Scheduling and Beamforming with Imperfect CSI in Fog Radio Access Networks

Nicolas Pontois

June 2017

1 Introduction

The fifth generation (5G) wireless system is expected to address many challenges, including an ever increasing number of communications and energy consumption limitations. To achieve such goals, ultra-dense small cell deployments and cloud computing are recognized as the two key enabling technologies. By deploying small cells, the energy consumption is reduced as users are closer to their base station (BS) and thus require a lower transmit power. However by deploying small cells, the inter-cell interferences are also enhanced as other BSs are also brought closer to users. To manage those interferences, the C-RAN architecture is often considered the key [1]. In C-RAN systems, the traditional BSs are replaced by three elements. The baseband units (BBUs) are clustered together into a BBU pool in a centralized cloud server. The two other elements are the distributively deployed remote radio heads (RRHs), and the fronthaul between the RRHs and the BBU pool. This architecture enables a joint optimization of network's messages encoding at the BBU level, using beamforming for instance to mitigate those interferences.

By moving the encoding decision farther away from users, it does enable an optimization using knowledge of the whole network. However it also means sharing channels' state information (CSIs) with the BBU pool. These CSIs are especially affected by network latency in C-RAN, as the fronthaul between eRRHs and the BBU pool does add latency. This means that an optimization performed at the BBU level is done using possibly outdated CSIs. Thus when moving the encoding decision to the BBU-level, there exists a trade-off between the number of CSIs available and the accuracy of these CSIs.

With regards to this trade-off, we can distinguish three possible allocation schemes. The first scheme is C-RAN, where everything is centralized and thus the allocation is performed globally but based on imperfect CSIs. The second scheme is RAN, where allocation is performed locally, at every BS, using exact CSI. The third scheme is called fog computing based radio access network (F-RAN).

The C-RAN architecture has received a lot of attention the last few years. Centralizing different BBUs into a BBU pool does provide many advantages

from a better use of resources to a better network scalability. The joint optimization of resources in particular has been studied extensively [2], [3]. The proposed algorithms do provide a noticeable performance upgrade over decentralized RAN algorithms with local optimization. While this shows the potential of global resource optimization, such algorithms see a steep decline in performances when also considering CSI uncertainty [4]. This prompted researchers to incorporate CSI imperfections in C-RAN and look for robust beamforming algorithms [5]. However the performances of the proposed beamformers are still quite far from those of the ideal situation where perfect CSI would be available at the BBU pool.

Taking into account this observation, the F-RAN architecture is a newly studied network architecture that could achieve better results in terms of resource optimization [6]. This architecture associates C-RAN with mobile edge computing (MEC). As in C-RAN, a BBU pool can perform joint baseband processing for RRHs they are connected to via fronthaul links. In addition in F-RAN the RRHs are equipped with a server, giving them additional capacities of caching and local cloud computing. These modified RRHs are refered to as enhanced RRHs (eRRHs) [7]. By providing MEC servers at eRRHs, this architecture adds caching potential to C-RAN. Cashing in F-RAN is important in order to reduce latency and especially to overcome fronthaul limitations, thus impacting rate allocation problems [7], [8].

In F-RAN, the allocation can be split into two parts: a first step that is performed at the BBU pool with access to global but outdated CSIs, and a second step, performed at each eRRH with access to local but perfect CSIs. This resource allocation split between the BBU pool and RRHs is especially appealing as the interest of control-data separation between the edge and the cloud in Fog-RAN has been proved [9]. In this article, looking at uplink communications and using imperfect CSI, it is shown that moving control functionalities to the edge, while performing joint decoding in the cloud yields potentially significant gains.

The difficulty then is to decide on a specific split, i.e. which functions should be performed at the BBU level and which ones should be put at the eRRH level. In this article, we present a specific split for a downlink F-RAN where global pre-scheduling is performed at the BBU pool, while each eRRH determines its own beamforming vectors.

In this split, the BBU pool decides of a users-to-eRRHs mapping. The idea behind this split is that the BBU computes a global optimization relying on its global but delayed CSIs. Then the pool assigns users to the eRRH they ended being served by in the globally optimal resource allocation. Then each eRRH computes its own beamforming vectors, using perfect CSIs. These beamformers could be identical to the ones computed at the BBU pool, netting no gain. However, eRRHs have access to exact CSIs and should be able to determine a better beamformer than the precomputed one.

Throughout this paper, vectors and matrices are denoted respectively by lower-case bold letters (e.g. $\mathbf{s}$) and upper-case bold letters (e.g. $\mathbf{W}$). We use $\mathbb{R}$ and $\mathbb{C}$ to denote real and complex domains respectively. The transpose, conjugate, conjugate transpose and trace of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^T$,

$\mathbf{A}^*$, $\mathbf{A}^H$ and $tr(\mathbf{A})$ respectively. In addition, $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The complex Gaussian distribution is represented by $\mathcal{CN}(.,.)$, while the expectation of a random variable is denoted by $\mathbb{E}[.]$. Calligraphy letters (e.g. $\mathcal{K}$) are used to denote sets, and $|.|$ represents the amplitude of a scalar. $max\_geig\big(\mathbf{A}, \mathbf{B}\big)$ returns the normalized dominant generalized eigenvector of matrix pair $\big(\mathbf{A}, \mathbf{B}\big)$.

