Theta \tag{A2}
$$

$$
\begin{aligned}
\text{s.t. } & \tilde{\psi}\left(\Theta,\tilde{\mathbf{A}},\tilde{\mathbf{W}},\Theta',\tilde{\mathbf{A}}',\tilde{\mathbf{W}}'\right) \leq \ln R,\\
& R \leq \tilde{f}_{U,g,k}\left(\tilde{\mathbf{W}},\tilde{\mathbf{W}}'\right),\ g\in\mathcal{G}_U,\ k\in\mathcal{N}_{U,g},\\
& R_{E,g} \leq \tilde{f}_{E,g,i}\left(\tilde{\mathbf{A}},\tilde{\mathbf{A}}'\right),\ g\in\mathcal{G}_E,\ i\in\mathcal{N}_{E,g},\\
& R \leq R_{E,g'},\ \text{if } \tilde{l}_g = \hat{l}_{g'},\ g\in\mathcal{G}_U,\ g'\in\mathcal{G}_E,\\
& \mathrm{tr}(\mathbf{E}_i^H\tilde{\mathbf{W}}_g\mathbf{E}_i) = 0,\ \text{if } c_{i,\tilde{l}_g}=u_{i,\tilde{l}_g}=0,\ i\in\mathcal{N}_E,\ g\in\mathcal{G}_U,\\
& p_C\left(\tilde{\mathbf{A}}\right) \leq P_C,\\
& p_{E,i}\left(\tilde{\mathbf{W}}\right) \leq P_{E,i},\ i\in\mathcal{N}_E.
\end{aligned}
$$

where we define the functions

$$
\begin{aligned}
&\tilde{\psi}\left(\Theta,\tilde{\mathbf{A}},\tilde{\mathbf{W}},\Theta',\tilde{\mathbf{A}}',\tilde{\mathbf{W}}'\right)\\
&\triangleq \varphi\left(\Theta,\Theta'\right) + \varphi\left(p_{\text{total}}\left(\tilde{\mathbf{A}},\tilde{\mathbf{W}}\right),p_{\text{total}}\left(\tilde{\mathbf{A}}',\tilde{\mathbf{W}}'\right)\right),\\
&\tilde{f}_{U,g,k}\left(\tilde{\mathbf{W}},\tilde{\mathbf{W}}'\right)\\
&\triangleq \log_2\det\left(\sum_{g'\in\mathcal{G}_U}\mathbf{G}_k\tilde{\mathbf{W}}_{g'}\mathbf{G}_k^H + \sigma_U^2\mathbf{I}_{n_{U,k}}\right)\\
&\quad - \varphi\left(\begin{array}{c}\sum_{g'\in\mathcal{G}_U\setminus\{g\}}\mathbf{G}_k\tilde{\mathbf{W}}_{g'}\mathbf{G}_k^H + \sigma_U^2\mathbf{I}_{n_{U,k}},\\ \sum_{g'\in\mathcal{G}_U\setminus\{g\}}\mathbf{G}_k\tilde{\mathbf{W}}'_{g'}\mathbf{G}_k^H + \sigma_U^2\mathbf{I}_{n_{U,k}}\end{array}\right),\\
&\tilde{f}_{E,g,i}\left(\tilde{\mathbf{A}},\tilde{\mathbf{A}}'\right)\\
&\triangleq \log_2\det\left(\sum_{m\in\mathcal{G}_E}\bar{c}_{i,\hat{l}_m}\mathbf{H}_i\tilde{\mathbf{A}}_m\mathbf{H}_i^H + \sigma_E^2\mathbf{I}_{n_{E,i}}\right)\\
&\quad - \varphi\left(\begin{array}{c}\sum_{m\in\mathcal{G}_E\setminus\{g\}}\bar{c}_{i,\hat{l}_m}\mathbf{H}_i\tilde{\mathbf{A}}_m\mathbf{H}_i^H + \sigma_E^2\mathbf{I}_{n_{E,i}},\\ \sum_{m\in\mathcal{G}_E\setminus\{g\}}\bar{c}_{i,\hat{l}_m}\mathbf{H}_i\tilde{\mathbf{A}}'_m\mathbf{H}_i^H + \sigma_E^2\mathbf{I}_{n_{E,i}}\end{array}\right).
\end{aligned}
$$

