

Content-Centric Multicast Beamforming in Cache-Enabled Cloud Radio Access Networks

Hao Zhou*, Meixia Tao*, Er kai Chen*, Wei Yu†

*Dept. of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

†Dept. of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

Email: {zhouhao_zh, mxtao, cek1006}@sjtu.edu.cn, weiyu@comm.utoronto.ca

Abstract—Multicast transmission and wireless caching are effective ways of reducing air and backhaul traffic load in wireless networks. This paper proposes to incorporate these two key ideas for content-centric transmission in a cloud radio access network (RAN) where multiple base stations (BSs) are connected to a central processor (CP) via finite-capacity backhaul links. Each BS has a cache with finite storage size and is equipped with multiple antennas. The BSs cooperatively transmit contents, either stored in the local cache or fetched from the CP, to multiple users in the network. Users requesting a same content form a multicast group and are served by a same cluster of BSs cooperatively using multicast beamforming. Assuming fixed cache placement, this paper investigates the joint design of multicast beamforming and content-centric BS clustering by formulating an optimization problem of minimizing the total network cost under the quality-of-service (QoS) constraints for each multicast group. The network cost involves both the transmission power and the backhaul cost. We model the backhaul cost using the mixed ℓ_0/ℓ_2 -norm of beamforming vectors. To solve this non-convex problem, we first approximate it using the semidefinite relaxation (SDR) method and concave smooth functions. We then propose a difference of convex functions (DC) programming algorithm to obtain suboptimal solutions and show the connection of three smooth functions. Simulation results validate the advantage of multicasting and show the effects of different cache size and caching policies in cloud RAN.

I. INTRODUCTION

Cloud radio access network (RAN) is an emerging network architecture capable of exploiting the advantage of multicell cooperation in the future fifth-generation (5G) wireless system [1]. In a cloud RAN, the base stations (BSs) are connected to a central processor (CP) via digital backhaul links, thus enabling joint data processing and precoding capabilities across multiple BSs. This paper proposes a *content-centric* view for cloud RAN design. We equip the BSs with finite-size cache, where popular contents desired by multiple users can be stored. We formulate and solve a network optimization problem while accounting for the finite-capacity backhaul links between the BSs and the CP.

To address the issue of limited backhaul, previous works on wireless cooperative networks [2]–[4] consider the problem of minimizing the backhaul traffic and transmission power by designing sparse beamformer and user-centric BS clustering. Further, [5] considers the weighted sum rate (WSR) optimization problem under per-BS backhaul constraints. However, all

these works focus on the unicast scenario and promote a user-centric view of system design without considering the effect of caching.

Recently, *wireless caching* has been investigated as an effective way of reducing peak traffic and backhaul load. By deploying caches at BSs and placing popular contents in them in advance, the issue of limited backhaul capacity can be addressed fundamentally. In [6], the authors show that with small or even no backhaul capacity, femto-caching can support high demand of wireless video distribution. In [7], the upper and lower bounds of the capacity of the caching system are derived, and it is shown that the network capacity could be further improved by using coded multicasting for content delivery. These studies motivate us to consider the *cache-enabled* cloud RAN, where each BS is equipped with a cache with finite storage size. Compared with cooperative networks without caching, cache-enabled cloud RAN can fundamentally reduce the backhaul cost and support more flexible BS clustering. We note that in [8], a similar wireless caching network has been considered, where the authors study the data assignment and unicast beamforming design.

Different from previous work on unicast [2]–[4], [8], where data is transmitted to each user individually no matter whether the actual contents requested by different users are the same or not, this paper focuses on the problem of multicast transmission. We assume that multiple users can request the same content, and the content is delivered using multicast beamforming to these users on the same resource block. Compared with traditional unicast, multicast can improve energy and spectral efficiency. In addition, since the popular contents cached in the BSs are possibly requested by multiple users, multicast could better exploit the potential of wireless caching.

This paper studies the joint design of multicast beamforming and content-centric BS clustering, which differs from the fixed BS clustering in coordinated multicell multicast networks [9] or user-centric BS clustering in unicast systems [3]. In each scheduling interval, the BS clustering is dynamically optimized with respect to each multicast group. We formulate an optimization problem with the objective of minimizing the total power consumption as well as the backhaul cost under the quality-of-service (QoS) constraints for each multicast group. The backhaul cost is formulated as a function of the mixed ℓ_0/ℓ_2 -norm of the beamforming vectors. The challenge in solving such a problem is due to both the non-convex QoS

This work is supported by the National Natural Science Foundation of China under grants 61322102 and 61221001.

constraints and the ℓ_0 -norm in the backhaul cost. In this paper, we first use the semidefinite relaxation (SDR) method [10] to handle the non-convex QoS constraints. We then adopt the smooth function approach in sparse signal processing to approximate the ℓ_0 -norm with concave smooth functions.

