Joint Beamforming and Clustering for Energy Efficient Multi-Cloud Radio Access Networks

Robert-Jeron Reifert*, Alaa Alameer Ahmad*, Hayssam Dahrouj†, Anas Chaaban‡, Aydin Sezgin*, Tareq Y. Al-Naffouri† and Mohamed-Slim Alouini†
*Institute of Digital Communication Systems, Ruhr-Universität Bochum, Germany
†Communication Theory Lab, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
‡School of Engineering, The University of British Columbia, Kelowna, Canada

***Abstract*—The tremendous growth of data traffic in mobile communication networks (MCNs) and the associated exponential increase in mobile devices' numbers necessitate the use of multi-cloud radio access networks (MC-RANs) as a viable solution to cope with the requirements of next-generation MCNs (6G). In MC-RANs, each central processor (CP) manages the signal processing of its own set of base stations (BSs), and so the system performance becomes a function of the joint intra-cloud and inter-cloud interference mitigation techniques. To this end, this paper considers the problem of maximizing the network-wide energy efficiency (EE) subject to user-to-cloud association, fronthaul capacity, maximum transmit power, and achievable rate constraints, so as to determine the joint beamforming vector of each user and the user-to-cloud association strategy. The paper tackles the non-convex and mixed discrete-continuous nature of the problem formulation using fractional programming (FP) and inner-convex approximation (ICA) techniques, as well as l_0-norm relaxation heuristics, and shows how the proposed approach can be implemented in a distributed fashion via a reasonable amount of information exchange across the CPs. The paper simulations highlight the appreciable algorithmic efficiency of the proposed approach over state-of-the-art schemes.**

I. Introduction

In a single year, mobile network data traffic grew by 51% and reached 60 Exabytes (EB) per month in 2020. Similar developments are predicted for 2026, with a projection of mobile traffic reaching 226 EB per month [1]. Along with this dramatic increase in data traffic, the need for deploying a multitude of power-hungry transceivers emerges to meet different requirements of beyond-the-fifth-generation (B5G) mobile communication networks (MCNs) [2]. Concerned with the energy consumption of such networks, the energy efficiency (EE) metric, an indicator for the sustainability of future MCNs, is of growing interest in literature [2]–[4]. The EE constitutes a promising metric for keeping CO_2 emissions and electromagnetic pollution caused by MCNs manageable. It also enables providers to lower maintenance and power costs [4]. Cloud radio access networks (C-RAN) emerge as a promising network architecture for future networks as a means to improve the EE of current MCNs [5]. In C-RAN, the majority of baseband processing tasks are performed centrally using a pool of computing resources referred to as the central processor (CP). The CP is connected to a large set of geographically distributed base stations (BSs) via high-speed digital fronthaul links [6]. Such architecture allows for deploying many low-complex BSs, with fewer functionalities than traditional BSs. It also allows for smart interference management and efficient radio resource allocation using the CP. The former is of particular interest from an EE perspective since joint interference management allows for reducing the overall transmit power at the BSs [7]. However, given the exponential increase in devices and applications, a single CP would no longer be able to manage the large-scale interference. Therefore, deploying several CPs, each managing their associated resources and set of BSs, all while utilizing a minimal communication overhead between the CPs, becomes a necessity [8].

This paper considers a multi-cloud radio access network (MC-RAN), where each CP manages the signal processing of its own set of BSs. Resource allocation in MC-RANs gained attention in recent literature [9]–[11]. In [9], the authors study the user-to-cloud association problem in an MC-RAN by developing an iterative-auction algorithm than can be implemented in a distributed manner. The network-wide sum-rate maximization problem of an MC-RAN is analyzed in [10], [11]. Both works [10], [11], besides utilizing different network architectures, emphasize the necessity of managing both inter- and intra-cloud interference, as this is a major bottleneck in MC-RANs. Another recent related work to the current paper is reference [12], which addresses the EE problem in a cache-enabled MC-RAN system and proposes an iterative algorithm for distributed resource management. The algorithm in [12] remains, however, of a relatively high-complexity, which partially stems from its edge caching problem considerations.

Unlike the aforementioned references, the current paper considers the downlink of an MC-RAN and addresses the problem of maximizing the EE by jointly determining the user-to-cloud association and the beamforming vectors. We formulate such a mixed discrete-continuous non-convex optimization problem subject to per-BS power and fronthaul capacity constraints, per-user achievable rate constraints, and discrete association variables constraints. The paper proposes solving such a complex problem using an iterative algorithm that relies on fractional programming (FP), ℓ_0-norm relaxation, and inner-convex approximation (ICA). The proposed algorithm can be implemented in a distributed fashion across the multiple CPs. Numerical simulations highlight the superiority

[1]This work was partially funded by the Federal Ministry of Education and Research (BMBF) of the Federal Republic of Germany (Förderkennzeichen 01IS18063A, ReMiX). This research is in part supported by the Center of Excellence for NEOM Research at KAUST.

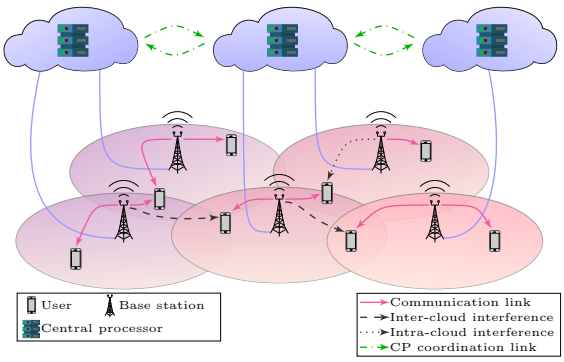


Fig. 1: MC-RAN consisting of 3 CPs, 5 BSs, and 8 users.

of our solution under different circumstances, show the advantages compared to baseline approaches, and illustrate the fast convergence speed and running time of the proposed algorithm as compared to the state-of-the-art leveraged techniques.

