Joint Content Placement, RRH Clustering and Beamforming for Cache-Enabled Cloud-RAN

Ming-Min Zhao $^{\#1}$, Yunlong Cai $^{\#2}$, Min-Jian Zhao $^{\#3}$, and Benoit Champagne $^{\star 4}$

$^{\#}$ *College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, 310027*
* *Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, H3A 0E9*
^{1,2,3} {*zmmblack, ylcai, mjzhao*}*@zju.edu.cn,* ^{4} *benoit.champagne@mcgill.ca*

***Abstract*—This work studies the joint problem of optimal content placement, RRH clustering and beamformer design, in a cache-enabled cloud-radio access network (C-RAN). In the considered system, multiple remote radio heads (RRHs) connected to a centralized baseband unit (BBU) pool via fronthaul links, cooperatively serve the downlink users by grouping them into potentially overlapping clusters. Each RRH is equipped with a local cache from which it can directly acquire the requested user contents, without the need to occupy the fronthaul links. We aim to jointly optimize the caching placement, user association and downlink beamforming vector at each RRH, in order to strike a balance between fronthaul traffic reduction and transmission power minimization. To this end, we propose to employ the ratio between these two important system utilities as the objective function, referred to as *caching efficiency*. A penalty dual decomposition (PDD) based algorithm is presented to address the resulting nonconvex optimization problem, which features coupling constraints and mixed-integer variables. Simulation results validate the efficiency of the proposed algorithm.**

***Index Terms*—Transceiver design, cloud-RAN, content placement, caching, RRH clustering, beamforming.**

I. Introduction

With the increasing demands for high-speed data traffic, especially content sharing and video streaming, wireless network operators are now faced with striking challenges in providing high throughput and low latency services to large communities of mobile users. To meet these new requirements, local caching of popular data at base stations (BSs) has been recently proposed as a promising solution for massive content delivery [1]–[6]. This approach essentially brings key information contents closer to the users and in turn reduces fronthaul utilization costs. Furthermore, as service providers move favored contents to intermediate nodes in the network, the access delay is reduced which improves the quality of experience for users.

To support the ever growing data traffic and computational demands of mobile users, another important technology is the cloud radio access network (C-RAN), which refers to an emerging network architecture that can improve the spectrum and energy efficiency compared to existing wireless networks [7]–[10]. In C-RAN, several low-cost low-power remote radio heads (RRHs) are deployed to replace the traditional high-cost BSs. Since most of the signal processing tasks are handled by a centralized baseband unit (BBU) pool that connects to the RRHs via digital fronthaul links, joint data processing and precoding are possible to improve system performance.[1]

In the literature, several works have investigated the joint problem of user-centric BS clustering and cooperative beamforming under dynamic conditions [8], [9], [11], [12]. With regard to fronthaul traffic reduction, this approach is attractive since the popular data of each user only need to be assigned to a small cluster of potentially overlapping BSs, instead of all BSs. This content delivery service can be carried out by carefully designing content placement such that the users can seize various transmission opportunities and fully exploit the caching gain [13]. The potential benefits of distributing and storing popular contents across the whole network have been investigated by many researchers [1], [5], [13]–[21]. In the context of C-RAN, to further improve the delivery rate and decrease backhaul/fronthaul costs and latency for mobile users, a promising solution is to cache popular contents directly at the RRHs [18].

While making significant advances, the aforementioned studies do not approach the design of content placement, RRH association and RRH transceiver by considering all three aspects jointly. In this work hence, we study the joint optimization of the content-aware C-RAN along these three critical design dimensions, aiming to strike a more favorable balance between fronthaul traffic reduction and transmission power minimization. To this end, we propose to maximize the ratio between the fronthaul traffic reduction and the total transmission power, termed *caching efficiency*, subject to quality of service (QoS), clustering and caching constraints. The joint design problem is quite challenging and difficult to handle due to the facts that the objective function and constraints are nonconvex, the optimization variables are tightly coupled, and the latter contain nontrivial discrete variables. By exploiting the problem structure, taking advantage of the Dinkelbach method [22] and embracing the penalty dual decomposition (PDD) framework [23], we show that the joint design problem can be solved by iterating over a sequence of simple and very efficient updates in the individual design variables.

The rest of the paper is organized as follows. In Section II, we present the system model of the content-aware C-RAN and formulate the associated joint optimization problem. In Section III, we develop the PDD-based algorithm for the solution

This work was supported in part by the National Natural Science Foundation of China under Grant 61831004, in part by the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under Grant LR19F010002, and in part by the Fundamental Research Funds for the Central Universities under Grant 2018QNA5005.

[1] Note that in the C-RAN context, the backhaul portion of the network comprises the intermediate links between the core network and the BBU pool, while the links between the BBUs and the RRHs at the edge of the network are usually referred to as fronthaul links. In general, content caching at the RRHs would save both backhaul and fronthaul costs. However, for conciseness, we will only mention fronthaul in the following.

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

of this problem and discuss its convergence. In Section IV, we present simulation results to characterize the performance of the proposed algorithm. Finally, conclusions are drawn in Section V.