2 System Model

2.1 Centralized Cloud RAN (C-RAN) Architecture

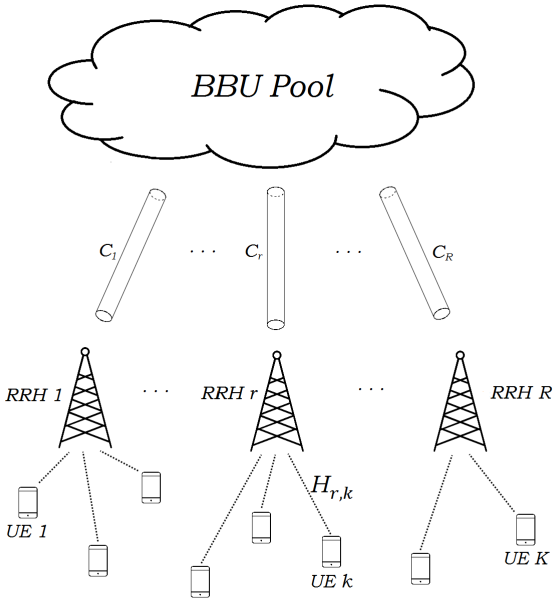

Figure 1: Cloud RAN architecture with optical fiber fronthaul

An architecture that has been considered for 5G is Centralized RAN. As depicted in figure 2, the system consists of one BBU pool linked to its R RRHs through optical fibers, of respective capacities C_r. This BBU pool has K users to serve within its coverage. Each RRH r is equipped with M_r transmit antennas. On the other side, each UE is considered to have only one unique receiving antenna.

In C-RAN, clustering and beamforming is jointly determined at the BBU pool level. This optimization thus is global, using all the CSIs available at the BBU pool.

In this article we will consider downlink communications, $\mathcal{K}$ will represent the set of all UEs and $\mathcal{R}$ the set of all eRRHs. Let's consider a message s_k received by a UE k. For all communications, the BBU pool transmits to each RRH their beamformers $\mathbf{w}_{rk} \in \mathbb{C}^{M_r \times 1}$ that were computed at the BBU pool. Let's write $M = \sum_{r \in \mathcal{R}} M_r$ the total number of transmit antennas at the RRHs within the pool. The message sent by all RRHs combined can then be formulated as $\mathbf{w}_k s_k$, where $\mathbf{w}_k = \left[\mathbf{w}_{1k} \mathbf{w}_{2k} \cdots \mathbf{w}_{Rk}\right] \in \mathbb{C}^{M \times 1}$ is the concatenation of all beamformers for UE k.

This formulation stands for all users. Vectors $\mathbf{h}_{rk} \in \mathbb{C}^{M_r \times 1}$ then denote the channel coefficient between RRH r and UE k. The signal y_k received at UE k is:

$$y_k = \mathbf{h}_k^H \sum_{k' \in \mathcal{K}} \mathbf{w}_{k'} s_{k'} + n_k \tag{1}$$

where $\mathbf{h}_k = \left[\mathbf{h}_{1k} \mathbf{h}_{2k} \cdots \mathbf{h}_{Rk}\right] \in \mathbb{C}^{M \times 1}$ is the concatenation of RRHs to UEs CSI vectors, and $n_k \sim \mathcal{CN}(0, n_0^2)$ denotes an AWGN noise at UE k.

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \mathbf{h}_k^H \sum_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathbf{w}_{k'} s_{k'} + n_k \tag{2}$$

The first term in (2) is the desired signal, while the second and third terms represent respectively the interference and noise components.

2.2 Fog-RAN Architecture

The architecture studied here is a downlink Fog-RAN. This architecture is similar to C-RAN in that it is also using the centralization of BBUs into a pool. The main difference is that intelligence is pushed toward the edge of the network, using the Mobile Edge Computing paradigm. To that extent traditional RRHs are enhanced with cloud and cashing capacities. These are called eRRHs (Enhanced Remote Radio Heads) in Fog RAN. The added cloud computing capacities allow for new resource allocation methods. In Fog RAN resource optimization is split into two problems: pre-scheduling performed at the BBU pool and local beamforming performed at each eRRH. The motivation for using this architecture is to have both a lower complexity and a lower latency compared to C-RAN, while still benefiting from a centralized architecture design. An important difference with the model presented in the previous section is that in Fog RAN users are scheduled, by the BBU pool, to be served by a maximum of one eRRH, as beamformers are not determined globally. Let's consider a message s_k received by a UE k, associated to eRRH r. For such communication, the BBU pool's transmit signal to eRRH r can be presented as:

$$\mathbf{s}_r = \begin{pmatrix} s_{r1} \\ \vdots \\ s_{rK_r} \end{pmatrix} \tag{3}$$

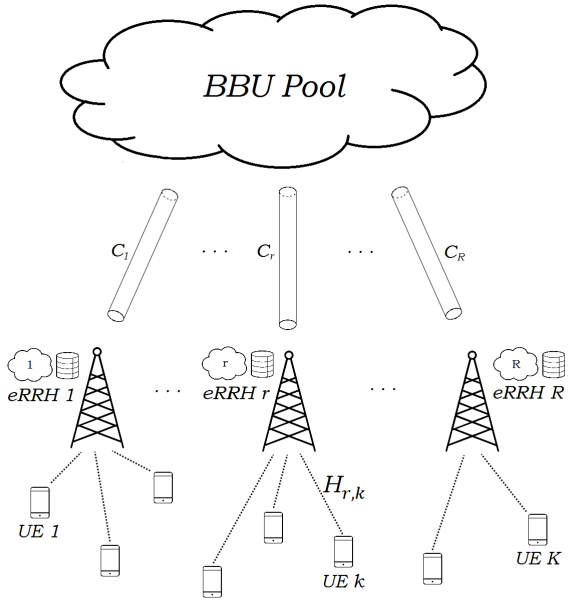

Figure 2: Fog RAN architecture with optical fiber fronthaul

where K_r denotes the number of UEs covered by eRRH r. Going further, $\mathcal{K}_r$ will denote the set of users associated to eRRH r.