In sparse signal processing, one approach to handle the ℓ_0 -norm minimization problem is to approximate the ℓ_0 -norm with its reweighted ℓ_1 -norm [11] and update the weight factors iteratively. Another approach is the smooth function method [4], where the authors use Gaussian family functions to approximate the ℓ_0 -norm. The smooth function method is a better approximation to the ℓ_0 -norm but its performance highly depends on the smoothness factor of the approximation function. In this paper, we adopt the smooth function approach and solve the approximated problem with a difference of convex functions (DC) algorithm [12]. We explore the use of three different smooth functions, the logarithmic function, the exponential function, and the arctangent function, and show that with a particular weight updating rule, all three are equivalent to the reweighted ℓ_1 -norm minimization [11]. Simulation results are presented to illustrate the performance of proposed algorithm and the benefit of wireless caching.

Notations: Boldface uppercase letters denote matrices and boldface lowercase letters denote column vectors. The sets of complex numbers and binary numbers are denoted as \mathbb{C} and \mathbb{B} respectively. The statistical expectation, transpose and Hermitian transpose are denoted as $\mathbb{E}(\cdot)$, $(\cdot)^T$ and $(\cdot)^H$ respectively. The Frobenius norm and the ℓ_0 -norm are denoted as $\|\cdot\|_2$ and $\|\cdot\|_0$ respectively. An all-one vector of length M is denoted as $\mathbf{1}_M$. An all-zero vector of length M is denoted as $\mathbf{0}_M$. The inner product of matrices \mathbf{X} and \mathbf{Y} is defined as $\langle \mathbf{X}, \mathbf{Y} \rangle = \text{Tr}(\mathbf{X}^H \mathbf{Y})$. For a square matrix $\mathbf{S}_{M \times M}$, $\mathbf{S} \succeq \mathbf{0}$ means that \mathbf{S} is positive semidefinite.

II. SYSTEM MODEL

A. Signal Model

We consider a cache-enabled cloud RAN with one CP, L BSs and K mobile users. Each BS is equipped with N_t antennas and each user has a single antenna. Scheduling and beamformer design are done at the CP. Each BS is connected to the CP via a finite-capacity backhaul link. The total number of contents is F ; different contents are independent. The CP stores all the contents and there is a cache at each BS, which stores finite number of contents. Each user requests a content according to the content popularity, and users requesting the same content form a multicast group. We assume that the total number of multicast groups is M . The set of users in group m is denoted as \mathcal{K}_m with $\sum_{m=1}^M |\mathcal{K}_m| = K$.

We study the cooperative downlink multicast transmission and dynamic content-centric BS clustering. Each group m is served by a cluster of BSs cooperatively, denoted as \mathcal{Q}_m . In each scheduling interval, the BS clustering $\{\mathcal{Q}_m\}_{m=1}^M$ is dynamically optimized by the CP. For example, in Fig. 1, the instantaneous BS clusters for different groups are $\mathcal{Q}_1 = \{1, 2, 3\}$, $\mathcal{Q}_2 = \{2\}$ and $\mathcal{Q}_3 = \{2, 3\}$, respectively. For BS 3, since it serves both group 1 and group 3, it should acquire

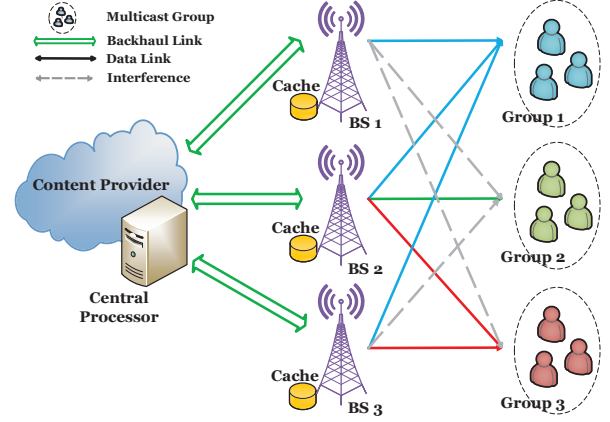


Fig. 1: An example of downlink cloud RAN with $M = 3$ groups and $L = 3$ cache-enabled BSs connected to a CP via digital backhaul links, where each multicast group is served by a cluster of BSs.

the contents for these two groups either from its local cache or through backhaul.