II. System Model and Problem Formulation

In this section, we first describe a content-aware C-RAN system, wherein multiple RRHs serve mobile users by forming potentially overlapping clusters. We then formulate the joint optimization problem for the content placement, RRH clustering and beamformer design.

A. System Model

We consider a content-ware C-RAN, which consists of N multi-antenna RRHs, indexed by $n \in \mathcal{N} \triangleq \{1, \cdots, N\}$, K single-antenna mobile users, indexed by $k \in \mathcal{K} \triangleq \{1, \cdots, K\}$, and a centralized BBU pool. The RRHs, each equipped with a common number L of antennas for simplicity, are individually connected to the BBUs via high-speed fronthaul links. We assume that the BBU pool has access to the information contents that can be potentially requested by all the users, and distributes each user's content to an individually selected cluster of RRHs via the fronthaul links. Each user is then cooperatively served by the associated RRH cluster through joint beamforming.

Let $\mathbf{w}_{k,n} \in \mathbb{C}^{L\times 1}$ denote the dowlink beamforming vector from RRH n to user k, and let $\mathbf{w}_k = [\mathbf{w}_{k,1}^H, \mathbf{w}_{k,2}^H, \cdots, \mathbf{w}_{k,N}^H]^H$ denote the aggregate, network wide beamforming vector form all RRHs to user k. The received signal at user k, at a given instance of symbol transmission (or time slot), can then be written as

$$y_k = \mathbf{h}_k^H \mathbf{w}_k x_k + \sum_{j\neq k}^{K} \mathbf{h}_k^H \mathbf{w}_j x_j + n_k, \tag{1}$$

where x_k (x_j) is the information symbol transmitted to user k $(j \neq k)$ and n_k represents an additive noise term. Modeling these quantities as zero-mean, mutually independent random variables, the signal-to-interference-plus-noise ratio (SINR) of user k can be defined as

$$\text{SINR}_k \triangleq \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum\limits_{j\neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2}, \tag{2}$$

Consequently, the achievable data rate of user k is given by $R_k = B\log(1 + \text{SINR}_k)$, where B denotes the total available channel bandwidth.

Different from the conventional C-RAN systems, we here assume that each RRH can cache a certain amount of content objects within a local storage device. At regular time intervals, referred to as transmission times, each user submits a content request according to a certain probability distribution specific to that user. If the requested content has already been cached locally at a serving RRH, then this RRH can access the content directly and transmit it to the user without the need for fronthaul data transfer.[2] It is assumed that enough time slots are available within a transmission time interval to complete the content delivery to the users, prior to the next transmission time. Without significant loss in generality, let us assume that the complete set of available user contents is represented by F binary files, indexed by $f \in \mathcal{F} = \{1, 2, \cdots, F\}$, each with normalized size of unity. The local storage size of RRH n is denoted as $Y_n \leq F$, which means that RRH n can cache Y_n content files at most. Let $c_{f,n} = 1$ indicate that content f is cached in RRH n and $c_{f,n} = 0$ otherwise, with the constraint that $\sum_{f=1}^{F} c_{f,n} \leq Y_n$. Considering that a request for a content file that is not locally cached leads to a fronthaul utilization of one unit per serving RRH, the total fronthaul traffic reduction of the cache-enabled C-RAN can be expressed as [20]

$$C_B = \sum_{k=1}^{K}\sum_{n=1}^{N} s_{k,n} \sum_{f=1}^{F} P_{k,f} c_{f,n}, \tag{3}$$

where $P_{k,f}$ denotes the probability that user k requests content file f and $s_{k,n}$ is the user-RRH association indicator, where $s_{k,n} = 1$ means that RRH n belongs to the serving cluster for user k and $s_{k,n} = 0$ otherwise.. The cost of the transmission of the requested contents by all of the users from their serving RRHs can be assessed in terms of the total transmission power, defined here as

$$C_P = \sum_{k=1}^{K}\sum_{n=1}^{N} \|\mathbf{w}_{k,n}\|^2. \tag{4}$$

In this work, we introduce a new objective function, termed *caching efficiency* and defined as

$$C \triangleq C_B / C_P, \tag{5}$$

which measures the amount of fronthaul traffic reduction that can be achieved per unit of consumed transmission power. The main motivation to employ the fronthaul efficieny as the objective function in system design is based on the observation that with increasing transmission power budget, larger serving clusters can be formed for each user, which further reduces the fronthaul utilization. Hence, the objective function (5) is intuitively pleasing since it takes into account the proportionality relationship between the available power budget and the fronthaul reduction.