The eRRH then performs beamforming to allocate resources towards its users. To do so, eRRH r multiplies received messages $\mathbf{s}_{rk}$ by the associated beamformer $\mathbf{w}_{rk} \in \mathbb{C}^{M_r \times 1}$. These signals are superposed to form the signal $\mathbf{t}_r$ to be transmitted by the eRRH:

$$\mathbf{t}_r = \sum_{k \in \mathcal{K}_r} \mathbf{w}_{rk} \; s_{rk} \tag{4}$$

The vector $\mathbf{h}_{rk} \in \mathbb{C}^{M_r \times 1}$ then denotes the channel coefficient between eRRH r and UE k. The signal y_{rk} is received at UE k:

$$y_{rk} = \mathbf{h}_{rk}^H \; \mathbf{t}_r + n_k \tag{5}$$

where n_k denotes the AWGN noise at UE $k \in \mathcal{K}_r$. We suppose that $n_k \sim \mathcal{CN}(0, n_0^2)$ for $r \in \mathcal{R}, k \in \mathcal{K}$. We can then rewrite (5) as follows:

$$y_{rk} = \mathbf{h}_{rk}^H \mathbf{w}_{rk} s_{rk} + \mathbf{h}_{rk}^H \sum_{\substack{k' \in \mathcal{K}_r \\ k' \neq k}} \mathbf{w}_{rk'} s_{rk'} + n_k \tag{6}$$

As we did for the CRAN architecture in section 2.1, we can define concatenated CSI vectors $\mathbf{h}_k$ and beamformers $\mathbf{w}_k$ as follows: $\mathbf{h}_k = [\mathbf{h}_{1k}\mathbf{h}_{2k}\cdots\mathbf{h}_{Rk}]$ and $\mathbf{w}_k = [\mathbf{w}_{1k}\mathbf{w}_{2k}\cdots\mathbf{w}_{Rk}] = [0\cdots0\ \mathbf{w}_{rk}\ 0\cdots0]$. This concatenated beamformer $\mathbf{w}_k$ is thus filled with zeros, having only one non-zero vector $\mathbf{w}_{rk}$. However using these concatenated vectors allows us to have compatible notations between the C-RAN and Fog-RAN architectures. We will thus rewrite the signal received by a user $k \in \mathcal{K}$ (6), but this time by all the eRRHs:

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \mathbf{h}_k^H \sum_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathbf{w}_{k'} s_{k'} + n_k \tag{7}$$

2.3 Imperfect Chanel State Information

The centralized architecture of C-RAN composed of a BBU pool allows for optimization of resources at a network-wide scale. To perform resource allocation optimization, the BBU pool needs to have access to global CSI, i.e. every eRRH needs to feed back its CSIs (i.e. vectors $\mathbf{h}_{rk}$) to the BBU pool via the fronthaul links. This feedback mechanism via the fronthaul does introduce delays, meaning that at time slot t the BBU pool has access to outdated CSIs $\mathbf{h}_{rk}(t-d),\ \forall r,k$. These delays thus need to be considered and incorporated in the aforementioned model.

As stated in [10], the CSI errors can be modelled in two ways. The first model is called the stochastic error (SE) model, where the probability distribution of the CSI errors is Gaussian. This model is applicable when CSI imperfections are mainly due to channel estimation errors. The other model, named the norm-bounded error (NBE) model, specifies an uncertainty set. This model can be applicable when CSI errors are dominated by quantization errors.

In the C-RAN architecture, the imperfection of CSI available at the BBU pool being mainly due to fronthaul delays, we will consider the SE model for this study. The SE model was also used to model CSI errors in existing articles looking for beamforming algorithms in C-RAN or MIMO architectures which are robust to the CSI imperfections [5], [11], [12].

Taking as model the stochastic error model, we can thus rewrite the estimated wireless channel state information matrix $\widehat{\mathbf{h}}_{rk} \in \mathbb{C}^{M_r \times 1}$ available at the BBU as follows:

$$\widehat{\mathbf{h}}_{rk} = \mathbf{h}_{rk} + \mathbf{e}_{rk} \tag{8}$$

where $\widehat{\mathbf{h}}_{rk}$ denotes the estimated delayed CSI vector available at the BBU, and $\mathbf{e}_{rk}$ represents the error, or imperfection. This error vector $\mathbf{e}_{rk}$ is assumed to be Gaussian, with zero mean and a variance of $\sigma_e^2 \mathbf{I}_{M_r}$:

$$\mathbf{e}_{rk} \sim \mathcal{CN}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{M_r}) \tag{9}$$

In addition, we will suppose that eRRHs have access to exact CSIs on their part, considering that such imperfections are small compared to the imperfec-

tions at the BBU pool. This means that this CSI imperfections model should be considered only for optimization problems performed at the BBU pool.

Remark: Using a wireless fronthaul has been discussed as a possibility to decrease deployment costs of C-RAN [5]. However a wireless fronthaul would only add to CSI imperfections as inter-RRH interferences would also appear during CSI signalling.

2.4 Performance Metrics: SINR, SLNR and Sum-Rate

To formulate the resource optimization problems, we first need to introduce the different performance metrics. We first define the Signal to Interference and Noise Ratio (SINR). This metric can be either local or global. For global SINR γ_k^{Glo}, interference with all other users are considered and thus can be formulated as follows:

$$\gamma_k^{Glo} = \frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\left|\sum\limits_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathbf{h}_k^H \mathbf{w}_{k'}\right|^2 + n_0^2} \tag{10}$$

Remark: For this equation as for the rest of this article, we consider symbol amplitude $\mathbb{E}\{|s_{rk}|^2\} = 1$, as well as the average noise power $\mathbb{E}\{|n_k|^2\} = n_0^2$. This allows for more concise expressions and thus a better readability.