We denote the aggregate beamforming vector of group m from all BSs as $\mathbf{w}_m \in \mathbb{C}^{LN_t \times 1} = [\mathbf{w}_{1,m}^H, \mathbf{w}_{2,m}^H, \dots, \mathbf{w}_{L,m}^H]^H$, where $\mathbf{w}_{l,m} \in \mathbb{C}^{N_t \times 1}$ is the beamforming vector for group m at BS l . Note that the BS clustering is implicitly defined by the beamforming vectors. If the beamforming vector $\mathbf{w}_{l,m}$ is $\mathbf{0}_{N_t}$, then BS l does not serve group m and is thus not in \mathcal{Q}_m . On the other hand, if $\mathbf{w}_{l,m} \neq \mathbf{0}_{N_t}$, BS l is part of the serving cluster of group m . Thus, the size of the BS cooperation cluster for group m can be expressed as a mixed ℓ_0/ℓ_2 -norm of the beamforming vector $\{\mathbf{w}_{l,m}\}_{l=1}^L$, i.e. $|\mathcal{Q}_m| = \sum_{l=1}^L \|\mathbf{w}_{l,m}\|_2^2$.

We denote the data symbol of the content requested by group m as $s_m \in \mathbb{C}$, with $\mathbb{E}[|s_m|^2] = 1$. For user $k \in \mathcal{K}_m$, its received downlink signal y_k can be written as

$$y_k = \mathbf{h}_k^H \mathbf{w}_m s_m + \sum_{n \neq m} \mathbf{h}_k^H \mathbf{w}_n s_n + z_k, \quad (1)$$

where $\mathbf{h}_k \in \mathbb{C}^{LN_t \times 1}$ is the network-wide channel vector from all BSs to user k and $z_k \sim \mathcal{CN}(0, \sigma^2)$ is the additive noise.

The received signal-to-interference-plus-noise ratio (SINR) at user $k \in \mathcal{K}_m$ is

$$\text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_m|^2}{\sum_{n \neq m} |\mathbf{h}_k^H \mathbf{w}_n|^2 + \sigma^2}. \quad (2)$$

We define the target SINR vector as $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \dots, \gamma_M]$ with each element γ_m being the target SINR to be achieved by the users in group m . In this paper, we consider the fixed rate transmission as in [3], where the transmission rate for group m is set as $R_m = \log_2(1 + \gamma_m)$. Thus, to successfully decode the message, for any user $k \in \mathcal{K}_m$, its achievable data rate should be larger than R_m , that is, for $\forall m, k \in \mathcal{K}_m$, $\log_2(1 + \text{SINR}_k) \geq R_m$.

B. Cache Model

We assume that each content has normalized size of 1 and the local storage size of BS l is F_l ($F_l < F$), which is also the maximum number of contents it can store. Therefore, we define a cache placement matrix $\mathbf{C} \in \mathbb{B}^{L \times F}$, where $c_{l,f} = 1$ means the content f is cached in BS l and $c_{l,f} = 0$ means the opposite. Note that $\forall l, \sum_{f=1}^F c_{l,f} \leq F_l$.

We assume that the cache placement is static, that is, matrix \mathbf{C} is fixed and is known at the CP (similar assumptions have been made in previous literature, e.g., [8]). This assumption is based on the fact that the optimization of cache placement is performed on a large time scale; while the beamforming design is done on a small time scale of channel coherence time. Hence, it is reasonable to assume that during a short scheduling interval, the cache placement policy remains unchanged.

C. Cost Model

We consider the total network cost which consists of both the transmission power and the backhaul cost. Let f_m denote the content requested by users in multicast group m . For BS l in \mathcal{Q}_m , if content f_m is in its cache, it can access the content directly without costing backhaul. On the contrary, if content f_m is not cached, BS l needs to first fetch this content from the CP via the backhaul link. Since the data rates of fetching the contents from the CP need to be as large as the content-delivery rate, the backhaul cost in this case is modeled as the transmission rates of multicast groups.

The total backhaul cost at all BSs can be written as

$$C_B = \sum_{m=1}^M \sum_{l=1}^L \|\mathbf{w}_{l,m}\|_2^2 R_m (1 - c_{l,f_m}). \quad (3)$$

The total network cost can be written as

$$C_N = \underbrace{\eta \sum_{m=1}^M \|\mathbf{w}_m\|_2^2}_{\text{Power Consumption}} + \underbrace{\sum_{m=1}^M \sum_{l=1}^L \|\mathbf{w}_{l,m}\|_2^2 R_m (1 - c_{l,f_m})}_{\text{Backhaul Cost}}, \quad (4)$$

where η is a weight parameter. By adjusting the value of η , we can emphasize on one cost versus the other.

Note that in a network without caching, there is a tradeoff between power and backhaul cost. To reduce power consumption, each group can be served by more BSs, which increases backhaul cost. However, in cache-enabled cloud RAN, for each group, the BSs caching the requested content can be involved in the cooperative cluster of the group without costing extra backhaul.