B. Problem Formulation

In this work, we aim to jointly optimize content placement, RRH clustering and cooperative beamforming at each transmission time interval, so as to maximize the caching efficiency, which can be formulated as the following optimization problem:

$$\max_{\{\mathbf{w}_{k,n}\},\, \{s_{k,n}\},\, \{c_{f,n}\}} \; C \tag{6a}$$

$$\text{s.t. } \text{SINR}_k \geq \gamma_k,\ \forall k, \tag{6b}$$

$$s_{k,n} = 0 \text{ or } 1,\ \forall k, n, \tag{6c}$$

$$\sum_{k=1}^{K} s_{k,n} \leq X_n,\ \forall n, \tag{6d}$$

$$(1 - s_{k,n})\mathbf{w}_{k,n} = \mathbf{0},\ \forall k, n, \tag{6e}$$

$$c_{f,n} = 0 \text{ or } 1,\ \forall f, n,\ \sum_{f=1}^{F} c_{f,n} \leq Y_n,\ \forall n. \tag{6f}$$

The QoS constraint (6b) requires that the SINR of user k should be no smaller than a given positive target threshold γ_k. Constraint (6c) means that the values of the user association

[2] In a C-RAN without caching capabilities, the RRHs need to fetch the requested content from the BBU pool via fronthaul links, and possibly from the cloud content cache via backhaul links.

indices $s_{k,n}$ can only be 0 or 1. Constraint (6d) indicates that the maximum number of users that RRH n can serve is limited by X_n. Finally, constraint (6e) forces the beamforming vector $\mathbf{w}_{k,n}$ to be an all-zero vector if user k is not served by RRH n. Note that problem (6) is highly nonconvex and features both continuous and discrete variables which are coupled together in (6e) due to the RRH clustering operation. Consequently, problem (6) is quite challenging and it does not appear possible to obtain a globally optimal solution.

III. Proposed PDD-based Algorithm

In this section, we present the detailed derivation of the proposed PDD-based algorithm, which relies on the Dinkelbach method [22]. The proposed algorithm exhibits a twin-loop structure: the inner loop (approximately) solves the augmented Lagrangian (AL) problem [24], [25], while the outer loop updates the Dinkelbach variable, and the dual variables or the penalty parameter, depending on a constraint violation status. In the proposed iterative algorithm, we show that each subproblem in the inner loop can be solved either in closed-form or by the bisection method [26].

A. Reformulation of Problem (6)

By employing the Dinkelbach method, problem (6) can be reformulated as

$$\begin{aligned}&\min_{\{\mathbf{w}_{k,n}\},\ \{s_{k,n}\},\ \{c_{f,n}\},\ \varsigma} \ -C_B+\varsigma C_P\\ &\text{s.t. (6b)} - \text{(6f)}.\end{aligned} \tag{7}$$

The main motivation behind the use of the Dinkelbach variable $\varsigma \in \mathbb{R}$ is to decouple the fractional objective in (6) into a subtractive form that can be tackled more easily. It can be shown that there exists ς such that the optimal solution of (7) corresponds to that of (6).

Next, we introduce auxiliary variables $\{\hat{s}_{k,n}\}$ and $\{\mathbf{w}_k^j\}_{j\in\mathcal{K}\backslash\{k\}}$ which satisfy

$$\mathbf{w}_k^j=\mathbf{w}_k,\ \forall j\in\mathcal{K}\backslash\{k\},\ \forall k, \tag{8}$$

$$s_{k,n}=\hat{s}_{k,n},\ \forall k,n. \tag{9}$$

Note that (8) and (9) can be equivalently interpreted as introducing $K-1$ and 1 redundant copies of variables $\mathbf{w}_k$ and $s_{k,n}$, respectively.[3] Then, problem (7) can be equivalently expressed as

$$\min_{\bar{\mathcal{W}},\ \varsigma} \ -C_B+\varsigma C_P \tag{10a}$$

$$\text{s.t. (6d)} - \text{(6f)},\ \text{(8) and (9)}, \tag{10b}$$

$$\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum\limits_{j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_j^k|^2+\sigma_k^2}\geq\gamma_k,\ \forall k, \tag{10c}$$

$$s_{k,n}(1-\hat{s}_{k,n})=0,\ \forall k,n, \tag{10d}$$

$$0\leq\hat{s}_{k,n}\leq 1,\ \forall k,n, \tag{10e}$$

where $\bar{\mathcal{W}}\triangleq\{\{\mathbf{w}_k\},\ \{\mathbf{w}_k^j\}_{j\in\mathcal{K}\backslash\{k\}},\ \{s_{k,n}\},\ \{\hat{s}_{k,n}\},\ \{c_{f,n}\}\}$. The introduction of these auxiliary variables represents a critical step in developing the proposed PDD-based algorithm. Indeed, by adopting these new variables, we can partition the complete set of optimization variables into smaller non-overlapping subsets, or blocks, that can be optimized separately. Specifically, the joint optimization problem (7) can be decomposed into a number of subproblems which either admit closed-form solutions or can be solved via simple iterative approaches. Hence, through the introduction of auxiliary variables and judiciously exploiting the block structure, low-complexity algorithms can be devised for the optimization of each block of variables.