A local SINR γ_{rk}^{Loc} differs from γ_k^{Glo} in that only signals from an RRH r are considered to compute both useful and interference signals received by user k. This metric is especially adapted when only limited knowledge is assumed to be available, and it can be written as follows:

$$\gamma_{rk}^{Loc} = \frac{\left|\mathbf{h}_{rk}^H \mathbf{w}_{rk}\right|^2}{\left|\sum\limits_{\substack{k' \in \mathcal{K}_r \\ k' \neq k}} \mathbf{h}_{rk}^H \mathbf{w}_{rk'}\right|^2 + n_0^2} \tag{11}$$

The SINR is difficult to optimize directly, hence an alternative criterion called Signal to Leakage and Noise Ratio (SLNR) has been used [11]. This ratio is based on the concept of power leakage, defined as the total power leaked from a user k to other users:

$$\left|\sum_{k' \neq k} \mathbf{h}_{k'}^H \mathbf{w}_k\right|^2$$

Something that is worth noting when looking at this expression is that, compared to the interference whose computation involves all beamformers $\mathbf{w}_{k'}$ $\forall k'$, only user k's beamformer $\mathbf{w}_k$ does appear in the expression of the leakage. This remark is what makes the SLNR metric especially suited to Fog-RAN, as the

beamforming vectors are computed and thus known locally, by one eRRH. However, unlike the SINR, the SLNR is not directly linked to users' data rates. Thus, in general, performing an optimization on an SLNR metric will not guarantee a high sum-rate.

Using this notion of leakage a global SLNR ζ_k^{Glo} can be formulated, taking into account the leakage induced by user k towards all other users in the network:

$$\zeta_k^{Glo} = \frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\left|\sum\limits_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathbf{h}_{k'}^H \mathbf{w}_k\right|^2 + n_0^2} \tag{12}$$

In the Fog-RAN architecture, intelligence is pushed to the edge, to the eRRHs where resource allocation is performed. These eRRHs only having access to limited information motivated the introduction of a new metric called local SLNR ζ_{rk}^{Loc}. For this local SLNR, only users in the neighborhood of an eRRH r are considered when determining the leakage induced by user k. This area of knowledge is represented by a radius of knowledge D, and can be formulated as follows:

$$\zeta_{rk}^{Loc} = \frac{\left|\mathbf{h}_{rk}^H \mathbf{w}_{rk}\right|^2}{\left|\sum\limits_{\substack{d(r,k') \leq D \\ k' \neq k}} \mathbf{h}_{rk'}^H \mathbf{w}_{rk}\right|^2 + n_0^2} \tag{13}$$

Another way to measure the efficiency of an allocation is by computing its total sum-rate SR. It consists in summing all users' data-rates R_k, every user $k \in \mathcal{K}$ having a maximum achievable rate defined by:

$$R_k \leq \log\left(1 + \gamma_k^{Glo}\right) \tag{14}$$

$$SR = \sum_{k \in \mathcal{K}} R_k \tag{15}$$

In this paper we consider the case where each user has only a single data stream, and assume that messages s_{rk} are independent and distributed according to $s_{rk} \sim \mathcal{CN}(0, \sigma^2)$.

3 Problem Formulation

In section 2, we presented the system model of both a C-RAN and a Fog-RAN architecture-based network. The problem studied in this article is about resource allocation in such architectures, looking to optimize certain performance metrics presented in section 2.4.

We presented both the SINR and SLNR metrics, which have been directly used to determine the beamforming vectors in previous studies [5], [11], [13].

Another approach to measuring the efficiency of an allocation is to directly look at users' data rates [7], [8]. A metric called Weighted Sum-Rate (WSR), using those data rates, is commonly used when doing resource optimization [2]. It aims at maximizing the total sum of data rates for users that are being scheduled. These data rates are affected of coefficients α_k which denote the priority weight associated with user k at the current scheduling slot. This priority can be updated to the proportional fairness criterion for instance. This WSR metric is defined as:

$$WSR = \sum_k \alpha_k R_k$$

3.1 C-RAN with Perfect CSI

In C-RAN, clustering and beamforming are jointly determined at the BBU pool level. This optimization thus is global, using all the CSIs available at the BBU pool. The system model for this architecture was presented in section 2. The BBU pool is responsible for the resource allocation of all RRHs. The major advantage of doing joint signal processing resides in the resulting interference management. As there is no need to worry about interfering with signals from other stations, which is problematic in decentralized architectures, one can directly maximize users' sum-rate, as performed in [2]:

$$\max_{\mathcal{K}_r, \mathbf{w}_k \ \forall k \in \mathcal{K}, \forall r \in \mathcal{R}} \quad \sum_{k \in \mathcal{K}} \alpha_k R_k \tag{16}$$

where

$$s.t. \quad \sum_{k \in \mathcal{K}_r} \left|\left| \mathbf{w}_{rk} \right|\right|_2^2 \leq P_r \tag{17}$$

$$\sum_{k \in \mathcal{K}_r} R_k \leq C_r \tag{18}$$

$$R_k \leq \log\left(1 + \gamma_k^{Glo}\right) \tag{19}$$

This centralized architecture, with perfect CSIs, represents an ideal scenario in terms of performances. We are able to optimize resources based on network-wide knowledge of perfect channel states. However the major drawback is the need to feedback to the BBU pool all the CSIs, available at the RRH level, which can represent a significant burden over bandwidth-limited fronthaul links.

