# Joint Cloud and Edge Processing for Latency Minimization in Fog Radio Access Networks

Seok-Hwan Park[1], Osvaldo Simeone[2] and Shlomo Shamai (Shitz)[3]
[1]Division of Electronic Engineering, Chonbuk National University, Jeonju, 54896 Korea
[2]CWCSPR, New Jersey Institute of Technology, 07102 Newark, New Jersey, USA
[3]Department of Electrical Engineering, Technion, Haifa, 32000, Israel
Email: seokhwan@jbnu.ac.kr, osvaldo.simeone@njit.edu, sshlomo@ee.technion.ac.il

***Abstract*—This work studies the joint design of cloud and edge processing for latency minimization in the downlink of a fog radio access network (F-RAN). In an F-RAN, edge processing allows the low-latency delivery of popular multimedia content by leveraging caching at enhanced remote radio heads (eRRHs). Cloud processing, instead, enables the transmission of arbitrary content at high spectral efficiencies thanks to the centralized control at a baseband processing unit (BBU), but at the cost of a potentially larger latency due to BBU-to-eRRHs communication over fronthaul links. For an arbitrary caching, or pre-fetching, strategy, the design of the delivery phase is studied based on the use of the fronthaul links in either or both *hard-transfer* and *soft-transfer* modes. The problem of minimizing the delivery latency, encompassing both fronthaul and wireless transmissions, of the requested contents from the BBU to the requesting user equipments (UEs) is tackled with respect to channel precoding and fronthaul compression strategies. Numerical results are provided to compare the latency performance of the hard- and soft-transfer fronthauling schemes in terms of delivery latency, offering new insights as compared to existing studies that focus solely on the transmission rate of the wireless segment.**

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

## I. Introduction

The cloud radio access network (C-RAN) architecture benefits from the centralization gains, in terms of spectral efficiency and reduced cost, that come from the virtualization of the baseband processing functionalities of remote radio heads (RRHs) at a centralized baseband signal processing unit (BBU) [1][2]. A key downside of this solution is its reliance on high-rate fronthaul links between the BBU and the RRHs. With state-of-the-art technologies, these links may become the main performance bottleneck of a C-RAN system per critical indicators such as spectral efficiency and latency. To address these limitations, an evolved network architecture, referred to as *Fog Radio Access Network* (F-RAN), has been recently proposed, which enhances a C-RAN system by equipping the RRHs with caching and signal processing functionalities [3]-[5] (see also [6]), as shown in Fig. 1. We refer to the resulting RRHs as *enhanced RRHs* (eRRHs).

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

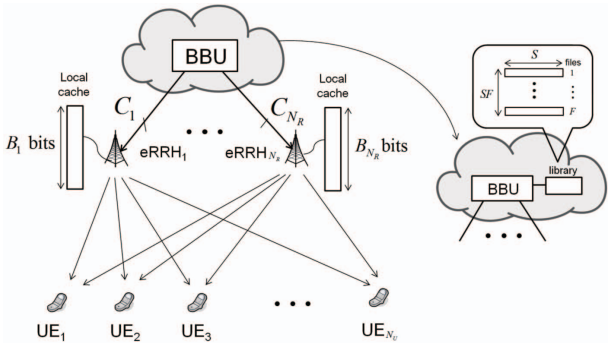


Figure 1. Illustration of an F-RAN, which has both cloud and edge processing capabilities: the BBU in the cloud can perform joint baseband processing and the eRRHs are equipped with local caches.

An F-RAN follows the standard operating phases of cache-aided systems (see, e.g., [7]-[14]), namely the pre-fetching and the delivery phases. Pre-fetching operates at a large time scale, which encompasses multiple transmission intervals, corresponding to the period in which content popularity remains constant. Instead, the delivery phase operates separately on each transmission interval based on the file messages cached during the pre-fetching phase. The fronthaul-aware design of the pre-fetching or delivery phases can be studied under the assumption that the fronthaul links are leveraged in a *hard-transfer mode* as in [8]-[12] or *soft-transfer mode* as in [5]. The hard-transfer mode amounts to the transmission to the eRRHs of the requested content that is not present in the local caches. Instead, the soft-transfer mode delivers a quantized version of the baseband signals encoded at the BBU to the eRRHs as in C-RAN systems (see, e.g., [1]). Information-theoretic considerations on F-RANs can be found in [14] (see also [9][13]).