B. Algorithm Design

In this subsection, we aim to develop an efficient PDD-based algorithm to solve problem (10). To this end, the AL of problem (10) is first formulated as

$$\begin{aligned}&\min_{\bar{\mathcal{W}},\ \varsigma} \ -C_B+\varsigma\sum_{k=1}^{K}\|\mathbf{w}_k\|^2+P_\rho\\ &\text{s.t. (6d), (6f), (10c) and (10e)},\end{aligned} \tag{11}$$

where the penalty term

$$\begin{aligned}P_\rho\triangleq&\tfrac{1}{2\rho}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big((s_{k,n}(1-\hat{s}_{k,n})\\ &+\rho\lambda_{k,n})^2+(s_{k,n}-\hat{s}_{k,n}+\rho\hat{\lambda}_{k,n})^2\Big)\\ &+\tfrac{1}{2\rho}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\|\mathbf{w}_k-\mathbf{w}_k^j+\rho\boldsymbol{\mu}_{j,k}\|^2\\ &+\tfrac{1}{2\rho}\sum_{k=1}^{K}\sum_{n=1}^{N}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho\boldsymbol{\xi}_{k,n}\|^2,\end{aligned} \tag{12}$$

$\{\lambda_{k,n}\}$, $\{\hat{\lambda}_{k,n}\}$, $\{\boldsymbol{\mu}_{j,k}\}$ and $\{\boldsymbol{\xi}_{k,n}\}$ denote the dual variables corresponding to the constraints (10d), (9), (8) and (6e), respectively. $\mathbf{J}_n\triangleq[\mathbf{0}_{L\times(n-1)L},\mathbf{I}_{L\times L},\mathbf{0}_{L\times(N-n)L}]\in\{0,1\}^{L\times NL}$ is a binary selection matrix. The coefficient $\rho>0$ is used to control the size of the penalty (i.e., decreasing ρ increases the penalty).

Our proposed algorithm mainly consists of two embedded loops. In the outer loop, indexed by positive integer m, we update the Dinkelbach variable ς_m and either the dual variables $\{\lambda_{k,n}^m,\hat{\lambda}_{k,n}^m,\boldsymbol{\mu}_{j,k}^m,\boldsymbol{\xi}_{k,n}^m\}$ or the penalty parameter ρ_m, where the dependence of these variables on the outer iteration index m is now made explicit. In the inner loop, we employ the block successive upper-bound minimization (BSUM) method [27] to iteratively optimize the primal variables $\bar{\mathcal{W}}$ over selected blocks of variables while keeping the other variables fixed. In the following, we first develop the BSUM method in details, then present the update of the dual variables, the Dinkelbach variable and the penalty parameter, and finally summarize the overall algorithm.

In the inner loop, with fixed dual variables, penalty parameter and Dinkelbach variable, we propose to divide the primal variables into four blocks that will be treated separately, i.e., $\{s_{k,n}\}$, $\{\hat{s}_{k,n}\}$, $\{\mathbf{w}_k,\mathbf{w}_k^j\}$ and $\{c_{f,n}\}$. We now proceed with the optimization of each block.

1. Block $\{s_{k,n}\}$: The optimization problem of $\{s_{k,n}\}$ can be expressed as

$$\begin{aligned}&\min_{\{s_{k,n}\}} \ -\sum_{k=1}^{K}s_{k,n}\sum_{f=1}^{F}P_{k,f}c_{f,n}\\ &+\tfrac{1}{2\rho_m}\sum_{k=1}^{K}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho_m\boldsymbol{\xi}_{k,n}^m\|^2\\ &+\tfrac{1}{2\rho_m}\sum_{k=1}^{K}\Big((s_{k,n}(1-\hat{s}_{k,n})+\rho_m\lambda_{k,n}^m)^2+(s_{k,n}-\hat{s}_{k,n}+\rho_m\hat{\lambda}_{k,n}^m)^2\Big)\\ &\text{s.t. }\sum_{k=1}^{N}s_{k,n}\leq X_n,\ \forall n.\end{aligned} \tag{13}$$

[3] The introduction of these redundant copies may seem, at first glance, artificial. However, it will be clear later that with their help, the optimal solutions of certain subproblems can be obtained in nearly closed-form and consequently the corresponding, underlying problem (6) can be easily solved. We emphasize that in contrast to $s_{k,n}$ which only takes on binary values, its copy $\hat{s}_{k,n}$ is a continuous variable.

It can be seen that for each n, $\{s_{k,n}\}_{k=1}^{K}$ can be optimized separately in a parallel manner. In addition, problem (13) is a quadratic programming (QP) problem with linear constraint, whose optimal solution can be obtained in closed-form by resorting to its Lagrangian dual problem. The detailed derivation is omitted here for brevity.

2. Block $\{\mathbf{w}_k, \mathbf{w}_j^k\}_{j\in\mathcal{K}\setminus\{k\}}$: The corresponding optimization problem can be expressed as[4]

$$\begin{aligned}&\min_{\{\mathbf{w}_k,\mathbf{w}_j^k\}_{j\neq k}} \varsigma_m\|\mathbf{w}_k\|^2 + \tfrac{1}{2\rho_m}\sum_{n=1}^{N}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k + \rho_m\boldsymbol{\xi}_{k,n}^m\|^2\\ &+\tfrac{1}{2\rho_m}\sum_{j=1,j\neq k}^{K}\Big(\|\mathbf{w}_k - \mathbf{w}_k^j + \rho_m\boldsymbol{\mu}_{j,k}^m\|^2 + \|\mathbf{w}_j^j - \mathbf{w}_j^k + \rho_m\boldsymbol{\mu}_{k,j}^m\|^2\Big)\\ &\text{s.t. } \frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum\limits_{j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_j^k|^2+\sigma_k^2} \geq \gamma_k,\end{aligned} \tag{14}$$

which can be efficiently solved by resorting to the Lagrangian dual problem and employing the bisection method. The detailed derivation is included in Appendix A.