II. System Model and Problem Formulation

This paper considers a downlink MC-RAN with a total number of C CPs. Let B denote the number of BSs in the network and K the number of users. Then, let $\mathcal{B} = \{1, \cdots, B\}$ be the set of BSs, $\mathcal{K} = \{1, \cdots, K\}$ be the set of single-antenna users, and $\mathcal{C} = \{1, \cdots, C\}$ be the set of CPs. Each BS b is connected to one specific CP with a fronthaul link of capacity F_b. We assume that each user can be served by one and only one CP, albeit by many of the BSs connected to that CP. These BSs are denoted by $\mathcal{B}_c$. In particular, the paper assumes that there are disjoint BS-clusters, each of size B_c, where $\mathcal{B}_c = \{1, \cdots, B_c\}$, $\cup_{c\in\mathcal{C}}\mathcal{B}_c = \mathcal{B}$, $\mathcal{B}_c \cap \mathcal{B}_{c'} = \varnothing, \forall c \neq c'$, where B_c is the number of BSs coordinated by CP c. Fig. 1 shows an example of the studied system model. The channel vector from BS b to user k is denoted by $\mathbf{h}_{b,k} \in \mathbb{C}^L$, where L is the number of BS antennas. Let $\tilde{\mathbf{h}}_{c,k} = [\mathbf{h}_{i_1,k}^T, \cdots, \mathbf{h}_{i_{B_c},k}^T]^T$ denote the aggregate channel vector from all the BSs managed by CP c to user k, where $\{i_1, \cdots, i_{B_c}\} \in \mathcal{B}_c$ are the indices of all BSs connected to CP c. For convenience and mathematical tractability of the system, we assume that each CP has knowledge of the channel state information (CSI) of all users assigned to it. As a trade-off between computational complexity and required fronthaul capacity, we consider the following functional split between the CP and the set of BSs: the corresponding CP performs encoding of the requested data for user k, then the encoded symbol at each time slot, defined as s_k, is forwarded to a set of BSs. These BSs then perform modulation and precoding tasks and cooperate in transmitting the data to the respective user, a strategy known as data-sharing transmission strategy [13]. We define the costs associated with the baseband processing tasks at the CP as $\{P_k^{\text{proc}}, \forall k \in \mathcal{K}\}$. We further define the aggregate transmit beamforming vector from the BSs in $\mathcal{B}_c$ to user k as $\tilde{\mathbf{w}}_{c,k} = \left[\mathbf{w}_{i_1,k}^T, \cdots, \mathbf{w}_{i_{B_c},k}^T\right]^T$, where $\mathbf{w}_{b,k} \in \mathbb{C}^{L\times 1}$ is the beamforming vector from BS $b \in \mathcal{B}_c$ to user k. Since some BSs do not participate in serving user k, some beamforming vectors become $\mathbf{w}_{b,k} = \mathbf{0}_L$ and thus $\tilde{\mathbf{w}}_{c,k}$ is expected to be group-sparse by design. It is further assumed that sharing the beamforming coefficients over the fronthaul links with the corresponding BSs is negligible compared to sharing the user data [14]. At last, we define a binary association variable $z_{c,k}$, that determines the association between CP c and user k, such that $z_{c,k} = 1$ if user k is served by CP c, and 0 otherwise. To this end, we define the signal to interference plus noise ratio (SINR) of user k served by CP c as

$$\text{SINR}_{c,k} = \frac{|\tilde{\mathbf{h}}_{c,k}^\dagger \tilde{\mathbf{w}}_{c,k}|^2}{\sum_{(c',k')\neq(c,k)} |z_{c',k'}\, \tilde{\mathbf{h}}_{c',k}^\dagger \tilde{\mathbf{w}}_{c',k'}|^2 + \sigma^2}, \quad (1)$$

where σ^2 is the additive white Gaussian noise variance. The interference term can be split into the intra-cloud and inter-cloud interference, i.e., $c' = c$, $k' \neq k$, and $c' \neq c$, $k' \neq k$, respectively. The achievable rate of user k served by cloud c, denoted by $R_{c,k}$, is bounded by

$$R_{c,k} \leq \tau \log_2(1 + \text{SINR}_{c,k}), \quad (2)$$

where τ is the channel bandwidth. The required fronthaul capacity at BS b is then given by

$$C_b = \sum\nolimits_{k\in\mathcal{K}} \mathbb{1}\{\|\mathbf{w}_{b,k}\|_2^2\} R_{c,k}. \quad (3)$$

Note that we use the indicator function $\mathbb{1}\{\cdot\}$ to decide whether the rate of user k contributes to BS b's fronthaul traffic load exclusively based on beamforming vectors. In general, we have $\mathbb{1}\{x\} = 1$, only if the argument is non-zero, i.e., $x \neq 0$. In our MC-RAN model, the EE at each cloud is defined as the ratio of the sum-rate of all users and the total network power consumption. The power consumption constitutes three parts: transmit power, processing power, and the power consumption required for circuitry, cooling, and other necessary operations. Thus, the EE at CP c, denoted by Ψ_1, is defined as

$$\Psi_1 = \frac{\sum_{k\in\mathcal{K}} R_{c,k}}{\sum_{k\in\mathcal{K}} \mathbb{1}\{\|\tilde{\mathbf{w}}_{c,k}\|_2^2\} P_k^{\text{proc}} + \sum_{b\in\mathcal{B}_c} P_b^{\text{tx}} + P_c^{\text{cir}}}, \quad (4)$$

where the transmit power is

$$P_b^{\text{tx}} = \frac{1}{\eta_b} \sum\nolimits_{k\in\mathcal{K}} \|\mathbf{w}_{b,k}\|_2^2, \quad (5)$$

and where η_b is the transmit amplifier's efficiency at BS b, and P_c^{cir} is the fixed operational cost.