3.2 C-RAN with Imperfect CSI

However, as seen in section 2.3, in C-RAN the BBU pool does not have access to perfect CSIs, due to feedback delays. Thus we will now incorporate the imperfect CSI model presented in section 2.3 into the resource optimization problem (16)-(19).

Algorithm 1 WSR Maximization With Per-BS Backhaul Constraints Under Dynamic BS Clustering [2]

Initialization: CSIs $\widehat{\mathbf{h}}_{rk}, \forall r, k$
Repeat: Users-to-eRRHs mapping $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$
  Initialization: $\widehat{\mathbf{w}}_{rk} = \mathbf{0}, \forall r, k$
1: Apply algorithm BCD [3]
2: Get $\widehat{\mathbf{w}}_{rk}$
3: Retrieve $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$ with $k \in \mathcal{K}_r \Leftrightarrow \widehat{\mathbf{w}}_{rk} \neq 0$
4: **return** $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$

For greater clarity, in the following we will denote by $\widehat{\gamma}_k$, $\widehat{\zeta}_k$ and $\widehat{R}_k$ respectively the SINR, SLNR and user data rate R_k obtained with imperfect CSI (i.e. with $\widehat{\mathbf{h}}_k$ instead of $\mathbf{h}_k$). The optimization problem can now be rewritten as follows:

$$\max_{\mathcal{K}_r, \widehat{\mathbf{w}}_k \ \forall k \in \mathcal{K}, \forall r \in \mathcal{R}} \quad \sum_{k \in \mathcal{K}} \alpha_k \widehat{R}_k \tag{20}$$

$$s.t. \quad \sum_{k \in \mathcal{K}_r} \left|\left| \widehat{\mathbf{w}}_{rk} \right|\right|_2^2 \leq P_r \tag{21}$$

$$\sum_{k \in \mathcal{K}_r} \widehat{R}_k \leq C_r \tag{22}$$

$$\widehat{R}_k \leq \log\left(1 + \widehat{\gamma}_k^{Glo}\right) \tag{23}$$

These delayed CSIs have for impact the choice of suboptimal beamformers for the RRHs. While some robust beamforming algorithms have been designed to balance this effect in C-RAN [5], resource allocation performances in C-RAN are still greatly deteriorated by imperfect CSIs. These performances represent a minimum goal to achieve for our Fog-RAN solution to be an upgrade.

3.3 D-RAN

Decentralized RAN represents the architecture that is currently deployed, with 4G LTE for instance, and thus is a benchmark showing how significantly we can improve resource allocation performances in 5G. From a communications perspective, compared to the system model presented in section 2, the difference here is the absence of BBU pool. In D-RAN, the e-Node B directly computes the local beamformers without global optimization. For each e-Node B r, the local beamformer is computed by solving:

$$\max_{\mathbf{w}_{rk} \ \forall k \in \mathcal{K}_r} \quad \sum_{k \in \mathcal{K}_r} \alpha_k R_k \tag{24}$$

$$s.t. \quad \sum_{k \in \mathcal{K}_r} \left|\left| \mathbf{w}_{rk} \right|\right|_2^2 \leq P_r \tag{25}$$

$$R_k \leq \log\left(1 + \gamma_{rk}^{Loc}\right) \tag{26}$$

The difference then with both C-RAN and Fog-RAN is the absence of any global coordination among all eNode Bs, as D-RAN is a fully distributed architecture.

3.4 Fog RAN

In this study we attempt to address the resource allocation problem in Fog-RAN. We propose to split the resource allocation decision into two problems carried out at both the BBU and eRRH levels. The first step, pre-scheduling, is performed at the BBU level. It consists in a user clustering, where the BBU pool decides to which eRRHs each user should be assigned for a given time slot. It thus consists in determining a partition $\left(\mathcal{K}_r\right)_{r \in \mathcal{R}}$ of $\mathcal{K}$:

$$\mathcal{K} = \coprod_{r \in \mathcal{R}} \mathcal{K}_r$$

Remark: This is not exactly a partition of the set $\mathcal{K}$ of users as here we do allow subsets $\mathcal{K}_r$ to be empty, meaning that some eRRH may not have any scheduled user for a given time slot.

3.4.1 Pre-Scheduling

This users to eRRH mapping $\left\{\mathcal{K}_1, ..., \mathcal{K}_R\right\}$ is the result of a global resource optimization. It is thus performed globally at the BBU pool, but with access to only imperfect CSIs given the fronthaul transport latency. To get this clustering, there are multiple ways to operate.

CSI-Based Pre-Scheduling Here we will present a clustering method using a CSI-based approach [2]. This clustering method, to have access to all CSIs, requires a feedback from each user to the BBU pool. In this method, the clustering is retrieved from the following sum-rate optimization problem:

$$\max_{\mathcal{K}_r, \widehat{\mathbf{w}}_{rk} \ \forall k \in \mathcal{K}_r, \forall r \in \mathcal{R}} \quad \sum_{k \in \mathcal{K}} \alpha_k \widehat{R}_k \tag{27}$$

$$s.t. \quad \sum_{k \in \mathcal{K}_r} \left|\left| \widehat{\mathbf{w}}_{rk} \right|\right|_2^2 \leq P_r \tag{28}$$