3. Block $\{\hat{s}_{k,n}\}$: We consider the following problem:

$$\begin{aligned}&\min_{\hat{s}_{k,n}} \tfrac{1}{2\rho_m}(s_{k,n}(1-\hat{s}_{k,n}) + \rho_m\lambda_{k,n}^m)^2\\ &+\tfrac{1}{2\rho_m}(s_{k,n} - \hat{s}_{k,n} + \rho_m\hat{\lambda}_{k,n}^m)^2,\\ &\text{s.t. } 0 \leq \hat{s}_{k,n} \leq 1.\end{aligned} \tag{15}$$

Problem (15) is also a QP problem with linear constraint, which therefore admits a closed-form solution.

4. Block $\{c_{f,n}\}$: The optimization problem, which is separable among different n, can be formulated as follows:

$$\begin{aligned}&\max_{\{c_{f,n}\}} \sum_{f=1}^{F}\kappa_{f,n}c_{f,n}\\ &\text{s.t. } c_{f,n} = 0 \text{ or } 1,\ \forall f,\\ &\sum_{f=1}^{F} c_{f,n} \leq Y_n,\end{aligned} \tag{16}$$

where $\kappa_{f,n} = \sum_{k=1}^{K} s_{k,n}P_{k,f}$. In essence, the aim of problem (16) is to determine which subset of Y_n files should be cached by RRH n. The optimal solution to such a problem is simply to cache the Y_n files that have the largest benefits, i.e.,

$$c_{f,n}^{\text{opt}} = \begin{cases} 1, \text{ if } f \in \mathcal{K}_n \\ 0, \text{ otherwise} \end{cases} \tag{17}$$

where $\mathcal{K}_n \triangleq \underset{\bar{\mathcal{K}}\subset\mathcal{F},|\bar{\mathcal{K}}|=Y_n}{\arg\max}\Big(\sum_{f\in\bar{\mathcal{K}}}\kappa_{f,n}\Big)$.

In the outer loop, the dual variables $\{\lambda_{k,n}^m,\ \hat{\lambda}_{k,n}^m,\ \boldsymbol{\mu}_{j,k}^m,\ \boldsymbol{\xi}_{k,n}^m\}$ can be updated by

$$\lambda_{k,n}^{m+1} = \lambda_{k,n}^m + \frac{1}{\rho_m}(s_{k,n}(1-\hat{s}_{k,n})), \tag{18a}$$

$$\hat{\lambda}_{k,n}^{m+1} = \hat{\lambda}_{k,n}^m + \frac{1}{\rho_m}(s_{k,n} - \hat{s}_{k,n}), \tag{18b}$$

$$\boldsymbol{\mu}_{j,k}^{m+1} = \boldsymbol{\mu}_{j,k}^m + \frac{1}{\rho_m}(\mathbf{w}_k - \mathbf{w}_k^j), \tag{18c}$$

$$\boldsymbol{\xi}_{k,n}^{m+1} = \boldsymbol{\xi}_{k,n}^m + \frac{1}{\rho_m}((1-s_{k,n})\mathbf{J}_n\mathbf{w}_k). \tag{18d}$$

The Dinkelbach variable and the penalty parameter can be updated as follows:

$$\varsigma_{m+1} = C_B^m/C_P^m, \tag{19}$$

$$\rho_{m+1} = q\rho_m, \tag{20}$$

where $q < 1$ is a control parameter used to increase the value of the penalty term P_ρ in (12) during each outer iteration.

Besides, we denote the maximum constraint violation among all the equality constraints in problem (10) as ϖ, whose formal definition appears in (21) at the top of the next page. This is an important quantity that can be employed to determine if the proposed algorithm converges, and whether we should update the dual variables or increase the penalty parameter.

The main steps of the proposed PDD-based algorithm are summarized in Algorithm 1. We observe that the complexity for solving problem (13), (15) and (16) is almost negligible compared with that of solving problem (14). Therefore, the overall complexity can be expressed as[5] $\mathcal{O}\left(M^{\max}i^{\max}N^3L^3K^4\log_2\left(\frac{\lambda_{\max}-\lambda_{\min}}{\varepsilon}\right)\right)$, where $\lambda_{\max} = \max\{\overline{\lambda}_k\}_{k\in\mathcal{K}}$ and $\lambda_{\min} = \min\{\underline{\lambda}_k\}_{k\in\mathcal{K}}$ denote the upper and lower bounds of the corresponding dual variables as detailed in Appendix A and ε denotes the precision of the bisection method.