Prior works [5],[7]-[12] investigated the optimization of precoding and fronthaul transmission strategies under different criteria, namely minimum delivery rate [5], degrees-of-freedom (DoF) [7][11] or compound network energy cost [9][10][12]. These performance metrics are suitable to assess the impact of caching or the spectral or energy efficiency of cloud-assisted systems under fronthaul capacity limitations.

However, they do not target the critical performance indicator of delivery latency. Specifically, in the formulation of [5], which assumes a fixed fronthaul capacity constraint, the amount of the information that the fronthaul link can deliver to eRRHs is given and not dependent on the allowed delivery latency, which is also not accounted for by the network cost considered in [9].

To fill the gap identified above, this work studies the design of the delivery phase in an F-RAN with the goal of minimizing the delivery latency by accounting for both the *fronthaul latency* for transmission between BBU and eRRHs and the *edge latency* for wireless transmission between eRRHs and user equipments (UEs). In the model under study of this work, the amount of the information transferred on the fronthaul links depends on the fronthaul latency, which is to be optimized by taking into account the cached content and the resulting latency on the wireless channel. As in [5], we consider a hybrid fronthauling mode which includes as special cases both hard- and soft-transfer fronthauling modes. An iterative algorithm is derived based on the concave-convex procedure (CCCP), which is used to provide numerical results that compare the latency performance of hard- and soft-transfer fronthauling.

## II. System Model

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. In addition to the functionalities performed by conventional RRHs in C-RAN, each eRRH $i$ in an F-RAN is equipped with a cache that can store $B_i$ bits. Furthermore, it also has baseband processing capabilities. Each eRRH $i$ is connected to the BBU with a fronthaul link of capacity $C_i$ bit per symbol of the downlink channel for $i \in \mathcal{N}_R \triangleq \{1, \ldots, N_R\}$. We denote the numbers of antennas of eRRH $i$ and UE $k$ by $n_{R,i}$ and $n_{U,k}$, respectively, and define the notation $n_R \triangleq \sum_{i \in \mathcal{N}_R} n_{R,i}$.

We consider communication for content delivery via the outlined F-RAN system. Accordingly, UEs request contents, or files, from a library of $F$ files, each of size $S$ bits, which are delivered by the network across a number of transmission intervals. Labeling the files in order of popularity, the probability $P(f)$ of a file $f$ to be selected is defined by Zipf's distribution $P(f) = cf^{-\gamma}$ for $f \in \mathcal{F} \triangleq \{1, \ldots, F\}$, where $\gamma \geq 0$ is a given popularity exponent and $c \geq 0$ is set such that $\sum_{f \in \mathcal{F}} P(f) = 1$. In any transmission interval, each UE $k$ requests file $f_k \in \mathcal{F}$ with probability $P(f = f_k)$, and the requested files $f_k$ are independent across the index $k$. Files are typically chunks of videos and can hence be assumed to be transmitted within a given transmission interval.

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

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

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

The system operates in two phases, namely pre-fetching and delivery (see, e.g., [7]). In the **pre-fetching phase**, that operates at a large time scale corresponding to the period in which file popularity remains constant, each eRRH $i$ downloads and stores up to $B_i$ bits from the library of files, which is of size $SF$ bits. We define the *fractional caching capacity* $\mu_i$ of eRRH $i$ as

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

Accordingly, each eRRH can potentially store a fraction $\mu_i$ of each file.