$$\sum_{k \in \mathcal{K}_r} \widehat{R}_k \leq C_r \tag{29}$$

$$\widehat{R}_k \leq \log\left(1 + \widehat{\gamma}_k^{Glo}\right) \tag{30}$$

$$\sum_{r \in \mathcal{R}} \left|\left| \widehat{\mathbf{w}}_{rk} \right|\right|_0^2 = 1 \tag{31}$$

where P_r represents the transmit power budget for RRH r. This optimization is similar to what is performed in full C-RAN architectures. The difference here is that the beamformers $\widehat{\mathbf{w}}_{rk}$ are computed only to retrieve the implicit scheduling:

$$if\ \widehat{\mathbf{w}}_{rk} \neq \mathbf{0},\ then\ k \in \mathcal{K}_r$$

Algorithm 2 CSI-Based User Pre-Scheduling

1: **Initialization:** $\widehat{\mathbf{w}}_k\ \forall k,\ \mathcal{K}_r\ \forall r$
2: **Repeat:**
3: Fix $\mathbf{w}_k\ \forall k$, compute the MMSE receiver $\mathbf{u}_k$ and the corresponding MSE e_k according to XXX and XXX
4: Update the MSE weight ρ_k according to XXX
5: Find the optimal transmit beamformer $\mathbf{w}_k$ under fixed $\mathbf{u}_k$ and $\rho_k\ \forall k$, by solving the QCQP problem XXX using convex optimization tool CVX [14]
6: Compute the achievable rate $R_k\ \forall k$, according to (14)
7: Update $\widehat{R}_k = R_k$ and $\beta_k^r\ \forall r, k$ according to XXX
8: **Until** Convergence
9: Get $\widehat{\mathbf{w}}_k\ \forall k$
10: Retrieve $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$ with $k \in \mathcal{K}_r \Leftrightarrow \widehat{\mathbf{w}}_{rk} \neq \mathbf{0}$

Path Loss-Based Pre-Scheduling Another clustering approach based on path loss is also interesting to study. In this method, only the path loss is used to determine the scheduling. As each user cannot be scheduled by more than one eRRH, each eRRH is associated to its strongest candidate (lower path loss). While this scheduling is not as sophisticated, it has the advantage to reduce fronthaul burden as the CSIs are no longer required at the BBU. In addition, it would also reduce significantly pre-scheduling computations and thus power consumption at the BBU. However performing such scheduling could result in unfair distributions between users and eRRHs. Especially in heterogeneous networks where eRRHs don't have the same beam power available, this pre-scheduling could overload some eRRHs while leaving others with almost no user to serve. Thus monitoring the tradeoff between pre-scheduling efficiency and bandwidth/power consumption seems interesting.

3.4.2 Local Beamforming

The beamforming vectors computed at the BBU are not actually used by eRRHs in Fog-RAN, unlike in C-RAN. After linking users to designated eRRHs, resource allocation is then reoptimized in a second step, at each eRRH level, by determining beamforming vectors. This beamforming problem is thus local, as the eRRH optimizes the resources dedicated to the communications of its assigned users. However this optimization is performed using exact CSI. As this optimization is performed locally at each eRRH though, there is no guarantee regarding inter-eRRH interference. To address this problem, we present two approaches, the first optimizing data rates under an interference constraint while the second solution maximizes SLNR.

Interference Temperature Method In this solution, an optimization similar to what was performed at the BBU is performed but with three important differences. The first difference is about the optimization variables. Here a local optimization is performed, thus only optimizing on beamforming for assigned users. The second difference is that this optimization is performed using exact CSIs. The last difference is the introduction of a third constraint, which is an interference threshold ε_{th} for all eRRHs. This constraint ensures that all leakage induced by a given eRRH is below this given threshold ε_{th}. The optimization problem can be formulated as follows, at each eRRH r:

$$\max_{\mathbf{w}_{rk} \ \forall k \in \mathcal{K}_r} \quad \sum_{k \in \mathcal{K}_r} \alpha_k R_k \tag{32}$$

$$s.t. \quad \sum_{k \in \mathcal{K}_r} \left|\left|\mathbf{w}_{rk}\right|\right|_2^2 \leq P_r \tag{33}$$

$$\sum_{k \in \mathcal{K}_r} \left| \sum_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathbf{h}_{rk'}^H \mathbf{w}_{rk} \right|^2 \leq \varepsilon_{th} \tag{34}$$

$$R_k \leq \log\left(1 + \gamma_{rk}^{Loc}\right) \tag{35}$$

Generalised SLNR Method Another approach, instead of still maximizing user data rates, would be to optimize the SLNR, ensuring an interference management. To do so we introduce a new metric called generalized SLNR Γ_r. This generalized SLNR takes into account the total signal emitted, sum of users' signals, and the total leakage, sum of all leakages created by all users' signals.

This can be displayed in the following definition:

$$\Gamma_r = \frac{\sum\limits_{k\in\mathcal{K}_r} \left|\mathbf{h}_{rk}^H \mathbf{w}_{rk}\right|^2}{\sum\limits_{\substack{k'\in\mathcal{K}\\k'\neq k}} \left|\sum\limits_{k\in\mathcal{K}_r} \mathbf{h}_{rk'}^H \mathbf{w}_{rk}\right|^2 + n_0^2}$$