A. Problem Formulation

This paper considers the problem of maximizing the constrained network-wide EE, which can be formulated mathematically as follows:

$$\begin{aligned}
\max_{\mathbf{w},\mathbf{z},\mathbf{r}} \quad & \sum\nolimits_{c\in\mathcal{C}} \Psi_1 && && (6a)\\
\text{s.t.} \quad & (2), \\
& P_b^{\text{tx}} \leq P_b^{\max}, && \forall b \in \mathcal{B}_c, \forall c \in \mathcal{C}, && (6b)\\
& C_b \leq F_b, && \forall b \in \mathcal{B}_c, \forall c \in \mathcal{C}, && (6c)\\
& \sum\nolimits_{c\in\mathcal{C}} z_{c,k} = 1, && \forall k \in \mathcal{K}, && (6d)\\
& z_{c,k} \in \{0,1\}, && \forall k \in \mathcal{K}, \forall c \in \mathcal{C}, && (6e)\\
& \sum\nolimits_{k\in\mathcal{K}} z_{c,k} \leq K_c^{\max}, && \forall c \in \mathcal{C}, && (6f)\\
& \|\mathbf{w}_{b,k}\|_2^2 \leq M z_{c,k}, && \forall k \in \mathcal{K}, \forall b \in \mathcal{B}_c, \forall c \in \mathcal{C}, && (6g)
\end{aligned}$$

where the optimization is over the beamforming vectors from all the C CPs to all users $\mathbf{w} = \text{vec}(\{\tilde{\mathbf{w}}_{c,k} | \forall (c,k) \in \mathcal{C} \times \mathcal{K}\})$, the association variables $\mathbf{z} = \text{vec}(\{z_{c,k} | \forall (c,k) \in \mathcal{C} \times \mathcal{K}\})$, and the rates $\mathbf{r} = \text{vec}(\{R_{c,k} | \forall (c,k) \in \mathcal{C} \times \mathcal{K}\})$. Note that $\text{vec}(\mathcal{H})$ defines a column vector consisting of all N elements in set $\mathcal{H}$. Here, the notations $\text{vec}(\mathcal{H}) \equiv [h_1, \cdots, h_N]^T$ or $\text{vec}(\mathcal{H}) \equiv [\mathbf{h}_1^T, \cdots, \mathbf{h}_N^T]^T$ infer whether $\mathcal{H}$ consists of scalars or vectors, respectively. $K_c^{\max}$ is the maximum number of users

CP c can serve, and M is a positive real constant related to the big-M constraint. In problem (6), the transmit power and fronthaul capacity limits at the BSs are defined as $P_b^{\max}$ and F_b, respectively. When a user k is connected to CP c, its rate is bounded by the achievable rate constraint (2). The transmit power for serving all associated users of BS b in cloud c is bounded by (6b), while BS b's available fronthaul capacity is represented by (6c). Further, a user k is connected to only one CP c, which is ensured by (6d) and (6e). Constraint (6f) limits the number of users that can connect to a CP c and thus balances the load. At last, the big-M constraint (6g) reads as follows: When k and c are associated, i.e., $z_{c,k} = 1$, the beamformers are unconstrainted since M is large. Otherwise, when $z_{c,k} = 0$, the beamforming vectors from all BSs $b \in \mathcal{B}_c$ to user k are forced to zero. The above problem (6) is a mixed-integer non-convex optimization problem, which is hard to solve in general. This paper, therefore, proposes solving the problem through well-chosen problem reformulations using tactful techniques from optimization theory, as entailed next.

III. Proposed Algorithm

This section proposes solving problem (6) using a two-step approach. Firstly, we propose finding a feasible user-to-cloud association solution, i.e., $\mathbf{z}$, using a generalized assignment problem formulation. Then, in a second step, we revisit problem (6) with fixed $\mathbf{z}$ and solve it using a series of optimization techniques, namely, l_0-norm relaxation, FP, and ICA. Such heuristic approach only provides suboptimal yet reasonable solutions. However, measuring the gap between optimal and proposed solution is impossible, since (6) is NP-hard. A highlight of the proposed approach is that it can be implemented in a distributed fashion across the different CPs.