The pre-fetching policy chooses $nB_i$ bits out of the library of $SF$ bits to be stored in the cache of eRRH $i$. We focus on uncoded caching strategies, and, 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 $S_l$ bits with $\sum_{l \in \mathcal{L}} S_l = S$ and $\mathcal{L} \triangleq \{1, \ldots, L\}$ (see, e.g., [7, Sec. III]). Then, the pre-fetching strategy can be modeled by defining binary caching variables $\{c^i_{f,l}\}_{f \in \mathcal{F}, l \in \mathcal{L}, i \in \mathcal{N}_R}$ as

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

while satisfying the cache memory constraint at eRRH $i$ as

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

In the numerical results in Sec. V, we will focus on a baseline fractional randomized pre-fetching strategy, in which each file $f$ is split into $N_R$ disjoint fragments of equal size, i.e., $L = N_R$ and $S_l = S/L$ for all $l \in \mathcal{L}$, that are distributed over eRRHs chosen randomly without replacement (see [5, Sec. III-C]).

In the **delivery phase**, that operates separately on each transmission interval, 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. As will be detailed in Sec. III, we are interested in minimizing the duration of transmission intervals, which we refer to as delivery latency.

## III. Delivery Phase With Hybrid Fronthauling

In this section, we consider the design of the delivery phase in each transmission interval under a hybrid fronthauling mode, whereby the capacity of each fronthaul link is used to carry both hard and soft information about the uncached files.

### A. Hybrid Fronthauling

We first describe precoding under hybrid fronthauling mode. Hard-mode fronthauling requires the determination of the set of eRRHs to which each subfile $(f,l)$ is transferred on the fronthaul link. As in [5], this is done by defining the binary variable $d^i_{f,l}$ as

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

Allowing for both hard- and soft-transfer fronthauling, the signal $\mathbf{x}_i$ transmitted by eRRH $i$ on the downlink channel is given as the superposition of a locally encoded signal, based on the files in the cache or received via hard-transfer fronthauling, and of a signal obtained from the BBU by means of soft-transfer fronthauling. This yields

$$\mathbf{x}_i = \sum_{f\in\mathcal{F}_{\rm req}} \sum_{l\in\mathcal{L}} \left(1 - \bar{c}^i_{f,l}\bar{d}^i_{f,l}\right) \mathbf{V}^i_{f,l}\mathbf{s}_{f,l} + \hat{\mathbf{x}}_i, \tag{6}$$

where $\mathbf{V}^i_{f,l} \in \mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix for the baseband signal $\mathbf{s}_{f,l} \in \mathbb{C}^{n_{S,f,l}\times 1}$, which encodes the subfile $(f,l)$ available at the eRRH, from the cache or the fronthaul, which is distributed as $\mathbf{s}_{f,l} \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, while $\hat{\mathbf{x}}_i$ represents the quantized baseband signal received from the BBU on the fronthaul link via soft-transfer mode. For a binary variable $a \in \{0,1\}$, $\bar{a}$ is defined as $1-a$. Note that hard-transfer fronthauling is obtained by setting $\hat{\mathbf{x}}_i = \mathbf{0}$ for all $i \in \mathcal{N}_R$, and soft-transfer mode is obtained by setting $d^i_{f,l} = 0$ for all $f \in \mathcal{F}_{\rm req}$, $l \in \mathcal{L}$ and $i \in \mathcal{N}_R$.

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

$$\tilde{\mathbf{x}}_i = \sum_{f\in\mathcal{F}_{\rm req}} \sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}\mathbf{s}_{f,l}, \tag{7}$$

where $\mathbf{U}^i_{f,l} \in \mathbb{C}^{n_{R,i}\times n_{S,f,l}}$ is the precoding matrix for the baseband signal $\mathbf{s}_{f,l}$ that encodes the fragment $(f,l)$ not available at eRRH $i$. The signal $\tilde{\mathbf{x}}_i$ is quantized, obtaining the signal $\hat{\mathbf{x}}_i$ in the right-hand side of (6) as

$$\hat{\mathbf{x}}_i = \tilde{\mathbf{x}}_i + \mathbf{q}_i, \tag{8}$$

where $\mathbf{q}_i$ denotes the quantization noise, which is independent of $\tilde{\mathbf{x}}_i$ and distributed as $\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{\Omega}_i)$, with covariance matrix $\mathbf{\Omega}_i \succeq \mathbf{0}$. We assume that the signals $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ for different eRRHs $i \neq j$ are quantized independently (see also [2] for more general strategies).