III. PROBLEM FORMULATION AND APPROXIMATION

In this section, we present the optimization problem of minimizing the total network cost by jointly designing multicast beamforming and BS clustering. We show that this problem is a non-convex problem and further approximate it with two steps.

A. Problem Formulation

Our objective is to minimize the total network cost, under the constraints of the peak transmission power at each BS and the SINR requirement of each multicast group.

The optimization problem is formulated as

$$\mathcal{P}_0 : \quad \underset{\{\mathbf{w}_m\}_{m=1}^M}{\text{minimize}} \quad C_N \quad (5a)$$

$$\text{subject to} \quad \text{SINR}_k \geq \gamma_m, \forall m, k \in \mathcal{K}_m \quad (5b)$$

$$\sum_{m=1}^M \|\mathbf{w}_{l,m}\|_2^2 \leq P_l, \forall l \quad (5c)$$

where P_l is the peak transmission power at BS l .

Problem \mathcal{P}_0 is a non-convex problem, where the non-convexity comes from both the ℓ_0 -norm in the objective function and the SINR constraints. Unlike traditional unicast beamforming problem where the non-convex SINR constraints can be transformed to a second-order cone programming (SOCP) problem and the optimal solutions can be obtained with convex optimization, the multicast beamforming problem is NP-hard [10]. In this paper, we use two techniques to approximate problem \mathcal{P}_0 , namely *SDR relaxation* and *ℓ_0 -norm approximation*. The overall procedure to solve \mathcal{P}_0 is shown in Fig. 2, with each step elaborated in following sections.

B. Step 1 – SDR Relaxation

In both single-cell [10] and multicell [9] scenarios, the semidefinite relaxation (SDR) method has been used to deal with the non-convex SINR constraints in multicast beamforming design problems.

We define two sets of matrices $\{\mathbf{W}_m \in \mathbb{C}^{LN_t \times LN_t}\}_{m=1}^M$ and $\{\mathbf{H}_k \in \mathbb{C}^{LN_t \times LN_t}\}_{k=1}^K$ as

$$\mathbf{W}_m = \mathbf{w}_m \mathbf{w}_m^H \quad \text{and} \quad \mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H, \quad \forall m, k. \quad (6)$$

We further define a set of selecting matrices $\{\mathbf{J}_l\}_{l=1}^L$, where each matrix $\mathbf{J}_l \in \mathbb{B}^{LN_t \times LN_t}$ is a diagonal matrix defined as

$$\mathbf{J}_l = \text{diag} \left(\left[\mathbf{0}_{(l-1)N_t}^H, \mathbf{1}_{N_t}^H, \mathbf{0}_{(L-l)N_t}^H \right] \right), \quad \forall l. \quad (7)$$

Therefore, we have

$$\|\mathbf{w}_{l,m}\|_2^2 = \text{Tr}(\mathbf{W}_m \mathbf{J}_l), \forall l, m. \quad (8)$$

By adopting the SDR method, problem \mathcal{P}_0 can be relaxed and rewritten as

\mathcal{P}_{SDR} :

$$\underset{\{\mathbf{W}_m\}_{m=1}^M}{\text{minimize}} \quad \sum_{m=1}^M \eta \text{Tr}(\mathbf{W}_m) + \sum_{m=1}^M \sum_{l=1}^L \alpha_{l,m} \|\text{Tr}(\mathbf{W}_m \mathbf{J}_l)\|_0 \quad (9a)$$

$$\text{subject to} \quad \frac{\text{Tr}(\mathbf{W}_m \mathbf{H}_k)}{\sum_{n \neq m}^M \text{Tr}(\mathbf{W}_n \mathbf{H}_k) + \sigma^2} \geq \gamma_m, \forall m, k \in \mathcal{K}_m \quad (9b)$$

$$\sum_{m=1}^M \text{Tr}(\mathbf{W}_m \mathbf{J}_l) \leq P_l, \quad \forall l \quad (9c)$$

$$\mathbf{W}_m \succeq \mathbf{0}, \quad \forall m \quad (9d)$$

where, to further simplify the mathematical representation, we have defined a set of constants $\{\alpha_{l,m}\}_{l=1,\dots,L}^{m=1,\dots,M}$ with $\alpha_{l,m} = R_m(1 - c_{l,f_m})$.