A. Generalized Assignment Problem

To find $\mathbf{z}$ in the initial step, we first characterize an EE-like utility function, which depicts the benefit of assigning a user k to CP c, as follows:

$$\mathcal{U}(c,k) = \frac{R_{c,k}}{\sum_{b\in\mathcal{B}_c} \frac{1}{\eta_b}\left\|\mathbf{w}'_{b,k}\right\|_2^2 + P_k^{\text{proc}}}. \tag{7}$$

Such a benefit is mainly dependent on two factors, 1) the CP c to which user k is assigned and 2) the characteristic of the vector $\mathbf{w}'_{b,k}$, which is the fixed beamforming vector from BS b to user k. Note that we only fix the beamforming vectors in this part of the algorithm. More specifically, at this step, we assume the beamformers to have a maximum ratio transmitter (MRT) structure, i.e., $\left\{\mathbf{w}'_{b,k} = \frac{\mathbf{h}_{b,k}}{\|\mathbf{h}_{b,k}\|_2^2}, \forall k \in \mathcal{K}\right\}$. The generalized assignment problem can then be written as:

$$\begin{aligned} \max_{\mathbf{z}} \quad & \sum\nolimits_{(c,k)\in(\mathcal{C},\mathcal{K})} z_{c,k}\ \mathcal{U}(c,k) \\ \text{s.t.} \quad & \text{(6d), (6e), (6f).} \end{aligned} \tag{8}$$

Problem (8) maximizes the utility function $\mathcal{U}(c,k)$ with respect to (w.r.t.) the association variable $\mathbf{z}$ and is a relaxed formulation of (6). We note that the generalized assignment problem (8) can be solved efficiently, e.g., using the distributed iterative-auction algorithm [9].

B. Problem Reformulation

With the discrete user-to-cloud association $\mathbf{z}$ found via solving problem (8), the challenge in solving (6) remains in determining the beamforming vectors under per-BS fronthaul and power constraints, as well as the achievable rate constraints, i.e., constraints (2), (6b), (6c), and (6g). The resulting problem remains non-convex since the objective is non-concave and the remaining constraints are non-convex. We, therefore, now reformulate the indicator function as an ℓ_0-norm, which allows approximating the discrete ℓ_0-norm with a weighted ℓ_1-norm, similar to [14]. More precisely, we rewrite the indicator functions in (6a) and (6c) as follows:

$$\mathbb{1}\left\{\left\|\mathbf{w}_{b,k}\right\|_2^2\right\} \triangleq \left\|\left\|\mathbf{w}_{b,k}\right\|_2^2\right\|_0 = \beta_{b,k}\left\|\mathbf{w}_{b,k}\right\|_2^2, \tag{9}$$

$$\mathbb{1}\left\{\left\|\tilde{\mathbf{w}}_{c,k}\right\|_2^2\right\} \triangleq \left\|\left\|\tilde{\mathbf{w}}_{c,k}\right\|_2^2\right\|_0 = \tilde{\beta}_{c,k}\left\|\tilde{\mathbf{w}}_{c,k}\right\|_2^2, \tag{10}$$

where the constant weights are defined as

$$\beta_{b,k} = \left(\delta + \left\|\mathbf{w}_{b,k}\right\|_2^2\right)^{-1}, \quad \tilde{\beta}_{c,k} = \left(\delta + \left\|\tilde{\mathbf{w}}_{c,k}\right\|_2^2\right)^{-1}, \tag{11}$$

with regularization constant $\delta > 0$. Note that the ℓ_1-norm is omitted in (9) and (10), since the argument is scalar. As BS b assigns low transmit powers to user k, the weight $\beta_{b,k}$ increases. This burdens b's fronthaul link (6c) and, at some point, the algorithm might exclude b from serving k. This ensures k being only served by BSs with reasonable transmit power. Note that the weighted ℓ_1-norm is applied to a quadratic function of the beamforming vector, which yields a smooth, convex, and continuous function, which is more traceable from an optimization perspective.

Making use of the ℓ_0-norm approximations in (9), the fronthaul constraint (6c) can now be formulated as

$$\sum\nolimits_{k\in\mathcal{K}} \beta_{b,k}\left\|\mathbf{w}_{b,k}\right\|_2^2 R_{c,k} \leq F_b, \quad \forall b \in \mathcal{B}_c, \forall c \in \mathcal{C}. \tag{12}$$

As this formulation is still complicated from an optimization perspective, we introduce slack variables for the relaxed ℓ_1-norm terms, i.e., $\mathbf{t} = \text{vec}(\{t_{k,b} | \forall (k,b) \in \mathcal{K} \times \mathcal{B}\})$ and $\tilde{\mathbf{t}} = \text{vec}(\{\tilde{t}_{c,k} | \forall (c,k) \in \mathcal{C} \times \mathcal{K}\})$. For all users, BSs, and CPs, we define

$$\beta_{b,k}\left\|\mathbf{w}_{b,k}\right\|_2^2 \leq t_{k,b}, \tag{13}$$

$$\tilde{\beta}_{c,k}\left\|\tilde{\mathbf{w}}_{c,k}\right\|_2^2 \leq \tilde{t}_{c,k}, \tag{14}$$

$$\sum\nolimits_{k\in\mathcal{K}} t_{k,b} R_{c,k} \leq F_b. \tag{15}$$

Here (15) is the reformulated fronthaul capacity constraint (i.e., the one associated with (12)), which is bilinear in the optimization variables and is thus suitable for ICA applications. Now we reformulate the non-convex maximum achievable rate constraint (2). Introducing the variable $\boldsymbol{\gamma} = \text{vec}(\{\gamma_{c,k} | \forall (c,k) \in \mathcal{C} \times \mathcal{K}\})$, we define

$$R_{c,k} \leq \tau \log_2(1 + \gamma_{c,k}), \qquad \forall k \in \mathcal{K}, \forall c \in \mathcal{C}, \tag{16}$$

$$\gamma_{c,k} \leq \text{SINR}_{c,k}, \qquad \forall k \in \mathcal{K}, \forall c \in \mathcal{C}. \tag{17}$$