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

$$\begin{aligned} R_{f_k,l} &= q_{k,l}\left(\bar{\mathbf{V}},\mathbf{\Omega}\right) \triangleq I\left(\mathbf{s}_{f_k,l};\mathbf{y}_k|\mathbf{s}_{f_k,1},\ldots,\mathbf{s}_{f_k,l-1}\right) \\ &= \Phi\left(\begin{array}{c} \mathbf{H}_k\bar{\mathbf{V}}_{f_k,l}\bar{\mathbf{V}}^\dagger_{f_k,l}\mathbf{H}^\dagger_k, \\ \sum_{m=l+1}^{L}\mathbf{H}_k\bar{\mathbf{V}}_{f_k,m}\bar{\mathbf{V}}^\dagger_{f_k,m}\mathbf{H}^\dagger_k + \\ \sum_{f\in\mathcal{F}_{\rm req}\setminus\{f_k\}}\sum_{m\in\mathcal{L}}\mathbf{H}_k\bar{\mathbf{V}}_{f,m}\bar{\mathbf{V}}^\dagger_{f,m}\mathbf{H}^\dagger_k \\ +\mathbf{H}_k\bar{\mathbf{\Omega}}\mathbf{H}^\dagger_k + N_0\mathbf{I} \end{array}\right) \end{aligned} \tag{9}$$

where we defined the aggregated precoding matrix $\bar{\mathbf{V}}_{f,l} \triangleq [\bar{\mathbf{V}}^1_{f,l};\ldots;\bar{\mathbf{V}}^{N_R}_{f,l}]$ for subfile $(f,l)$ with $\bar{\mathbf{V}}^i_{f,l} \triangleq (1-\bar{c}^i_{f,l}\bar{d}^i_{f,l})\mathbf{V}^i_{f,l} + \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}$, and the notations $\bar{\mathbf{V}} \triangleq \{\bar{\mathbf{V}}_{f,l}\}_{f\in\mathcal{F}_{\rm req},\,l\in\mathcal{L}}$, $\bar{\mathbf{\Omega}} \triangleq \mathrm{diag}(\mathbf{\Omega}_1,\ldots,\mathbf{\Omega}_{N_R})$ and $\Phi(\mathbf{A},\mathbf{B}) \triangleq \log\det(\mathbf{A}+\mathbf{B}) - \log\det(\mathbf{B})$. We note that, if a file $f'$ is requested by multiple UEs, it is multicast to all the requesting UEs, and hence the rate $R_{f',l}$ of each subfile $(f',l)$ is limited by the minimum of the functions $q_{k,l}(\bar{\mathbf{V}},\mathbf{\Omega})$ for all requesting UEs $k$.

### B. Delivery Latency

Let us define as $n_{E,f,l}$ the number of symbols, or channel uses, needed to transmit a subfile $(f,l)$ on the wireless channel to all requesting UEs. Assuming that $n_{E,f,l}$ is large enough to enable the use of the information-theoretic metrics in (9), this can be computed as

$$n_{E,f,l} = \frac{S_l}{R_{f,l}} = \max_{k\in\mathcal{N}_{U,f}} \frac{S_l}{q_{k,l}\left(\bar{\mathbf{V}},\mathbf{\Omega}\right)}, \tag{10}$$

where $\mathcal{N}_{U,f} \triangleq \{k \in \mathcal{N}_U | f_k = f\}$ is the set of the UEs requesting the $f$th content, and the second equality comes from (9). Considering all subfiles, we obtain the total number of channel uses on the wireless channel as

$$n_E = \max_{f\in\mathcal{F}_{\rm req},l\in\mathcal{L}} n_{E,f,l}, \tag{11}$$

where the maximization is over all requested subfiles $(f,l)$.