The SINR constraints in problem \mathcal{P}_{SDR} are convex. We denote the resulting optimal $\{\mathbf{W}_m\}$ after solving problem \mathcal{P}_{SDR} as $\{\mathbf{W}_m^*\}$. If \mathbf{W}_m^* is already rank-one, then for group m the optimal aggregate beamformer \mathbf{w}_m^* of problem \mathcal{P}_0 can be obtained by applying eigen-value decomposition to \mathbf{W}_m^* as $\mathbf{W}_m^* = \lambda_m^* \hat{\mathbf{w}}_m \hat{\mathbf{w}}_m^H$ and taking $\mathbf{w}_m^* = \sqrt{\lambda_m^*} \hat{\mathbf{w}}_m$. Otherwise, the beamforming vectors $\{\mathbf{w}_m\}$ can be generated with the *randomization* method used in [9] and [10].

C. Step 2 – ℓ_0 -norm Approximation

To solve the problem \mathcal{P}_{SDR} , we further approximate the non-convex ℓ_0 -norm in the objective with a continuous function denoted as $f(x)$. Specifically, we consider three frequently used smooth functions: logarithmic function, exponential function and arctangent function [13], defined as

$$f(\mathbf{X}) = \begin{cases} \log\left(\frac{\text{Tr}(\mathbf{X}) + \theta}{\theta}\right), & \text{for log-function} \\ 1 - e^{-\frac{\text{Tr}(\mathbf{X})}{\theta}}, & \text{for exp-function} \\ \frac{2}{\pi} \arctan\left(\frac{\text{Tr}(\mathbf{X})}{\theta}\right), & \text{for atan-function} \end{cases} \quad (10)$$

where θ is a parameter to adjust the smoothness of the functions. In all three cases, with larger θ , the function is smoother but is a worse approximation to the ℓ_0 -norm.

In [4], the authors use the Gaussian family smooth functions, where the ℓ_0 -norm of $\|\mathbf{w}\|_2$ is approximated with $f(\|\mathbf{w}\|_2) = 1 - \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\theta^2}\right)$. In this work, by adopting the exponential smooth function, the ℓ_0 -norm is approximated with $f_{exp}(\mathbf{W}) = 1 - \exp\left(-\frac{\text{Tr}(\mathbf{W})}{\theta}\right)$. Comparing $f(\|\mathbf{w}\|_2)$ and $f_{exp}(\mathbf{W})$, we can see that the exponential smooth function in (10) has the same approximation effect as the Gaussian smooth function in [4].

With smooth function, \mathcal{P}_{SDR} can be rewritten as

\mathcal{P}_{SF} :

$$\text{minimize}_{\{\mathbf{W}_m\}_{m=1}^M} \sum_{m=1}^M \eta \text{Tr}(\mathbf{W}_m) + \sum_{m=1}^M \sum_{l=1}^L \alpha_{l,m} f(\mathbf{W}_m \mathbf{J}_l) \quad (11a)$$

$$\text{subject to} \quad (9b), (9c), (9d). \quad (11b)$$

For ease of presentation, we express the objective as the summation of two functions $G(\mathbf{W})$ and $F(\mathbf{W})$, defined as

$$G(\mathbf{W}) = \sum_{m=1}^M \eta \text{Tr}(\mathbf{W}_m) \quad \text{and} \quad F(\mathbf{W}) = \sum_{m=1}^M \sum_{l=1}^L \alpha_{l,m} f_{l,m}, \quad (12)$$

where $f_{l,m} = f(\mathbf{W}_m \mathbf{J}_l), \forall l, m$.

We see that $G(\mathbf{W})$ and $F(\mathbf{W})$ are an affine and a concave function of \mathbf{W} , respectively, so problem \mathcal{P}_{SF} can be viewed as the difference of two continuous convex functions with convex constraints. Therefore, this problem can be solved with the DC algorithm, which falls in the category of majorization minimization (MM) algorithms [12].

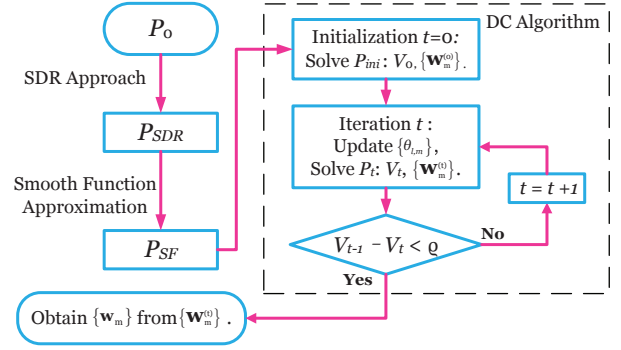


Fig. 2: Overall procedure for solving problem \mathcal{P}_0

IV. DC BASED ALGORITHM AND ANALYSIS

In this section, we first present the DC based algorithm for solving the problem \mathcal{P}_{SF} using the logarithmic smooth function. We then show that the resulting algorithm can also be interpreted as a DC algorithm with the other two smooth functions and a different θ updating rule.