Algorithm 1 The Proposed PDD-based Algorithm

```
Initialize $\{\mathbf{w}_k^j\}^0, \{c_{f,n}\}^0, \{s_{k,n}\}^0 = \{\hat{s}_{k,n}\}^0, \eta_0, \varrho_0, q$ and $\varsigma_0$.
Set the outer iteration index $m = 0$.
repeat
    Set the inner iteration index $i = 0$.
    repeat
        Update $\{\mathbf{w}_k, \mathbf{w}_j^k\}_{j\neq k}^{i+1}$ by solving problem (14).
        Update $\{\hat{s}_{k,n}\}^{i+1}$ by solving problem (15).
        Update $\{c_{f,n}\}^{i+1}$ by (17).
        Update $\{s_{k,n}\}^{i+1}$ by solving problem (13).
        $i \leftarrow i + 1$.
    until some convergence condition is met.
    Assign $\bar{\mathcal{W}}^i$ to $\bar{\mathcal{W}}^0$. Calculate $\varpi$ via (21).
    If $\varpi \leq \eta_m$, update the dual variables via (18), otherwise set $\rho_{m+1} =$
    $q\rho_m$. Set $\varrho_{m+1} = q\varrho_m$, $\eta_{m+1} = \varrho_{m+1}^{1/6}$, update the Dinkelbach
    parameter via (19) and $m \leftarrow m + 1$.
until some convergence condition is met.
```

Remark 1: A complete characterization of the convergence properties of Algorithm 1 is rather involved and falls outside the scope of the present work, where the focus is on algorithm design and performance study. Nevertheless, as a future research direction, we here provide a schematic outline of the various steps involved in the convergence proof of Algorithm 1. Firstly, we should prove that the caching efficiency sequence obtained by the successive iterations are non-decreasing after a finite number of iterations due to the property of the Dinkelbach and the PDD methods. Then, resorting to the convergence of the subsequence and the optimality condition of the Dinkelbach subproblem, we could show that the limit points of the subsequence are stationary points of the Dinkelbach subproblem. Subsequently, depending on the convergence of the caching efficiency sequence, we could infer that the convergent point of the Dinkelbach variable ς is actually equal to the caching efficiency value evaluated at the limit point of the subsequence. Finally, we would show that the limit

[4] Due to the additive nature of the AL, we only need to consider a single value of k at a time, i.e., optimization for other values of k can be done separately in parallel.

[5] It is worth noting that the main component that constitutes the complexity of Algorithm 1 is performing the eigenvalue decomposition of a $KNL\times KNL$ matrix multiple times. For general matrices, the computational complexity would be $\mathcal{O}(K^3N^3L^3)$. However, since $\mathbf{A}_k$ and $\mathbf{D}_k$ (which are defined in Appendix A) are sparse matrices, the corresponding complexity can be significantly reduced by further exploiting the special structure of $\mathbf{A}_k$ and $\mathbf{D}_k$, which we do not detail in this work.

$$\varpi = \max_{\forall k,j,n} \left\{ |s_{k,n}(1-\hat{s}_{k,n})|,\ |s_{k,n}-\hat{s}_{k,n}|,\ \|\mathbf{w}_k - \mathbf{w}_k^j\|_\infty,\ \|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k\|_\infty \right\} \tag{21}$$

points of the subsequence are stationary points of the original problem (6) on the basis of the first order optimality conditions of the Dinkelbach subproblem and the convergence property of the Dinkelbach variable ς.

IV. Simulation Results

In this section, the performance of the proposed PDD-based algorithm is evaluated numerically. The following system parameter values are used throughout unless otherwise specified: $N = 7$, $K = 6$, $L = 2$, $F = 700$ and $\sigma_k^2 = -70$ dBm, $\forall k$. Each RRH is located at the center of an hexagonal cell, with the distance between adjacent RRHs set to 100 meters. The users are uniformly and independently distributed within the service area of the RRHs. We consider Rayleigh fading channels with large-scale pathloss (in dB) modeled as $-147.3 - 43.3\log_{10}D$, where the propagation distance D is measured in kilometers. For simplicity, we assume that all users have the same SINR requirements, and that each RRH has the same local storage size and can support the same number of users, i.e.: $\gamma_k = \gamma, \forall k \in \mathcal{K}$, $X_n = X = 3$, $Y_n = Y$, $\forall n \in \mathcal{N}$. In Algorithm 1, we use the following parameter values: $M^{\max} = 500$, $i^{\max} = 20$, $\eta_0 = 100$, $\varrho_0 = 100$, $q = 0.95$ and $\varsigma_0 = 1$. The simulations are done on a computer with Intel (i7-6700HQ) CPU running at 2.60GHz and 8GB RAM.

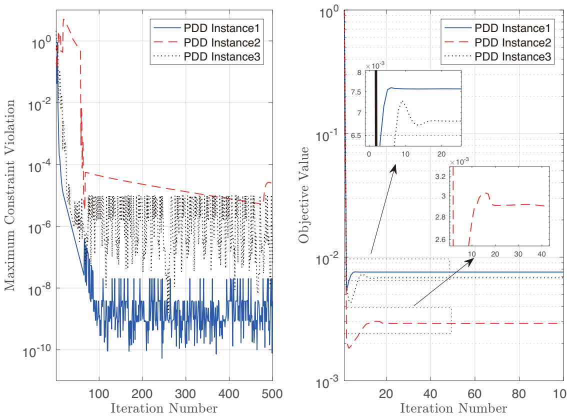

Fig. 1. Maximum constraint violation and objective value versus outer iteration number for 3 problem instances.