References

1. Checko, A.; Chrisiansen, H.L.; Yan, Y.; Scolari, L.; Kardaras, G.; Berger, M.S.; Dittmann, L. Cloud RAN for mobile networks - A technology overview. *IEEE Commun. Suv. Tutorials* **2015**, *17*, 405–426.
2. Simeone, O.; Maeder, A.; Peng, M.; Sahin, O.; Yu, W. Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems. *J. Commun. Netw.* **2016**, *18*, 135–149.
3. Del la Oliva, A.; Hernandez, J.A.; Larrabeiti, D.; Azcorra, A. An overview of the CPRI specification and its application to C-RAN-based LTE scenarios. *IEEE Commun. Mag.* **2016**, *54*, 152–159. [CrossRef]
4. Park, S.-H.; Simeone, O.; Sahin, O.; Shamai, S. Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks. *IEEE Trans. Signal Process.* **2013**, *61*, 5646–5658. [CrossRef]
5. Park, S.-H.; Simeone, O.; Sahin, O.; Shamai, S. Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory. *IEEE Signal Process. Mag.* **2014**, *31*, 69–79. [CrossRef]
6. Park, S.; Chae, C.-B.; Bakh, S. Large-scale antenna operation in heterogeneous cloud radio access networks: A partial centralization approach. *IEEE Wirel. Commun.* **2015**, *22*, 32–40. [CrossRef]
7. Zhou, Y.; Yu, W. Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN. *IEEE Trans. Signal Process.* **2016**, *64*, 4138–4151. [CrossRef]
8. Liu, L.; Yu, W. Cross-layer design for downlink multihop cloud radio access networks with network coding. *IEEE Trans. Signal Process.* **2017**, *65*, 1728–1740. [CrossRef]

9. Park, S.-H.; Simeone, O.; Shamai, S. Multi-tenant C-RAN with spectrum pooling: Downlink optimization under privacy constraints. *IEEE Trans. Veh. Technol.* **2018**, *67*, 10492–10503. [CrossRef]
10. Park, S.-H.; Simeone, O.; Shamai, S. Robust baseband compression against congestion in packet-based fronthaul networks using multiple description coding. *Entropy* **2019**, *21*, 1–15. [CrossRef]
11. Sanguinetti, L.; Moustakas, A.L.; Debbah, M. Interference management in 5G reverse TDD HetNets with wireless backhaul: A large system analysis. *IEEE J. Sel. Areas Commun.* **2015**, *33*, 1187–1200. [CrossRef]
12. Vu, T.X.; Nguyen, T.V.; Quek, T.Q.S. Power optimization with BLER constraint for wireless fronthauls in C-RAN. *IEEE Commun. Lett.* **2016**, *20*, 602–605. [CrossRef]
13. Park, S.-H.; Lee, K.-J.; Song, C.; Lee, I. Joint design of fronthaul and access links for C-RAN with wireless fronthauling. *IEEE Signal Process. Lett.* **2016**, *23*, 1657–1661. [CrossRef]
14. Stephen, R.G.; Zhang, R. Joint millimeter-wave fronthaul and OFDMA resource allocation in ultra-dense CRAN. *IEEE Trans. Commun.* **2017**, *65*, 1411–1423. [CrossRef]
15. Park, S.-H.; Song, C.; Lee, K.-J. Inter-cluster design of wireless fronthaul and access links for the downlink of C-RAN. *IEEE Wireless Commun. Lett.* **2017**, *6*, 270–273. [CrossRef]
16. Hu, B.; Hua, C.; Zhang, J.; Chen, C.; Guan, X. Joint fronthaul multicast beamforming and user-centric clustering in downlink C-RANs. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 5395–5409. [CrossRef]
17. Hu, B.; Hua, C.; Chen, C.; Guan, X. Joint beamformer design for wireless fronthaul and access links in C-RANs. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2869–2881. [CrossRef]
18. Anokye, P.; Ahiadormey, R.K.; Song, C.; Lee, K.-J. Achievable sum-rate analysis of massive MIMO full-duplex wireless backhaul links in heterogeneous cellular networks. *IEEE Access* **2018**, *6*, 23456–23469. [CrossRef]
19. Kalfas, G.; Vagionas, C.; Antonopoulos, A.; Kartsakli, E.; Mesodiakaki, A.; Papaioannou, S.; Maniotis, P.; Vardakas, J.S.; Verikoukis, C.; Pleros, N. Next generation fiber-wireless fronthaul for 5G mmWave networks. *IEEE Commun. Mag.* **2019**, *57*, 138–144. [CrossRef]
20. Beyene, Y.D.; Jantti, R.; Tirkkonen, O.; Ruttik, K.; Iraji, S.; Larmo, A.; Tirronen, T.; Torsner, J. NB-IoT technology overview and experience from cloud-RAN implementation. *IEEE Wirel. Commun.* **2017**, *24*, 26–32. [CrossRef]
21. Liu, D.; Yang, C. Energy efficiency of downlink networks with caching at base stations. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 907–922. [CrossRef]
22. Pedarsani, R.; Maddah-Ali, M.A.; Niesen, U. Online coded caching. *IEEE/ACM Trans. Netw.* **2016**, *24*, 836–845. [CrossRef]
23. Park, S.-H.; Simeone, O.; Shamai, S. Joint cloud and edge processing for latency minimization in fog radio access networks. In Proceedings of the 2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Edinburgh, UK, 3–6 July 2016; pp. 1–5.
24. Zhou, B.; Cui, Y.; Tao, M. Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 6284–6297. [CrossRef]
25. Tao, M.; Chen, E.; Zhou, H.; Yu, W. Content-centric sparse multicast beamforming for cache-enabled cloud RAN. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 6118–6131. [CrossRef]
26. Park, S.-H.; Simeone, O.; Shamai, S. Joint optimization of cloud and edge processing for fog radio access networks. *IEEE Trans. Wireless Commun.* **2016**, *15*, 7621–7632. [CrossRef]
27. Sengupta, A.; Tandon, R.; Simeone, O. Fog-aided wireless networks for content delivery: Fundamental latency trade-offs. *IEEE Trans. Inf. Theory* **2017**, *63*, 6650–6678. [CrossRef]
28. Chen, D.; Al-Shatri, H.; Mahn, T.; Klein, A.; Kuehn, V. Energy efficient robust F-RAN downlink design for hard and soft fronthauling. In Proceedings of the 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal, 3 June 2018; 1–5.
29. Tran, T.X.; Hajisami, A.; Pompili, D. Cooperative hierarchical caching in 5G cloud radio access networks. *IEEE Netw.* **2017**, *31*, 35–41. [CrossRef]
30. Kim, J. Energy-efficient dynamic packet downloading for medical IoT platforms. *IEEE Trans. Ind. Infor.* **2015**, *11*, 1653–1659. [CrossRef]
31. Dai, B.; Yu, W. Energy efficiency of downlink transmission strategies for cloud radio access networks. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 1037–1050. [CrossRef]
32. Lee, S.-R.; Jung, J.; Park, H.; Lee, I. A new energy-efficient beamforming strategy for MISO interfering broadcast channels based on large systems analysis. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 2872–2882. [CrossRef]