For fronthaul transmission, as explained, the BBU needs to send $S_{H,i} = \sum_{f\in\mathcal{F}_{\rm req}}\sum_{l\in\mathcal{L}} d^i_{f,l}S_l$ bits to eRRH $i$ to transmit part of the requested contents via hard-transfer fronthauling. Soft-transfer fronthauling, instead, requires the BBU to send a number $g_i(\mathbf{U},\mathbf{\Omega})$ of bits per sample for the signal $\hat{\mathbf{x}}_i$ in (8) that depends on the precoding $\mathbf{U} \triangleq \{\mathbf{U}^i_{f,l}\}_{f\in\mathcal{F}_{\rm req},l\in\mathcal{L},i\in\mathcal{N}_R}$ and on the quantization covariance matrices $\mathbf{\Omega} \triangleq \{\mathbf{\Omega}_i\}_{i\in\mathcal{N}_R}$. To obtain the rate $g_i(\mathbf{U},\mathbf{\Omega})$, we use standard rate-distortion arguments as in, e.g., [15, Ch. 3], to write

$$\begin{aligned} g_i\left(\mathbf{U},\mathbf{\Omega}\right) &\triangleq I\left(\tilde{\mathbf{x}}_i;\hat{\mathbf{x}}_i\right) \\ &= \Phi\left(\sum_{f\in\mathcal{F}_{\rm req}}\sum_{l\in\mathcal{L}} \bar{c}^i_{f,l}\bar{d}^i_{f,l}\mathbf{U}^i_{f,l}\mathbf{U}^{i\dagger}_{f,l},\mathbf{\Omega}_i\right). \end{aligned} \tag{12}$$

The latency $n_{F,i}$ on the fronthaul link to eRRH $i$ is then given as

$$n_{F,i} = \frac{S_{H,i} + n_E g_i(\mathbf{U},\mathbf{\Omega})}{C_i}, \tag{13}$$

which is normalized to the duration of a wireless channel symbol. Therefore, the latency $n_F$ of the fronthaul transmission from the BBU to all the eRRHs can be written as $n_F = \max_{i\in\mathcal{N}_R} n_{F,i}$ and the total latency as

$$n_T = n_E + n_F = \max_{f\in\mathcal{F}_{\rm req}, l\in\mathcal{L}} n_{E,f,l} + \max_{i\in\mathcal{N}_R} n_{F,i}. \quad (14)$$

## IV. Problem Definition and Optimization

In this work, we aim at minimizing the total latency $n_T$ in (14) subject to per-eRRH fronthaul capacity and transmit power constraints. The problem is stated as

$$\underset{n_E,n_F,\bar{\mathbf{V}},\mathbf{R},\mathbf{\Omega}}{\text{minimize}} \ n_E + n_F \quad (15a)$$

$$\text{s.t. } n_E \geq \frac{S_l}{R_{f,l}}, \ f\in\mathcal{F}_{\rm req}, l\in\mathcal{L}, \quad (15b)$$

$$n_F \geq \frac{n_E g_i\left(\bar{\mathbf{V}},\mathbf{\Omega}\right) + S_{H,i}}{C_i}, \ i\in\mathcal{N}_R, \quad (15c)$$

$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}},\mathbf{\Omega}\right), \ l\in\mathcal{L}, k\in\mathcal{N}_U, \quad (15d)$$

$$\sum_{f\in\mathcal{F}_{\rm req}}\sum_{l\in\mathcal{L}} \mathrm{tr}\left(\mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^\dagger \mathbf{E}_i + \mathbf{\Omega}_i\right) \leq P_i, \ i\in\mathcal{N}_R, \quad (15e)$$

where the function $g_i(\bar{\mathbf{V}},\mathbf{\Omega})$ in (15c) is defined by substituting $\mathbf{U}_{f,l}^i = \mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}$ into $g_i\left(\mathbf{U},\mathbf{\Omega}\right)$ in (12). Note that the prefetching variables (3), the fronthaul transfer variables (5) and the parameters $\{S_{H,i}\}_{i\in\mathcal{N}_R}$ are fixed.