Note that both constraints (16) and (17) are still non-convex; however, (16) can be addressed through ICA and (17) can be tackled using FP techniques, as shown next. Based on the above reformulations, problem (6) can now be re-written as:

$$\begin{aligned} \max_{\mathbf{w},\mathbf{r},\mathbf{t},\tilde{\mathbf{t}},\boldsymbol{\gamma}} \quad & \sum\nolimits_{c\in\mathcal{C}} \Psi_2 \\ \text{s.t.} \quad & \text{(6b), (6g), (13), (14), (15), (16), (17),} \end{aligned} \tag{18}$$

where

$$\Psi_2 = \frac{\sum_{k\in\mathcal{K}} R_{c,k}}{\sum_{k\in\mathcal{K}} \tilde{t}_{c,k} P_k^{\text{proc}} + \sum_{b\in\mathcal{B}_c} P_b^{\text{tx}} + P_c^{\text{cir}}}. \tag{19}$$

Solving problem (18) remains challenging since constraints

$$\Psi_3 = 2y_c\sqrt{\sum\nolimits_{k\in\mathcal{K}} R_{c,k}} - y_c^2\left[\sum\nolimits_{k\in\mathcal{K}} \tilde{t}_{c,k}P_k^{\text{proc}} + \sum\nolimits_{b\in\mathcal{B}_c} P_b^{\text{tx}} + P_c^{\text{cir}}\right] \tag{20}$$

$$g_1(\mathbf{w}) = \gamma_{c,k} - 2\text{Re}\left\{u_{c,k}^{\dagger}\tilde{\mathbf{w}}_{c,k}^{\dagger}\tilde{\mathbf{h}}_{c,k}\right\} + |u_{c,k}|^2\left[\sigma^2 + \sum\nolimits_{(c',k')\neq(c,k)} |z_{c',k'}\,\tilde{\mathbf{h}}_{c',k}^{\dagger}\tilde{\mathbf{w}}_{c',k'}|^2\right] \tag{21}$$

(15)-(17) are non-convex and the objective (19) has a non-convex fractional form. We, therefore, address such challenges using FP and ICA, as illustrated in the next subsection.

C. FP and ICA

To solve problem (18), we first use the quadratic transform, first introduced in [15, Theorem 1], to transform (19) into its quadratic transform associate (20), where the y_c's are the corresponding auxiliary variables and $\mathbf{y}$ is defined as $\mathbf{y} = \text{vec}(\{y_c|\forall c \in \mathcal{C}\})$. Note that, for fixed $\mathbf{y}$, the first term in (20) is the square root of a sum of linear variables, which is nondecreasing and concave, and the second term is also concave. We can obtain the optimal auxiliary variables fixing all other variables $\mathbf{r}$, $\tilde{\mathbf{t}}$, and $\mathbf{w}$ and computing the partial derivative of (20) w.r.t. $\mathbf{y}$. To this end, the optimal y_c can then be expressed as

$$y_c^* = \frac{\sqrt{\sum_{k\in\mathcal{K}} R_{c,k}}}{\left[\sum_{k\in\mathcal{K}} \tilde{t}_{c,k}P_k^{\text{proc}} + \sum_{b\in\mathcal{B}_c} P_b^{\text{tx}} + P_c^{\text{cir}}\right]}. \tag{22}$$

Now, we focus on the non-convex constraint (17), which contains a fraction of the signal and the interference plus noise, and is dependent on the beamforming vectors $\mathbf{w}$. We then transform (17) using the quadratic transform as typically handled in a multidimensional and complex case, using [15, Theorem 2]. We define $g_1(\mathbf{w})$ in (21), which is the result of the quadratic transform tailored to our problem. Here $\mathbf{u} = \text{vec}(\{u_{c,k}|\forall (c,k) \in \mathcal{C}\times\mathcal{K}\})$ is a vector containing complex-valued auxiliary variables. The optimal $u_{c,k}$ for fixed $\mathbf{w}$ is computed as

$$u_{c,k}^* = \frac{\tilde{\mathbf{w}}_{c,k}^{\dagger}\tilde{\mathbf{h}}_{c,k}}{\left[\sigma^2 + \sum_{(c',k')\neq(c,k)} |z_{c',k'}\,\tilde{\mathbf{h}}_{c',k}^{\dagger}\tilde{\mathbf{w}}_{c',k'}|^2\right]}, \tag{23}$$

which is obtained by setting the partial derivative of $g_1(\mathbf{w})$ w.r.t. $u_{c,k}$ to zero and then solving for $u_{c,k}$.

The constraint (15) is bilinear in the optimization variables. To tackle such incurring non-convexity, we equivalently rewrite (15) as

$$t_{k,b}R_{c,k} = \frac{1}{4}\left(\,(t_{k,b} + R_{c,k})^2 - (t_{k,b} - R_{c,k})^2\,\right). \tag{24}$$

This formulation is now in the form of a difference of two convex functions (convex plus concave function), which allows for applying ICA methods. The idea of ICA is to find a convex surrogate upper-bound to the non-convex function (24). We keep the convex part of (24) and linearize the concave part using the first-order Taylor expansion

$$g_2(\mathbf{t},\mathbf{r},\mathbf{t}',\mathbf{r}') \triangleq \sum\nolimits_{k\in\mathcal{K}} \left(\,(t_{k,b} + R_{c,k})^2\right. \tag{25}$$
$$\left. - 2\left(t'_{k,b} - R'_{c,k}\right)(t_{k,b} - R_{c,k}) + \left(t'_{k,b} - R'_{c,k}\right)^2\,\right) - 4F_b.$$