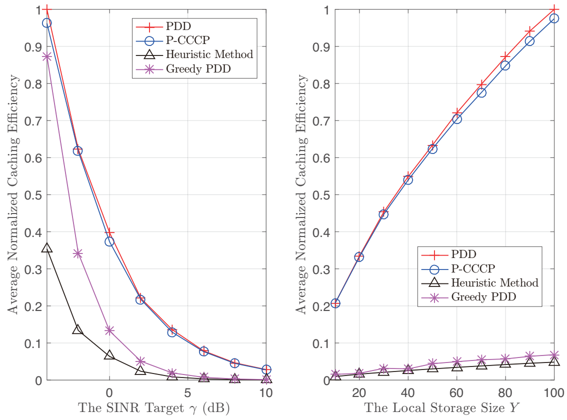

Fig. 2. Average normalized caching efficiency versus the SINR target threshold of each user and the local storage size of each RRH.

In the simulations, we assume that the popularity of the files can be measured based on the number and behavior of the requests. Specifically, two types of files can be requested, both with similar Zipf distribution with parameter 0.4 [20]. Also, two types of users with different file preferences are considered, i.e.: Type 1 users prefer Type 1 files with probability 0.8 and Type 2 files with probability 0.2; Type 2 users prefer Type 2 files with probability 0.8 and Type 1 files with probability 0.2.

In Fig. 1, we investigate the typical convergence behavior of Algorithm 1 in the case $Y = 100$ and $\gamma = 6$dB. The results show that the proposed PDD-based algorithm converges in a few hundreds of outer iterations. In Fig. 2, we examine the average normalized caching efficiency as a function of the SINR target threshold γ (with $Y = 100$) and the local storage size Y with $\gamma = 5$dB fixed. For comparison, we also provide the performance of an algorithm based on the penalty concave-convex procedure (P-CCCP), a separate PDD-based algorithm and a heuristic method. The P-CCCP algorithm aims to solve problem (6) but is obtained by introducing suitable auxiliary variables and penalizing certain constraints to the objective function, while in the separate PDD-based algorithm, the fronthaul traffic reduction and total transmission power are separately optimized. In the heuristic method, the RRH clustering is simply determined based on the distances between the RRHs and users. As can be seen from Fig. 2, the performance of the proposed PDD-based algorithm is very close to that of the P-CCCP algorithm, and they both outperform the heuristic method and the separate PDD-based algorithm in terms of cache efficiency. The achieved caching efficiency is in inverse proportion to the SINR target γ of each user. This is because as γ increases, the RRHs need to increase their transmission power to satisfy the more stringent SINR requirements, while the fronthaul reduction barely changes with different γ. Besides, as the SINR target γ decreases, the achieved caching efficiency of the separate PDD-based algorithm becomes comparable to that of the joint design algorithm, mainly due to the fact that in this case the transmission power is not the dominant factor in achieving high caching efficiency. Also, with the increase of the local storage size in each RRH, more fronthaul reduction can be achieved, which results into higher caching efficiency.

V. Conclusions

In this work, we have studied the problem of joint transceiver design for a content-aware C-RAN system. An optimization framework was presented in which the ratio between fronthaul reduction and transmission power cost (i.e. caching efficiency) was employed as the objective function. A new design algorithm which utilizes the PDD framework was proposed to jointly optimize the RRH downlink beamforming vectors, the RRH clustering and the caching placement. Simulation results were presented, showing that the proposed algorithm exhibits very good performance.

Appendix A
Solution of Problem (14)

We first introduce the following auxiliary variables:

$$\mathbf{x}_k = \left[(\mathbf{w}_1^k)^H, \cdots, (\mathbf{w}_k^k)^H, \cdots, (\mathbf{w}_K^k)^H\right]^H, \tag{22}$$

$$\mathbf{P}_k = [\mathbf{0}_{NL\times(k-1)NL}, \mathbf{I}_{NL\times NL}, \mathbf{0}_{NL\times(K-k)NL}] \in \{0,1\}^{NL\times KNL}, \tag{23}$$

such that $\mathbf{P}_j\mathbf{x}_k = \mathbf{w}_j^k$ holds. We then observe that problem (14) can be equivalently formulated as follows:

$$\begin{aligned} &\min_{\mathbf{x}_k} \ \mathbf{x}_k^H\mathbf{A}_k\mathbf{x}_k + \mathbf{x}_k^H\mathbf{b}_k + \mathbf{c}_k\mathbf{x}_k \\ &\text{s.t. } \mathbf{x}_k^H\mathbf{D}_k\mathbf{x}_k \geq \sigma_k^2, \end{aligned} \tag{24}$$

where

$$\begin{aligned} \mathbf{A}_k \triangleq &\left(\tfrac{K-1}{2\rho_m} + \eta\right)\mathbf{P}_k^H\mathbf{P}_k + \tfrac{1}{2\rho_m}\sum_{j=1,j\neq k}^{K}\mathbf{P}_j^H\mathbf{P}_j \\ &+ \tfrac{1}{2\rho_m}\sum_{n=1}^{N}(1-s_{k,n})^2\mathbf{P}_k^H\mathbf{J}_n^H\mathbf{J}_n\mathbf{P}_k, \end{aligned} \tag{25}$$