A. DC Based Algorithm with Log-Function

The DC algorithm iteratively optimizes an approximated convex function of the original concave objective function and produces a sequence of improving $\{\mathbf{W}_m\}$. The algorithm converges to some global/local optimal solution.

The initial $\{\mathbf{W}_m^{(0)}\}$ is found by solving the following power minimization problem

$$\mathcal{P}_{ini} : V_{ini} \triangleq \text{minimize}_{\{\mathbf{W}_m\}_{m=1}^M} \sum_{m=1}^M \text{Tr}(\mathbf{W}_m) \quad (13a)$$

$$\text{subject to} \quad (9b), (9c), (9d). \quad (13b)$$

In the t -th iteration, $\{\mathbf{W}_m^{(t)}\}$ is generated as the solution of the approximated convex optimization problem,

$$\mathcal{P}_t : V_t \triangleq \text{minimize}_{\{\mathbf{W}_m\}_{m=1}^M} G(\mathbf{W}) + \sum_{m=1}^M \sum_{l=1}^L \alpha_{l,m} \left\langle \left(\nabla_{\mathbf{W}_m^{(t-1)}} f_{l,m} \right), \left(\mathbf{W}_m - \mathbf{W}_m^{(t-1)} \right) \right\rangle \quad (14a)$$

$$\text{subject to} \quad (9b), (9c), (9d). \quad (14b)$$

where $\nabla_{\mathbf{W}_m^{(t-1)}} f_{l,m}$ is the gradient matrix of $f_{l,m}$ at $\mathbf{W}_m^{(t-1)}$.

Specifically, for log-function, the gradient matrix $\nabla_{\mathbf{W}_m} f_{l,m} \in \mathbb{C}^{LN_t \times LN_t}$ at $\{\mathbf{W}_m^{(t)}\}$ is

$$D_{log}(l, m) = \nabla_{\mathbf{W}_m^{(t)}} f_{l,m} = \frac{\mathbf{J}_l}{\text{Tr}\left(\mathbf{W}_m^{(t)} \mathbf{J}_l\right) + \theta_{l,m}}, \quad (15)$$

If we let the smoothness factor $\theta_{l,m} = \epsilon$, where ϵ is a very small positive constant. Then we have

$$D_{log}(l, m) = \frac{\mathbf{J}_l}{\text{Tr}\left(\mathbf{W}_m^{(t)} \mathbf{J}_l\right) + \epsilon}. \quad (16)$$

Note that (16) has the similar form as the weight factor of the reweighted ℓ_1 -norm approach in [11]. Thus, the DC algorithm with log-function and smoothness factor $\theta_{l,m} \rightarrow 0$ is just the reweighted ℓ_1 -norm minimization algorithm of [3].

In (14a), function $F(\mathbf{W})$ is approximated with its first-order Taylor expansion, which provides an upper bound. This algorithm terminates when the sequence of $\{\mathbf{W}_m^{(t)}\}$ converges to some stationary point, and the objective value V_t converges, that is, $V_{t-1} - V_t < \varrho$, where ϱ is a small constant.

B. Updating Rule of θ for Other Smooth Functions

The performance of ℓ_0 -norm approximation algorithms depends on the smoothness factor θ . Intuitively, when x is large, θ should be large so that the approximation algorithm can explore the entire parameter space; when x is small, θ should be small so that $f(x)$ has behavior close to ℓ_0 -norm. In [14], the authors propose to use a decreasing sequence of θ , but the updating rule does not depend on x .

In this paper, we explore a novel θ updating rule that achieves the above effect automatically using a sequence of θ which depends on specific x in each iteration. More specifically, we propose to set θ to be the one that maximizes the gradient of the approximation function.

Note that the gradient matrices of the exponential and arctangent functions in (10) are, respectively,

$$\mathbf{D}_{exp}(l, m) = \frac{\mathbf{J}_l}{\theta_{l,m}} e^{-\frac{\text{Tr}(\mathbf{W}_m \mathbf{J}_l)}{\theta_{l,m}}}, \quad (17)$$

and

$$\mathbf{D}_{atan}(l, m) = \frac{2}{\pi} \cdot \frac{\mathbf{J}_l}{\theta_{l,m} \left(\frac{\text{Tr}(\mathbf{W}_m \mathbf{J}_l)}{\theta_{l,m}} \right)^2 + \theta_{l,m}}. \quad (18)$$