In (25), we introduce the feasible fixed values $\mathbf{t}' = \text{vec}(\{t'_{k,b}|\forall (k,b) \in \mathcal{K}\times\mathcal{B}\})$ and $\mathbf{r}' = \text{vec}(\{R'_{c,k}|\forall (c,k) \in \mathcal{C}\times\mathcal{K}\})$, which satisfy the constraints (6c) and (13). The feasible fixed values are updated iteratively by refining the feasible set in each iteration. This works well with the FP approach, which updates the optimal auxiliary variables in a similar manner. In the last step, we shift our focus to the non-convex constraint (16). We approximate its associated non-convex set via linearizing the logarithm around $\boldsymbol{\gamma}'$ using the first-order Taylor expansion. In this context, $\boldsymbol{\gamma}' = \text{vec}(\{\gamma'_{c,k}|\forall (c,k) \in \mathcal{C}\times\mathcal{K}\})$ are feasible fixed values. We thus define a convex upper-bound of (16) as

$$g_3(\boldsymbol{\gamma},\mathbf{r},\boldsymbol{\gamma}') \triangleq \frac{R_{c,k}}{\tau} - \log_2\left(1+\gamma'_{c,k}\right) - \left(\ln(2)\left(1+\gamma'_{c,k}\right)\right)^{-1}\left(\gamma_{c,k} - \gamma'_{c,k}\right). \tag{26}$$

Based on all the above reformulations and mathematical manipulations, the original complex optimization problem (6) can be approximately reformulated as the following computationally traceable optimization problem

$$\max_{\mathbf{w},\mathbf{r},\boldsymbol{\gamma},\mathbf{t},\tilde{\mathbf{t}}} \quad \sum\nolimits_{c\in\mathcal{C}} \Psi_3 \tag{27a}$$
$$\text{s.t.} \quad (6b),(6g),(13),(14),$$
$$g_1(\mathbf{w}) \leq 0, \qquad \forall k \in \mathcal{K}, \forall c \in \mathcal{C}, \tag{27b}$$
$$g_2(\mathbf{t},\mathbf{r},\mathbf{t}',\mathbf{r}') \leq 0, \qquad \forall b \in \mathcal{B}_c, \forall c \in \mathcal{C}, \tag{27c}$$
$$g_3(\boldsymbol{\gamma},\mathbf{r},\boldsymbol{\gamma}') \leq 0, \qquad \forall k \in \mathcal{K}, \forall c \in \mathcal{C}. \tag{27d}$$

The objective function in (27) is concave in the optimization variables and the constraints define a convex set, which makes problem (27) convex and efficiently solvable using standard tools (e.g., cvx [16]).

D. Distributed Resource Management in MC-RAN

We now illustrate how to solve (27) using a distributed implementation across the multiple CPs, which allows the user-to-cloud assignment and beamforming design to be done at the individual clouds with a reasonable amount of information exchange. Taking a closer look at constraint (27b), we note that it is sufficient that the CPs would exchange the interference information, i.e., $\sum_{(c',k')\neq(c,k)} \left|z_{c',k'}\,\tilde{\mathbf{h}}_{c',k}^{\dagger}\tilde{\mathbf{w}}_{c',k'}\right|^2$, from all other clouds $c \neq c'$ to solve the problem locally at each CP c. The detailed steps for such distributed resource management in MC-RANs are then listed in Algorithm 1. The first pre-optimization step finds feasible fixed values for the optimization variables, i.e., beamforming vectors $\mathbf{w}$, fixed rates $\mathbf{r}'$, auxiliary variables $\boldsymbol{\gamma}'$, $\mathbf{t}'$, and $\tilde{\mathbf{t}}$. Here the beamforming vectors are initialized as random feasible values. Next, we solve problem (8), which results in a feasible user-to-cloud association $\mathbf{z}$. For finding optimal $\mathbf{y}^*$ and $\mathbf{u}^*$ in the first iteration, $\mathbf{w}$ and $\tilde{\mathbf{t}}$ are required. In all remaining iterations, however, these variables can be taken from the output of problem (27). Then problem (27) is solved using cvx, the output of which is used to update the feasible fixed values. The above steps are repeated until convergence. The simulation results indeed validate the fast

Algorithm 1 Distributed Resource Management

1: Initialize $\mathbf{w}$, $\mathbf{r}'$, $\boldsymbol{\gamma}'$, $\mathbf{t}'$ and $\tilde{\mathbf{t}}$ to feasible values
2: Compute association $\mathbf{z}$, see III-A
Repeat: until convergence
3: Update $\mathbf{y}^*$ and $\mathbf{u}^*$ using (22) and (23)
4: Solve optimization problem (27)
5: Update $\mathbf{r}' = \mathbf{r}$, $\boldsymbol{\gamma}' = \boldsymbol{\gamma}$ and $\mathbf{t}' = \mathbf{t}$
6: **End**

TABLE I: Simulation Parameters

Antennas per BS	$L = 2$
Maximum transmit power	$P_b^{\max} = 32$dBm
Processing power	$P_b^{\text{proc}} = 15$dBm
Fixed operation costs	$P_c^{\text{cir}} = 38$dBm
Channel bandwidth	$\tau = 10$MHz
Transmit amplifier's efficiency	$\eta_b = 0.25$
Noise power	$\sigma^2 = -102 + 10\log_{10}(\tau) + 15$dBm
Path loss from BS b to user k	$\text{PL}_{b,k} = 128.1 + 37.6\log_{10}(d_{b,k})$
Log-normal shadowing	8dB
Rayleigh small scale fading	0dB

convergence of both the centralized and distributed versions of the algorithm, as illustrated later in the paper.