Thus we can formulate the optimization problem as follows:

$$\max_{\mathbf{w}_{rk}\ \forall k\in\mathcal{K}_r} \quad \Gamma_r \tag{36}$$

$$s.t. \quad \sum_{k\in\mathcal{K}_r} \left|\left|\mathbf{w}_{rk}\right|\right|_2^2 \leq P_r \tag{37}$$

$$R_k \leq \log\left(1+\gamma_{rk}^{Loc}\right) \tag{38}$$

SLNR Method In this method, SLNR optimization problems are solved in parallel at each eRRH. For this to be possible, we assume equal power allocation among users as in [13]. Each problem can be formulated as follows, for each eRRH r and each user $k \in \mathcal{K}_r$:

$$\max_{\mathbf{w}_{rk}} \quad \zeta_{rk}^{Loc} \tag{39}$$

$$s.t. \quad \left|\left|\mathbf{w}_{rk}\right|\right|_2^2 \leq \frac{P_r}{K} \tag{40}$$

$$R_k \leq \log\left(1+\gamma_{rk}^{Loc}\right) \tag{41}$$

Under this assumption, problem (39) was analytically solved [13], [15]. The optimal beamformer can thus be directly computed, for a user $k \in \mathcal{K}_r$ using:

$$\mathbf{w}_{rk}^{opt} = \sqrt{\frac{P_r}{K_r}} \cdot \max\ eig\left(\left(\sum_{k'\neq k} \mathbf{h}_{rk'}^H \mathbf{h}_{rk'} + \frac{K_r\sigma_n^2}{P_r}\mathbf{I}\right)^{-1} \mathbf{h}_{rk}^H \mathbf{h}_{rk}\right) \tag{42}$$

where $\max\ eig(M)$ represents the eigenvector corresponding to the largest eigenvalue of matrix M. This beamforming method, while requiring stronger assumption (equal-power allocation among users), presents the advantage to be very light in terms of computations.

While this approach has been studied in the context of equal power repartition among users [13], [15], the Fog-RAN architecture could overcome these restrictions. In a case of CSI-based pre-scheduling being performed at the BBU level, a centralized resource allocation is actually performed thus giving beamforming vectors for scheduled users. A power-level for each scheduled user could

Algorithm 3 SLNR-Based Beamforming Vectors with Equal Power

For $k \in \mathcal{K}$ **do**
  Find r such that $k \in \mathcal{K}_r$
  $\mathbf{w}_{rk} = \sqrt{\frac{P_r}{K_r}} \cdot \max\ eig\Big(\Big(\sum_{k' \neq k} \mathbf{h}_{rk'}^H \mathbf{h}_{rk'} + \frac{K_r \sigma_n^2}{P_r} \mathbf{I} \Big)^{-1} \mathbf{h}_{rk}^H \mathbf{h}_{rk} \Big)$
End

Algorithm 4 CSI-Based User Pre-Scheduling with Power Pre-Allocation

Initialization: $\widehat{\mathbf{w}}_k\ \forall k,\ \mathcal{K}_r\ \forall r$
Repeat:
  Fix $\widehat{\mathbf{w}}_k\ \forall k$, compute the MMSE receiver $\mathbf{u}_k$ and the corresponding MSE e_k according to XXX and XXX
  Update the MSE weight ρ_k according to XXX
  Find the optimal transmit beamformer $\widehat{\mathbf{w}}_k$ under fixed $\mathbf{u}_k$ and $\rho_k\ \forall k$, by solving the QCQP problem XXX using convex optimization tool CVX [14]
  Compute the achievable rate $R_k\ \forall k$, according to (14)
  Update $\widehat{R}_k = R_k$ and $\beta_k^r\ \forall r, k$ according to XXX
Until Convergence
Retrieve $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$ with $k \in \mathcal{K}_r \Leftrightarrow \widehat{\mathbf{w}}_{rk} \neq \mathbf{0}$
$\forall k,\ P_{rk} = ||\widehat{\mathbf{w}}_k||_2^2$

then be forwarded to eRRHs, alongside their user scheduling. This user power-level corresponds to the power of the user's beamforming vector determined during pre-scheduling.

This pre-allocated power P_{rk} would then change the optimal beamforming vector as follows:

$$\mathbf{w}_{rk}^{opt} = \sqrt{P_{rk}} \cdot \max\ eig\Big(\Big(\sum_{k' \neq k} \mathbf{h}_{rk'}^H \mathbf{h}_{rk'} + \frac{\sigma_n^2}{P_{rk}} \mathbf{I} \Big)^{-1} \mathbf{h}_{rk}^H \mathbf{h}_{rk} \Big) \tag{43}$$

Algorithm 5 SLNR-Based Beamforming Vectors with Power Pre-Allocation

For $k \in \mathcal{K}$ **do**
  Find r such that $k \in \mathcal{K}_r$
  Compute $\mathbf{w}_{rk} = \sqrt{P_{rk}} \cdot \max\ eig\Big(\Big(\sum_{k' \neq k} \mathbf{h}_{rk'}^H \mathbf{h}_{rk'} + \frac{\sigma_n^2}{P_{rk}} \mathbf{I} \Big)^{-1} \mathbf{h}_{rk}^H \mathbf{h}_{rk} \Big)$
End

Remark: In all the aforementioned local beamforming methods, we consider interference (or leakage) information to be available at each eRRH. In practice, to determine these interferences we could either use network statistics or take

advantage of pilot coverage areas. The latter means using pilot signals to determine interference channel states to UEs that are not assigned to a specific eRRH but are within its pilot coverage area.

4 Proposed Algorithms

Idea: Apply the algorithm of [3] once to extract users' RRH attachment clustering from the implicit resource allocation obtained. This action is the first step, performed at the BBU level. Then we solve the local resource optimization problem at each RRH to get all beamformers $\mathbf{w}_{rk}$. To solve this problem we use CVX on the SDP relaxed problem. Then vectors $\mathbf{w}_{rk}$ are retrieved by taking the principal eigenvector associated to the only eigenvalue of the maybe non-rank-one matrix $\mathbf{w}_{rk}\mathbf{w}_{rk}^H$.