An interesting observation is that for all three functions in (10), if we maximize their gradients, we get expressions of the same form. Specifically, for the log-function, we get (16) with optimal $\theta_{l,m}^* \rightarrow 0$. For the exp-function and atan-function, we get

$$\mathbf{D}_{exp}^*(l, m) = \max_{\theta_{l,m}} \mathbf{D}_{exp}(l, m) = \frac{\mathbf{J}_l}{e \cdot \text{Tr}(\mathbf{W}_m^{(t)} \mathbf{J}_l)}, \quad (19)$$

$$\mathbf{D}_{atan}^*(l, m) = \max_{\theta_{l,m}} \mathbf{D}_{atan}(l, m) = \frac{\mathbf{J}_l}{\pi \cdot \text{Tr}(\mathbf{W}_m^{(t)} \mathbf{J}_l)}, \quad (20)$$

respectively with the optimal $\theta_{l,m}^* = \text{Tr}(\mathbf{W}_m^{(t)} \mathbf{J}_l)$.

We see that the gradient matrices for all three approximation functions in (10) differ by a constant multiple only. Therefore, they lead to the same algorithm if we update $\theta_{l,m}$ such that $\nabla_{\mathbf{W}_m^{(t)}} f_{l,m}$ is maximized in the t -th iteration, i.e.,

$$\theta_{l,m}^* \triangleq \arg \max_{\theta_{l,m}} \nabla_{\mathbf{W}_m^{(t)}} f_{l,m}. \quad (21)$$

In this proposed algorithm, the approximation functions are adjusted dynamically to achieve a good tradeoff between smoothness and approximation to ℓ_0 -norm. Further, similarity between (16), (19), and (20) means that with proposed θ updating rule, these three approximation functions lead to almost the same performance.

V. SIMULATION RESULTS

We consider a cache-enabled cloud RAN covering an area of circle with the radius of 1.2km, where 7 BSs ($L = 7, N_t = 3$) are placed in a equilateral triangular lattice with the distance between adjacent BSs of 0.8km. The total number of contents is $F = 100$. A total number of 140 users are distributed in this network with uniform distribution and they are scheduled in a round-robin manner. In each scheduling interval, $K = 14$ users are scheduled. We assume half of the scheduled users request a common content (e.g., a live video) and each of the rest randomly requests one content according to the content popularity, which is modeled as Zipf distribution with skewness parameter 1. Users requesting the same content participate in the same multicast group. We assume all BSs have the same cache size. The channels between BSs and users are generated with a normalized Rayleigh fading component and a distance-dependent path loss, modeled as $PL(\text{dB}) = 148.1 + 37.6 \log_{10}(d)$ with 8dB log-normal shadowing, where d is the distance from the user to the BS. The transmit antenna power gain at each BS is 10 dBi. The power spectral density of downlink noise is -172dBm/Hz with the channel bandwidth of 10MHz. The peak transmission power of each BS is $P_l = 10\text{W}$, $\forall l$. The target SINR of each content is 10dB. We set $\varrho = 10^{-6}$ for the convergence condition in the DC algorithm and $\epsilon = 10^{-7}$ in (16). Each simulation result is averaged over 300 scheduling intervals.

In Fig. 3, we show the effects of wireless caching and the cache size. The popularity-aware cache refers to the policy where each BS caches the contents with highest popularity. The figure shows that compared with the network without cache, the cache-enabled network can reduce the backhaul cost by more than 50% when each BS only caches 5% of the total contents. The backhaul reduction is up to 75% when each BS can cache 30% of the total contents.

In Fig. 4, we compare the effects of different cache strategies. In the random cache policy, the contents are randomly cached with equal probability in each BS. The result shows that the popularity-aware cache has better power-backhaul efficiency in general. In specific, with the same transmission power cost of 38dBm, and the same cache size of 10, popularity-aware cache only costs about 1/4 backhaul cost of the random cache policy. However, in the extreme case when we do not consider the total transmission power, the minimum backhaul costs of the two caching strategies are about the same if the cache size is 30.

In Fig. 5, we compare the power-backhaul tradeoff of multicast and unicast transmission in the same scenario. We use popularity-aware caching policy with the cache size of 10. In unicast transmission, users of the same group are served by different beamformers and we use the algorithm proposed in [3], where in each iteration, for $\forall k, l$, the weight factor ρ_k^l is set to 0 if the content requested by user k is cached in BS l . If multiple users in the same group are served by a same BS, the backhaul cost is counted only once. The figure shows that in the power-limited system, the total power consumption

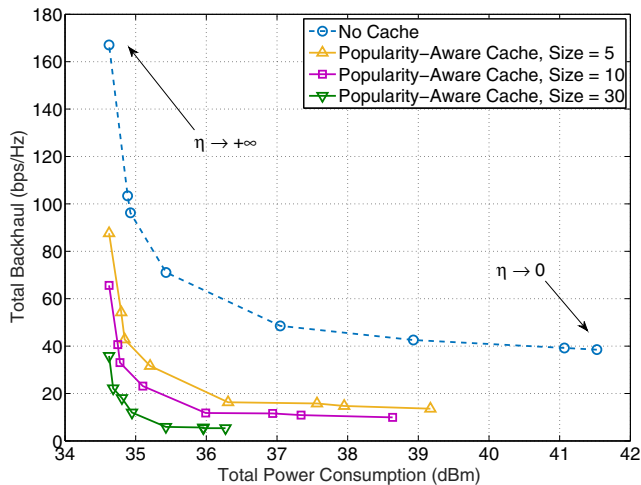


Fig. 3: Power-backhaul tradeoff for different cache size.