E. Comparison with Dinkelbach and ICA Algorithm

In [12], the authors propose an iterative algorithm for distributed resource management in cache-enabled MC-RAN. We refer to the solution in [12] as Dinkelbach and ICA Algorithm (DICA), which, when leveraged to the current paper setup, becomes a two-step procedure consisting of outer and inner loops. Feasible fixed values for ICA are updated in the outer loop, while the inner loop utilizes a Dinkelbach-like algorithm as a FP approach. The computational complexity of the DICA is in the order of $\mathcal{O}(\text{V}_{1,\max}\text{V}_{2,\max}(d_1)^{3.5})$, where $\text{V}_{1,\max}$ and $\text{V}_{2,\max}$ are the worst-case fixed number of iterations for convergence of the outer and inner loop, respectively, and $d_1 = (\text{K}(\text{B}(\text{L}+1)+3))$ is the total number of variables. Our proposed Algorithm 1 is rather a major simplification of DICA since we avoid the additional loop. This leads to a significant complexity saving, which is in the order of $\mathcal{O}(\text{V}_{1,\max}(d_1)^{3.5})$ in our case.

IV. Numerical Simulation

In this section, we evaluate the performance of our proposed algorithm. We consider an MC-RAN occupying a square area of $[-400\ 400] \times [-400\ 400]$ m^2, where 14 BSs serve 28 users randomly placed in the network. All parameters related to our simulations are listed in Table I. Algorithm 1 stops when the EE variations between iterations become sufficiently small. We assume that all BSs share the same fronthaul capacity and maximum transmit power. Initially, the beamformers are set randomly to feasible values. For illustration, the processing powers are assumed to be equal for all users and the fixed operating costs are equal for all clouds. A procedure to predetermine fixed clusters before optimizing only the beamforming vectors is considered as a baseline to benchmark the performance of our method. We refer to this method as fixed clustering, which is determined using a load balancing algorithm, e.g., see [14]. Also, we distinguish between two different implementations of our algorithm. On one hand, a centralized implementation treats all CPs as one logical CP, i.e., it rhetorically allows for unlimited communication between clouds. On the other hand, a distributed implementation is computed at each respective CP, with a reasonable amount of information exchange.

A. EE vs. Fronthaul Capacity

Fig. 2 shows the EE as a function of the fronthaul capacity for the two clustering schemes, each using both distributed and centralized implementations. Generally, we observe increased EEs when the fronthaul capacity gets bigger. This growth

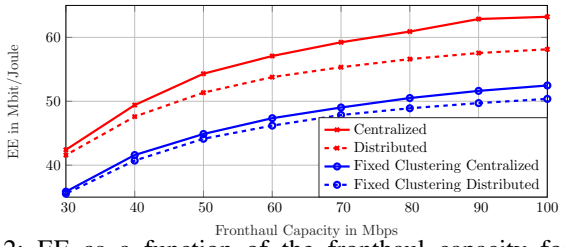
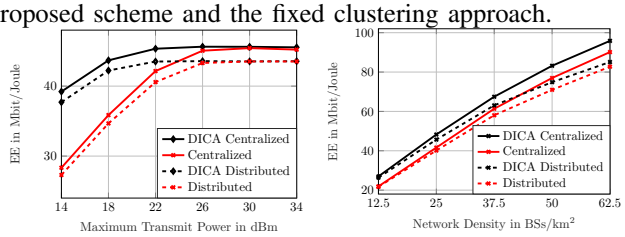


Fig. 2: EE as a function of the fronthaul capacity for the proposed scheme and the fixed clustering approach.



(a) $P_b^{\max}$. (b) Network density.

Fig. 3: EE as a function of different parameters for the proposed scheme and the DICA.

saturates at high fronthaul capacities. We further note that, regardless of being implemented using a centralized or distributed implementation, our proposed algorithm outperforms the baseline scheme. This is due to the fact that situations where BSs dropping out of serving clusters due to power constraints can hardly be compensated using the fixed clustering approach. Such fact highlights the necessity of deploying the proposed dynamic clustering approach for achieving a good EE. Fig. 2 further shows that the performance of both central and distributed implementations in the low-fronthaul regime is relatively similar. This promotes the role of our proposed algorithm in networks with high fronthaul congestion.

B. EE vs. Maximum Transmit Power

In another simulation set, we plot the EE as a function of the maximum transmit power in Fig. 3a. We distinguish between Algorithm 1 and the DICA. We fix the fronthaul capacity to $F_b = 50$Mbps for all BSs. While the DICA outperforms the proposed algorithm at low maximum transmit powers, both algorithms perform similarly at high transmit powers. The EE of the distributed DICA and our distributed implementation levels up at $P_b^{\max} = 26$dBm and higher, while our proposed centralized approach achieves a constant gap to the centralized DICA experiencing only reasonable loss. While assigning a small value to $P_b^{\max}$ significantly degrades the performance of the proposed algorithm, the more computationally complex DICA does not suffer the same EE loss at low powers, albeit with similar gradients at 14dBm. Also, we note that the centralized and distributed implementation achieve similar performance in terms of EE in low transmit powers, while the EE gap at high transmit powers is bigger.

C. EE vs. Network Density

The above figures assume that BSs are randomly placed throughout the network. In the next set of simulations, we generate an equal number of BSs at every cloud. That is, while the number of BSs per cloud is fixed, the BSs are randomly placed in the area of their respective serving cloud. Fig. 3b shows the EE versus the network density in BSs/km^2.