Algorithm 6 BBU Pre-Scheduling Algorithm

Input: CSIs $\widehat{\mathbf{h}}_{rk}, \forall r, k$
Output: Users-to-eRRHs mapping $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$
Initialization: $\widehat{\mathbf{w}}_{rk} = \mathbf{0}, \forall r, k$
1: Apply algorithm BCD [3]
2: Get $\widehat{\mathbf{w}}_{rk}$
3: Retrieve $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$ with $k \in \mathcal{K}_r \Leftrightarrow \widehat{\mathbf{w}}_{rk} \neq 0$
4: **return** $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$

Algorithm 7 eRRH Beamforming Algorithm

Input: $\{\mathcal{K}_1, ..., \mathcal{K}_R\}$, CSIs $\mathbf{h}_{rk}, \forall r, k$
Output: Beamforming vectors $\mathbf{w}_{rk}, \forall r, k$
Initialization: $\mathbf{w}_{rk} = \mathbf{0}, \forall r, k$
1: **for** $r \in \mathcal{R}$ **do**
2: Solve (36) for r, using CVX on the SDP relaxed problem
3: Obtain $\mathbf{w}_{rk}, \forall k \in \mathcal{K}_r$
4: **end for**
5: **return** $\mathbf{w}_{rk}, \forall r, k$

5 Appendix

$$\max_{\mathcal{K}_r, \widehat{\mathbf{w}}_{rk} \ \forall k \in \mathcal{K}_r, \forall r \in \mathcal{R}} \quad \sum_{k \in \mathcal{K}} \alpha_k \widehat{R}_k \tag{44}$$

$$s.t. \quad \sum_{k\in\mathcal{K}_r} \left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_2^2 \leq P_r \tag{45}$$

$$\sum_{k\in\mathcal{K}_r} \widehat{R}_k \leq C_r \tag{46}$$

$$\widehat{R}_k \leq \log\left(1+\widehat{\gamma}_k^{Glo}\right) \tag{47}$$

$$\sum_{r\in\mathcal{R}} \left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_0^2 = 1 \tag{48}$$

The constraint in norm zero (48) that forces every user to be attached to one and only one eRRH is non convex. In order to obtain a convex problem, we first relax this constraint into an inequality:

$$\sum_{r\in\mathcal{R}} \left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_0^2 = 1 \tag{49}$$

This relaxation allows for some users to not be scheduled for a given time slot. This relaxation is also considered in C-RAN, especially in the WSR maximization under dynamic BS clustering algorithm of B. Dai and W. Yu [2], used throughout this article as a C-RAN reference. Then, to get a convex constraint we use a $\updownarrow_2$-norm approximation technique of the $\updownarrow_0$-norm **TBD** In particular, we rewrite constraint (48) as follows:

$$\sum_{r\in\mathcal{R}} \left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_0^2 \approx \sum_{r\in\mathcal{R}} \beta_k^r \left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_2^2 \tag{50}$$

where β_k^r is a constant weight associated with eRRH r and user k. This constant is updated iteratively as:

$$\beta_k^r = \frac{1}{\left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_2^2 + \tau} \tag{51}$$

with $\tau > 0$ being a small regularization constant, and $\left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_2^2$ corresponding to the resulting beamforming vectors of the previous iteration. The heuristic behind this constant β_k^r (51) is that by being inversely proportional to the transmit power level $\left|\left|\widehat{\mathbf{w}}_{rk}\right|\right|_2^2$, the product gives results close to the zero-norm, as

References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks-a technology overview," in *IEEE Communication Surveys & Tutorials Vol. 17*, 2015.

[2] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," in *IEEE Access Vol. 2*, 2014.

[3] ——, "Backhaul-aware multicell beamforming for downlink cloud radio access network," in *IEEE IWCPM*, 2015.

[4] G. C. Alexandropoulos, P. Ferrand, and C. B. Papadias, "On the robustness of coordinated beamforming to uncoordinated interference and csi uncertainty," 2017.

[5] D. Wang, Y. Wang, R. Sun, and X. Zhang, "Robust c-ran precoder design for wireless fronthaul with imperfect channel state information," 2017.

[6] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," in ..., 2015.

[7] S.-H. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," in *IEEE Transactions on Wireless Communications Vol. 15*, 2016.

[8] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (c-ran) for mobile cloud computing systems," in *IEEE Transactions on Vehicular Technology Vol. 65*, 2016.

[9] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Control-data separation across edge and cloud for uplink communications in c-ran," 2017.

[10] P. Ubaidulla and A. Chockalingam, "Relay precoder optimization in mimo-relay networks with imperfect csi," in *IEEE Transactions on Signal Processing Vol. 59*, 2011.

[11] H. Du and P.-J. Chung, "A probabilistic approach for robust leakage-based mu-mimo downlink beamforming with imperfect channel state information," in *IEEE Transactions on Wireless Communications Vol. 11*, 2012.

[12] Y. Li, A. C. K. Soong, Y. Du, and J. Lu, "Beamforming with imperfect csi," in *IEEE WCNC*, 2007.

[13] H. Shen, W. Xu, A. L. Swindlehurst, and C. Zhao, "Transmitter optimization for per-antenna power constrained multi-antenna downlinks: An slnr maximization methodology," in *IEEE Transactions on Signal Processing Vol. 64*, 2016.

[14] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.1," `http://cvxr.com/cvx`, Mar. 2014.

[15] M. Sadek, A. Tarighat, and A. H. Sayed, "Active antenna selection in multiuser mimo communications," vol. 55, 2007.