Problem (15) is non-convex due to the non-convexity of the constraints (15b)-(15d). However, we first observe that the constraint (15b) is equivalent to the convex constraint

$$\log n_E + \log R_{f,l} \geq \log S_l, \ f\in\mathcal{F}_{\rm req}, l\in\mathcal{L}, \quad (16)$$

and that (15c) can be written as

$$\log n_F + \log C_i \geq \log\left(n_E g_i\left(\bar{\mathbf{V}},\mathbf{\Omega}\right) + S_{H,i}\right). \quad (17)$$

We can further see that the constraint (17) is satisfied if and only if there exist numbers $\lambda_i, \beta_i$ satisfying the conditions

$$\log n_F + \log C_i \geq \log\left(\lambda_i + S_{H,i}\right), \quad (18)$$

$$\log \lambda_i \geq \log n_E + \log\left(\beta_i\right), \quad (19)$$

$$\text{and } \beta_i \geq g_i\left(\bar{\mathbf{V}},\mathbf{\Omega}\right). \quad (20)$$

Therefore, the problem (15) can be restated equivalently as

$$\underset{n_E,n_F,\bar{\mathbf{V}},\mathbf{R},\mathbf{\Omega},\boldsymbol{\lambda},\boldsymbol{\beta}}{\text{minimize}} \ n_E + n_F \quad (21a)$$

$$\text{s.t. } \log n_E + \log R_{f,l} \geq \log S_l, \ f\in\mathcal{F}_{\rm req}, \ l\in\mathcal{L}, \quad (21b)$$

$$\log n_F + \log C_i \geq \log\left(\lambda_i + S_{H,i}\right), \ i\in\mathcal{N}_R, \quad (21c)$$

$$\log \lambda_i \geq \log n_E + \log\left(\beta_i\right), \ i\in\mathcal{N}_R, \quad (21d)$$

$$\beta_i \geq g_i\left(\bar{\mathbf{V}},\mathbf{\Omega}\right), \ i\in\mathcal{N}_R, \quad (21e)$$

$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}},\mathbf{\Omega}\right), \ l\in\mathcal{L}, k\in\mathcal{N}_U, \quad (21f)$$

$$\sum_{f\in\mathcal{F}_{\rm req}}\sum_{l\in\mathcal{L}} \mathrm{tr}\left(\mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^\dagger \mathbf{E}_i + \mathbf{\Omega}_i\right) \leq P_i, \ i\in\mathcal{N}_R, \quad (21g)$$

where we defined the notations $\boldsymbol{\lambda} \triangleq \{\lambda_i\}_{i\in\mathcal{N}_R}$ and $\boldsymbol{\beta} \triangleq \{\beta_i\}_{i\in\mathcal{N}_R}$. Noting that the constraints (21c)-(21f) have the difference-of-convex (DC) structure when stated in terms of the covariance matrices $\mathbf{W}_{f,l} \triangleq \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^\dagger \succeq \mathbf{0}$ and $\mathbf{\Omega}$, as in [2][5][9], we can adopt the concave-convex procedure (CCCP) to derive an alternative algorithm, whose convergence to local minimum was studied in [9]. The derivation follows standard steps and is omitted here due to space restriction.