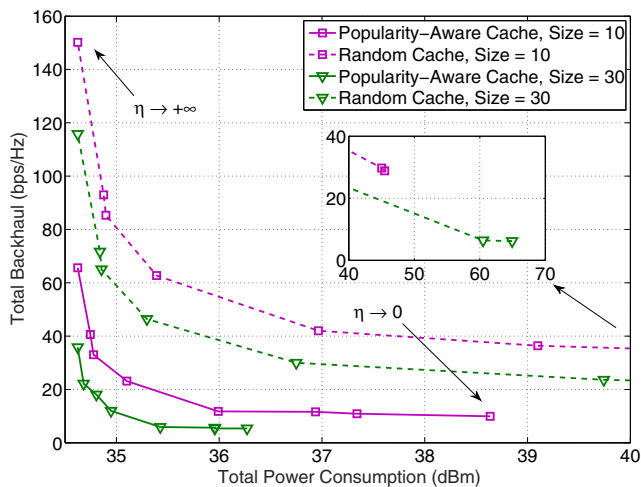


Fig. 4: Power-backhaul tradeoff for different cache policies.

of multicast transmission is about 2dB less than the unicast scenario. When the total power consumption is 38dBm, the backhaul cost of multicast transmission is only 1/3 of unicast transmission. These observations validates the advantage of caching and multicasting in such a scenario.

VI. CONCLUSION

This paper investigates the joint design of multicast beamforming and content-centric BS clustering in a cache-enabled cloud RAN. The optimization problem is formulated as the minimization of the total network cost, including the power consumption and the backhaul cost, under the QoS constraint of each multicast group. We adopt the SDR method and the smooth function approach, introduced in sparse signal processing, to approximate this non-convex problem as a DC programming problem. We then propose a DC algorithm to obtain sub-optimal solutions. Further, we propose a new smoothness updating method and give insight into its connection to reweighted ℓ_1 -norm minimization. Simulation results show that, compared with unicast transmission, multicast transmission can achieve better power-backhaul tradeoff, and

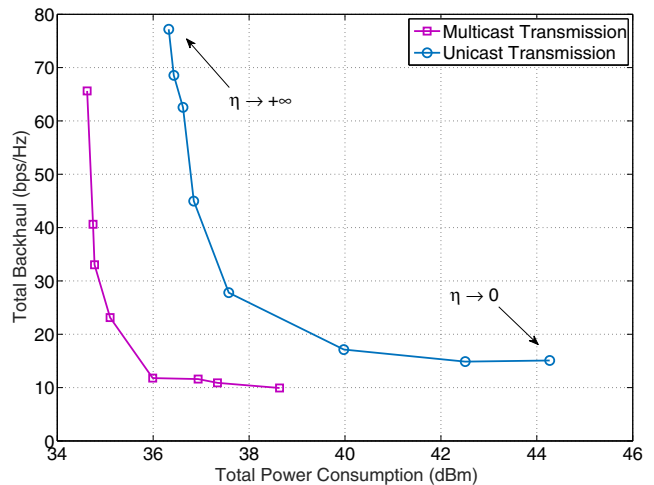


Fig. 5: Power-backhaul tradeoff for multicast and unicast transmission.

the backhaul cost can be further reduced with larger cache size.

REFERENCES

- [1] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.
- [2] J. Zhao, T. Q. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, 2013.
- [3] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network MIMO system via reweighted power minimization," in *Proc. IEEE Global Telecommun. Conf.*, 2013.
- [4] F. Zhuang and V. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb 2014.
- [5] 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.
- [6] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, April 2013.
- [7] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. on Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [8] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2014.
- [9] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–21, January 2013.
- [10] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, 2008.
- [11] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [12] R. Horst and N. V. Thoai, "DC programming: overview," *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.
- [13] F. Rinaldi, F. Schoen, and M. Sciacicone, "Concave programming for minimizing the zero-norm over polyhedral sets," *Computational Optimization and Applications*, vol. 46, no. 3, pp. 467–486, 2010.
- [14] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, 2009.