$$\begin{aligned} \mathbf{b}_k \triangleq &\tfrac{1}{2\rho_m}\Big(\sum_{n=1}^{N}(1-s_{k,n})\mathbf{P}_k^H\mathbf{J}_n^H\rho_m\boldsymbol{\xi}_{k,n}^m \\ &+ \sum_{j=1,j\neq k}^{K}\Big(\mathbf{P}_k^H\left(\rho_m\boldsymbol{\mu}_{j,k}^m - \mathbf{w}_k^j\right) - \mathbf{P}_j^H\left(\mathbf{w}_j^j + \rho_m\boldsymbol{\mu}_{k,j}^m\right)\Big)\Big), \end{aligned} \tag{26}$$

$$\begin{aligned} \mathbf{c}_k \triangleq &\tfrac{1}{2\rho_m}\Big(\sum_{n=1}^{N}\rho_m\boldsymbol{\xi}_{k,n}^m(1-s_{k,n})\mathbf{J}_n\mathbf{P}_k \\ &+ \sum_{j=1,j\neq k}^{K}\Big(\left(\rho_m\boldsymbol{\mu}_{j,k}^m - \mathbf{w}_k^j\right)^H\mathbf{P}_k - \left(\mathbf{w}_j^j + \rho_m\boldsymbol{\mu}_{k,j}^m\right)^H\mathbf{P}_j\Big)\Big), \end{aligned} \tag{27}$$

$$\mathbf{D}_k \triangleq \frac{\mathbf{P}_k^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{P}_k}{\gamma_k} - \sum_{j\neq k}^{K}\mathbf{P}_j^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{P}_j. \tag{28}$$

Since $\mathbf{A}_k$ is a full-rank matrix, we can decompose it as $\mathbf{A}_k = \mathbf{A}_k^{\frac{1}{2}}\mathbf{A}_k^{\frac{1}{2}}$. Furthermore, by introducing the substitution $\mathbf{y}_k = \mathbf{A}_k^{\frac{1}{2}}\mathbf{x}_k$, problem (24) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{y}_k} \ \mathbf{y}_k^H\mathbf{y}_k + \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k + \mathbf{c}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k \\ &\text{s.t. } \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k \geq \sigma_k^2. \end{aligned} \tag{29}$$

Next, we focus on the optimal solution of problem (29). Its Lagrange function can be expressed as

$$\mathcal{L} = \mathbf{y}_k^H\mathbf{y}_k + \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k + \mathbf{c}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k + \lambda_k\left(\sigma_k^2 - \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k\right), \tag{30}$$

where λ_k denotes the dual variable. Employing the eigenvalue decomposition, we can write $\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}} = \mathbf{VSV}^{-1}$, where $\mathbf{V}$ is unitary and $\mathbf{S}$ is diagonal. Note that in order for the problem to be feasible, the dual variable should satisfy $\mathbf{I} - \lambda_k\mathbf{VSV}^{-1} \succeq \mathbf{0}$, which is equivalent to $\mathbf{I} - \lambda_k\mathbf{S} \succeq \mathbf{0}$. Taking the derivative of $\mathcal{L}$ with respect to $\mathbf{y}_k^*$, we obtain

$$\mathbf{y}_k + \mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k - \lambda_k\mathbf{VSV}^{-1}\mathbf{y}_k = \mathbf{0}, \tag{31}$$

which is equivalent to

$$\mathbf{y}_k = \mathbf{V}(\mathbf{I} - \lambda_k\mathbf{S})^{-1}\mathbf{V}^{-1}(-\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k). \tag{32}$$

The Lagrange dual variable λ_k can be obtained by the bisection method, and the corresponding upper bound $\overline{\lambda}_k$ and lower bound $\underline{\lambda}_k$ can be found by resorting to $\mathbf{I} - \lambda_k\mathbf{S} \succeq 0$, which results into $\overline{\lambda}_k = \frac{1}{\max(0,\, \mathrm{diag}(\boldsymbol{S}))}$ and $\underline{\lambda}_k = 0$.

References

[1] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[2] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[3] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118–125, Dec. 2014.

[4] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.

[5] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 41, Feb. 2015.

[6] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

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

[8] M. Hong, R.-Y. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogenous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

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

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

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

[12] B. Hu, C. Hua, J. Zhang, C. Chen, and X. Guan, "Joint fronthaul multicast beamforming and user-centric clustering in downlink C-RANs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5395–5409, Aug. 2017.

[13] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[14] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[15] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *IEEE Int. Symp. on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Sep. 2014, pp. 1370–1374.

[16] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *IEEE Int. Symp. on Inform. Theory (ISIT)*, Jun. 2015, pp. 809–813.

[17] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *IEEE Int. Conf. on Commun. (ICC)*, May 2016, pp. 1–6.

[18] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.

[19] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[20] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2016, pp. 3521–3525.

[21] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[22] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, pp. 492–498, Mar. 1967.

[23] Q. Shi, M. Hong, X. Fu, and T.-H. Chang, "Penalty dual decomposition method for nonsmooth nonconvex optimization," *arXiv preprint*, 2017. [Online]. Available: https://arxiv.org/abs/1712.04767v1

[24] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, no. 5, pp. 303–320, 1969.

[25] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," *in Optimization, R. Fletcher, Ed., Academic Press, New York*, pp. 283–298, 1972.

[26] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[27] M. Hong, M. Razaviyayn, Z. Q. Luo, and J. S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.