As special cases, we obtain:

*1) Hard-Transfer Fronthauling:* To implement the hard-transfer fronthauling mode, we set the quantization covariance matrices $\mathbf{\Omega}_i = \mathbf{0}$ for all $i\in\mathcal{N}_R$, and fix the latency $n_F$ of the fronthaul link to $n_F = \max_{i\in\mathcal{N}_R} S_{H,i}/C_i$. We hence obtain the problem

$$\underset{n_E,\bar{\mathbf{V}},\mathbf{R}}{\text{minimize}} \ n_E + n_F \quad (22a)$$

$$\text{s.t. } n_E \geq \frac{S_l}{R_{f,l}}, \ f\in\mathcal{F}_{\rm req}, l\in\mathcal{L}, \quad (22b)$$

$$R_{f_k,l} \leq q_{k,l}\left(\bar{\mathbf{V}},\mathbf{\Omega}=\mathbf{0}\right), \ l\in\mathcal{L}, k\in\mathcal{N}_U, \quad (22c)$$

$$\sum_{f\in\mathcal{F}_{\rm req}}\sum_{l\in\mathcal{L}} \mathrm{tr}\left(\mathbf{E}_i^\dagger \bar{\mathbf{V}}_{f,l}\bar{\mathbf{V}}_{f,l}^\dagger \mathbf{E}_i\right) \leq P_i, \ i\in\mathcal{N}_R. \quad (22d)$$

*2) Soft-Transfer Fronthauling:* For the soft-transfer fronthauling mode, the fronthaul transfer variables $d_{f,l}^i$ are set to $d_{f,l}^i = 0$ for all $f\in\mathcal{F}_{\rm req}$ and $l\in\mathcal{L}$ so that $S_{H,i} = 0$ for all $i\in\mathcal{N}_R$. The optimization of the remaining variables $n_E$, $n_F$, $\bar{\mathbf{V}}$, $\mathbf{R}$ and $\mathbf{\Omega}$ can be tackled similar to the problems (15) and (22).

## V. Numerical Results

For numerical results presented in this section, we assume that the positions of eRRHs and UEs are uniformly distributed within a circular cell of radius 500m. We consider the channel model $\mathbf{H}_{k,i} = \sqrt{\rho_{k,i}}\tilde{\mathbf{H}}_{k,i}$, where the channel power $\rho_{k,i}$ is given as $\rho_{k,i} = 1/(1+(d_{k,i}/50)^3)$ with $d_{k,i}$ denoting the distance between eRRH $i$ and UE $k$ and the elements of $\tilde{\mathbf{H}}_{k,i}$ are independent and identically distributed as $\mathcal{CN}(0,1)$. We consider a symmetric setting where the eRRHs have the same transmit power and fronthaul capacity, i.e., $P_i = P$ and $C_i = C$ for $i\in\mathcal{N}_R$ and are equipped with caches of equal size, i.e., $\mu_i = \mu$ for $i\in\mathcal{N}_R$. The fronthaul transfer variables $\{d_{f,l}^i\}_{f\in\mathcal{F}_{\rm req},l\in\mathcal{L}}$ of the hard-transfer mode are set such that the subfile $(f_k,l)$ requested by UE $k$ is transferred on the fronthaul links to the $N_F$ eRRHs that have the largest channel gains $||\mathbf{H}_{k,i}||_F^2$ to the UE and have not stored the subfile. The parameter $N_F$ is optimized to minimize the average latency. We focus on the case with $N_R = 3$, $N_U = 6$, $F = 6$, $S = 1$ MB, $\gamma = 0.5$ $n_{R,i} = n_{U,k} = 1$ and $N_0 = 1$.

We first study the performance comparison among delivery strategies with hard- and soft-transfer fronthauling modes by plotting in Fig. 2 the average latency $n_T$ versus the fronthaul capacity $C$ for an F-RAN system with $\mu\in\{0,1/3,1\}$ and $P = 30$ dB. For reference, we also plot the latency obtained with no caching ($\mu = 0$) or full caching ($\mu = 1$).