33. Rostampoor, J.; Razavizadeh, S.M.; Lee, I. Energy efficient precoding design for SWIPT in MIMO two-way relay networks. *IEEE Trans. Veh. Technol.* **2017**, *66*, 7888–7896. [CrossRef]

34. Vu, T.T.; Ngo, D.T.; Dao, M.N.; Durrani, S.; Nguyen, D.H.N.; Middleton, R.H. Energy efficiency maximization for downlink cloud radio access networks with data sharing and data compression. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 4955–4970. [CrossRef]

35. You, C.; Huang, K.; Chae, H. Energy efficient mobile cloud computing powered by wireless energy transfer. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 1757–1771. [CrossRef]

36. Vazifehdan, J.; Prasad, R.V.; Niemegeers, I. Energy-efficient reliable routing considering residual energy in wireless ad hoc networks. *IEEE Trans. Mob. Comput.* **2014**, *13*, 434–447. [CrossRef]

37. Antonopoulos, A.; Verikoukis, C. Multi-player game theoretic MAC strategies for energy efficient data dissemination. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 592–603. [CrossRef]

38. Darsena, D.; Gelli, G.; Verde, F. Cloud-aided cognitive ambient backscatter wireless sensor networks. *IEEE Access.* **2019**, *7*, 57399–57414. [CrossRef]

39. Lee, S.-R.; Park, S.-H.; Lee, I. NOMA systems with content-centric multicast transmission for C-RAN. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 828–831. [CrossRef]

40. Park, S.-H.; Simeone, O.; Lee, W.; Shamai, S. Coded multicast fronthauling and edge caching for multi-connectivity transmission in fog radio access networks. In Proceedings of the 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Sapporo, Japan, 3–6 July 2017; pp. 1–5.

41. Shen, K.; Yu, W. Fractional programming for communication systems—Part I: Power control and beamforming. *IEEE Trans. Signal Process* **2018**, *66*, 2616–2630. [CrossRef]

42. Sun, Y.; Babu, P.; Palomar, D.P. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.* **2017**, *65*, 794–816. [CrossRef]

43. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.