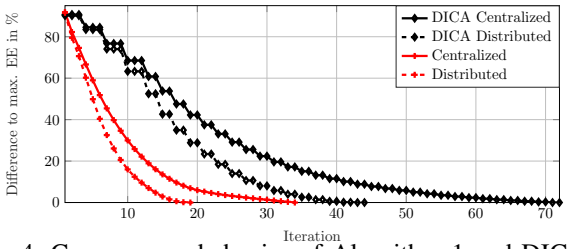


Fig. 4: Convergence behavior of Algorithm 1 and DICA.

We observe better results while implementing DICA, matching with previous observations. However, in dense networks (> 40BSs/km^2), Algorithm 1's EE approaches the EE of the DICA, especially while implementing both algorithms in a distributed fashion.

D. Convergence and Run Time

Finally, we plot the normalized difference of the objective value and optimal EE in % as a function of the number of iterations executed until converging in Fig. 4, i.e., $\frac{f^*-f^i}{f^*}$, where $f = \sum_{c\in\mathcal{C}} \Psi_1$, f^i is the objective value at iteration i, and f^* is the optimal value after convergence. We compare the distributed and centralized implementations at $F_b = 60$Mbps. Fig. 4 shows that the maximum iterations required for convergence are comparatively low, which highlights a key advantage of Algorithm 1. Fig. 4 also highlights how our distributed algorithm converges faster than the centralized implementation. This pronounces the significance of the proposed distributed resource management technique in MC-RAN, especially w.r.t. the available computational resources. Comparing our proposed algorithm to DICA, we observe the superiority of our proposed algorithm for both centralized and distributed implementations. Note that for DICA, the iteration counter captures the iterations of consecutive inner loops, i.e., we can observe a step-wise convergence behavior.

For completeness, we also show in Table II the running times of the different algorithms normalized to our proposed distributed solution's run time. In networks with 12.5BSs/km^2, the centralized and distributed versions of Algorithm 1 outperform DICA. The table shows that the centralized DICA takes about 6.5 times more run time than the proposed algorithm. Interestingly, both centralized algorithms become substantially slower with increasing network density, as their relative run times increase significantly. However, the distributed DICA keeps a constant time gap to the distributed Algorithm 1. Compared to their centralized counterparts, both approaches scale better with increasing network size, which makes them applicable in 6G. DICA, in particular, fails to compute solutions in an appropriate time frame as compared to the paper proposed approach, which highlights the numerical efficiency of the proposed algorithm in the context of MC-RANs.

TABLE II: Normalized Run Times

Density in BSs/km^2	Centralized	Distributed	DICA Centralized	DICA Distributed
12.5	1.62	1	6.51	2.33
25	2.56	1	8.33	2.49
37.5	3.49	1	10.47	2.35
50	4.15	1	12.75	2.37
62.5	4.85	1	14.07	2.50

V. Conclusion

The practical realization of B5G networks is expected to heavily rely on advanced resource management techniques, specifically designed for addressing the emerging large-scale interference. To this end, this paper considers an MC-RAN and addresses an EE maximization problem subject to user-to-cloud association, fronthaul capacity, maximum transmit power, and achievable rate constraints. Utilizing FP and ICA techniques, an iterative algorithm for distributed resource management in MC-RANs is developed. Numerical simulations investigate the proposed algorithm versus a fixed clustering baseline scheme and the DICA, all while accounting for centralized and distributed implementations. The paper results show that, when the fronthaul capacity or the number of BSs is limited, the proposed distributed algorithm is on par with the centralized implementation. Also, compared to the more complex DICA, the proposed algorithm converges faster than the DICA solution, especially in dense networks and high transmit power regimes, which further highlights the numerical merits of the proposed solution.

References

[1] "Ericsson mobility report november 2020," Ericson, Tech. Rep., Nov. 2020. [Online]. Available: https://www.ericsson.com/en/mobility-report/reports/november-2020

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[3] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wirel. Commun.*, vol. 24, no. 3, pp. 86–94, 2017.

[4] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, pp. 185–396, 01 2015.

[5] China Mobile, "C-RAN the road towards green RAN, ver. 3.0," *White Paper*, Dec. 2013.

[6] C. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[7] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE JSAC*, vol. 34, no. 4, pp. 1037–1050, 2016.

[8] Y. Zhang, B. Di, Z. Zheng, J. Lin, and L. Song, "Distributed multi-cloud multi-access edge computing by multi-agent reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2565–2578, 2021.

[9] H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini, "Distributed cloud association in downlink multicloud radio access networks," in *CISS*, 2015, pp. 1–3.

[10] A. A. Ahmad, H. Dahrouj, A. Chaaban, A. Sezgin, T. Y. Al-Naffouri, and M. Alouini, "Distributed cloud association and beamforming in downlink multi-cloud radio access networks," in *2020 IEEE ICC Workshops*, 2020, pp. 1–6.

[11] S. Park, O. Simeone, O. Sahin, and S. Shamai, "Inter-cluster design of precoding and fronthaul compression for cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 369–372, Aug. 2014.

[12] A. A. Ahmad, R.-J. Reifert, H. Dahrouj, A. Chaaban, A. Sezgin, T. Y. Al-Naffouri, and M.-S. Alouini, "Distributed resource management in downlink cache-enabled multi-cloud radio access networks," 2021. [Online]. Available: https://arxiv.org/abs/2104.03664

[13] L. Liu and W. Yu, "Cross-layer design for downlink multihop cloud radio access networks with network coding," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1728–1740, Apr. 2017.

[14] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

[15] K. Shen and W. Yu, "Fractional programming for communication systems part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[16] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014. [Online]. Available: http://cvxr.com/cvx