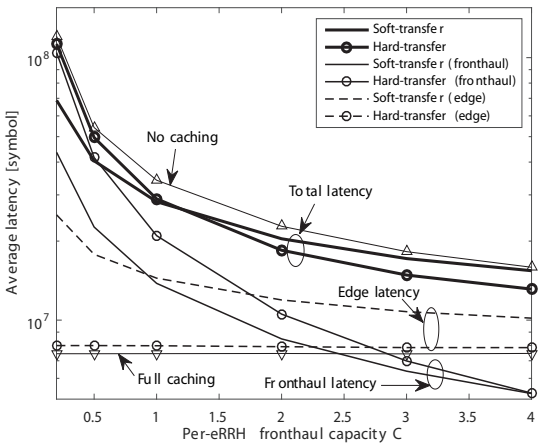


Figure 2. Average latency $n_T$ versus the fronthaul capacity $C$ for an F-RAN downlink ($\mu = 0, 1/3, 1$ and $P = 30$ dB).

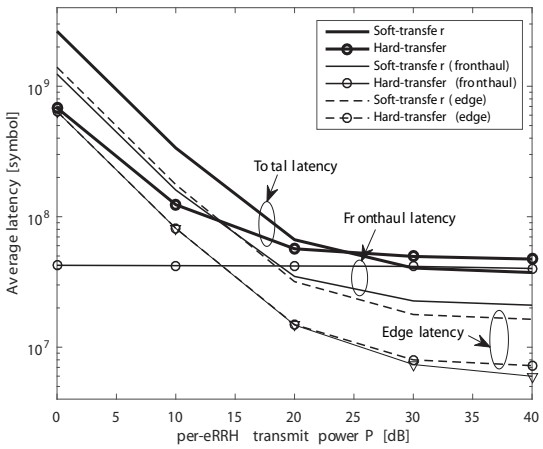


Figure 3. Average latency $n_T$ versus the SNR $P$ for an F-RAN downlink ($\mu = 1/3, 1$ and $C = 0.5$).

For the case of no caching, the hybrid fronthauling mode is assumed, whereas this is not shown for partial caching in order to enhance the legibility of the figures, and since the gains as compared to the best between soft- and hard-transfer fronthauling were seen to be minor. It is also noted that, for the case of full caching, the latency only encompasses the edge contribution.

It is observed from the figure that the soft-transfer mode outperforms the hard-transfer mode when the fronthaul capacity $C$ is small. Specifically, as $C$ decreases, the contribution to the latency due to fronthaul transmission is seen to grow more rapidly than the edge latency for both hard- and soft-transfer modes, and soft-transfer fronthauling is demonstrated to be more efficient in the use of fronthaul resources by means of baseband compression. We also note that the performance gap between the partial and full caching schemes decreases with the fronthaul capacity $C$, since the lack of cooperation opportunities on the cached files can be compensated for by increasing $C$.

We then investigate in Fig. 3 the effect of the signal-to-noise ratio (SNR) $P$ of the wireless link on the average latency performance. Comparing between hard- and soft-transfer modes, we can see that increasing $P$ for fixed fronthaul capacity $C$ yields a regime, similar to the case of low $C$, in which the fronthaul latency becomes the most relevant contribution to the overall latency. Under these conditions, it is again observed that soft-transfer fronthauling outperforms hard-transfer mode.

## VI. Conclusion

In this work, we have studied the joint design of cloud and edge processing for latency minimization in an F-RAN architecture. It is concluded that the soft-transfer fronthauling used in C-RAN implementation is to be preferred over the hard-transfer mode typical of backhaul links when the fronthaul latency becomes main bottleneck of the overall latency, that is for the cases with small fronthaul capacity or high SNR. In other regimes, hard-transfer fronthauling may provide a lower-latency solution. Interestingly, these conclusions are partly in contrast with the outcome of the analysis in [5], which shows that, when the goal is maximizing the minimum delivery rate on the wireless channel alone, soft-transfer fronthauling provides a more effective way to use fronthaul resources in most operating regimes.

## References


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

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

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

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

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

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

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

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

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

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

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

[12] D. Chen, S. Schedler and V. Kuehn, "Backhaul traffic balancing and dynamic content-centric clustering based on beamforming in the downlink of fog radio access network," arXiv:1602.05536, 2016.

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

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

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